Glossário Unificado de IA

Metodologia, implementação e acompanhamento de projetos de IA

Junho 2026

MINISTÉRIO DA
GESTÃO E DA INOVAÇÃO
EM SERVIÇOS PÚBLICOS

SUMÁRIO

1. Introdução

Este anexo é parte integrante do **GuIA** (Guia unificado de Inteligência Artificial). Este documento é uma referência consolidada dos termos, siglas e conceitos técnicos mencionados no Guia unificado de IA - Metodologia de Desenvolvimento de Soluções de IA e anexos técnicos.

Este anexo não substitui as referências, termos e glossários do Guia unificado de IA ou outros anexos, apenas consolida para facilidade de acesso.

2. Siglas e acrônimos

Sigla	Expansão
AIE	Autoavaliação de Impacto Ético
ANN	*Approximate Nearest Neighbors*
ANPD	Autoridade Nacional de Proteção de Dados
APF	Administração Pública Federal
APQC	*American Productivity & Quality Center*
ART	*Adversarial Robustness Toolbox*
ATLAS	*Adversarial Threat Landscape for AI Systems* (MITRE)
AUC-ROC	*Area Under the Receiver Operating Characteristic Curve*
BLEU	*Bilingual Evaluation Understudy*
BM25	*Best Match 25*
CMS	*Content Management System*
CoT	*Chain-of-Thought*
CVSSv3	*Common Vulnerability Scoring System version 3*
DAM	*Database Activity Monitoring*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
DCMI	*Dublin Core Metadata Initiative*
DITA	*Darwin Information Typing Architecture*
DMBOK	*Data Management Body of Knowledge (DAMA)*
DMS	*Document Management System*
DoS	*Denial of Service*
DPIA	*Data Protection Impact Assessment (equivalente internacional; não é o termo legal brasileiro)*
DPO	*Data Protection Officer*
DQ	*Data Quality*
ECM	*Enterprise Content Management*
ENAP	Escola Nacional de Administração Pública

Sigla	Expansão
eMAG	Modelo de Acessibilidade em Governo Eletrônico
ePING	Padrões de Interoperabilidade de Governo Eletrônico
ETP	Estudo Técnico Preliminar
FinOps	*Financial Operations*
GMUD	Gestão de Mudança
GSI/PR	Gabinete de Segurança Institucional da Presidência da República
GuIA	Guia unificado de Inteligência Artificial
HSM	*Hardware Security Module*
HyDE	*Hypothetical Document Embeddings*
IA	Inteligência Artificial
IoU	*Intersection over Union*
ISO	*International Organization for Standardization*
ITSM	*Information Technology Service Management*
KB	*Knowledge Base (Base de Conhecimento)*
KCS	*Knowledge-Centered Service*
KM	*Knowledge Management* (Gestão do Conhecimento)
KMS	Knowledge Management System
KPI	*Key Performance Indicator*
KV cache	*Key-Value cache*
LAI	Lei de Acesso à Informação (Lei 12.527/2011)
LGPD	Lei Geral de Proteção de Dados Pessoais (Lei 13.709/2018)
LLM	*Large Language Model*
LoRA	*Low-Rank Adaptation*
MAE	*Mean Absolute Error*
mAP	*Mean Average Precision*
MFA	*Autenticação Multifator*
MLOps	*Machine Learning Operations*
MRR	*Mean Reciprocal Rank*
MVP	*Minimum Viable Product*
NCSC	*National Cyber Security Centre*
NER	*Named Entity Recognition*
NIA	Núcleo de IA
NIST	*National Institute of Standards and Technology*
NLP	*Natural Language Processing*
OCI	*Open Container Initiative*
OASIS	*Organization for the Advancement of Structured Information Standards*

Sigla	Expansão
OWASP	*Open Web Application Security Project*
OWL	*Web Ontology Language*
PBIA	Plano Brasileiro de Inteligência Artificial (2024–2028)
PEFT	*Parameter-Efficient Fine-Tuning*
PHI	*Protected Health Information*
PII	*Personally Identifiable Information*
PPSI	*Política de Privacidade e Segurança da Informação*
PROV	*Provenance (W3C PROV-DM)*
QLoRA	*Quantized Low-Rank Adaptation*
RACI	*Responsible, Accountable, Consulted, Informed*
RAG	*Retrieval-Augmented Generation*
RAGAS	*Retrieval-Augmented Generation Assessment*
RBAC	*Role-Based Access Control*
RDF	*Resource Description Framework*
RIPD	Relatório de Impacto à Proteção de Dados Pessoais
RMF	*AI Risk Management Framework (NIST)*
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*
RPO	*Recovery Point Objective*
RSGI	Responsável Setorial pela Gestão de Integridade
RTO	*Recovery Time Objective*
SAST	*Static Application Security Testing*
SBOM	*Software Bill of Materials*
SCA	*Software Composition Analysis*
SCF	*Secure Controls Framework*
SDLC	*Software Development Life Cycle*
SECI	Socialização, Externalização, Combinação, Internalização
SFT	*Supervised Fine-Tuning*
SGIA	Sistema de Gestão de Inteligência Artificial (ISO/IEC 42001)
SGIP	Sistema de Gestão de Informações de Privacidade (ISO/IEC 27701)
SGSI	Sistema de Gestão de Segurança da Informação (ISO/IEC 27001)
SHACL	*Shapes Constraint Language*
SKOS	*Simple Knowledge Organization System*
SLA	*Service Level Agreement*
SME	*Subject Matter Expert*
SSO	*Single Sign-On*
SSOT	*Single Source of Truth*

Sigla	Expansão
STRIDE	*Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege*
SVM	*Support Vector Machine*
TCU	Tribunal de Contas da União
TF-IDF	*Term Frequency–Inverse Document Frequency*
TLS	*Transport Layer Security*
TR	Termo de Referência
TTFT	*Time to First Token*
USN	Unidade de Serviços de Nuvem
UUID	*Universally Unique Identifier*
UX	*User Experience*
VLM	*Vision-Language Model*
W3C	*World Wide Web Consortium*
WCAG	*Web Content Accessibility Guidelines*
XAI	*Explainable Artificial Intelligence*

3. Papéis e atores

Alta Gestão (órgão). Nível superior de liderança e decisão na administração pública, responsável por definir diretrizes estratégicas, tomar decisões institucionais e representar o órgão.

CISC gov.br (Centro Integrado de Segurança Cibernética): Equipe principal para prevenção, tratamento e resposta a incidentes cibernéticos do Governo Digital, sendo que devem ser notificados imediatamente, por parte dos órgãos, caso ocorra algum problema (como um vazamento de dados gerado por uma IA). Também é responsabilidade dessa equipe executar os testes de intrusão (*penetration tests)* e os testes estáticos e dinâmicos de segurança nas aplicações; realizar análises contínuas e investigações de inteligência sobre novas ameaças cibernéticas e monitorar padrões maliciosos no tráfego de rede; elaborar e publicar alertas, e emitir determinações com prazos rígidos para que os órgãos corrijam falhas e vulnerabilidades que sejam classificadas como de alta criticidade.

Comitê de IA / Ética. Instância responsável por orientar, avaliar e supervisionar o uso de IA, assegurando conformidade legal, ética, técnica e institucional.

Curador de dados (*Data owner*). Agente público responsável pela gestão de ativos de dados, internos ou externos ao órgão ou entidade, designado por liderança na estrutura organizacional [2]. Responsável por garantir a qualidade, confiabilidade, conformidade e uso adequado dos dados do órgão, ao longo de todo o seu ciclo de vida. Atua como um guardião dos dados, assegurando que sejam corretos, atualizados, bem documentados, acessíveis a quem precisa e utilizados de forma ética e conforme normas legais. Suas principais responsabilidades são: • Manter atualizada a catalogação de dados e metadados sob sua custódia nas unidades negociais; • Classificar os dados quanto ao nível de acesso definidos conforme legislações vigentes;

• Assegurar a proteção dos dados pessoais, observando as orientações do encarregado e conforme o disposto pela lei 13.709, de 14 de agosto de 2018 [1].

Curador técnico *(Data steward* - Administrador de Dados*).* Responsável por gerenciar os ativos de dados em nome de outrem (curador de dados) e no melhor interesse da organização, garantindo que os dados empresariais sejam de alta qualidade e possam ser usados de forma eficaz [5]. **Cria e gerencia Metadados principais**, sendo muitas vezes os responsáveis pela manutenção do Glossário de Negócios da organização. **Documenta regras de negócios, padrões de dados e regras de Qualidade de Dados (DQ)**, garantindo que o que se define como "dado de alta qualidade" tenha consenso na organização. **Gerencia problemas de qualidade**, atuando na identificação e resolução (ou facilitação da resolução) de falhas nos dados. Enquanto o curador de dados (*data owner* – especialista de negócio) indica o que é o dado, como, quando e por quem deve ser processado / usado, o **curador técnico** (Administrador de dados) é o elo entre o curador e o custodiante do dado (TI ou órgão responsável pelo processamento do dado). Para uso de RAG, há necessidade de uma curadoria especializada que necessita ser desempenhada por um perfil mais técnico. Um **curador técnico *(Adminstrador de dados)***, neste caso o perfil é mais adequado como responsável pela curadoria e revisão periódica de documentos na base de conhecimento.

Demandante. Unidade de um órgão ou órgão que origina a demanda por uma solução de IA.

Encarregado pelo tratamento de dados pessoais, doravante denominado ETDP (**DPO -** *Data Protection Officer*) - pessoa indicada pelo controlador como braço técnico consultivo para atuar como canal de comunicação (intermediador), orientador sobre boas práticas, fiscalizador de normas e relator sobre riscos e necessidades de melhorias na proteção de dados pessoais, conforme a Lei Geral de Proteção de Dados (LGPD – Lei nº 13.709/2018). Não toma decisões sobre as finalidades do tratamento de dados.

Executor (Parceiro executor). Entidade, órgão ou empresa responsável por desenvolver uma solução de IA. Denominado "Produtor de IA" na ABNT NBR ISO/IEC 22989:2023.

Facilitador. Pessoa jurídica, servidor público ou parceiro com capacidade de executar a fase de descoberta de hipóteses e de solução na etapa de enquadramento do desafio.

Gestor de Segurança da Informação: Responsável por orientar, conduzir diagnósticos, planejar e monitorar todas as medidas de segurança da informação do órgão. Neste guia, ele atua como a principal autoridade técnica que aprova o mapeamento de riscos cibernéticos, valida os testes de segurança antes da implantação e garante a supervisão ininterrupta do sistema durante a operação.

Gestor de TIC (tecnologia da Informação e Comunicação): responsável por planejar, executar e monitorar projetos tecnológicos, alinhando-os à Estratégia de Governo Digital (EGD) e ao Plano Diretor de TIC (PDTIC) do órgão.

Núcleo de IA (NIA): iniciativa estratégica destinada a coordenar o conjunto de ações que visam a transformar a administração pública por meio da adoção de inteligência artificial ética e responsável (PBIA, 2025, p.39). No contexto deste guia, o NIA é consultado quando necessário para alinhamento estratégico ou informado via sistema de organização e registro dos sistemas de IA sobre o andamento dos projetos.

RSGI (Responsável Setorial pela Gestão de Integridade). Papel responsável por coordenar e

gerir riscos para a integridade institucional relacionados ao projeto de IA. Atua como guardião ético nas fases iniciais, prevenindo vieses discriminatórios, falhas de *compliance* e corrupção algorítmica. Sigla usada nas matrizes RACI da Metodologia.

Sustentador (Parceiro de sustentação). Entidade, órgão ou empresa responsável por operar e manter a solução de IA após sua implantação.

4. Ciclo de vida e metodologia

Ciclo de vida de soluções de IA. Conjunto estruturado de etapas que cobre o desenvolvimento, implantação, operação e desativação de uma solução de IA: Prospecção → Estruturação → Experimentação → Implementação → Implantação (*rollout*) → Sustentação → Desativação.

Prospecção de desafios. Etapa inicial que identifica, mapeia e prioriza desafios do serviço público endereçáveis por IA, alinhando-os à estratégia federal.

Estruturação do desafio. Etapa que aprofunda a compreensão do problema, avalia viabilidade técnica, ética e financeira, e formula a proposta de solução a ser experimentada.

Experimentação. Etapa de desenvolvimento de protótipos e realização de testes técnicos para validar desempenho, segurança e viabilidade econômica da solução.

Implementação. Etapa de desenvolvimento ou contratação da solução, incluindo construção da camada aplicacional, integração ao ambiente de produção e documentação técnica.

Implantação (*rollout*). Etapa de execução do plano de entrada em produção, com validação da gestão de mudança e transferência da solução à equipe de sustentação.

Sustentação. Etapa que garante a continuidade da geração de valor pela solução, mantendo estabilidade técnica, segurança e precisão do modelo por meio de monitoramento e retreino.

Desativação. Etapa que garante a remoção controlada de um modelo ou algoritmo obsoleto, preservando rastreabilidade, memória técnica e prestação de contas.

Ponto de decisão. Momento formal ao final de uma fase onde os entregáveis são avaliados e se decide se o projeto avança, retorna para ajustes ou é descontinuado.

***Baseline*.** Estado atual mensurável do processo ou problema, tomado como referência para avaliar o impacto da solução de IA.

***Make or Buy*.** Decisão estratégica de desenvolver internamente ou adquirir uma solução de IA, considerando interoperabilidade, ciclo de vida e dependência de fornecedor.

Matriz RACI. Ferramenta de atribuição de papéis por atividade, definindo quem é Responsável executor (R), quem tem Autoridade de decisão (A), quem deve ser Consultado (C) e quem deve ser Informado (I).

***Threshold* (limiar).** Valor mínimo ou máximo definido como critério de aceitação de uma métrica de desempenho do modelo.

***Vendor lock-in* (aprisionamento tecnológico).** Dependência excessiva de um único fornecedor de tecnologia ou nuvem que impede a portabilidade da solução e limita a autonomia do órgão.

5. Entregáveis

Briefing (detalhamento do desafio). Documento produzido na fase de aprofundamento que detalha o contexto e o desafio validado pelo órgão para embasar a proposta de solução.

Dossiê de arquivamento (*model registry*). Repositório contendo a versão final do código, os dados de treino, os logs de decisão e as razões do *retirement*, para consulta em futuras auditorias.

ETP (Estudo Técnico Preliminar). Principal entregável da fase de aprofundamento, consolidando a análise técnica de viabilidade da solução.

***Model Card*.** Documento obrigatório antes da entrada em produção de um modelo LLM. Contempla intenção de uso, usos fora de escopo, dados de treinamento, métricas de desempenho, limitações conhecidas, vieses identificados, knowledge cutoff, riscos residuais e responsável pela manutenção.

***One Page* (Formulário de Desafio Preliminar).** Documento preenchido pelo demandante com o desafio, a abordagem atual, o potencial de IA, os benefícios esperados e o panorama de dados disponíveis.

Plano de Experimentação. Documento que detalha a abordagem de IA, a arquitetura de alto nível, as hipóteses a testar e o cronograma da fase de experimentação.

Relatório de Estruturação. Documento final da etapa de Estruturação com recomendação positiva ou negativa de avanço, incluindo as métricas de sucesso acordadas.

RIPD (Relatório de Impacto à Proteção de Dados Pessoais). Documento exigido pela LGPD para avaliar o impacto do tratamento de dados pessoais, conforme orientações da ANPD.

***Ticket* de *retirement*.** Documento que formaliza o motivo da desativação de um modelo: queda de acurácia, obsolescência de dados, novos requisitos regulatórios, entre outros.

6. Conceitos técnicos de IA e ML

Alucinação (LLM). Geração de informação factualmente incorreta ou não suportada pelos documentos de contexto por parte de um LLM, apresentada como se fosse verdadeira.

ANN (*Approximate Nearest Neighbors*). Busca aproximada de vizinhos mais próximos em bases vetoriais de grande escala. Estruturas como HNSW, IVF e FAISS reduzem latência e custo com trade-off em recall e tempo de indexação.

Aprendizado de máquina supervisionado. Abordagem de ML em que o modelo é treinado com dados rotulados para aprender a mapear entradas às saídas esperadas.

Aprendizado de máquina não supervisionado. Abordagem de ML em que o modelo identifica padrões em dados não rotulados, como agrupamentos (clusterização).

***Catastrophic forgetting* (esquecimento catastrófico).** Falha de *fine-tuning* em que o modelo perde capacidades gerais ao ser ajustado em *dataset* específico. Requer conjunto de avaliação separado e comparação contra o modelo base.

***Chain-of-Thought* (CoT).** Técnica de *prompt engineering* que instrui o modelo a produzir simulação de inferência passo a passo antes de concluir. Variantes incluem *zero-shot CoT, self-consistency CoT e Tree of Thought*.

***Concept drift* (deriva de conceito).** Mudança no relacionamento entre variáveis de entrada e saída ao longo do tempo, que torna o modelo desatualizado em relação ao fenômeno modelado.

Conhecimento paramétrico. Informações codificadas nos pesos do modelo durante o treinamento, acessíveis sem consulta a fontes externas. Contraponto à recuperação via RAG.

Corte de conhecimento (knowledge cutoff). Data a partir da qual o modelo não tem acesso a eventos, legislações ou dados publicados, pois seu treinamento foi encerrado antes disso. Limitação inerente do LLM puro.

***Data augmentation* (enriquecimento de dados).** Técnica para ampliar e diversificar o conjunto de dados de treinamento, aumentando qualidade e representatividade sem violar a confidencialidade.

***Data drift* (derivação de dados).** Mudança na distribuição estatística dos dados de entrada ao longo do tempo, que podem degradar a performance do modelo.

***Deep learning*.** Subcampo do aprendizado de máquina baseado em redes neurais profundas com múltiplas camadas de representação.

DPO *(Direct Preference Optimization)*. Técnica de alinhamento que otimiza diretamente sobre pares de preferências humanas, sem modelo recompensa intermediário. Mais estável e menos custoso que RLHF.

Explicabilidade (XAI). Capacidade de um sistema de IA de justificar suas predições ou decisões de modo compreensível por humanos. Exigida em sistemas de alto risco.

***Fairness* (equidade algorítmica).** Propriedade de um modelo de IA de não produzir resultados discriminatórios ou tendenciosos em relação a grupos populacionais.

***Feature engineering* (engenharia de atributos).** Processo de seleção, transformação e criação de variáveis a partir de dados brutos para alimentar modelos de ML.

Few-shot. Técnica em que dois a cinco exemplos de entrada-saída são incluídos no prompt antes da consulta real, calibrando formato, estilo e nível de detalhe esperado sem exigir ajuste fino.

***Fine-tuning*.** Técnica de ajuste de um modelo pré-treinado com dados de um domínio específico para melhorar seu desempenho naquele contexto. Alternativa ou complemento ao RAG quando o conhecimento é estável.

***Function calling / Tool use*.** Capacidade de um sistema LLM de selecionar uma função externa, preencher argumentos estruturados e utilizar o resultado na resposta final. Exige validação de argumentos fora do modelo e autorização por escopo mínimo.

Ground Truth. Conjunto de dados de referência com respostas corretas conhecidas, usado para validação experimental e calibração de métricas automáticas. Componente obrigatório do Plano de Experimentação.

Guardrail. Controle técnico e processual que restringe entradas, saídas e ações do sistema LLM de acordo com políticas de segurança, privacidade, qualidade e conformidade. Podem ser determinísticos, baseados em ML ou *in-context.*

***Human-in-the-loop*.** Abordagem que mantém supervisão humana em etapas críticas do ciclo de vida do modelo, como validação de retreino ou escalonamento de respostas de baixa confiança.

IA generativa. Abordagem de IA capaz de gerar conteúdo novo (texto, imagem, código) a partir de padrões aprendidos durante o treinamento.

***Instruction tuning* / *Domain tuning*.** *Instruction tuning* adapta o modelo a seguir formatos e instruções específicas; *domain tuning* adapta vocabulário e padrões de um domínio. O segundo pode memorizar documentos e reduzir capacidade geral se mal calibrado.

***Jailbreak*.** Técnica de manipulação de um LLM para contornar suas restrições de segurança e políticas de uso.

LLM (Modelo de Linguagem de Grande Escala). Modelo de IA treinado em vastos volumes de texto, capaz de gerar e compreender linguagem natural. Responsável pela geração de texto em sistemas RAG.

Janela de contexto *(context window)*. Limite máximo de *tokens* que o modelo processa em uma única inferência, abrangendo system *prompt*, histórico de conversa e consulta. Varia de 4K a 200K+ *tokens* conforme o modelo.

KV *cache* (key-value cache). Mecanismo de reuso de representações de atenção computadas para prefixos de *prompt* repetidos, reduzindo custo de inferência. Exige controle cuidadoso de isolamento entre usuários.

Linhagem de dados (*data lineage*). Rastreamento documentado da origem, transformações e destino dos dados utilizados no desenvolvimento do modelo, garantindo auditabilidade.

***Model registry*.** Repositório centralizado para catalogação, versionamento e rastreabilidade de modelos de IA ao longo do ciclo de vida.

Modelo fundacional (*foundation model*). Modelo de IA treinado em larga escala, responsável pela geração de texto natural. Recebe a consulta combinada com system prompt e histórico e produz resposta coerente. Pode ser obtido via API, auto-hospedado ou desenvolvido internamente.

PEFT / LoRA (*Parameter-Efficient Fine-Tuning* / *Low-Rank Adaptation*). Técnicas que ajustam apenas uma fração dos parâmetros do modelo, adicionando matrizes de baixo posto às camadas de atenção. Reduzem custo computacional e de armazenamento sem perda significativa de qualidade.

***Pipeline de retreino*.** Fluxo automatizado que atualiza o modelo com novos dados, mantendo sua precisão ao longo do tempo. Inclui etapa de validação humana (*human-in-the-loop*).

***Prompt Engineering*.** Técnica de formulação de instruções ao LLM para otimizar as respostas sem necessidade de retreinamento ou acesso a bases externas.

Quantização / QLoRA. Quantização reduz a precisão numérica dos pesos (ex.: INT8, INT4, GGUF) permitindo executar modelos em hardware de menor custo. QLoRA combina LoRA com carregamento quantizado do modelo-base para *fine-tuning* eficiente.

Reconhecimento e síntese de voz. Abordagem de IA para transcrever fala em texto e converter texto em fala.

***Red teaming adversarial*.** Atividade técnica reprodutível de testes adversariais planejados, cobrindo *jailbreak*, injeção direta e indireta de *prompt*, extração de system *prompt*, uso indevido de ferramentas e geração de desinformação. Obrigatório antes de produção.

RLHF (*Reinforcement Learning from Human Feedback*). Técnica de alinhamento que treina o modelo com um modelo recompensa intermediário, baseado em avaliações humanas de utilidade, segurança e correção.

Robustez adversarial. Capacidade de um modelo de IA de manter desempenho correto diante de entradas manipuladas ou ataques adversariais.

***Sandbox*.** Ambiente de execução isolado para testar modelos ou componentes sem risco de comprometimento do sistema de produção.

SFT (*Supervised Fine-Tuning*). Ajuste fino com pares instrução-resposta curados por especialistas, adaptando estilo, vocabulário técnico e formato de saída do modelo.

Sistema agentivo. Sistema que combina LLM, ferramentas, memória e ciclos de planejamento/execução para resolver tarefas em múltiplas etapas. Requer controles explícitos de orçamento de passos, tempo, ferramentas e política de confirmação humana.

Sistema de recomendação. Sistema de IA que sugere itens relevantes a usuários com base em comportamento ou preferências.

Solução aplicacional ("casca"). Camada de interface e funcionalidade construída sobre o modelo de IA — *chatbots*, assistentes virtuais ou sistemas inteligentes — distinta do desenvolvimento do modelo em si.

***Speculative decoding*.** Técnica de aceleração de geração que usa um modelo auxiliar menor para propor *tokens* candidatos, validados pelo modelo principal. Reduz latência, mas adiciona complexidade e requer validação de compatibilidade.

Temperatura. Parâmetro que reescala os *logits* antes do *softmax*, alterando a entropia da distribuição amostrada. Valores próximos de zero tornam a saída mais determinística; valores acima de 1,0 aumentam diversidade. Para sistemas governamentais recomenda-se entre 0,0 e 0,3.

Token / Tokenizador / Tokenização. Unidade discreta de texto processada pelo modelo. O tokenizador converte texto em sequências de *tokens* e reconverte *tokens* de saída em texto legível. Afeta diretamente janela de contexto, custo de inferência e latência.

Top-K. Parâmetro que restringe a amostragem aos K *tokens* mais prováveis em cada passo, descartando opções de baixa probabilidade.

Top-p (*nucleus sampling*). Parâmetro que seleciona o menor conjunto de *tokens* cuja probabilidade acumulada atinge o limiar p. Adapta dinamicamente o número de candidatos conforme a entropia da distribuição.

Visão computacional. Campo da IA dedicado à interpretação e análise de imagens e vídeos por máquinas.

***Zero-shot*.** Técnica de *prompt engineering* em que o modelo executa a tarefa descrita sem exemplos prévios, aplicando diretamente seu conhecimento paramétrico.

7. Arquitetura RAG

Base vetorial (*vector database*). Sistema de armazenamento projetado para indexar e consultar *embeddings* por similaridade semântica, com suporte à filtragem por metadados.

Busca vetorial / busca semântica. Método de recuperação que utiliza *embeddings* para encontrar documentos por proximidade de significado, mesmo quando os termos exatos diferem.

Busca esparsa / lexical. Método de recuperação baseado em correspondência direta de termos (TF-IDF, BM25), com alta precisão para siglas, números e nomes próprios.

Busca híbrida (*hybrid search*). Combinação de busca vetorial e busca lexical, mesclando resultados por média ponderada para aproveitar as vantagens de ambas as abordagens.

Camada de Indexação. Componente arquitetural responsável por receber, validar, extrair texto e gerar *embeddings* dos documentos fonte.

Camada de Geração. Componente arquitetural que constrói o *prompt* combinando instrução, contexto recuperado e consulta, invoca o LLM e extrai citações.

Camada de Governança. Componente arquitetural que centraliza controle de acesso, auditoria de interações, monitoramento de qualidade e gestão de incidentes.

Camada de Recuperação. Componente arquitetural que processa a consulta, gera o *embedding* correspondente e executa a busca com filtragem de metadados.

***Chunking* (segmentação).** Divisão de documentos em trechos menores (tipicamente 256–1024 *tokens*) para viabilizar a indexação e a recuperação eficiente.

Contexto recuperado. Conjunto de *chunks* selecionados pela etapa de recuperação e fornecidos ao LLM para fundamentar a geração da resposta.

***Cross-encoder*.** Modelo de re-*ranking* que analisa conjuntamente a consulta e cada documento candidato, oferecendo avaliação de relevância mais precisa que *bi-encoders*, porém com maior latência.

Distância euclidiana. Métrica de distância geométrica direta entre vetores no espaço multidimensional. Alternativa à similaridade por cosseno em bases vetoriais.

***Embedding* (codificação semântica).** Representação vetorial densa de texto que captura significado semântico em um espaço matemático multidimensional, permitindo busca por similaridade.

Filtragem por metadados. Restrição da recuperação de *chunks* com base em atributos estruturados, como classificação de segurança, data de validade ou unidade responsável.

HyDE (*Hypothetical Document Embeddings*). Técnica em que o modelo gera uma resposta hipotética para a consulta, usando-a como alvo de busca vetorial. A resposta hipotética tende a ser semanticamente mais próxima dos documentos reais que a pergunta original.

Inferência do LLM. Processo de geração de resposta pelo modelo a partir do *prompt* estruturado (instrução + contexto recuperado + consulta do usuário).

Metadados (em RAG). Informações adicionais associadas a cada *chunk*: identificador único (UUID), documento fonte, localização, data, classificação de segurança, responsável e palavras-chave.

***Multi-Query Retrieval*.** Técnica em que o sistema gera múltiplas variações da consulta original para ampliar a cobertura na base vetorial, aumentando a probabilidade de recuperar documentos relevantes.

Naive RAG. Implementação básica de RAG com recuperação direta por similaridade vetorial, sem técnicas de expansão de consulta ou reranking. Ponto de partida arquitetural para avaliação de melhorias.

Orquestrador. Componente do sistema RAG que coordena a interação entre todos os demais: valida a entrada, invoca o pipeline de recuperação, constrói o *prompt* e pós-processa a saída do LLM.

Pipeline de indexação. Fluxo *offline* que processa documentos fonte (carregamento, limpeza, *chunking*, geração de *embeddings*, armazenamento vetorial) para preparar a base de conhecimento.

***Query rewriting* (reescrita de consulta).** Transformação automática de perguntas mal estruturadas ou informais em termos técnicos alinhados ao vocabulário do corpus documental.

***Query routing* (roteamento de consulta).** Classificação da intenção do usuário antes da busca, determinando se deve usar RAG, consultar uma API externa, responder com conhecimento geral ou encaminhar para atendimento humano.

RAG (*Retrieval-Augmented Generation*). Arquitetura que combina um mecanismo de recuperação de documentos com um LLM generativo, fundamentando as respostas em documentos institucionais verificáveis e atualizáveis sem re-treinamento do modelo.

RAG multimodal. Extensão da arquitetura RAG para processar documentos com conteúdo não textual (tabelas, gráficos, formulários, diagramas) usando modelos de visão-linguagem (VLMs).

Re-*ranking*. Etapa adicional de reordenação dos resultados da busca inicial, usando um modelo *cross-encoder* que avalia cada par consulta-documento de forma mais precisa.

***Semantic drift* (degradação semântica).** Degradação gradual da relevância dos *embeddings* ao longo do tempo, causada pela evolução de terminologia institucional e surgimento de novos conceitos não capturados no modelo de *embedding* original.

Similaridade por cosseno. Métrica de distância entre *embeddings* que mede o ângulo entre vetores, ignorando sua magnitude. Escolha mais comum em aplicações de linguagem natural.

***System prompt*.** Instrução fornecida ao LLM antes da consulta do usuário, definindo diretrizes de comportamento, restrições e o papel do modelo no sistema.

8. Métricas de avaliação

***Answer Relevancy* (relevância da resposta).** Mede o grau em que a resposta gerada aborda diretamente a consulta do usuário.

AUC-ROC. Área sob a curva ROC (*Receiver Operating Characteristic*). Mede a capacidade discriminativa de um classificador binário independente do limiar de decisão.

BERTScore. Métrica semântica que compara texto gerado e referência por similaridade entre *embeddings* contextuais, capturando equivalência semântica além da sobreposição superficial de n-gramas.

BLEU. Métrica de sobreposição de n-gramas entre texto gerado e texto de referência, medindo precisão. Usada em tradução automática e sumarização com gabarito. Não captura equivalência semântica.

Conjunto *golden* (*golden test set*). Conjunto de pares consulta-resposta esperada estável ao longo do tempo, reexecutado a cada release e em ciclos periódicos para detecção de regressões de qualidade.

***Context Precision* (precisão do contexto).** Proporção do contexto recuperado que foi efetivamente útil para gerar a resposta.

***Context Recall* (cobertura do contexto)**. Grau em que o contexto recuperado cobre todas as informações necessárias para responder adequadamente à consulta.

F1-Score. Média harmônica entre precisão e revocação. Métrica padrão para avaliação de modelos de classificação com classes desbalanceadas.

***Faithfulness* (fundamentação).** Mede o grau em que uma resposta gerada é suportada pelo contexto recuperado, sem adicionar informação ausente nos documentos. Principal indicador de alucinação.

***Groundedness* / Aderência ao *prompt*.** Grau em que a resposta gerada é fundamentada no conteúdo fornecido no *prompt*, sem acrescentar informação ausente. Distinto de *faithfulness*: mede aderência ao contexto da conversa, não a documentos recuperados.

***Hit Rate* (taxa de acerto).** Percentual de consultas que têm ao menos um resultado relevante entre os k primeiros retornados.

LLM-as-a-Judge. Paradigma de avaliação em que um modelo LLM avaliador pontua respostas em dimensões como utilidade, segurança e correção factual. O avaliador deve ser distinto do modelo avaliado. Requer calibração com avaliadores humanos antes do uso em produção.

mAP / IoU. *Mean Average Precision e Intersection over Union*. Métricas padrão para avaliação de modelos de detecção de objetos e segmentação em visão computacional.

***Model drift* (deriva comportamental do modelo).** Mudança no comportamento do sistema LLM causada por atualização silenciosa do modelo-base pelo fornecedor, sem alteração do system *prompt*. Distinto de *concept drift* e *data drift*.

MRR (*Mean Reciprocal Rank*). Média do inverso da posição do primeiro resultado relevante para cada consulta. Mede a rapidez em encontrar informação relevante.

Observabilidade semântica. Capacidade de inspecionar não apenas métricas operacionais (latência, erro, custo), mas as decisões linguísticas e arquiteturais que produziram uma resposta: intenção classificada, rota escolhida, versão do *prompt, guardrails* ativados, motivo de *fallback*.

Perplexidade (*Perplexity*). Medida de quão bem o modelo prevê uma sequência de texto. Valores menores indicam texto mais previsível e fluente. Não implica correção factual; usada na avaliação do modelo base durante seleção e pós-*fine-tuning*.

***Precision*@k.** Proporção de resultados relevantes entre os k primeiros retornados pelo sistema de recuperação.

RAGAS. Framework de avaliação automatizada de sistemas RAG que mede *faithfulness, answer relevancy, context precision e context recall* por análise de sobreposição semântica.

***Recall*@k.** Proporção de todos os resultados relevantes existentes que estão presentes entre os k primeiros retornados.

RMSE / MAE. *Root Mean Square Error e Mean Absolute Error*. Métricas de erro para modelos de regressão, medindo a distância média entre previsão e valor real.

ROUGE. Métrica de sobreposição de n-gramas entre texto gerado e referência, medindo revocação. ROUGE-L avalia a maior subsequência comum. Complementar ao BLEU em tarefas de sumarização.

***Shadow evaluation* (avaliação em modo sombra).** Avaliação automática de amostras de consultas reais de produção por LLM-*as-a-Judge* ou classificadores, sem intervenção do usuário, alimentando *dashboards* de qualidade em tempo real.

Taxa de escalação para humanos. Percentual de interações transferidas para atendimento humano. Sinal de limitação do sistema ou de consultas fora do escopo.

Taxa de reformulação. Percentual de consultas seguidas por uma reformulação similar do mesmo usuário em curto intervalo, indicador indireto de insatisfação com a resposta.

TTFT (*Time to First Token*). Tempo decorrido entre o envio da requisição e a geração do primeiro *token* de resposta. Métrica operacional crítica para experiência do usuário em sistemas de atendimento ao cidadão.

9. Segurança, privacidade e governança

***Adversarial* ML.** Categoria de riscos de segurança cibernética específica de sistemas de aprendizado de máquina, incluindo envenenamento de dados, ataques de evasão, inversão de modelo e extração. Referenciada na fase de Avaliação de Riscos da Metodologia.

***Adversarial poisoning* (envenenamento da base de conhecimento).** Inserção deliberada de documentos falsos ou enganosos na base de conhecimento de um sistema RAG por ator malicioso, para que o sistema gere respostas incorretas.

Anonimização. Processo irreversível de transformação de dados pessoais para que não permitam identificação dos titulares. Técnicas incluem k-anonimato, l-diversidade e t-*closeness*.

Ataques de inversão de *embeddings*. Técnica *adversarial* em que um atacante tenta reconstruir documentos originais a partir dos vetores de *embeddings* armazenados na base vetorial.

Backdoor em modelos pré-treinados. Comprometimento de pesos de modelos obtidos de repositórios públicos, ativados por entradas específicas (backdoor triggers) para produzir comportamentos maliciosos em produção.

CVSSv3. Common Vulnerability Scoring System versão 3. Sistema padronizado de pontuação de severidade de vulnerabilidades (crítico, alto, médio, baixo), usado para definir SLAs de remediação pós-deploy.

CycloneDX. Padrão aberto de Software Bill of Materials (SBOM) mantido pela OWASP. Exigido como formato mandatório na Metodologia para garantir visibilidade de componentes, rastreabilidade de vulnerabilidades e conformidade com cadeia de suprimentos de software.

***Data masking / Obfuscation*.** Aplicação de técnicas de ocultação de dados sensíveis (PII, PHI) ao criar conjuntos de dados em ambientes não produtivos.

***Defense in depth* (defesa em profundidade).** Princípio de segurança em que múltiplas camadas de controle independentes são combinadas, de modo que a falha de uma camada não comprometa o sistema inteiro.

***Due diligence* de segurança.** Avaliação sistemática da postura de segurança de fornecedores, bibliotecas e componentes externos antes de sua adoção.

Envenenamento de dados (*data poisoning*). Inserção de dados maliciosos no conjunto de treinamento ou na base de conhecimento para degradar ou manipular o comportamento do modelo.

Fairness audit (auditoria de viés/equidade). Auditoria metodológica do modelo para verificar ausência de discriminação por grupos populacionais. Deve ser realizada antes da implantação e periodicamente em produção, com metodologia documentada: grupos avaliados, métricas de paridade demográfica e igualdade de oportunidade.

Injeção de *prompt* (*prompt injection*). Ataque em que instruções maliciosas são inseridas na entrada do modelo, diretamente pelo usuário ou indiretamente via documentos indexados, para subverter seu comportamento.

***Indirect prompt injection* (injeção indireta de *prompt*).** Variante de *prompt injection* em que as instruções maliciosas estão embutidas em documentos indexados, não na consulta direta do usuário.

Jailbreak — *ver seção 5.*

***Membership inference attack*.** Ataque em que adversário determina se um dado específico foi utilizado no treinamento do modelo, explorando diferenças de comportamento entre dados vistos e não vistos.

***Model stealing / extraction* (extração de modelo).** Ataque em que adversário reconstrói o comportamento do modelo por meio de consultas massivas à API, obtendo aproximação funcional sem acesso direto aos pesos.

Nuvem de governo. Infraestrutura de nuvem com armazenamento físico em solo brasileiro. Dados sensíveis e restritos do setor público devem ser hospedados nessa modalidade para garantir soberania operacional, conforme requisito da Metodologia

Pseudonimização. Substituição de identificadores diretos de dados pessoais por pseudônimos, mantendo a possibilidade de re-identificação com uso de chave adicional.

RBAC (*Role-Based Access Control*). Controle de acesso baseado em funções, em que permissões são concedidas a papéis organizacionais, não diretamente a usuários individuais.

RTO / RPO (*Recovery Time / Point Objective*). Objetivos de recuperação de desastre: RTO define o tempo máximo tolerável de indisponibilidade; RPO define o volume máximo tolerável de perda de dados. Requisito para backups de modelos *fine-tuned* auto-hospedados.

SAST (*Static Application Security Testing*). Análise estática de segurança de código-fonte, integrada ao pipeline de CI/CD. Requisito obrigatório a cada *commit* em sistemas LLM.

SCA (*Software Composition Analysis*). Varredura automatizada de composição de software para identificar vulnerabilidades em bibliotecas e dependências externas. Requisito obrigatório junto ao SAST.

Segurança por design (*security by design*). Abordagem em que controles de segurança são incorporados desde a concepção da arquitetura, não adicionados retroativamente.

Soberania digital. Princípio de que dados sensíveis ou sigilosos devem ser processados e armazenados em infraestruturas nacionais, preferencialmente a nuvem de governo.

STRIDE. Metodologia de modelagem de ameaças que categoriza riscos em seis classes: *Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service e Elevation of Privilege*. Adaptada nos anexos técnicos para ameaças específicas de LLM e RAG.

TLS (*Transport Layer Security*). Protocolo criptográfico para comunicação segura entre componentes do sistema. TLS 1.2 como mínimo obrigatório, com preferência por TLS 1.3.

***Vault* / HSM (*Hardware Security Module*).** Cofres de credenciais para gestão segura de *tokens* de API e chaves criptográficas. Uso obrigatório para credenciais de acesso a modelos; vedado armazenamento em código-fonte ou arquivos de configuração.

10. Termos técnicos relacionados à base de conhecimento [10]

Termo	Definição
Article Confidence Score	Índice composto que combina recência, volume de visualizações, avaliação positiva e taxa de abandono para estimar a confiabilidade de um artigo.
BM25	Algoritmo de ranqueamento de relevância em recuperação de texto completo, variante ponderada de TF-IDF que normaliza por comprimento do documento.
Card sorting	Técnica de design de taxonomia em que usuários agrupam itens conceituais para revelar modelos mentais de classificação.
Conteúdo canônico	Versão oficial e autoritativa de um tópico, da qual outras referências derivam sem duplicar. Implementa o princípio SSOT.
Coreference resolution	Processo de resolução de referências pronominais e anafóricas em texto, necessário para que modelos de linguagem mantenham o contexto ao consumir trechos isolados.
Data poisoning	Ingestão deliberada ou acidental de conteúdo incorreto ou tendencioso que degrada a qualidade da base e dos sistemas que a consomem.
Deflection rate	Taxa de sessões na base de conhecimento que resultam em resolução sem abertura de chamado de suporte.
DITA	*Darwin Information Typing Architecture*. Arquitetura de conteúdo estruturado que organiza informação em tópicos reutilizáveis e tipados (*concept, task, reference, troubleshooting*).
Domain Owner	Responsável pela propriedade e qualidade do conteúdo de um domínio temático específico da base.

Termo	Definição
Dublin Core	Conjunto de 15 elementos de metadados interoperáveis mantido pelo DCMI, base para descrição de recursos digitais com suporte a *Linked* Data.
Findability rate	Taxa de sessões de busca em que o usuário localiza e seleciona um resultado pertinente sem reformular a consulta.
Folksonomia	Sistema de classificação por *tags* geradas livremente pelos usuários, sem hierarquia formal ou controle de vocabulário.
Freshness boost	Ajuste de ranqueamento de busca que favorece artigos mais recentes ou recém-revisados por função de decaimento temporal.
Grounding rate	Proporção de respostas geradas por um sistema de IA que citam ao menos um artigo da base de conhecimento como fonte de fundamentação.
KCS	*Knowledge-Centered Service*. Metodologia operacional que integra captura e manutenção de conhecimento ao fluxo de resolução de demandas de atendimento.
Knowledge graph	Implementação em banco de grafos de uma ontologia formal, representando entidades como nós e relações como arestas rotuladas.
KMS	*Knowledge Management System*. Ecossistema de processos, práticas e plataformas que suporta criação, compartilhamento, uso e gestão do conhecimento organizacional. A base de conhecimento é a camada de dados do KMS.
Ontologia	Modelo formal que descreve conceitos, atributos e relações de um domínio com restrições lógicas, permitindo inferência automatizada. Implementada tipicamente em OWL sobre RDF.
OWL	*Web Ontology Language*. Linguagem de ontologia do W3C baseada em lógica de descrição, que permite definir restrições e inferência formal sobre grafos RDF.
PROV-DM	*Provenance Data Model*. Padrão W3C para representação interoperável de proveniência: quem produziu o quê, quando e em que contexto.
RAG	*Retrieval-Augmented Generation*. Arquitetura de IA que recupera trechos relevantes de uma base de conhecimento para fundamentar a geração de respostas por um modelo de linguagem.
RDF	*Resource Description Framework*. Modelo de dados do W3C para representar afirmações como triplas sujeito-predicado-objeto.
RBAC	Role-Based Access Control. Modelo de controle de acesso em que permissões são atribuídas a papéis, e papéis são atribuídos a usuários.
SECI	Modelo de conversão do conhecimento de Nonaka e Takeuchi: Socialização, Externalização, Combinação e Internalização.
SHACL	*Shapes Constraint Language*. Padrão W3C para validação de grafos RDF contra restrições estruturais definidas.
SKOS	*Simple Knowledge Organization System*. Padrão W3C para publicação e mapeamento de vocabulários controlados, taxonomias e tesauros como Linked Data.
SSOT	*Single Source of Truth*. Princípio arquitetural em que cada fato ou instrução reside em um único local canônico, eliminando inconsistências por duplicação.
Snapshot versioning	Estratégia de versionamento que congela e arquiva cada versão de um artefato com identificador cronológico, preservando o histórico sem deletar versões anteriores.

Termo	Definição
Tesauro	Vocabulário controlado que vai além da taxonomia hierárquica ao mapear relações de equivalência (sinônimos e termos preferidos), hierarquia e associação entre conceitos. Padrão ISO 25964.
Topic Maps	Padrão ISO/IEC 13250 que separa a camada de conhecimento lógico da camada de recursos físicos, permitindo sobrepor uma rede semântica a repositórios não estruturados sem modificar os dados brutos.
Tree testing	Técnica de validação de taxonomia em que usuários navegam por uma estrutura hierárquica para localizar itens específicos, revelando falhas de organização sem o viés visual do design da interface.
Zero-result rate	Taxa de consultas de busca que não retornam nenhum resultado. Indicador primário de lacunas de cobertura ou desalinhamento de vocabulário.

11. Normas, padrões e frameworks

ABNT NBR ISO/IEC 22989:2023. Define vocabulário e conceitos de IA, incluindo a denominação "Produtor de IA" para o executor do desenvolvimento.

ABNT NBR ISO/IEC 25010. Modelo de qualidade de produto de software aplicável a sistemas de IA, abordando adequação funcional, confiabilidade, segurança, manutenibilidade e outros atributos.

ABNT NBR ISO/IEC 27001 (SGSI). Norma para Sistema de Gestão de Segurança da Informação.

ABNT NBR ISO/IEC 27701 (SGIP). Extensão da ISO 27001 voltada ao Sistema de Gestão de Informações de Privacidade, relevante para conformidade com LGPD.

ABNT NBR ISO/IEC 42001:2024 (SGIA). Especifica requisitos para estabelecer, implementar, manter e melhorar continuamente um Sistema de Gestão de Inteligência Artificial.

CIS *Controls*. *Center for Internet Security Controls* — controles de segurança da informação com *Privacy Companion Guide* para alinhamento com princípios de privacidade.

DAMA-DMBOK. *Data Management Body of Knowledge — framework* abrangente para gestão de dados, cobrindo governança, qualidade, arquitetura, segurança e metadados.

Decreto nº 7.845/2012. Regulamenta o tratamento de informações sigilosas no âmbito da Administração Pública Federal.

ePING. Padrões de Interoperabilidade de Governo Eletrônico — premissas, políticas e especificações técnicas para uso de TIC no governo federal.

***FinOps Framework*.** Framework para gestão financeira de nuvem com foco em visibilidade de custos em tempo real e otimização contínua de recursos.

GSI/PR. Gabinete de Segurança Institucional da Presidência da República — emite diretrizes de segurança da informação e cibernética aplicáveis ao uso de tecnologias no setor público.

ISO/IEC 23894. Norma de orientações sobre gestão de riscos de IA (identificação, análise e mitigação). Cobre todo o ciclo de vida do sistema. Citada na Metodologia e no Anexo RAG sem entrada no glossário.

ISO/IEC 42005:2025. Norma de Avaliação de Impacto de Sistemas de IA. Referenciada no Anexo RAG como padrão para análise de impacto, complementar ao NIST AI RMF.

ISO/IEC 5338. Define processos de ciclo de vida para sistemas de IA, cobrindo planejamento, desenvolvimento, implantação, operação, manutenção e descontinuação.

MITRE ATLAS. *Adversarial Threat Landscape for AI Systems* — catálogo de táticas e técnicas de ataques específicos a sistemas de IA.

NIST AI RMF. *Framework* voluntário estruturado em quatro funções (*Govern, Map, Measure, Manage)* para incorporar confiabilidade e gestão de riscos em sistemas de IA.

OWASP Top 10 for LLM *Applications*. Lista das dez principais vulnerabilidades em aplicações baseadas em modelos de linguagem, incluindo injeção de *prompt*, envenenamento de dados e vazamento de contexto. Métricas de sucesso e KPIs.

12. Referências bibliográficas

[1] BRASIL. Ministério da Gestão e da Inovação em Serviços Públicos. Governo Digital. *Cartilha de governança de dados*. Disponível em: https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/governancadedados/glossario-de-termos-de-dados. Acesso em: 08 abr. 2026.

[2] BRASIL. Ministério da Gestão e da Inovação em Serviços Públicos. Governo Digital. *Glossário de termos de dados*. Disponível em: https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/governancadedados/arquivos/CartilhaGovDadosvol3.pdf. Acesso em: 08 abr. 2026.

[3] OWASP Foundation. Threat Modeling. OWASP Community – The Open Web Application Security Project. Disponível em: https://owasp.org/www-community/Threat_Modeling. Acesso em: 10 dez. 2025.

[4] MITRE. AI Security 101. ATLAS – Adversarial Threat Landscape for Artificial-Intelligence Systems. Disponível em: https://atlas.mitre.org/resources/ai-security-101. Acesso em: 10 dez. 2025.

[5] DAMA International. *DAMA-DMBOK: Data Management Body of Knowledge*. 2. ed. Bradley Beach, NJ: Technics Publications, 2017.

[6] National Cyber Security Centre (NCSC). *Guidelines for Secure AI System Development: Secure Design.* NCSC – National Cyber Security Centre, 2023. Disponível em: https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development/guidelines/secure-design. Acesso em: 11 dez. 2025.

[7] **CPQD. Metodologia de Desenvolvimento de Soluções de IA:** *Guia Unificado de Inteligência Artificial (GuIA).* Frente M3.2 – Subfrente M3.2.1. Meta 3 – Projeto INSPIRE. Convênio nº 01.25.0728.00. MGI/Finep. Campinas-SP, 2025.

[8] **CPQD. Metodologia de Desenvolvimento de Soluções de IA:** *Anexo técnico – Implementação de IA com Geração Aumentada por Recuperação (RAG)***.** Frente M3.2 – Subfrente M3.2.1. Meta 3 – Projeto INSPIRE. Convênio nº 01.25.0728.00. MGI/Finep. Campinas-SP, 2026.

[9] **CPQD. Metodologia de Desenvolvimento de Soluções de IA:** *Anexo técnico – Implementação de IA com grandes modelos de linguagem (LLM - Large Language Models)***.** Frente M3.2 – Subfrente M3.2.1. Meta 3 – Projeto INSPIRE. Convênio nº 01.25.0728.00. MGI/Finep. Campinas-SP, 2026.

[10] **CPQD. Metodologia de Desenvolvimento de Soluções de IA:** *Anexo técnico – Base de conhecimento***.** Frente M3.2 – Subfrente M3.2.1. Meta 3 – Projeto INSPIRE. Convênio nº 01.25.0728.00. MGI/Finep. Campinas-SP, 2026.