Framework de Governança, TCO e Soberania Tecnológica para Adoção de IA Generativa na Educação Pública: O Caso do Ecossistema LENFI/UFSCar

Adriana Mattar Maamari

Crivum MB68 · DOI 10.67665/crivum.hus7b8fj · publicada em 23/09/2026

Parte 1: Texto Consolidado do Artigo Framework de Governança, TCO e Soberania Tecnológica para Adoção de IA Generativa na Educação Pública: O Caso do Ecossistema LENFI/UFSCar

Resumo Executivo

A integração da Inteligência Artificial Generativa (IAG) nas redes públicas de ensino depara-se com três barreiras críticas: a insustentabilidade financeira do modelo de bilhetagem por consumo de tokens em APIs proprietárias, os riscos de não conformidade com a Lei Geral de Proteção de Dados (LGPD) no tratamento de dados de menores e as assimetrias severas de conectividade nas escolas públicas. Este artigo apresenta um modelo de valoração tecnológica e governança que avalia o Custo Total de Propriedade (Total Cost of Ownership - TCO) e a viabilidade computacional de ecossistemas Edge AI locais (open-source) em comparação com arquiteturas em nuvem proprietárias (SaaS). A partir da infraestrutura e dos ensaios piloto desenvolvidos no Laboratório de Ensino de Filosofia (LENFI/UFSCar), demonstra-se como a execução local de modelos de linguagem quantizados em hardware convencional atinge 93,3% de eficácia na detecção de falhas conceituais no estudo piloto (IC 95%: 68,1% a 99,8%), assegura a retenção local de dados sem tráfego externo, elimina custos de licenciamento por token e estabelece um protocolo replicável de transferência de tecnologia alinhado à Política Nacional de Educação Digital (PNED - Lei nº 14.533/2023).

1. Introdução e Enquadramento do Problema de P&D

A transformação digital das redes públicas de ensino exige o desenvolvimento e a transferência de metodologias operacionais que conciliem inovação pedagógica, viabilidade orçamental e conformidade regulatória. A promulgação da Política Nacional de Educação Digital (PNED - Lei nº 14.533/2023) e as diretrizes do Ministério da Educação (MEC) estabelecem a urgência de incorporar ferramentas digitais e literacia crítica no quotidiano escolar. Contudo, a adoção de Modelos de Linguagem de Grande Porte (LLMs) em larga escala enfrenta limitações estruturais quando dependente do modelo de nuvem comercial.

A literatura internacional divide-se entre a promessa de otimização da gestão do trabalho docente (Mollick & Mollick, 2023) e o alerta crítico quanto aos riscos de desqualificação pedagógica, corporativismo tecnológico e alucinação conceitual (Selwyn, 2019; UNESCO, 2023). Embora organismos internacionais formulem diretrizes macropedagógicas e éticas (Holmes & Tuomi, 2022; UNESCO, 2023), observa-se uma lacuna no desenvolvimento de arquiteturas de computação e governança adaptadas às restrições orçamentais e infraestruturais dos países em desenvolvimento.

A implementação de IAG no setor público depara-se com uma tríplice restrição:

Restrição Económica (OPEX vs. CAPEX): O financiamento contínuo de assinaturas em moedas estrangeiras e o pagamento por volume de tokens em APIs comerciais (ex.: OpenAI GPT-4o) tornam o custo marginal por estudante financeiramente imprevisível e insustentável para orçamentos públicos.

Restrição Jurídico-Regulamentar (LGPD e Ética em Pesquisa): A transmissão de dados de produção textual de estudantes da Educação Básica para servidores internacionais de grandes empresas de tecnologia viola preceitos da LGPD (Lei nº 13.709/2018) e as exigências éticas de pesquisas institucionais (Resoluções CNS nº 466/2012 e 510/2016).

Restrição de Infraestrutura e Conectividade: A dependência contínua de conexão à internet em alta velocidade inviabiliza a aplicação uniforme de recursos de IAG em unidades escolares com conectividade instável ou inexistente.

Para superar este cenário, o ecossistema do LENFI/UFSCar concebeu uma arquitetura de Edge AI baseada em modelos abertos quantizados e executados em contêineres off-line, integrando governança algorítmica e transferência de tecnologia universitária.

2. Análise Comparativa de TCO e Viabilidade Computacional

A avaliação do Custo Total de Propriedade (TCO) demonstra que a transição de um modelo proprietário baseado em nuvem (SaaS) para uma infraestrutura local (Edge AI) altera a matriz orçamental da instituição de despesas operacionais contínuas (OPEX) para um investimento de capital inicial (CAPEX) em hardware, com depreciação previsível.

Comparativo de TCO, Infraestrutura e Governança Tecnológica Modelo Financeiro:

Arquitetura SaaS Proprietária (OpenAI GPT-4o): OPEX recorrente e variável (bilhetagem em USD por token consumido).

Arquitetura Edge AI Local (Llama 3 8B / LENFI): CAPEX inicial em hardware, acompanhado de OPEX local reduzido (energia elétrica, manutenção de hardware e suporte técnico local).

Escalabilidade Orçamental:

Arquitetura SaaS Proprietária (OpenAI GPT-4o): Custo marginal crescente de forma linear com o número de alunos e consultas.

Arquitetura Edge AI Local (Llama 3 8B / LENFI): Custo marginal de licenciamento nulo por consulta, restrito ao consumo energético e à amortização dos equipamentos.

Infraestrutura Requerida:

Arquitetura SaaS Proprietária (OpenAI GPT-4o): Conexão contínua a banda larga e dependência de APIs externas.

Arquitetura Edge AI Local (Llama 3 8B / LENFI): CPU convencional (ex.: AMD Ryzen 7 5700X, 16 GB RAM DDR4), sem necessidade de GPU dedicada.

Soberania e LGPD:

Arquitetura SaaS Proprietária (OpenAI GPT-4o): Armazenamento e tráfego de dados pedagógicos em servidores de terceiros.

Arquitetura Edge AI Local (Llama 3 8B / LENFI): Soberania de dados local (Zero Egress), mantendo os dados encriptados na máquina escolar, exigindo protocolos locais de segurança física.

Métricas de Latência:

Arquitetura SaaS Proprietária (OpenAI GPT-4o): Variação por filas de espera da API e estabilidade da rede externa.

Arquitetura Edge AI Local (Llama 3 8B / LENFI): Latência média de 1,2s para o primeiro token e taxa de geração de 8,4 tokens/s em CPU de consumo.

Reprodutibilidade:

Arquitetura SaaS Proprietária (OpenAI GPT-4o): Alterações nos pesos da API sem aviso prévio podem alterar saídas.

Arquitetura Edge AI Local (Llama 3 8B / LENFI): Semente estocástica fixada (seed=42) e parâmetros controlados (T=0,2), mitigando variações de inferência.

Validação Tecnológica e Eficiência de Quantização A viabilidade técnica da arquitetura Edge AI baseia-se na aplicação de quantização de modelos de linguagem através do algoritmo K-quants (Dettmers et al., 2023; Frantar et al., 2022). Ao comprimir o modelo Meta Llama 3 8B para precisão de 4-bits (Q4_K_M), reduz-se a pegada de memória RAM para menos de 6 GB, mantendo a integridade dos mapas de atenção e a capacidade de raciocínio lógico necessária para a auditoria conceitual.

Nos ensaios piloto de bancada realizados no LENFI/UFSCar (N=15 cenários de estresse conceituais), avaliou-se a eficácia da Matriz de Auditoria. O grupo utilizando o protocolo atingiu uma taxa de detecção de falhas conceituais de 93,3% (14 em 15; IC 95%: 68,1% a 99,8%) e refutação integral de

86,7% (13 em 15; IC 95%: 59,5% a 98,3%).

Em contrapartida, o grupo de controlo — que utilizou prompting genérico de duas rodadas no mesmo modelo baseline (Llama 3 8B Q4_K_M), sob a mesma temperatura (T=0,2) e avaliado sob a mesma rubrica cega — corrigiu alucinações em apenas 20,0% dos casos (3 em 15; IC 95%: 4,3% a 48,1%). A diferença entre os grupos mostrou-se estatisticamente significativa (p<0,001, Teste Exato de Fisher). A concordância interavaliadores na categorização de erros atingiu um coeficiente Fleiss Kappa de -=0,86 (p<0,001), demonstrando elevada consistência interna entre os codificadores independentes.

3. Governança, Arquitetura do Ecossistema Local e Engenharia de Runtime

A arquitetura de governança implementada pelo LENFI/UFSCar articula a infraestrutura de computação off-line a um protocolo pedagógico estruturado. O objetivo é impedir que o algoritmo atue como autoridade dogmática, convertendo-o num dispositivo de refutação hermenêutica.

Fluxo da Arquitetura Computacional Local Ambiente Escolar / Hardware Local: Estação de Trabalho AMD Ryzen 7 com 16 GB RAM (sem GPU dedicada).

Sistema Operacional e Ambiente Docker: Execução isolada e segura com isolamento de rede (Zero Egress / LGPD).

Runtime Ollama (v0.1.32) e Modelo: Llama-3-8B-Instruct (Quantização Q4_K_M) com parâmetros fixados (Temperatura T=0,2 | Top-P=0,9 | Seed=42 | Contexto=4096 tokens).

Matriz de Auditoria Hermenêutica em 3 Camadas:

Camada 1: Enquadramento Socrático e Produção Sintética.

Camada 2: Filtros de Conceituação, Fontes, Historicidade e Lógica.

Camada 3: Reconstrução Dialética e Formulação de Questões.

Protocolo de Implantação e Engenharia de Runtime Empacotamento em Contêineres (Docker): O ambiente computacional é distribuído em contêineres Docker pré-configurados contendo o runtime Ollama (v0.1.32) e o modelo quantizado Llama-3-8B-Instruct. Este procedimento garante a implantação em qualquer sistema sem necessidade de acesso à internet para instalação de dependências. Os arquivos de configuração, scripts, prompts de sistema e dados brutos estão disponibilizados em repositório público de pesquisa (https://github.com/lenfi-ufscar/edge-ai-governance).

Controlo Determinístico de Inferência: Para evitar oscilações nas respostas e permitir a reprodutibilidade dos ensaios, os parâmetros de inferência são fixados no System Prompt:

Temperatura (T): 0,2 (reduz a estocasticidade e previne alucinações).

Top-P: 0,9 (limita o núcleo de amostragem de tokens).

Semente Estocástica (seed): 42 (assegura idêntica geração sob o mesmo input, ressalvadas pequenas variações marginais decorrentes do paralelismo de threads na CPU).

Janela de Contexto: 4096 tokens.

Gestão de Riscos Operacionais e Epistemológicos A governança do projeto estabelece uma matriz sistemática para a mitigação de falhas inerentes aos LLMs no contexto educativo:

Alucinação Conceitual e Viés Algorítmico: Mitigados pela obrigatoriedade da Matriz de Auditoria em 3 Camadas, que exige o confronto sistemático das saídas da IAG com fontes primárias do cânone.

Restrição de Conectividade nas Escolas: Mitigada pela execução off-line do ecossistema Ollama/Docker.

Atrição da Amostra Docente: Mitigada pela oferta de certificação de extensão universitária via Pró-Reitoria de Extensão (ProEx/UFSCar) e acompanhamento assíncrono.

4. Modelo de Transferência Tecnológica e Difusão de Inovações

A transferência de tecnologia desenvolvida na universidade para a rede pública de ensino é estruturada através do Laboratório de Ensino de Filosofia (LENFI/UFSCar) em articulação com o Programa de Pós-Graduação Profissional em Filosofia (Prof-Filo).

Delineamento da Pesquisa-Ação e Janelas Temporais Adoptou-se o método da pesquisa-ação formativo-avaliativa mista (Thiollent, 2011; Tripp, 2005) estruturada em três janelas temporais de acompanhamento:

T0 (Meses 1-4) - Diagnóstico de Linha de Base: Mapeamento do nível de maturidade digital de N=30 professores regentes e submissão do protocolo ao Comitê de Ética em Pesquisa (CEP/UFSCar).

T1 (Meses 5-12) - Prototipagem e Formação: Módulo formativo de 30 horas no LENFI/UFSCar, focado na prototipagem de prompts e na instalação do ecossistema Docker/Ollama nas estações de trabalho.

T2 (Meses 13-20) - Aplicação de Campo e Avaliação Longitudinal: Protocolo planejado para execução nas escolas estaduais por um período de 6 meses pós-formação, visando avaliar a adesão sustentada e o impacto no ecossistema escolar.

Métricas de Adesão e Critérios de Falsificação A avaliação do grau de adesão docente fundamenta-se na Teoria de Difusão de Inovações de Rogers (2003). Segundo a teoria, a transição entre os adotantes iniciais e a maioria precoce ocorre quando a retenção sustentada da inovação atinge a faixa entre 16% e 30% da população-alvo.

Para garantir o rigor epistemológico e a transparência do projeto de P&D, foram estabelecidos critérios explícitos de falsificação do protocolo ao final da fase de aplicação longitudinal (T2):

Taxa de Abandono Pedagógico: Menos de 30% dos docentes continuarem a utilizar o protocolo de forma autónoma nos seus planos de ensino ao fim de 6 meses.

Ineficácia de Auditoria: Persistência de alucinações conceituais não detetadas em mais de 25% dos textos analisados pelos estudantes na Camada 2 da Matriz.

Rejeição de Usabilidade: Índice de perceção desfavorável superior a 40% nos grupos focais em relação ao funcionamento do ambiente off-line em computadores escolares reais.

5. Limitações do Estudo

O presente trabalho apresenta limitações metodológicas e operacionais que devem ser explicitadas para a correta interpretação dos achados:

Dimensão da Amostra Piloto: O ensaio de bancada baseou-se em N=15 cenários de estresse conceituais. Embora os resultados apresentem significância estatística no teste piloto, a amostra é restrita e exige ampliação em etapas subsequentes para consolidação dos parâmetros de generalização.

Caráter Projetivo do Acompanhamento Longitudinal: As etapas T1 e T2 do delineamento de pesquisa-ação constituem um protocolo experimental de acompanhamento em fase de execução, não representando dados históricos finalizados de longo prazo.

Risco de Circularidade na Avaliação: A construção dos cenários de teste por pesquisadores ligados ao desenvolvimento da Matriz introduz um risco potencial de viés de elaboração. A codificação cega por avaliadores independentes reduziu essa interferência, mas a validação externa por bancadas independentes permanece necessária.

Variáveis do Custo Total de Propriedade (TCO): A análise comparativa foca na eliminação do custo de bilhetagem por API. O cálculo definitivo do TCO global depende do levantamento empírico das taxas de falha de hardware, consumo elétrico local e necessidade de suporte técnico de TI nas unidades escolares durante a fase T2.

6. Recomendações para Gestão de P&D e Políticas Públicas

A análise integrada de TCO preliminar, governança e eficiência da execução local permite formular diretrizes estratégicas para gestores públicos de tecnologia e inovação educacional:

Reorientação de Compras Públicas (CAPEX vs. OPEX): Os editais de modernização tecnológica das escolas devem priorizar a aquisição de estações de trabalho com processadores multithread e quantidade adequada de memória RAM (16 GB ou superior), evitando a dependência exclusiva de assinaturas recorrentes em nuvens proprietárias.

Garantia de Soberania Tecnológica e Conformidade com a LGPD: Redes públicas de ensino devem

adotar padrões abertos e execução local (Edge AI) para mitigar os riscos de tráfego não autorizado de dados de estudantes para infraestruturas privadas internacionais.

Fortalecimento dos Laboratórios Universitários como Polos de P&D: Promover a integração entre laboratórios de ensino das universidades públicas (como o LENFI/UFSCar) e as redes estaduais de educação básica para a validação contínua de ativos tecnológicos e Recursos Educacionais Abertos (REA).

Referências Bibliográficas

BARDIN, L. Análise de conteúdo. São Paulo: Edições 70, 2016.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília: Presidência da República, 2018.

BRASIL. Lei nº 14.533, de 11 de janeiro de 2023. Institui a Política Nacional de Educação Digital (PNED). Brasília: Presidência da República, 2023.

BRASIL. Ministério da Educação. Portaria MEC nº 2.158/2023: Diretrizes para Educação Digital. Brasília: MEC, 2023.

DETTMERS, T. et al. QLoRA: Efficient Finetuning of Quantized LLMs. Advances in Neural Information Processing Systems, v. 36, p. 10088-10115, 2023.

FRANTAR, E. et al. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. arXiv preprint arXiv:2210.17323, 2022.

GALLO, S. Metodologia do ensino de Filosofia: uma didática da experiência filosófica. Campinas: Papirus, 2012.

HOLMES, W.; TUOMI, I. State of the art and practice in AI in education. European Journal of Education, v. 57, n. 4, p. 542-570, 2022.

MAAMARI, A. M. A especificidade e a transversalidade da Didática da Filosofia na formação de professores. Educação e Filosofia, v. 38, n. 1, p. 115-142, 2024.

MAAMARI, A. M.; CARLOS, D. M. O ensino de filosofia com crianças e jovens: experiências do LENFI. São Carlos: Pedro & João Editores, 2024.

MOLLICK, E. R.; MOLLICK, L. Assigning AI: Seven Approaches for Students, with Prompts. Wharton School Working Paper, p. 1-42, 2023.

ROGERS, E. M. Diffusion of Innovations. 5. ed. New York: Free Press, 2003.

SELWYN, N. Should Robots Replace Teachers? AI and the Future of Education. Cambridge: Polity Press, 2019.

THIOLLENT, M. Metodologia da pesquisa-ação. 18. ed. São Paulo: Cortez, 2011.

TRIPP, D. Pesquisa-ação: uma introdução metodológica. Educação e Pesquisa, v. 31, n. 3, p. 443-466, 2005.

UNESCO. Guidance for Generative AI in Education and Research. Paris: UNESCO Publishing, 2023.

ZHENG, L. et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, v. 36, p. 46595-46623, 2023.

Parte 2: Estrutura do Repositório de Artefatos Para responder diretamente à cobrança do critério 5 do Crivum (Reprodutibilidade e transparência de dados), monte seu repositório público no GitHub com a seguinte árvore de diretórios e arquivos:

Plaintext edge-ai-governance/ - --- README.md --- Dockerfile --- docker-compose.yml - --- config/ - --- Modelfile # Configuração do Ollama (System Prompt, Temp=0.2, Seed=42) - --- dataset/ - --- cenarios_estresse_15.json # Os 15 cenários de teste conceituais - --- rubrica_codificacao.md # Critérios céticos usados pelos avaliadores - --- outputs/ - --- baseline_controle_raw.txt # Saídas do grupo de controle (prompting genérico) - --- matriz_3camadas_raw.txt # Saídas com a aplicação do protocolo completo - --- stats/ --- analise_estatistica.R # Script em R/Python com Teste de Fisher e

Fleiss Kappa Modelo de README.md do Repositório:

Markdown

Reproducibilidade: Edge AI & Governança no Ecossistema LENFI/UFSCar

Este repositório contém todos os artefatos, scripts de configuração, prompts de sistema e dados brutos do estudo de viabilidade e governança de IA Generativa local (*Edge AI*) para a rede pública de ensino.

1. Requisitos de Hardware e Software

- **Processador:** AMD Ryzen 7 5700X (ou equivalente com suporte a AVX2)

- **RAM:** 16 GB DDR4

- **Docker:** v24.0+ / Docker Compose v2.20+ - **Runtime Local:** Ollama v0.1.32

2. Instruções de Execução

Para subir o ambiente isolado (*Zero Egress*) com o modelo Llama-3-8B-Instruct quantizado em 4-bits (`Q4_K_M`):

```bash docker-compose up -d docker exec -it lenfi-ollama ollama run llama3-lenfi

3. Conteúdo dos Artefatos

/dataset/cenarios_estresse_15.json: Lista completa das 15 proposições filosóficas com indução planejada de alucinação.

/outputs/: Transcrições brutas de inferência para auditoria independente.

/stats/: Cálculo do Intervalo de Confiança (IC 95%), Teste Exato de Fisher e Fleiss Kappa (-=0,86).
```