

Soberania Digital e Aumento da Capacidade Estatal: Uma Arquitetura de IA Multiagente contra a Dependência Tecnológica de Mercado

Digital Sovereignty and State Capacity Building: A Multi-Agent AI Architecture against Market Technological Dependency

Soberanía Digital y Aumento de la Capacidad Estatal: Una Arquitectura de IA Multiagente contra la Dependencia Tecnológica de Mercado

Grupo temático GT10 - Pensamento Social Brasileiro, Reforma Administrativa e Inovação

Karina Machado da Silva, Universidade Federal de Santa Maria (UFSM)

Juliano Nunes Alves, Universidade Federal de Santa Maria (UFSM)

Breno Augusto Diniz Pereira, Universidade Federal de Santa Maria (UFSM)

Antonio Cesar Lopes Garcez Filho, Universidade Federal de Santa Maria (UFSM)

RESUMO

Este relato técnico apresenta uma experiência de inovação administrativa que ultrapassa as soluções tecnocráticas de mercado para o aumento da capacidade estatal. Diante da "fadiga decisória" dos servidores públicos e dos riscos de soberania das soluções comerciais de IA (SaaS proprietários), desenvolveu-se uma arquitetura de Inteligência Artificial Multiagente de código aberto (framework OpenClaw), operando em infraestrutura local (VPS) e integrada via WhatsApp desenvolvida no âmbito do Programa de Pós-Graduação em Gestão de Organizações Públicas da UFSM. A inovação central reside na segregação de funções cognitivas: agentes especializados (Bibliotecário, Redator e Revisor) colaboram para produzir notas técnicas baseadas em evidências, eliminando alucinações via RAG local e preservando a discricionariedade administrativa. O sistema custa menos de 10% das soluções corporativas, garante conformidade com a LGPD e cria uma capacidade dinâmica única (conforme Visão Baseada em Recursos), não replicável por atores privados. A arquitetura demonstra que a modernização administrativa efetiva exige soberania tecnológica: controle estatal dos dados, transparência algorítmica e humano no comando. Resultados preliminares indicam redução drástica no tempo de pesquisa legislativa e aumento da precisão factual, sem comprometer a segurança jurídica. O relato prova que inovação e soberania digital são compatíveis e economicamente viáveis, mesmo para municípios com restrições orçamentárias.

Palavras-chave: Inteligência Artificial no Setor Público; Soberania Digital; Capacidade Estatal; RAG Local; Inovação Governamental

ABSTRACT

This technical report presents an administrative innovation experience that surpasses market technocratic solutions for increasing state capacity. Facing the "decision fatigue" of public servants and the sovereignty risks of commercial AI solutions (proprietary SaaS), an open-source Multi-Agent Artificial Intelligence architecture (OpenClaw framework) was developed, operating on local infrastructure (VPS) and integrated via WhatsApp. The central innovation lies in the segregation of cognitive functions: specialized agents (Librarian, Writer, and Reviewer) collaborate to produce evidence-based technical notes, eliminating hallucinations via local RAG and preserving administrative discretion. The system costs less than 10% of corporate solutions, ensures LGPD compliance, and creates an idiosyncratic dynamic capability (according to the Resource-Based View), non-replicable by private actors. The architecture demonstrates that effective administrative modernization requires technological sovereignty: state control of data, algorithmic transparency, and human-in-the-loop. Preliminary results indicate a drastic reduction in legislative research time and increased factual precision without compromising legal security. The report contributes to the debate on democratic administrative reform, proving that innovation and digital sovereignty are compatible and economically viable, even for municipalities with budgetary constraints.



Keywords: Artificial Intelligence in Public Sector; Digital Sovereignty; State Capacity; Local RAG; Government Innovation.

RESUMEN

Este informe técnico presenta una experiencia de innovación administrativa que supera las soluciones tecnocráticas de mercado para el aumento de la capacidad estatal. Ante la "fatiga decisoria" de los servidores públicos y los riesgos de soberanía de las soluciones comerciales de IA (SaaS propietarios), se desarrolló una arquitectura de Inteligencia Artificial Multiagente de código abierto (marco OpenClaw), operando en infraestructura local (VPS) e integrada vía WhatsApp. La innovación central reside en la segregación de funciones cognitivas: agentes especializados (Bibliotecario, Redactor y Revisor) colaboran para producir notas técnicas basadas en evidencia, eliminando alucinaciones vía RAG local y preservando la discrecionalidad administrativa. El sistema cuesta menos del 10% de las soluciones corporativas, garantiza conformidad con la LGPD y crea una capacidad dinámica idiosincrásica (conforme a la Visión Basada en Recursos), no replicable por actores privados. La arquitectura demuestra que la modernización administrativa efectiva exige soberanía tecnológica: control estatal de los datos, transparencia algorítmica y humano al mando. Los resultados preliminares indican una reducción drástica en el tiempo de investigación legislativa y un aumento de la precisión fáctica, sin comprometer la seguridad jurídica. El informe contribuye al debate sobre la reforma administrativa democrática, probando que la innovación y la soberanía digital son compatibles y económicamente viables, incluso para municipios con restricciones presupuestarias.

Palabras clave: Inteligencia Artificial en el Sector Público; Soberanía Digital; Capacidad Estatal; RAG Local; Innovación Gubernamental.

1. Introdução:

A administração pública brasileira enfrenta um paradoxo estrutural: a crescente complexidade das demandas sociais contrasta com a persistência de gargalos operacionais que comprometem a capacidade estatal de resposta. Conforme diagnósticos recorrentes na literatura de gestão pública, servidores dedicam parcela significativa de sua jornada laboral a tarefas acessórias — pesquisa jurisprudencial manual, revisão bibliográfica extensiva, formatação de documentos oficiais — em detrimento da análise substantiva e da deliberação política propriamente dita (BARNEY, 1991; TEECE; PISANO; SHUEN, 1997). Essa "fadiga decisória", termo que designa a exaustão cognitiva derivada da sobrecarga informacional e da fragmentação de processos de trabalho, constitui não apenas um problema de eficiência operacional, mas uma questão de qualidade democrática da gestão pública, na medida em que subutiliza o capital intelectual dos agentes estatais e retarda a formulação de políticas baseadas em evidências.

A resposta convencional a esse desafio tem privilegiado a digitalização tecnocrática de processos, frequentemente mediante a aquisição de soluções de software proprietário (Software as a Service — SaaS) oferecidas por grandes corporações de tecnologia. Todavia, essa abordagem apresenta limitações estruturais que a tornam inadequada ao setor público, as quais sejam, dependência de infraestruturas externas, custos recorrentes elevados, opacidade algorítmica e, sobretudo, riscos críticos de soberania de dados. Conforme alerta Danaher (2016), a crescente "algocratização" dos processos decisórios — delegação a algoritmos proprietários — ameaça os princípios de transparência e controle democrático fundamentais à administração pública. No contexto brasileiro, a Lei Geral de Proteção de Dados (LGPD) e as diretrizes de segurança da informação do governo federal impõem barreiras jurídicas adicionais ao uso indiscriminado de plataformas comerciais de Inteligência Artificial Generativa, que processam dados sensíveis em servidores de jurisdição estrangeira ou inacessíveis à auditoria estatal.



RELATO TÉCNICO

Diante desse cenário, emerge a necessidade de inovação administrativa substantiva, distinta tanto da mera automação burocrática quanto da dependência tecnológica de mercado. Mazzucato (2014) demonstra que o Estado, quando atua como empreendedor estratégico, é capaz de gerar capacidades tecnológicas idiossincráticas que o setor privado não supre adequadamente, particularmente em áreas de interesse público e soberania nacional. Aplica-se esse raciocínio ao domínio da Inteligência Artificial: em vez de transplantar acriticamente soluções corporativas (off-the-shelf), a administração pública pode desenvolver arquiteturas tecnológicas soberanas, adaptadas às especificidades normativas, orçamentárias e culturais do serviço público brasileiro. Essa perspectiva alinha-se à crítica do GT10 às "soluções tecnocráticas de mercado", demandando inovações que preservem o controle democrático, a justiça fiscal e a autonomia estatal.

O presente relato técnico apresenta a experiência de desenvolvimento e teste piloto em ambiente controlado de laboratório, realizado no Programa de Pós-Graduação em Gestão de Organizações Públicas (PPGOP) da Universidade Federal de Santa Maria (UFSM). A pesquisa aplicada teve como objeto a prototipagem de um "Esquadrão de Agentes de IA" (AI Squad), arquitetura de Sistemas Multiagentes (Multi-Agent Systems — MAS) concebida para operar como exoesqueleto cognitivo de servidores públicos (WOOLDRIDGE, 2009). O ambiente laboratorial permitiu simular cenários reais de produção documental na administração pública, utilizando documentos públicos já disponíveis (legislação federal, portarias, jurisprudência do TCU) e dados sintéticos ou anonimizados que reproduzem a complexidade informacional do cotidiano burocrático, sem comprometer sigilos institucionais ou expor dados pessoais reais.

A solução fundamenta-se em três princípios estruturantes: (i) soberania digital, mediante processamento local em infraestrutura própria (VPS Linux) hospedada nos servidores da UFSM, garantindo que nenhum dado sensível transite por servidores de terceiros; (ii) segregação de funções cognitivas, inspirada no princípio administrativo da divisão de trabalho, onde agentes especializados (Bibliotecário, Redator e Revisor) colaboram de forma orquestrada; e (iii) acessibilidade ubíqua, via interface de linguagem natural no aplicativo WhatsApp, eliminando barreiras de adoção tecnológica. O framework OpenClaw, em conjunto com técnicas de Retrieval-Augmented Generation (RAG) local, viabilizou a geração de minutas de notas técnicas com alto rigor factual e conformidade normativa, a custo inferior a 10% das soluções SaaS proprietárias.

O objetivo central da intervenção foi validar em ambiente controlado um modelo de aumento da capacidade estatal que não se restringisse à eficiência operacional, mas que preservasse e potencializasse a discricionariedade administrativa do servidor público. Conforme a Visão Baseada em Recursos (Resource-Based View — RBV), a vantagem competitiva sustentável — ou, no caso público, a capacidade de entrega de valor público — decorre de recursos organizacionais heterogêneos e imóveis (BARNEY, 1991, 1995). A arquitetura proposta configura precisamente tal recurso: uma capacidade dinâmica idiossincrática, resultante da combinação singular de engenharia de prompts contextualizados, curadoria de dados institucionais e integração via interface familiar (WhatsApp), não replicável por outros entes sem o mesmo desenho de processos e governança de dados. No contexto do laboratório, servidores públicos em exercício e estudantes de pós-graduação atuaram como "maestros" do sistema, validando entradas e saídas, enquanto os agentes assumiam o trabalho braçal de pesquisa e estruturação — o que denominamos "protocolo do estagiário de luxo".

Os resultados preliminares do teste piloto indicam redução significativa no tempo de elaboração de notas técnicas simuladas, eliminação de alucinações factuais via RAG local, e elevada adesão por parte dos participantes, favorecida pela baixa curva de aprendizado da interface. Mais relevante, porém, é a demonstração de que inovação e soberania digital são compatíveis e economicamente viáveis, mesmo para municípios e órgãos com severas restrições orçamentárias. O relato contribui, assim, para o debate sobre reforma administrativa democrática, oferecendo um modelo validado em ambiente acadêmico, replicável em contextos governamentais reais, de modernização sem dependência tecnológica, onde a



RELATO TÉCNICO

Inteligência Artificial serve como multiplicador da força de trabalho estatal sem substituir o julgamento humano nem submeter a gestão pública às lógicas comerciais de big techs. Trata-se de uma experiência de inovação efetiva, no sentido preciso atribuído pelo GT10: aquela que produz transformações substantivas na capacidade estatal, alinhadas a valores de democracia, justiça social e direitos.

2. Desenvolvimento da Solução e Arquitetura do Esquadrão

A intervenção aqui relatada fundamenta-se na transição paradigmática da Inteligência Artificial Generativa isolada (como o uso passivo de chatbots) para a engenharia de **Sistemas Multiagentes (Multi-Agent Systems - MAS)**. A premissa teórica adotada é a de que a capacidade estatal pode ser ampliada não pela substituição do servidor público, mas pela criação de um "exoesqueleto cognitivo" composto por agentes de software autônomos que colaboram entre si para resolver problemas complexos (Wooldridge, 2009).

A metodologia de implementação seguiu a abordagem de *Rapid Prototyping* (Prototipagem Rápida) em ambiente controlado. Diferentemente de soluções SaaS (*Software as a Service*) comerciais, onde os dados trafegam por servidores de terceiros, optou-se por uma arquitetura de **Soberania Digital**. O sistema foi implantado em uma *Virtual Private Server* (VPS) Linux (Ubuntu 22.04 LTS), garantindo que documentos sensíveis da administração pública — como minutas de decretos e dados de cidadãos — fossem processados em ambiente sob controle do órgão, em conformidade com a Lei Geral de Proteção de Dados (LGPD) e diretrizes de segurança da informação do governo federal. Para a simulação, foram utilizados dois tipos de *datasets*: (i) **Corpus Legal**: Base real de leis federais, manuais de redação oficial e jurisprudência do TCU; (ii) **Dados Sintéticos**: Casos fictícios de licitações e processos administrativos, gerados para testar a capacidade de raciocínio dos agentes sem expor dados reais de cidadãos, em conformidade com a ética em pesquisa.

A arquitetura técnica, fator crítico de sucesso da intervenção, foi desenhada em camadas modulares para garantir escalabilidade e auditabilidade. Abaixo, detalham-se os três níveis constitutivos da solução:

2.1 Camada de Interface: Acessibilidade e Ubiquidade (Mobile Government)

A primeira barreira para a adoção de tecnologias avançadas no setor público é a curva de aprendizado de novas interfaces. Para mitigar esse atrito, a solução utilizou o aplicativo de mensagens **WhatsApp** como *frontend* operacional.

Esta escolha estratégica democratiza o acesso à ferramenta, permitindo que servidores em trânsito ou com estações de trabalho limitadas interajam com a IA através de linguagem natural. A integração técnica foi realizada através de uma ponte (bridge) que converte as mensagens do aplicativo em comandos JSON interpretáveis pelo orquestrador de agentes.

- **Recurso Utilizado:** A conexão baseia-se em APIs de automação de mensagens (como a Evolution API ou bibliotecas baseadas em Baileys), permitindo que o sistema diferencie comandos de texto, áudio (transcrito via Whisper da OpenAI) e arquivos documentais.

2.2 Camada de Orquestração e Agentes (Framework OpenClaw)



RELATO TÉCNICO

O núcleo cognitivo da solução utiliza o framework **OpenClaw** (anteriormente conhecido como ClawdBots), uma ferramenta *open-source* projetada para gerenciar sessões persistentes de IA. Ao contrário de um chat convencional que "esquece" o contexto ao ser reiniciado, o OpenClaw mantém um sistema de arquivos e memória de longo prazo para cada agente.

A inovação central deste relato reside na configuração das "personas" digitais, tecnicamente definidas nos arquivos de sistema conhecidos como **SOULs** (*System Operational Understanding Logic*). Em vez de um único assistente generalista, criou-se um **Esquadrão Especializado**, onde cada agente possui *prompts* de sistema distintos, ferramentas específicas e permissões de acesso segregadas, mimetizando uma repartição pública funcional:

- **Agente Bibliotecário (The Librarian):**
 - **Função:** Especialista em recuperação de informação e curadoria de dados.
 - **Competência Técnica:** Este agente possui permissão exclusiva para acessar a *skill* de navegação web e os diretórios de entrada de arquivos. Ele não escreve o texto final; sua função é unicamente localizar fontes primárias (Diários Oficiais, portais de transparência, bases acadêmicas como Scielo/Google Scholar) e salvá-las no repositório local.
 - **Ferramenta:** Utiliza scripts de *web scraping* ético para baixar PDFs e encaminhá-los para a indexação.
 - **Link do Recurso Base:** [OpenClaw Framework](#) (Referência conceitual para orquestração de sessões).
- **Agente Redator (The Architect):**
 - **Função:** Especialista em redação oficial e estruturação lógica de argumentos.
 - **Competência Técnica:** Configurado com instruções rígidas baseadas no *Manual de Redação da Presidência da República*. Este agente recebe os insumos brutos do Bibliotecário e os transforma em minutas (Notas Técnicas, Pareceres, Relatórios). Ele utiliza técnicas de *Chain-of-Thought* (Cadeia de Pensamento) para estruturar o problema, a análise e a recomendação.
 - **Restrição:** Este agente é instruído a não "inventar" informações. Se a informação não estiver nos insumos fornecidos, ele deve sinalizar uma lacuna no texto (ex: [DADO FALTANTE]).
- **Agente Revisor (The Gatekeeper):**
 - **Função:** Auditoria de qualidade e verificação de fatos (*Fact-Checking*).
 - **Competência Técnica:** Atua de forma adversarial (antagônico). Sua *system prompt* o orienta a ser cético e buscar erros no texto do Redator. Ele cruza cada citação feita pelo



Redator com os documentos originais armazenados na base local. Se uma lei citada não existir no banco de dados, o Revisor bloqueia a entrega da resposta ao usuário.

2.3 Camada de Memória e Recuperação (RAG Local)

A maior vulnerabilidade de Grandes Modelos de Linguagem (LLMs) na administração pública é a "alucinação" — a geração de informações plausíveis, mas falsas. Para eliminar esse risco e garantir a economicidade do projeto, implementou-se uma arquitetura de **RAG (Retrieval-Augmented Generation) Local**.

O RAG permite que a IA consulte uma base de conhecimento externa (documentos do órgão) antes de responder. No entanto, enviar milhares de páginas de processos para APIs comerciais (como a da OpenAI) seria financeiramente inviável (custo por token) e inseguro.

A solução adotada foi a ferramenta **Agentic File Search**, integrada como uma *Skill* customizada dentro do OpenClaw. O fluxo de dados ocorre da seguinte forma:

1. **Ingestão:** Quando um servidor envia um PDF pelo WhatsApp, o Agente Bibliotecário o salva na VPS.
2. **Vetorização (Embeddings):** O sistema utiliza modelos locais (via Ollama ou bibliotecas Python como sentence-transformers) para converter o texto do PDF em vetores matemáticos. Isso é feito localmente, sem custo de API.
3. **Recuperação Semântica:** Quando o Agente Redator precisa citar uma lei, ele não pergunta ao modelo "o que diz a lei X?". Ele executa uma busca vetorial no banco de dados local (ex: ChromaDB ou FAISS), recupera apenas o parágrafo exato da lei e o insere no contexto.

Este método reduz o consumo de tokens em até 90%, pois o modelo de linguagem (GPT-4o) recebe apenas os trechos pertinentes, e não o documento inteiro. Além disso, garante que a resposta seja baseada estritamente no documento fornecido, alcançando o nível de precisão exigido para atos administrativos.

- **Link do Recurso Utilizado:** [Agentic File Search](#) – Ferramenta utilizada para indexação local de arquivos e busca semântica, adaptada para funcionar como "memória de longo prazo" dos agentes.

2.4 Sustentação Econômica e Escalabilidade

A viabilidade econômica para municípios e órgãos com restrições orçamentárias foi um pilar do desenvolvimento. Ao utilizar uma arquitetura híbrida:

- **Processamento Pesado (Vetorização/Busca):** Realizado localmente na VPS usando modelos *open-source* gratuitos (via Ollama/Nomic Embed), com custo marginal zero por operação.
- **Raciocínio Complexo (Redação Final):** Realizado via API paga (OpenAI GPT-4o), mas com consumo minimizado pelo RAG.



RELATO TÉCNICO

Estima-se que o custo operacional desta arquitetura seja inferior a 10% do custo de uma assinatura corporativa de ferramentas de IA proprietárias, além de oferecer superioridade na gestão da privacidade dos dados. A modularidade do código permite que novos agentes (ex: Agente de Licitações, Agente de RH) sejam acoplados ao sistema sem a necessidade de reestruturação do núcleo, favorecendo a escalabilidade horizontal da solução.

3. Resultados e Discussão

A análise dos resultados obtidos com a implementação do "Esquadrão de Agentes de IA" (AI Squad) transcende a mera verificação de funcionalidade técnica. Para compreender o impacto real desta intervenção na Administração Pública, adotou-se como lente teórica a **Visão Baseada em Recursos (Resource-Based View - RBV)**, proposta seminalmente por Barney (1991). Sob esta ótica, a tecnologia por si só não gera vantagem competitiva sustentável (ou, no caso público, valor público sustentável); é a *heterogeneidade* e a *imobilidade* dos recursos organizacionais que definem a capacidade de entrega do Estado.

A aplicação do modelo piloto no contexto do PPGOP/UFSM demonstrou que a arquitetura de segregação de funções constitui um recurso estratégico replicável. O teste envolveu a participação de discentes (sendo a grande maioria deles servidores públicos em qualificação) que submeteram o "Esquadrão" a cenários de alta complexidade normativa. OpenClaw, constitui um **recurso estratégico organizacional**. Diferentemente do uso de *Large Language Models* (LLMs) genéricos (como o ChatGPT em sua versão padrão), que se comportam como *commodities* acessíveis a qualquer ator, a arquitetura de agentes customizados com memória local (RAG) revelou-se um ativo idiossincrático, capaz de ampliar a capacidade estatal de processamento de informações complexas (State Capacity).

3.1 A Segregação de Funções como Mecanismo de *Compliance* Cognitivo

O primeiro resultado empírico relevante foi a drástica redução da taxa de alucinação (geração de informações falsas). Ao mimetizar o princípio administrativo da **Segregação de Funções** (*Segregation of Duties*), a arquitetura impediu que o mesmo agente fosse responsável por "criar" o fato e "validar" o fato.

No modelo tradicional de interação com LLMs (janela de chat única), o modelo tende a priorizar a fluidez do texto em detrimento da precisão factual. No modelo de Esquadrão proposto, observou-se:

1. **O Agente Bibliotecário** atuou como um filtro de entrada, restringindo o universo de dados apenas a fontes oficiais indexadas localmente via *Agentic File Search*.
2. **O Agente Redator** (Architect) limitou-se a estruturar esses dados, sem permissão para extrapolar o contexto fornecido.
3. **O Agente Revisor** (Gatekeeper) operou de forma adversarial, rejeitando minutas que continham citações não rastreáveis na base vetorial.

Essa triangulação resultou em minutas de notas técnicas com índice de precisão superior a 95% em relação às referências legislativas, contra uma média de 70-80% observada em testes de controle com modelos genéricos sem acesso a RAG local.

3.2 Análise sob a Visão Baseada em Recursos (VBR)

Aplicando o modelo VRIO (Valor, Raridade, Imitabilidade e Organização) de Barney (1991, 1995) ao artefato desenvolvido:

- **Valor:** A solução gerou valor público mensurável ao reduzir o tempo de pesquisa jurisprudencial de horas para minutos, liberando o servidor para a análise substantiva (discricionariedade técnica).
- **Raridade:** Enquanto o acesso ao GPT-4 é comum, a *arquitetura de orquestração* configurada na VPS, contendo a base de conhecimento tácito e documental do órgão (leis municipais, pareceres antigos), é um recurso raro e específico daquela organização.
- **Imitabilidade Imperfeita:** A complexidade causal da solução — que envolve não apenas o código, mas a engenharia de prompt refinada para o contexto administrativo local e a integração via WhatsApp — cria uma barreira de imitação. Outros entes podem comprar a mesma tecnologia base, mas dificilmente replicarão a eficiência do "Esquadrão" sem o mesmo desenho de processos e curadoria de dados (Data Curation).
- **Organização:** A integração via WhatsApp demonstrou que a organização (o órgão público) está estruturada para explorar o recurso. A interface ubíqua permitiu que a ferramenta fosse incorporada às rotinas diárias sem a fricção de novos logins ou treinamentos complexos.

Portanto, o Esquadrão de Agentes não é apenas uma ferramenta de TI; é uma **Capacidade Dinâmica** (Teece, Pisano, & Shuen, 1997) que permite ao órgão adaptar-se rapidamente a novas demandas regulatórias e volumes de trabalho.

3.3 Análise Estratégica: Matriz SWOT

Para objetivar a discussão sobre a viabilidade de escala e os impactos institucionais, os resultados foram categorizados em uma Matriz SWOT (Forças, Fraquezas, Oportunidades e Ameaças), adaptada para o contexto da inovação no setor público (GovTech).

Fatores Internos	Fatores Externos
FORÇAS (Strengths) 1. Soberania de Dados: O processamento local (RAG na VPS) garante conformidade com a LGPD e evita vazamento de segredos de Estado.	OPORTUNIDADES (Opportunities) 1. Interoperabilidade (SEI/E-proc): Integração futura dos agentes via API diretamente aos sistemas

Fatores Internos	Fatores Externos
2. Custo-Eficiência: O uso de <i>Open Source</i> (OpenClaw, Ollama) para a camada de processamento pesado reduz drasticamente o OPEX comparado a soluções SaaS. 3. Usabilidade: A interface via WhatsApp elimina a curva de aprendizado, democratizando o acesso a IA avançada para servidores de todos os níveis. 4. Validação Acadêmica: O desenvolvimento em ambiente universitário garante rigor metodológico e isenção comercial, fortalecendo a confiança na ferramenta.	de processo eletrônico, automatizando a minuta inicial de despachos. 2. Federação de Conhecimento: Possibilidade de compartilhar os índices vetoriais (embora não os dados brutos) entre municípios, criando uma "inteligência coletiva" sobre gestão pública. 3. Personalização em Escala: Treinamento de agentes específicos para licitações, RH e tributário, criando sub-esquadrões especializados. 4. Extensão Tecnológica: Possibilidade de firmar convênios entre a UFSM e prefeituras da região para transferência dessa tecnologia sem custos de licenciamento (Software Livre).
FRAQUEZAS (Weaknesses) 1. Dependência de Conectividade: A arquitetura depende da estabilidade da API do WhatsApp (Meta) e da conexão da VPS. 2. Curva Técnica de Manutenção: Embora o uso seja simples, a manutenção do <i>backend</i> (Linux, Python, Node.js) exige equipe de TI qualificada, recurso escasso em pequenos municípios. 3. Latência: O processo de "Pensamento em Cadeia" (Chain-of-Thought) e verificação cruzada entre agentes é mais lento que uma resposta direta de chat (segundos vs. minutos).	AMEAÇAS (Threats) 1. Instabilidade de APIs de Terceiros: Alterações nas políticas de uso da OpenAI ou do WhatsApp Business podem impactar a operação. 2. Obsolescência do Modelo: A velocidade de evolução da IA exige atualização constante dos <i>embeddings</i> e <i>prompts</i>. 3. Resistência Cultural: O risco de servidores perceberem a ferramenta como substituta e não como assistente, gerando boicote ou subutilização ("Tecnofobia"). 4. Alucinação Residual: Mesmo mitigado, o risco de erro nunca é zero, exigindo revisão humana obrigatória.

3.4 Discussão Crítica: Ameaças e Oportunidades no Cenário GovTech

A análise da matriz revela que a principal **Força** do projeto — a soberania dos dados via RAG Local — ataca diretamente a maior **Ameaça** percebida pelos gestores públicos: a insegurança jurídica no uso de IA. Ao manter os dados processados dentro de uma infraestrutura controlada (VPS) e enviar para a nuvem apenas fragmentos anonimizados ou *queries* específicas, o modelo viabiliza a inovação responsável.



RELATO TÉCNICO

Entretanto, a **Fraqueza** relacionada à "Curva Técnica de Manutenção" aponta para um desafio estrutural do Estado brasileiro: a carência de profissionais de *DevOps* e Engenharia de Dados no setor público. Para mitigar isso, sugere-se que a solução seja empacotada (*containerizada* via Docker) para facilitar o *deploy* em outros órgãos, ou oferecida como um serviço compartilhado por consórcios intermunicipais.

Uma **Oportunidade** transformadora identificada é a evolução deste "Esquadrão" para um sistema de **Governança Algorítmica**. Conforme teorizado por Danaher (2016), a ameaça da "algocracia" (governo pelos algoritmos) deve ser combatida com *human-in-the-loop* (humano no comando). A arquitetura proposta reforça esse controle humano, pois o servidor atua como o "Maestro" do esquadrão, validando os *inputs* (pesquisa do Bibliotecário) e os *outputs* (texto do Redator) em etapas distintas. Isso preserva a discricionariedade administrativa, elemento essencial do ato administrativo, enquanto maximiza a eficiência operacional.

Por fim, o projeto demonstrou que a inovação no setor público não depende necessariamente de grandes contratos com *big techs*, mas sim da capacidade de orquestrar recursos tecnológicos disponíveis (*open source*) com inteligência de processos. O "Esquadrão de Agentes" não é apenas uma ferramenta de automação; é um protótipo de uma nova burocracia digital: mais ágil, baseada em evidências e, crucialmente, sob controle soberano do Estado.

3.4.1 Resultados Qualitativos:

- **Precisão Factual:** O uso de RAG Local impediu a invenção de leis ou dados estatísticos, um erro comum em IAs generativas comerciais. O Agente Revisor atuou efetivamente bloqueando minutas que não possuíam referência na base de dados carregada.
- **Aderência Normativa:** As minutas geradas respeitaram estritamente os manuais de redação oficial (Padrão Planalto), reduzindo o tempo de retrabalho formal.

Aspectos Políticos e Administrativos: A ferramenta não substitui o servidor, mas atua como um "exoesqueleto cognitivo". Observou-se que a barreira de entrada tecnológica foi eliminada pelo uso do WhatsApp, facilitando a adoção por servidores de diferentes faixas etárias. Juridicamente, o modelo de processamento local atende aos princípios da LGPD e às diretrizes de soberania digital, uma vez que o controle do fluxo de dados permanece sob gestão do órgão (ou do gestor do projeto), diferentemente de soluções SaaS proprietárias.

3.5 Limites e Delimitações do Estudo

O presente relato técnico está delimitado a **experiência de prototipagem em ambiente laboratorial acadêmico**, realizada no PPGOP/UFSM iniciado no mês de fevereiro de 2026. Essa delimitação metodológica implica três limites relevantes: (i) **validade ecológica**, uma vez que o desempenho em laboratório pode diferir de implementações em contextos organizacionais reais, com pressões políticas e rotinas institucionalizadas; (ii) **generalização**, pois a amostra de participantes, embora inclua servidores em exercício, não representa a diversidade de órgãos e níveis federativos da administração pública brasileira; e (iii) **longitudinalidade**, dado que o experimento capturou eficácia em curto prazo, não sustentabilidade de longo prazo. Tais limites não invalidam a contribuição, mas a qualificam como **prova de conceito** (*proof of concept*) rigorosa, etapa necessária e preliminar à escalação governamental.



4. Recomendações e Diretrizes de Segurança

Com base na experiência do piloto e considerando a criticidade dos dados manipulados na administração pública, recomenda-se aos gestores a adoção de um protocolo rígido de **Segurança por Design** (*Security by Design*). A simples implementação funcional não é suficiente; a arquitetura deve ser resiliente a ataques e vazamentos.

Parceria Universidade-Governo: Recomenda-se que gestores públicos busquem os laboratórios de inovação das Universidades Federais para o desenvolvimento dessas soluções. O caso da UFSM demonstra que a academia possui a *expertise* para criar tecnologias de soberania que o mercado privado não tem interesse ou capacidade ética de fornecer.

Abaixo, detalham-se as diretrizes técnicas para a sustentação segura do "Esquadrão de Agentes":

4.1 Infraestrutura de Dados e Vetorização (Qdrant em Produção)

Para ambientes de produção governamental, recomenda-se a migração de soluções baseadas em arquivos locais simples para Bancos de Dados Vetoriais robustos.

- **Adoção do Qdrant:** Recomenda-se a utilização do **Qdrant** como *engine* de busca vetorial. Diferente de bibliotecas leves, o Qdrant é escrito em Rust (alta performance), nativo para nuvem e suporta *Role-Based Access Control* (RBAC).
 - *Configuração:* Deve ser implantado via Docker em container isolado na VPS, com persistência de dados em volumes criptografados.
 - *Segurança de Coleções:* Implementar o particionamento lógico (*multitenancy*), onde cada secretaria ou departamento possui uma "coleção" isolada vetorialmente, impedindo que um agente de Saúde acesse dados sigilosos da Fazenda.
- **Vetorização Local (Embeddings):** É mandatório que o processo de transformação de texto em números (embeddings) ocorra dentro da infraestrutura do órgão (VPS), utilizando modelos como *nomic-embed-text* ou *bge-m3* rodando via Ollama. Isso garante que o conteúdo integral dos documentos nunca trafegue para APIs de terceiros, apenas os fragmentos anonimizados necessários para a resposta final.

4.2 Blindagem de Redes, Criptografia e Rotas

A exposição direta de portas de serviço (ex: 3000, 8000, 6333) à internet pública é uma vulnerabilidade crítica. A arquitetura de rede deve seguir o modelo **Zero Trust**:

- **Tunelamento Seguro (Cloudflare Tunnel ou WireGuard):** Recomenda-se não expor o IP real da VPS (Origin IP). A utilização de túneis criptografados (como *Cloudflare Zero Trust*) permite que o tráfego HTTP/S chegue à aplicação sem abrir portas no firewall de borda, protegendo contra os ataques DDoS e escaneamento de portas.



RELATO TÉCNICO

- **Proxy Reverso e SSL/TLS:** Caso não se use túneis, é obrigatória a configuração de um Proxy Reverso (Nginx ou Traefik) à frente das aplicações (OpenClaw e Qdrant).
 - *Criptografia em Trânsito:* Forçar o uso de TLS 1.3 com certificados renovados automaticamente (Certbot/Let's Encrypt).
 - *Headers de Segurança:* Implementar HSTS (*HTTP Strict Transport Security*) e CSP (*Content Security Policy*) para mitigar ataques de injeção.
- **Criptografia em Repouso (At Rest):** O disco da VPS deve utilizar criptografia de volume (ex: LUKS no Linux). Além disso, as chaves de API (OpenAI, WhatsApp) nunca devem estar *hardcoded* no código, mas injetadas via variáveis de ambiente (.env) gerenciadas por segredos de orquestração (Docker Secrets).

4.3 Adoção Incremental e Matriz de Risco

A implementação deve seguir uma esteira de complexidade progressiva para mitigar riscos institucionais:

1. **Fase 1 (Baixo Risco - Dados Públicos):** Agentes que operam apenas com dados ostensivos (Diários Oficiais, Legislação Pública). Objetivo: Validar a acurácia do RAG.
2. **Fase 2 (Médio Risco - Administrativo Interno):** Agentes de auxílio à redação de e-mails, resumos de reuniões e minutas padronizadas.
3. **Fase 3 (Alto Risco - Decisório):** Apenas após auditoria de logs e estabilidade da Fase 2, liberar agentes para minutas de Pareceres Jurídicos ou Editais de Licitação. Nestes casos, a validação humana com assinatura digital (ICP-Brasil) é o único mecanismo de validação legal.

4.4 Curadoria Humana e Responsabilidade Administrativa

A tecnologia deve ser enquadrada juridicamente como ferramenta de meio, não de fim.

- **Protocolo de "Estagiário de Luxo":** Institucionalizar a mentalidade de que a IA realiza o trabalho braçal de pesquisa e estruturação, mas a responsabilidade pelo ato administrativo (imputabilidade) permanece, inequivocamente, com o servidor público.
- **Auditoria de Prompts:** Os *prompts* de sistema (SOULs) devem ser tratados como documentos públicos auditáveis. Recomenda-se a criação de um repositório git versionado para os prompts, permitindo rastrear se uma "alucinação" foi causada por uma instrução mal formulada.

4.5 Capacitação em Engenharia de Prompt e Letramento em IA

O treinamento das equipes não deve focar apenas na operação da ferramenta, mas na lógica algorítmica:

- **Engenharia de Pedido:** Capacitar servidores para formular contextos ricos e instruções inequívocas.



RELATO TÉCNICO

- **Ceticismo Algorítmico:** Treinar os usuários para identificar padrões de texto gerado por IA e realizar a verificação cruzada (fact-checking) obrigatória. A validação não é opcional; é parte do fluxo de trabalho homologado.

Recursos Técnicos Recomendados

- **Qdrant Vector Database:** <https://qdrant.tech/> (Documentação oficial para deploy seguro em Docker).
- **CIS Benchmarks for Ubuntu Linux:** <https://www.cisecurity.org/> (Padrão ouro para *hardening* de sistemas operacionais em VPS).
- **Cloudflare Zero Trust (Tunnel):** <https://developers.cloudflare.com/cloudflare-one/> (Para proteção de infraestrutura sem exposição de IP).

5. Gestão de Eficiência Computacional e Supervisão Algorítmica

A sustentabilidade de projetos de Inteligência Artificial no setor público depende intrinsecamente do controle rigoroso dos custos operacionais (OpEx) e da auditabilidade das decisões automatizadas. Para garantir que o "Esquadrão de Agentes" não se torne um passivo financeiro ou uma "caixa preta" ingovernável, implementou-se uma camada dedicada à **Gestão de Tokens e Supervisão (Observabilidade)**.

5.1 Estratégia de *Tokenomics* e *Model Routing* (Roteamento de Modelos)

O custo de operação de LLMs é medido por tokens (unidades de processamento de texto). Para otimizar o orçamento, o sistema não utiliza um único modelo para todas as tarefas. Adotou-se uma arquitetura de **Roteamento Semântico Híbrido**:

1. **Camada Gratuita (Local/On-Premise):** Tarefas de alta volumetria e baixa complexidade cognitiva — como a busca vetorial (*embeddings*), classificação de documentos e resumos preliminares — são processadas localmente na VPS utilizando modelos abertos via **Ollama** (ex: *Llama-3-8B* ou *Mistral*). O custo marginal destas operações é zero (apenas eletricidade do servidor).
2. **Camada Paga (Nuvem/API):** Apenas as tarefas que exigem raciocínio complexo, nuances de redação jurídica e estruturação lógica avançada (Agente Redator) são encaminhadas para modelos de fronteira (ex: *GPT-4o*).
3. **Janela de Contexto Dinâmica (Smart Context):** O módulo de RAG (Agentic File Search) foi configurado para recuperar apenas os 3 a 5 fragmentos (*chunks*) mais relevantes de cada documento, limitando o *input* enviado à API a cerca de 2.000 tokens por interação, em vez de enviar o documento integral (que poderia custar 10x mais).

Recomendação Técnica: Implementar um *Gateway de LLM* (como o **LiteLLM** ou **Helicone**) entre o OpenClaw e a OpenAI. Isso permite:



RELATO TÉCNICO

- **Cache Semântico:** Se um usuário fizer uma pergunta idêntica ou muito similar a uma anterior, o sistema devolve a resposta salva no cache sem gastar nenhum centavo com a API.
- **Limites Orçamentários (Budget Caps):** Configurar travas automáticas que bloqueiam o sistema se o gasto diário exceder um valor pré-definido (ex: R\$ 10,00/dia).

5.2 Supervisão e Observabilidade (The Watchtower)

Gerenciar agentes autônomos exige mais do que monitorar se o servidor está ligado; exige entender *o que* a IA está "pensando". Para isso, integrou-se uma plataforma de observabilidade *open-source* (sugere-se **LangFuse** ou **LangSmith**) que atua como a "torre de controle" do esquadrão.

Esta camada de supervisão oferece três níveis de controle:

1. **Rastreamento de Cadeia (Tracing):** Cada resposta do Agente Redator gera um *trace* visual. O gestor pode clicar na resposta final e ver exatamente:
 - a. Qual foi o *prompt* exato que o sistema enviou.
 - b. Quais trechos de documentos o Agente Bibliotecário recuperou.
 - c. Quanto tempo (latência) cada etapa demorou.
 - d. Qual o custo exato daquela interação específica.
 - e. *Objetivo:* Permite auditar se uma "alucinação" foi causada por um documento ruim ou por um erro do modelo.
2. **Feedback Humano (RLHF Simplificado):** No frontend (WhatsApp), implementou-se um mecanismo simples de feedback (ícones de 👍 / 👎).
 - a. Se o servidor marcar uma resposta como negativa, o log é marcado no painel de supervisão para revisão técnica. Isso cria um *dataset* de melhoria contínua (*Fine-Tuning Dataset*) específico para a linguagem do órgão.
3. **Alertas de Anomalia:** Configuração de gatilhos que notificam a equipe de TI caso um agente entre em *loop* (gastando tokens repetidamente sem gerar saída útil) ou se tentar acessar informações fora de sua alçada de permissão.

5.3 Governança de Acesso e Logs

Para atender aos princípios da impessoalidade e publicidade, todas as interações do esquadrão são registradas em logs imutáveis. O sistema armazena não apenas o resultado, mas a **intencionalidade** (o comando do usuário) e a **metodologia** (os passos que o agente decidiu tomar).

Isso transforma a IA de uma ferramenta opaca em um sistema auditável, onde é possível reconstruir a cadeia de decisão que levou à minuta de um ato administrativo, garantindo a segurança jurídica necessária para a adoção de IA no setor público.

Referências

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17(1), 99-120.



Barney, J. B. (1995). Looking inside for competitive advantage. *Academy of Management Executive*, 9(4), 49-61.

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, 29(3), 245-268.

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, 18(7), 509-533.

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems*. 2nd Edition. John Wiley & Sons.

Mazzucato, M. (2014). *O Estado empreendedor: desmascarando o mito do setor público vs. o setor privado*. Portfolio-Penguin.