INDÍCE

1	INTRODUÇÃO	2
2	OBJETO DA CONTRATAÇÃO	2
3	DO ENQUADRAMENTO NO PDTIC	6
4	DA JUSTIFICATIVA DA NECESSIDADE DA AQUISIÇÃO	7
5	DO LEVANTAMENTO DAS POSSÍVEIS SOLUÇÕES	8
6	DA JUSTIFICATIVA PARA A ESCOLHA DA SOLUÇÃO	10
7	DOS RESULTADOS PRETENDIDOS	10
8	DA JUSTIFICATIVA DE PARCELAMENTO OU NÃO	11
9	DO PARECER QUANTO A VIABILIDADE TÉCNICA DA CONTRATAÇÃO	13
10	DA LEI DE ACESSO A INFORMAÇÃO (LEI Nº 12.527/11)	13
11	DA CONCLUSÃO	14

ESTUDO TÉCNICO PRELIMINAR

1 INTRODUÇÃO

O presente estudo pretende assegurar a viabilidade técnica da aquisição de uma solução de processamento de Inteligência Artificial que atenda às necessidades da Procuradoria Geral do Estado do Rio de Janeiro, bem como fundamentar o Termo de Referência previsto no art. 6º, inciso XX, da Lei 14.133/21.

Nesse sentido, considerando a Nota Técnica SGE n.º 01/2015, de 11 de agosto de 2015, do TCE-RJ, e no art. 10, III do Decreto n.º 46.642/2019, este documento apresenta um modelo de contratação que, por meio da análise das soluções disponíveis no mercado, identifica aquela que, simultaneamente, contemple os procedimentos de Tecnologia da Informação (TI) indispensáveis ao suprimento das demandas da PGE-RJ, e que esteja de acordo com as normas vigentes e os princípios da Administração Pública.

2 OBJETO DA CONTRATAÇÃO

2.1 DA NECESSIDADE DE UTILIZAÇÃO DE UNIDADES DE PROCESSAMENTO GRÁFICO (PLACAS DE VÍDEO)

O atual perfil das atividades exercidas pela PGE-RJ enseja a aquisição de uma solução para processamento de Inteligência Artificial (IA), com foco na implementação de Modelos de Linguagem de Grande Escala (LLMs - *Large Language Models*).

Inicialmente, destaca-se o modelo LLM Llama3 da Meta, que se sobressai entre os modelos de linguagem aberta disponíveis atualmente. Sua notoriedade se fundamenta na capacidade avançada de processar dados complexos com precisão e eficiência. O Llama3 oferece um suporte robusto para uma variedade de domínios e contextos linguísticos, garantindo assim análises técnicas de alta qualidade e confiabilidade (https://llama.meta.com/llama3/).

O modelo Llama3 está disponível em algumas versões como: 8 bilhões de parâmetros e 70 bilhões de parâmetros. Os parâmetros em um modelo de linguagem de grande escala referem-se aos pesos ajustáveis que o modelo aprende durante o treinamento com grandes conjuntos de dados, determinando como o modelo processa e gera texto. Quanto mais parâmetros um modelo possui, maior sua capacidade de aprender e representar nuances linguísticas complexas.

A aquisição desta solução visa proporcionar análises avançadas, automação de processos e suporte à tomada de decisões jurídicas, melhorando a eficiência e a eficácia dos serviços prestados pela PGE-RJ.

A execução de Modelos de Linguagem de Grande Escala (LLMs), como o Llama3 da Meta, exige um poder de processamento significativo. Embora seja possível realizar tais operações usando apenas CPUs, a utilização de Unidades de Processamento Gráfico (GPUs),

contidas em dispositivos denominados “placas de vídeo”, é altamente recomendada devido às suas vantagens específicas para tarefas de aprendizado de máquina e inteligência artificial.

Para calcular a memória de vídeo (VRAM) necessária para a inferência em modelos como o Llama 3, é possível através do uso da seguinte fórmula:

$$M = \frac{(P * 4B)}{(32/Q)} * 1.2$$

Onde:

- **M:** Memória da GPU expressa em Gigabytes
- **P:** Número de parâmetros no modelo
- **4B:** 4 bytes, que representam os bytes usados para cada parâmetro
- **32:** Existem 32 bits em 4 bytes
- **Q:** Quantidade de bits usados para carregar o modelo
- **1.2:** Representa uma sobrecarga de 20% para carregamento adicional na memória da GPU

Referência: https://www.substratus.ai/blog/calculating-gpu-memory-for-llm

Exemplo de Cálculo: Para um modelo de 70 bilhões de parâmetros (70B) usando 5 bits de quantização:

*m = ((P * 4B) / (32 / Q)) * 1.2*

*m = ((70,000,000,000 * 4) / (32 / 5)) * 1.2*

*m = (280,000,000,000 / 6.4) * 1.2*

*m = 43,750,000,000 * 1.2*

m = 52,500,000,000 bytes

Convertendo para Gigabytes:

52,500,000,000 bytes / 1,000,000,000 = 52.5GB

Portanto, para um modelo de 70 bilhões de parâmetros usando quantização de 5 bits, seriam necessários aproximadamente 52.5GB de memória GPU para carregar o modelo, incluindo a sobrecarga de 20%.

2.2 DA QUANTIDADE MÍNIMA DE PLACAS DE VÍDEO

Para implementar o modelo Llama3 de 70 bilhões de parâmetros usando quantização de 5 bits, levando em consideração as demandas computacionais do modelo em ambientes de desenvolvimento e produção, é recomendável utilizar no mínimo oito placas de vídeo (GPUs). Esta configuração não apenas garante um desempenho otimizado, mas também proporciona redundância, assegurando a estabilidade operacional necessária na Procuradoria Geral do Estado do Rio de Janeiro (PGE-RJ).

Justificativa para a quantidade de oito GPUs:

a. **Desempenho e Velocidade de Processamento:** O treinamento e a inferência de modelos de linguagem de grande escala, como o Llama3, envolvem cálculos complexos e intensivos. O uso de oito GPUs permite uma divisão eficiente das tarefas, resultando em uma aceleração significativa do processamento. Isso é crucial para reduzir o tempo de treinamento, que de outra forma poderia levar semanas ou meses;

b. **Memória Suficiente:** Cada GPU contribui com uma quantidade significativa de VRAM. O uso de oito GPUs assegura que haja memória suficiente para carregar o modelo completo de 70 bilhões de parâmetros, além de oferecer espaço adicional para as operações necessárias durante o treinamento e a inferência. Isso evita gargalos e assegurar um processamento suave e eficiente;

c. **Redundância e Confiabilidade:** Ter oito GPUs proporciona uma margem de redundância. Caso uma ou mais GPUs falhem, o sistema pode continuar operando com as GPUs restantes, garantindo a continuidade do trabalho e evitando interrupções que poderiam afetar a produtividade e os prazos da PGE-RJ;

d. **Escalabilidade e Flexibilidade:** Uma configuração com oito GPUs oferece uma base sólida para futuros projetos e expansões. À medida que as necessidades da PGE-RJ crescem ou novos modelos mais complexos são desenvolvidos, o sistema estará preparado para lidar com essas demandas adicionais sem a necessidade de grandes investimentos adicionais em hardware;

e. **Compatibilidade com Frameworks de IA:** A maioria dos frameworks de aprendizado de máquina, como *TensorFlow* e *PyTorch*, são altamente otimizados para operar em configurações com múltiplas GPUs. Isso maximiza a utilização dos recursos disponíveis e assegura que o desempenho do hardware seja totalmente aproveitado;

f. **Fine-Tuning e Customização:** O processo de *fine-tuning*, que envolve ajustar um modelo pré-treinado para tarefas específicas com novos dados, é intensivo em termos de computação e pode ocupar uma ou mais GPUs por um longo período de tempo. O uso de oito GPUs acelera significativamente este processo, permitindo que a PGE-RJ ajuste rapidamente o Llama3 para suas necessidades específicas. Durante o fine-tuning, cada GPU pode ficar ocupada por várias horas ou até dias, dependendo da complexidade dos dados e das tarefas de ajuste fino. Isso melhora a precisão e a relevância dos modelos de IA para os contextos jurídicos específicos da PGE-RJ, resultando em análises e decisões mais precisas;

g. **Capacidade de Manutenção e Atualização Contínua:** A utilização de oito GPUs permite uma manutenção e atualização contínua dos modelos de IA. Com essa capacidade, a PGE-RJ pode iterar sobre seus modelos, incorporando novos dados e refinando os algoritmos de forma eficiente, sem interrupções significativas nos serviços.

Dessa forma, o uso de oito GPUs assegura que a PGE-RJ esteja equipada com uma solução robusta e de alta performance para o processamento de inteligência artificial, atendendo às demandas atuais e futuras com eficiência e confiabilidade. A capacidade de realizar *finetuning* de forma hábil garante que os modelos sejam altamente relevantes e adaptados às necessidades específicas da Procuradoria.

2.3 DAS ESPECIFICAÇÕES DE HARDWARE E DA DEFINIÇÃO DE SERVIDOR

Para garantir a execução eficiente de Modelos de Linguagem de Grande Escala (LLMs), é essencial que a solução de processamento de IA atenda a requisitos específicos de hardware. O equipamento necessário para essas demandas é denominado servidor, e deve atender a determinadas especificações técnicas para assegurar um desempenho ideal.

2.3.1 Definição e Diferença Entre Servidor e Estação de Trabalho:

Um servidor é um sistema computacional projetado para fornecer recursos, serviços e dados a vários computadores ou usuários simultaneamente em uma rede. Ele é otimizado para suportar altas cargas de trabalho, oferecendo alta disponibilidade e confiabilidade em ambientes corporativos e de data centers. Normalmente, servidores são equipados com hardware robusto que permite o processamento de grandes volumes de dados e a gestão eficiente de múltiplas tarefas e usuários ao mesmo tempo.

Por outro lado, uma estação de trabalho é um computador individual destinado ao uso por um único usuário, com foco em tarefas que exigem alto desempenho, como desenvolvimento de software, design gráfico e análise de dados. Embora a estação de trabalho

não seja projetada para lidar com a carga contínua e o atendimento a múltiplos usuários como um servidor, ela pode ser utilizada para desenvolver soluções e aplicações. Essas soluções podem posteriormente ser disponibilizadas como serviços para inúmeros usuários por meio de um servidor, que é capaz de gerenciar o processamento e a entrega desses serviços em escala.

2.3.2 Da Especificação mínima do Servidor:

a. **CPU:** O servidor deve ter no mínimo dois processadores com pelo menos 24 núcleos cada, além do suporte a *multi-threading*, necessário para gerenciar tarefas auxiliares e pré-processamento de dados. Processadores robustos garante, que o servidor possa lidar com cargas de trabalho intensivas sem gargalos;

b. **RAM:** O servidor deve ter no mínimo 1024 GB (1 TB) de RAM para suportar operações de dados intensivos e cargas de trabalho simultâneas. A RAM suficiente é crucial para garantir que o servidor possa processar grandes volumes de dados de forma eficiente, sem quedas de desempenho;

c. **Armazenamento:** O servidor deve ser equipado com um SSD de alta velocidade, ou tecnologia superior, com capacidade mínima de 10 TB. O armazenamento rápido é necessário para lidar com grandes conjuntos de dados e modelos, permitindo acesso ágil e processamento eficiente, reduzindo os tempos de espera e aumentando a eficiência geral do sistema;

Essas especificações são essenciais para atender às demandas intensivas de processamento e para garantir que o servidor possa suportar múltiplos usuários e operações simultâneas de forma eficaz, enquanto oferece os serviços ou aplicações de IA desenvolvidas em estações de trabalho.

3 DO ENQUADRAMENTO NO PDTIC

O Plano Diretor de Tecnologia da Informação e Comunicação (PDTIC) do biênio 2023/2024, instrumento essencial para a realização e planejamento das ações de TI, contempla a presente contratação em seus objetivos, nos seguintes termos:

Objetivo 1	**Garantir excelência em TI para atender às finalidades institucionais.**
Meta 1	Sistemas de informação institucionais alinhados aos processos organizacionais até dezembro de 2024.

Ação 7	Sustentar todos os sistemas implantados e em operação na PGE/RJ para atender atividades finalísticas e setoriais.
Ação 8	Desenvolver sistemas com o uso de tecnologias modernas (Sistemas em camadas, *APIs*, Microsserviços, etc.), para otimizar as atividades da PGE/RJ;
Ação 11	Prospectar e implantar iniciativas que otimizem a atuação da Dívida Ativa com o uso de novas tecnologias (*BI, IA, Big Data Analytics* e *Data Quality*);

4 DA JUSTIFICATIVA DA NECESSIDADE DA AQUISIÇÃO

A implementação de Modelos de Linguagem de Grande Escala (LLMs), como o Llama 3, apresenta diversas vantagens que justificam a aquisição de uma solução de IA para a Procuradoria Geral do Estado do Rio de Janeiro (PGE-RJ). A seguir, são elencados os principais benefícios:

a. **Eficiência:** A automação de tarefas repetitivas, como a análise de documentos e a extração de informações relevantes, se traduz em ganhos significativos de tempo e recursos. LLMs podem redigir minutas de documentos jurídicos, resumir processos extensos e realizar triagem de informações, otimizando o trabalho dos procuradores;

b. **Precisão:** A utilização de LLMs contribui para a redução de erros humanos em análises jurídicas complexas. A capacidade dos modelos de compreender e processar grandes volumes de texto resulta em maior precisão na interpretação de dados e na geração de relatórios, proporcionando uma base sólida para a tomada de decisões;

c. **Produtividade:** A liberação de recursos humanos para atividades estratégicas e de maior valor agregado é um dos grandes benefícios da implementação de LLMs. Com os modelos cuidando de tarefas mecânicas e repetitivas, os profissionais podem se concentrar em atividades que exigem julgamento crítico e expertise jurídica, elevando a qualidade do trabalho desenvolvido pela PGE-RJ;

d. **Inovação:** A capacidade de explorar novas abordagens e soluções para desafios jurídicos é ampliada com o uso de LLMs. Estes modelos podem identificar padrões e insights que não seriam facilmente percebidos por humanos, abrindo caminho para novas estratégias legais e contribuindo para a evolução das práticas jurídicas;

e. **Tomada de Decisão:** O suporte avançado para a tomada de decisões com base em grandes volumes de dados é uma vantagem crucial. Ao analisar vastos conjuntos de informações, LLMs fornecem uma base sólida para decisões mais informadas e precisas, aumentando a eficiência e a eficácia das atividades da PGE-RJ.

A implementação de LLMs como o Llama 3 é, portanto, fundamental para melhorar a eficiência, precisão, produtividade, inovação e suporte à tomada de decisão nas atividades da

Procuradoria Geral do Estado do Rio de Janeiro. Essa aquisição é essencial para o cumprimento das finalidades institucionais e para a continuidade do serviço público, em consonância com os princípios da administração pública e a probidade administrativa.

5 DO LEVANTAMENTO DAS POSSÍVEIS SOLUÇÕES

5.1 SERVIDOR COM GPUS DEDICADAS (ON PREMISE)

A solução de servidores com GPUs dedicadas on premise consiste na aquisição de hardware específico para processamento de inteligência artificial, instalado e mantido nas dependências da PGE-RJ. Esta opção permite maior controle sobre os dados e o processamento, além de reduzir a latência, uma vez que todo o processamento é realizado localmente.

Para estimar o valor desta possível solução, duas propostas de orçamentos compatíveis com as especificações supracitadas foram levantadas, são elas:

EMPRESA	MODELO DO SERVIDOR	VALOR
DECISION/UNITECH	DELL PE XE9680	R$ 3.784.659,74
NEWROUTE	HPE Cray Supercomputing XD670	U$ $577,303.00
F2IT	DELL PowerEdge R760XA	R$ 3.894.158,61

** As propostas detalhadas estão disponíveis em anexo*

Vantagens:

1. **Segurança de Dados:** Maior controle sobre os dados sensíveis e confidenciais, evitando possíveis riscos associados ao armazenamento e processamento em ambientes externos;

2. **Latência Reduzida:** Processamento local diminui a latência, proporcionando respostas mais rápidas, essenciais para análises em tempo real;

3. **Customização:** Possibilidade de configurar e otimizar o hardware e software de acordo com as necessidades específicas da PGE-RJ.

Desvantagens:

1. **Custo Inicial Elevado:** Aquisição e instalação de servidores e GPUs dedicadas representam um investimento inicial significativo;

2. **Manutenção:** Necessidade de equipe especializada para a manutenção e atualização do hardware e software, gerando custos adicionais;

3. **Escalabilidade:** Expansão da capacidade de processamento requer novos investimentos em hardware, o que pode ser menos flexível comparado a soluções na nuvem.

5.2 PROCESSAMENTO EM NÚVEM (CLOUD COMPUTING)

A solução em nuvem para processamento de IA envolve o uso de serviços de computação fornecidos por terceiros, como Amazon Web Services (AWS), Google Cloud Platform (GCP) ou Microsoft Azure. Esses serviços oferecem recursos escaláveis e flexíveis para processamento de grandes volumes de dados com modelos de linguagem de grande escala.

Para estimar o valor desta possível solução, duas propostas de orçamentos compatíveis com as especificações supracitadas e similares em hardware a solução on premise foram levantadas, são elas:

EMPRESA	INSTÂNCIA	CUSTO MENSAL	CUSTO ANUAL
AMAZON (AWS)	AMAZON EC2 p5.48xlarge	U$ 39.380,19	U$ 472.562,28
GOOGLE	a3-highgpu-8g	U$ 28,353.20	U$ 340.238,40

** As propostas detalhadas estão disponíveis em anexo*

Vantagens:

1. **Escalabilidade:** Capacidade de aumentar ou diminuir os recursos de processamento conforme a demanda, sem necessidade de investimento inicial em hardware;

2. **Custo Variável:** Modelo de pagamento conforme o uso, permitindo melhor gerenciamento dos custos operacionais;

3. **Manutenção:** Redução da necessidade de manutenção interna, pois os provedores de serviços na nuvem são responsáveis pela manutenção e atualização dos servidores.

Desvantagens:

1. **Segurança de Dados:** Dependência de políticas de segurança dos provedores de nuvem, o que pode representar riscos adicionais em comparação com soluções *on premise*;

Latência: Possibilidade de maior latência devido à necessidade de transferência de dados entre a PGE-RJ e os servidores na nuvem;

2. **Dependência de Terceiros:** Dependência de terceiros para a continuidade dos serviços, o que pode representar riscos em caso de interrupções ou mudanças nos termos de serviço;

3. **Processos Sigilosos:** Processamento de dados sensíveis e sigilosos em servidores de terceiros pode representar riscos adicionais em termos de confidencialidade e conformidade com regulamentações.

6 DA JUSTIFICATIVA PARA A ESCOLHA DA SOLUÇÃO

De acordo com as soluções apresentadas, observa-se que a solução on premise (5.1) com servidores dedicados com GPUs é a mais adequada para o atendimento das necessidades da PGE-RJ.

Os motivos que justificam a escolha são:

1. **Segurança de Dados:** A solução on premise oferece maior controle sobre os dados sensíveis e sigilosos, evitando os riscos associados ao armazenamento e processamento em ambientes externos. Este controle é essencial para garantir a confidencialidade e conformidade com regulamentações específicas da PGE-RJ.

2. **Latência Reduzida:** Com o processamento sendo realizado localmente, a latência é significativamente reduzida, proporcionando respostas mais rápidas, essenciais para análises em tempo real e eficiência nas operações diárias.

3. **Customização:** A solução on premise permite configurar e otimizar o hardware e software de acordo com as necessidades específicas da PGE-RJ, garantindo um desempenho alinhado às demandas internas.

4. **Continuidade e Controle:** A continuidade dos serviços não depende de terceiros, eliminando riscos de interrupções ou mudanças nos termos de serviço que poderiam impactar as operações da PGE-RJ. Além disso, o controle total sobre a infraestrutura facilita a gestão e manutenção dos recursos.

Diante da análise das necessidades e recursos disponíveis pela PGE-RJ, a solução on premise com servidores dedicados com GPUs se apresenta como a melhor estratégia de implementação para garantir a segurança, eficiência e eficácia das operações da Procuradoria, em consonância com os princípios da administração pública e a probidade administrativa.

7 DOS RESULTADOS PRETENDIDOS

A solução adquirida deverá permitir os seguintes resultados principais:

1. **Eficiência Operacional**:
   - Automação de tarefas repetitivas.
   - Redução de erros humanos.
2. **Tomada de Decisão**:
   - Processamento de grandes volumes de dados.
   - Geração de insights valiosos.
3. **Inovação**:
   - Desenvolvimento de novas soluções tecnológicas.
   - Fine-tuning personalizado dos modelos de IA.
4. **Segurança**:
   - Maior controle sobre dados sensíveis.
   - Alinhamento com normas vigentes.

Esses resultados garantirão a excelência das atividades da PGE-RJ.

7.1. Dos Possíveis Impactos Ambientais

O presente **Estudo Técnico Preliminar (ETP)** não efetuou análise de possíveis impactos ambientais e medidas mitigadoras, incluindo requisitos de baixo consumo de energia e logística reversa, pelos seguintes motivos:

7.1.1. Escala e Natureza da Contratação: A aquisição se limita a um único servidor de Inteligência Artificial, cuja escala reduzida não representa impacto ambiental significativo. Equipamentos desse tipo são projetados para operar de maneira eficiente, e o volume de resíduos gerados ao longo de seu ciclo de vida é mínimo, especialmente considerando sua longa durabilidade.

7.1.2. Conformidade com Padrões Ambientais de Mercado: Os servidores disponíveis no mercado já são fabricados em conformidade com regulamentações ambientais obrigatórias, como eficiência energética e controle de substâncias perigosas (RoHS - Restriction of Hazardous Substances).

7.1.3. Irrelevância de Impactos Ambientais no Contexto Geral: O impacto ambiental gerado por um único equipamento é insignificante quando comparado ao consumo energético e à geração de resíduos de toda a infraestrutura tecnológica da organização. Portanto, uma análise detalhada de mitigação de impactos seria desproporcional.

Diante das razões apresentadas, a análise de possíveis impactos ambientais e respectivas medidas mitigadoras foi considerada desnecessária para o presente Estudo Técnico Preliminar. Essa decisão assegura a proporcionalidade e a eficiência no planejamento da contratação.

8 DA JUSTIFICATIVA DE PARCELAMENTO OU NÃO

A configuração de um servidor de IA de alto desempenho envolve a integração de diversos componentes críticos, como chassis, processadores, unidades de processamento gráfico (GPUs), memória RAM, armazenamento, e sistemas de resfriamento e alimentação. A sinergia entre esses componentes é essencial para garantir a performance, estabilidade e segurança do sistema, conforme as seguintes considerações:

8.1 Integração e Compatibilidade: Os componentes de hardware e software devem ser compatíveis e otimizados para operar de forma conjunta. A compra de peças separadas de diferentes fornecedores pode resultar em problemas de compatibilidade, aumentando o risco de falhas no sistema;

8.2 Desempenho: A performance dos sistemas de IA é altamente dependente da arquitetura integrada. Por exemplo, a eficiência no processamento paralelo, crucial para operações de IA, requer uma configuração otimizada e testada entre CPUs e GPUs;

8.3 Suporte e Manutenção: A contratação de um único fornecedor para o sistema integrado facilita o suporte técnico e a manutenção, permitindo uma resolução de problemas mais rápida e eficaz;

8.4 Economia de Escala: A aquisição de todos os componentes de um único fornecedor possibilita a obtenção de melhores condições comerciais, incluindo descontos significativos devido à compra em volume, além de menores custos logísticos e de transporte;

8.5 Redução de Custos de Gestão: A gestão de contratos e fornecedores é simplificada ao tratar com um único fornecedor, o que reduz custos administrativos e a complexidade da gestão de múltiplos contratos;

8.6 Custos de Integração: A aquisição de componentes separados poderia necessitar de gastos adicionais com especialistas para integrar e testar os sistemas, além de possíveis custos extras em caso de necessidade de adaptações ou soluções para problemas de compatibilidade;

8.7 Justificativa Legal: Com base no artigo 40 da Lei 14.133/2021, o princípio do parcelamento é recomendado quando tecnicamente viável e economicamente vantajoso. No entanto, conforme estabelecido nos incisos do § 3º do mesmo artigo, o parcelamento não será adotado nas seguintes condições, todas aplicáveis ao presente caso:

- *Inciso I: A economia de escala e a redução de custos de gestão recomendam a compra do item de um mesmo fornecedor.*

- *Inciso II: O objeto a ser contratado configura um sistema único e integrado, cuja integridade pode ser comprometida pelo parcelamento, devido ao risco de incompatibilidade e complexidade na integração dos componentes.*

8.8 Conclusão: Diante das justificativas técnicas, econômicas e legais apresentadas, conclui-se que o **parcelamento do objeto da licitação para a aquisição de um servidor de IA com GPUs não é tecnicamente viável nem economicamente vantajoso**. Portanto, recomenda-se a contratação de um único fornecedor para fornecer a solução completa e integrada, assegurando a funcionalidade, desempenho e segurança do sistema.

9 DO PARECER QUANTO A VIABILIDADE TÉCNICA DA CONTRATAÇÃO

A análise técnica realizada neste estudo considerou que a contratação pretendida é viável desde que siga as premissas estabelecidas neste Estudo Técnico Preliminar.

No decorrer do processo de aquisição, novas informações técnicas poderão ser adicionadas ao processo. Neste caso, a equipe responsável por sua elaboração deverá ser comunicada e o ETP pode ser readequado se houver mudança de contexto ou o surgimento de novas tecnologias que invalidem o estudo atual. Esta nova versão do ETP deverá ser anexada ao processo administrativo da aquisição e deverá indicar claramente quais itens foram readequados.

Todavia, essas melhorias devem estar devidamente registradas e aderentes às instruções administrativas e jurídicas do órgão.

10 DA LEI DE ACESSO A INFORMAÇÃO (LEI Nº 12.527/11)

O Estudo Técnico Preliminar (ETP) para a aquisição de um servidor de Inteligência Artificial (IA) **não necessita de classificação** nos termos da **Lei nº 12.527/2011** (Lei de Acesso à Informação), pois seu conteúdo é essencialmente técnico e administrativo, sem informações estratégicas, sensíveis ou protegidas por sigilo.

O documento aborda especificações técnicas padronizáveis, análises de viabilidade e justificativas orçamentárias, todas de natureza pública e sem risco de comprometer a segurança institucional, comercial ou pessoal. Não há dados que se enquadrem nas hipóteses de restrição previstas nos arts. 23 e 24 da LEI, como informações que afetem a segurança nacional, a privacidade ou o sigilo industrial.

Dessa forma, o ETP deve ser amplamente divulgado, promovendo transparência, controle social e eficiência, em conformidade com os princípios constitucionais da publicidade e da eficiência administrativa.

11 DA CONCLUSÃO

Este estudo deverá ser submetido à apreciação da Gerente de Tecnologia da Informação para validar e aprovar este documento técnico, buscando atender aos interesses da PGE-RJ.

Rio de Janeiro, 27 de Novembro de 2024.

gov.br Documento assinado digitalmente
CELSO ARAUJO FONTES
Data: 27/11/2024 16:38:16-0300
Verifique em https://validar.iti.gov.br

Demandante: Celso Araujo Fontes
Id Funcional: 4334665

gov.br Documento assinado digitalmente
WESCLEY BARBOSA DE PAIVA DE CARVALHO
Data: 28/11/2024 08:40:47-0300
Verifique em https://validar.iti.gov.br

Responsável Técnico: Wescley Barbosa de Paiva de Carvalho
Id Funcional: 50286820