

## Especificação Técnica

### **Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços**

Elaborada Exclusivamente para a PROCempa junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

## Requisitos Técnicos

### 1. OBJETO

Registro de preços para eventual contratação de **plataforma integrada de Inteligência Artificial**, com fornecimento de **infraestrutura software** para execução de modelos de linguagem de grande porte, combinando:

- 1.1. Execução local (on-premise): Modelos com até 120 bilhões (120B) de parâmetros
- 1.2. Execução em nuvem (cloud): Modelos com 350 bilhões (350B) de parâmetros ou superiores
- 1.3. Acesso sob demanda a plataformas, modelos e ferramentas baseadas em IA
- 1.4. Ambiente integrado para desenvolvimento, teste e produção de soluções digitais

Visando atender às necessidades tecnológicas da PROCempa no apoio a processos internos, desenvolvimento de soluções digitais e modernização da gestão pública.

### 2. JUSTIFICATIVA

A crescente demanda por inovação, automação e eficiência nas atividades desempenhadas pela PROCempa exige a adoção de soluções tecnológicas baseadas em Inteligência Artificial de última geração.

A **arquitetura híbrida (local + cloud)** justifica-se por:

- 2.1. Soberania de dados: Processamento local de informações sensíveis e críticas
- 2.2. Baixa latência: Respostas em tempo real para aplicações internas
- 2.3. Escalabilidade: Acesso a modelos ultra-capazes (350B+) via cloud para tarefas complexas
- 2.4. Economicidade: Custo-benefício otimizado entre infraestrutura própria e serviços sob demanda (1/8)
- 2.5. Continuidade operacional: Redundância e resiliência entre ambientes

A contratação por meio de **Ata de Registro de Preços** permitirá aquisição sob demanda, de forma ágil, escalável e com melhor custo-benefício, garantindo aderência à dinâmica das necessidades institucionais.

|  |  |  |  |
|--|--|--|--|
|  |  |  |  |
|--|--|--|--|

## Especificação Técnica

### Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços

Elaborada Exclusivamente para a PROCEMPA junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

### 3. ESCOPO DOS SERVIÇOS

A solução deve fornecer acesso às seguintes capacidades:

#### 3.1. Infraestrutura Local (On-Premise)

##### Software de gerenciamento local:

- 3.1.1. Orquestração de modelos: Plataforma para deploy, versionamento e escalonamento de múltiplos modelos
- 3.1.2. APIs de inferência: Endpoints REST/gRPC compatíveis com OpenAI API, vLLM, ou TGI (Text Generation Inference)
- 3.1.3. Quantização: Suporte a GGUF (*GPT-Generated Unified Format*), GPTQ (Generative Pre-trained Transformer Quantization), AWQ (*Activation-aware Weight Quantization*), EXL2 (ExLlamaV2) para otimização de memória
- 3.1.4. Monitoring: Dashboard de uso de GPU, memória, throughput (tokens/s), latência
- 3.1.5. Security: Autenticação, autorização, logging de auditoria, criptografia em repouso e trânsito

#### 3.2. Infraestrutura em Nuvem (Cloud)

##### Acesso a modelos 350B+ parâmetros:

- 3.2.1. Modelos disponíveis: GPT-4o, Claude 3.5/4 Opus, Gemini Ultra, ou equivalentes com 350B+ parâmetros
- 3.2.2. Throughput garantido: Mínimo 1000 requisições/minuto (RPM) com rate limiting configurável (2/8)
- 3.2.3. Latência: P95 < 3 segundos para respostas de até 4096 tokens
- 3.2.4. Disponibilidade (SLA): Mínimo 99,9% uptime mensal
- 3.2.5. Região preferencial: Data centers localizados no Brasil ou América do Sul

##### Integração híbrida:

- 3.2.6. Failover automático: Roteamento inteligente entre local e cloud baseado em disponibilidade, custo e sensibilidade dos dados
- 3.2.7. Unified API: Interface única para acesso a modelos locais e cloud, abstraindo complexidade
- 3.2.8. Cache distribuído: Redução de chamadas à cloud via cache de respostas frequentes

|  |  |  |  |
|--|--|--|--|
|  |  |  |  |
|--|--|--|--|

## **Especificação Técnica**

### **Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços**

Elaborada Exclusivamente para a PROCempa junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

#### **3.3. Plataformas de IA Generativa**

- 3.3.1. Texto: Geração, sumarização, tradução, reescrita, análise de sentimentos
- 3.3.2. Código: Geração, refatoração, debug, documentação, revisão de código
- 3.3.3. Imagens: Geração, edição, análise, OCR inteligente
- 3.3.4. Áudio: Transcrição, síntese de voz, análise de sentimentos em áudio
- 3.3.5. Vídeo: Análise, sumarização, geração de legendas

#### **3.4. Interfaces de Acesso**

- 3.4.1. Web UI: Interface web responsiva para usuários não-técnicos
- 3.4.2. API REST/GraphQL: Para integração com sistemas existentes da PROCempa
- 3.4.3. SDKs: Python, JavaScript/TypeScript, Java, .NET
- 3.4.4. CLI: Ferramenta de linha de comando para automação e scripts
- 3.4.5. ChatOps: Integração com Slack, Microsoft Teams, Mattermost

#### **3.5. Automação e Agentes**

- 3.5.1. Assistentes virtuais: Chatbots para atendimento interno e externo
- 3.5.2. Agentes autônomos: Execução de tarefas multi-step com ferramentas (web search, database, APIs)
- 3.5.3. Workflow automation: Integração com ferramentas de RPA e BPM
- 3.5.4. Knowledge retrieval: RAG (Retrieval-Augmented Generation) com bases documentais da PROCempa

#### **3.6. Desenvolvimento de Software**

- 3.6.1. Code completion: Autocompletar inteligente em IDEs (VS Code, JetBrains)
- 3.6.2. Code review: Análise automática de segurança, qualidade, boas práticas
- 3.6.3. Test generation: Geração de testes unitários, de integração, E2E (3/8)
- 3.6.4. Documentation: Geração automática de documentação de código e APIs
- 3.6.5. Legacy modernization: Análise e sugestão de refatoração de código legado

#### **3.7. Análise de Dados**

- 3.7.1. Análise preditiva: Modelos para forecasting, classificação, clustering
- 3.7.2. Sumarização: Documentos longos, relatórios, transcrições de reuniões
- 3.7.3. Categorização: Classificação automática de documentos, tickets, solicitações
- 3.7.4. Extração de entidades: NER (Named Entity Recognition) para dados estruturados
- 3.7.5. Anomaly detection: Detecção de anomalias em logs, métricas, transações

## Especificação Técnica

### Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços

Elaborada Exclusivamente para a PROCempa junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

#### 3.8. Treinamento Customizado

- 3.8.1. Fine-tuning: Ajuste fino de modelos open-source com dados da PROCempa
- 3.8.2. RAG/Embeddings: Indexação de documentos corporativos para retrieval
- 3.8.3. Prompt engineering: Biblioteca de prompts otimizados para casos de uso da PROCempa
- 3.8.4. Evaluation: Framework para avaliação de qualidade, viés, alucinação

#### 3.9. Suporte e Acompanhamento

- 3.9.1. Suporte técnico: Remoto especializado, 24/7 para incidentes críticos
- 3.9.2. Treinamento: Capacitação de equipes técnicas e usuárias
- 3.9.3. Documentação: Manuais, tutoriais, exemplos de código, FAQs
- 3.9.4. Updates: Atualizações regulares de modelos, segurança, performance

## 4. ESPECIFICAÇÕES TÉCNICAS DETALHADAS

#### 4.1. Requisitos de Performance (Local)

| Métrica                 | Mínimo Requerido                   |
|-------------------------|------------------------------------|
| Throughput (120B, INT4) | $\geq 20$ tokens/segundo           |
| Latência P50 (120B)     | $\leq 2$ segundos (primeiro token) |
| Latência P95 (120B)     | $\leq 5$ segundos (primeiro token) |
| Context window (local)  | Mínimo 128K tokens                 |
| Context window (cloud)  | Mínimo 256K tokens                 |
| Concorrência (local)    | Mínimo 50 requisições simultâneas  |
| Uptime (local)          | 99,5% (infraestrutura própria)     |
| Uptime (cloud)          | 99,9% (SLA do provedor)            |

## **Especificação Técnica**

### **Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços**

Elaborada Exclusivamente para a PROCempa junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

#### **4.2. Requisitos de Segurança**

- 4.2.1. Criptografia: TLS 1.3 em trânsito, AES-256 em repouso (4/8)
- 4.2.2. Autenticação: OAuth 2.0, OIDC, SAML 2.0, integração com AD/LDAP da PROCempa
- 4.2.3. Autorização: RBAC (Role-Based Access Control) granular
- 4.2.4. Audit logging: Logs imutáveis de todas as requisições (quem, quando, o quê)
- 4.2.5. Network isolation: VPC dedicada, security groups, firewall de aplicação
- 4.2.6. Vulnerability scanning: Varredura regular de vulnerabilidades e patches
- 4.2.7. Compliance: Certificações SOC 2 Type II, ISO 27001, ou equivalentes

#### **4.3. Requisitos de Monitoramento e Observabilidade**

- 4.3.1. Metrics: Coleta de métricas de performance, uso, erro (Prometheus-compatible)
- 4.3.2. Logging: Centralização de logs com busca full-text (ELK/Loki-compatible)
- 4.3.3. Tracing: Distributed tracing para requisições entre local e cloud (OpenTelemetry)
- 4.3.4. Alerting: Alertas configuráveis via PagerDuty, OpsGenie, Slack, email
- 4.3.5. Dashboards: Painéis em tempo real de saúde do sistema (Grafana-compatible)

#### **4.4. Requisitos de Escalabilidade**

- 4.4.1. Auto-scaling (local): Escalonamento horizontal baseado em demanda (Kubernetes)
- 4.4.2. Auto-scaling (cloud): Burst para cloud em picos de demanda
- 4.4.3. Load balancing: Distribuição inteligente de carga entre instâncias
- 4.4.4. Rate limiting: Proteção contra abuso e garantia de QoS
- 4.4.5. Quota management: Limites de uso por usuário, projeto, departamento

### **5. POLÍTICA DE DADOS, PRIVACIDADE E SOBERANIA**

#### **5.1. Regime de Retenção (Zero Retention)**

- 5.1.1. Política Obrigatória: As ferramentas contratadas devem operar obrigatoriamente sob política de Retenção Zero para Treinamento.
- 5.1.2. Privacidade: Prompts, logs, documentos anexados e resultados gerados pela PROCempa não podem ser armazenados pelos provedores de modelos (OpenAI, Google, Anthropic, etc.) para fins de treinamento, melhoria de modelos públicos ou qualquer uso que não seja a prestação do serviço contratado.
- 5.1.3. Armazenamento Técnico: Permitido apenas em caráter técnico para fins de segurança e auditoria da própria contratante, com retenção máxima de 90 dias.

|  |  |  |  |
|--|--|--|--|
|  |  |  |  |
|--|--|--|--|

## **Especificação Técnica**

### **Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços**

Elaborada Exclusivamente para a PROCempa junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

#### **5.2. Data Residency (Residência de Dados)**

- 5.2.1. Preferência Nacional: Soluções com residência de dados no Brasil terão preferência na avaliação.
- 5.2.2. Soberania de Dados: O fornecedor deve priorizar armazenamento de logs e processamento de inferências em instâncias localizadas em território nacional.
- 5.2.3. Processamento Internacional: Caso o modelo específico exija processamento em data centers internacionais, a contratada deve comprovar que a transferência de dados possui:
  - 5.2.3.1. Criptografia de ponta-a-ponta (TLS 1.3+)
  - 5.2.3.2. Conformidade com LGPD e mecanismos de transferência internacional adequados (Cláusulas Contratuais Padrão, Binding Corporate Rules)
  - 5.2.3.3. Garantia de não retenção para treinamento

#### **5.3. Conformidade Legal**

- 5.3.1. LGPD: Adequação total à Lei Geral de Proteção de Dados (Lei 13.709/2018)
- 5.3.2. Segurança: Ambiente em nuvem com segurança de nível empresarial
- 5.3.3. Privacidade: Políticas de privacidade transparentes e acessíveis
- 5.3.4. Governança: Nomeação de DPO (Data Protection Officer) acessível à PROCempa

## Especificação Técnica

### Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços

Elaborada Exclusivamente para a PROCempa junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

#### 6. ESTIMATIVA DE CONSUMO

| Item | Descrição do Serviço                                                                             | Unidade de Medida     | Quantidade Estimada Anual | Observações sobre Métricas e Custos                                                                                                                                                                                                                                                                                        |
|------|--------------------------------------------------------------------------------------------------|-----------------------|---------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | Acesso a Plataforma de IA (API e interface web)<br>- Modelo base GPT, Claude, Gemini ou similar. | Assinatura Mensal     | 100 unidades              | Custo Fixo: Plano com custo fixo recorrente por acesso disponibilizado, independentemente do volume de uso individual, desde que respeitados os termos da plataforma contratada.                                                                                                                                           |
| 2    | Créditos adicionais de uso sob demanda (expansão além do pacote básico).                         | Pacote (mil unidades) | 500 pacotes               | Total Estimado: 500 mil unidades.<br>- Métricas Permitidas: As unidades podem ser medidas tanto em "tokens" quanto em "horas de processamento". A licitante deve definir o valor unitário para cobrir os custos da métrica escolhida (processamento ou volume de dados).<br>- Modelos: GPT, Claude, Gemini e/ou similares. |
| 3    | Suporte técnico especializado (atendimento remoto).                                              | Hora Técnica          | 120 horas                 | -                                                                                                                                                                                                                                                                                                                          |
| 4    | Treinamento de modelo customizado.                                                               | Serviço               | 6 serviços anuais         | Refere-se à customização (fine-tuning ou RAG) de modelos existentes com dados da PROCempa para tarefas específicas, e não à criação de um novo LLM do zero.                                                                                                                                                                |

## Especificação Técnica

### Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços

Elaborada Exclusivamente para a PROCempa junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

| Item                                      | Unidade  | Estimativa Mensal | Estimativa Anual |
|-------------------------------------------|----------|-------------------|------------------|
| <b>Tokens (local)</b>                     | milhões  | 50M               | 600M             |
| <b>Tokens (cloud)</b>                     | milhões  | 100M              | 1.2B             |
| <b>Horas de GPU (local)</b>               | horas    | 720h (24/7)       | 8.760h           |
| <b>Requisições API</b>                    | milhares | 500K              | 6M               |
| <b>Armazenamento<br/>,lol(embeddings)</b> | GB       | 100GB             | -                |
| <b>Usuários ativos</b>                    | usuários | 200               | -                |

*Nota: Valores estimados baseados em projeção de uso. A contratação será por demanda, conforme necessidade real.*

## 7. VIGÊNCIA da ATA e do CONTRATO

- 7.1. A Ata de Registro de Preços terá vigência de 12 (doze) meses, contados a partir da data de sua assinatura, podendo ser prorrogada por igual período, conforme previsto no art. 87, § único, inciso IV do Regulamento Interno de Licitações e Contratos da PROCempa.
- 7.2. O Contrato decorrente da Ata terá vigência inicial de 12 (doze) meses, podendo ser prorrogado por até 60 (sessenta) meses, conforme previsto no art. 71 da Lei 13.303/16

## 8. FORMA DE PAGAMENTO

- 8.1. Os pagamentos serão efetuados conforme a efetiva contratação de cada item, mediante:
  - 8.1.1. Apresentação de Nota Fiscal
  - 8.1.2. Relatório de execução (consumo de tokens, horas de GPU, requisições)
  - 8.1.3. Aprovação pela equipe fiscalizadora da PROCempa
  - 8.1.4. Respeito aos prazos estabelecidos no contrato decorrente da Ata
- 8.2. Faturamento: Mensal, com medição baseada em uso real (pay-as-you-go) ou pacotes pré-pagos com desconto volumétrico.

|  |  |  |  |
|--|--|--|--|
|  |  |  |  |
|--|--|--|--|

## Especificação Técnica

### **Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços**

Elaborada Exclusivamente para a PROCEMPA junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

## **9. CONDIÇÕES DE PARTICIPAÇÃO E EXECUÇÃO**

### **9.1. Capacidade Técnica**

A Contratada deverá comprovar capacidade técnica e operacional para:

- 9.1.1. Operação de plataformas de IA em escala enterprise
- 9.1.2. Integração de sistemas híbridos (local + cloud)
- 9.1.3. Suporte técnico especializado 24/7

### **9.2. Grupos Econômicos**

Para fins de pontuação e habilitação/qualificação técnica, serão aceitos atestados de capacidade técnica emitidos para empresas do **mesmo grupo econômico** da licitante, desde que comprovada a aptidão para prestar os serviços objeto do certame.

### **9.3. Fiscalização**

A execução dos serviços será acompanhada e fiscalizada por equipe designada da PROCEMPA, com:

- 9.3.1. Reuniões de acompanhamento mensais
- 9.3.2. Relatórios de performance e SLA trimestrais
- 9.3.3. Auditorias técnicas semestrais

### **9.4. Penalidades e Sanções**

Em caso de descumprimento de SLA, serão aplicadas penalidades conforme:

- 9.4.1. Uptime < 99%: Crédito de 10% sobre a mensalidade
- 9.4.2. Uptime < 95%: Crédito de 25% sobre a mensalidade
- 9.4.3. Uptime < 90%: Rescisão contratual por justa causa

## **10. RESULTADOS ESPERADOS**

- 10.1. Aumento de produtividade: Automação de  $\geq 40\%$  das atividades operacionais e técnicas (7/8)
- 10.2. Desenvolvimento acelerado: Redução de  $\geq 50\%$  no tempo de desenvolvimento de sistemas
- 10.3. Atendimento melhorado: IA em  $\geq 80\%$  dos canais de atendimento à população
- 10.4. Redução de retrabalho: Melhoria de  $\geq 60\%$  na qualidade da produção técnica
- 10.5. Soberania tecnológica:  $\geq 70\%$  das inferências processadas localmente (dados sensíveis)
- 10.6. Economia de custos: ROI positivo em  $\leq 12$  meses com arquitetura híbrida otimizada

|  |  |  |  |
|--|--|--|--|
|  |  |  |  |
|--|--|--|--|

## **Especificação Técnica**

### **Contratação de Plataforma de IA com Infraestrutura Híbrida– Registro de Preços**

Elaborada Exclusivamente para a PROCEMPA junto ao SEI 25.12.000001575-7

Validade: de 04/08/2026 até 04/08/2027

## **11. CRITÉRIO DE JULGAMENTO**

O julgamento das propostas será realizado pelo Menor Preço Total do Lote (Único), considerando:

- 11.1. Valor ofertado para infraestrutura local
- 11.2. Valor ofertado para acesso cloud (tokens/horas)
- 11.3. Custos de implantação e migração
- 11.4. Custos de suporte e manutenção

Fatores de desempate (na ordem):

- 11.5. Maior percentual de residência de dados no Brasil
- 11.6. Melhor SLA de uptime oferecido
- 11.7. Maior número de modelos disponíveis (local + cloud)
- 11.8. Melhor tempo de resposta técnico (suporte)

## **12. ANEXOS OBRIGATÓRIOS**

- 12.1. Anexo I: Declaração de Qualificação Técnica
- 12.2. Anexo II: Proposta Comercial Detalhada (CAPEX + OPEX)
- 12.3. Anexo III: Arquitetura Técnica Proposta (diagramas)
- 12.4. Anexo IV: Plano de Implantação e Cronograma
- 12.5. Anexo V: Política de Segurança e Privacidade
- 12.6. Anexo VI: SLA Proposto (métricas e penalidades)
- 12.7. Anexo VII: Lista de Modelos Disponíveis (local + cloud)