

LICITAÇÃO ELETRÔNICA 67/2025**Serviços de Inteligência Artificial****ESCLARECIMENTOS**

Questionamento 6: O processo contém duas versões da Especificação Técnica: a de 26/09/2025 ("Contratação de Serviço de Inteligência Artificial", 2 páginas) e a de 04/08/2026 ("Contratação de Plataforma de IA com Infraestrutura Híbrida", 10 páginas). Os esclarecimentos já publicados (Questionamentos 1 a 5) referem-se à numeração da versão de 2025 ("item 8", "Anexo VII").

Confirma-se que a versão de 04/08/2026 substitui integralmente a de 26/09/2025 e que as respostas aos Questionamentos 1 a 5 permanecem válidas quanto à habilitação por declaração (Anexo I.I), ao custo fixo do item 1, à métrica do item 2 e ao alcance do item 4?

Resposta 6: Sim. Para esta licitação prevalece a especificação técnica de 04/08/2026, que substitui a versão anterior para fins de elaboração das propostas. As respostas e esclarecimentos publicados anteriormente permanecem válidos somente naquilo que for compatível com o edital e seus anexos vigentes. Quando houver conflito, prevalecem seus anexos e esta resposta oficial.

Questionamento 7: O Anexo VI e o item 6 da Especificação definem o item 1 como "Assinatura Mensal", com "quantidade estimada anual: 100 unidades". A tabela de consumo do mesmo item 6 estima 200 usuários ativos. A quantidade de 100 refere-se a 100 assinaturas mantidas ao longo de 12 meses (1.200 assinaturas-mês) ou a 100 assinaturas-mês no total do ano? Qual a base a ser utilizada para o valor total do lote?

Resposta 7: A unidade de medição é "assinatura mensal". A quantidade anual estimada de 100 corresponde a 100 unidades de assinatura mensal para fins de composição do preço registrado; não se deve multiplicá-la automaticamente por 12. Os 200 usuários ativos são referência de dimensionamento e não alteram a quantidade estimada nem constituem cobrança adicional. A utilização poderá ser solicitada durante a vigência da ata/contrato, respeitados os quantitativos, os valores registrados e as regras de contratação sob demanda.

Questionamento 8: A Resposta 3 estabelece que o item 1 tem custo fixo "independentemente do volume de uso individual, desde que respeitados os termos da plataforma contratada". A estimativa de consumo prevê 1,2 bilhão de tokens em nuvem por ano. Admite-se que a licitante estabeleça, nos termos da plataforma, política de uso justo por assinatura (limite de mensagens por período, janela de contexto e franquia de geração multimídia), sendo o consumo excedente remunerado pelo item 2? Em caso afirmativo, a política deverá constar da proposta?

Resposta 8: É admitida política de uso justo, desde que apresentada de forma objetiva na proposta e aprovada na contratação. A política deverá informar franquia, limites por período, recursos incluídos, forma de medição, aviso de atingimento e tratamento de excesso. O item 1 não poderá sofrer cobrança automática fora dos termos publicados. O item 2 somente será faturado quando houver solicitação ou autorização da PROCempa, com quantidade e métrica verificáveis; o fornecedor não poderá bloquear ou cobrar excedente sem aviso e autorização conforme o procedimento contratual.

Questionamento 9: O item 3.1 exige software de gerenciamento local para execução de modelos de até 120B de parâmetros, o item 4.1 fixa métricas de desempenho local (20 tokens/s em INT4, 50 requisições simultâneas) e o item 6 estima 8.760 horas de GPU por ano. O Anexo II (Proposta Comercial Detalhada) contém a linha "Hardware / Outros: equipamentos locais" no CAPEX. Os servidores com GPU, a rede e o ambiente Kubernetes para a execução local serão fornecidos e operados pela PROCempa, cabendo à contratada o software de orquestração e inferência? Ou a contratada deverá fornecer também o hardware? Em qualquer caso, qual a configuração de GPU disponível ou esperada?

Resposta 9: Os servidores com GPU, a rede e o ambiente Kubernetes para a execução local serão fornecidos e operados pela PROCempa, cabendo à contratada exclusivamente o software de orquestração e inferência. Portanto, não há necessidade da contratada incluir no CAPEX do Anexo I.II rubrica para aquisição de hardware físico, devendo-se preencher a linha 'Hardware / Outros: equipamentos locais' com valor zero ou omiti-la, conforme a nova configuração de responsabilidade definida neste esclarecimento.

Questionamento 10: Os itens 3.1.1 a 3.1.4 descrevem orquestração, endpoints compatíveis com OpenAI API/vLLM/TGI, quantização e dashboard de GPU. Admite-se o atendimento do item 3.1 por meio de componentes de código aberto (por exemplo vLLM, Kubernetes, Prometheus e Grafana), instalados, configurados e mantidos pela contratada na infraestrutura da PROCempa?

Resposta 10: Sim. É permitida a utilização de software open source, desde que a licença seja compatível com o uso contratado e que a contratada responda por instalação, configuração, atualização, segurança, suporte, documentação, continuidade e correção de vulnerabilidades. A solução deverá manter os requisitos de autenticação, RBAC, auditoria, criptografia, disponibilidade e desempenho. A composição da solução e seus custos deverão constar da proposta técnica e comercial.

Questionamento 11: O item 4.2.7 exige "Certificações SOC 2 Type II, ISO 27001, ou equivalentes", sem indicar o titular da certificação nem o documento a apresentar. O SOC 2 é relatório emitido sobre a organização que opera o serviço; a ISO 27001 certifica o sistema de gestão de uma organização em escopo definido. Atende ao item 4.2.7 a comprovação de que a solução é operada sobre infraestrutura de nuvem certificada SOC 2 Type II e ISO 27001 (Microsoft Azure, região Brazil South), acompanhada da Política de Segurança e

Privacidade da contratada (Anexo V) e das evidências dos controles dos itens 4.2.1 a 4.2.6? Ou exige-se certificação emitida em nome da própria licitante? Em que momento deverá ser apresentada?

Resposta 11: Será aceita certificação própria da licitante, da infraestrutura utilizada ou combinação documental equivalente, desde que o escopo cubra os serviços ofertados e as responsabilidades de cada parte estejam demonstradas. A certificação da nuvem não comprova, sozinha, os controles operados pela contratada. A licitante deverá informar titular, número, escopo, validade, ambiente abrangido e matriz de responsabilidades, além de apresentar evidências quando solicitada. A comprovação será verificada na habilitação, em diligência e durante a execução. O Anexo I.V deve ser completado com o responsável de privacidade e os canais de incidente antes da assinatura.

Questionamento 12: O item 3.2.5 define região Brasil ou América do Sul como "preferencial"; o item 5.2.1 diz que soluções com residência no Brasil "terão preferência na avaliação"; o item 5.2.3 Alameda Vicente Pinzon, 54 - Vila Olímpia, São Paulo - SP, 04547-130 adm@grupodata.com.br www.datagroup.global Código do Documento: XXX-GD-2026-0000 Revisão do Template: 3.0 Data do Template: 04/09/2026 Elaborado por: Erika Diniz Aprovado por: Robson Siqueira Este documento é oficial somente em meio eletrônico. Quando impresso, é uma cópia não controlada, não podendo ser utilizada para realização das atividades. Página: 3/4 admite processamento internacional com salvaguardas. Parte dos modelos citados no item 3.2.1 (Claude, Gemini) não é oferecida pelos respectivos provedores em região brasileira. Confirma-se que a residência de dados no Brasil é critério de preferência e desempate (item 11.5), e não requisito eliminatório, admitindo-se modelos processados em outras regiões desde que atendido o item 5.2.3 (criptografia, LGPD, cláusulas de transferência e não retenção)?

Resposta 12: Sim. A residência no Brasil é preferência técnica e critério de desempate, não é requisito eliminatório geral. A contratada deverá indicar previamente o país, provedores, categorias de dados, finalidade, retenção e medidas complementares.

Questionamento 13: O item 3.2.2 exige "throughput garantido: mínimo 1.000 requisições/minuto" e o item 3.2.3 latência P95 inferior a 3 segundos para respostas de até 4.096 tokens. O throughput de 1.000 RPM é agregado da plataforma ou por modelo? A latência P95 refere-se ao tempo até o primeiro token ou à resposta completa? Como serão medidos para fins do SLA?

Resposta 13: Os 1.000 RPM são capacidade agregada da plataforma contratada, considerando o conjunto de modelos e canais disponibilizados, com possibilidade de rate limiting por usuário, aplicação e modelo. A métrica P95 de até 3 segundos será apurada como tempo até o primeiro token, para requisições de até 4.096 tokens, em janela mensal e ambiente de produção, excluídos indisponibilidade da rede da PROCempa, manutenção previamente comunicada e requisições fora do limite contratado. O fornecedor deverá disponibilizar logs e relatório que permitam reproduzir a medição. O tempo total de geração será acompanhado como indicador adicional, sem substituir o requisito de primeiro token.

Questionamento 14: Os itens 3.4.3 (SDKs em Python, JavaScript/TypeScript, Java e .NET), 3.4.4 (CLI), 3.4.5 (ChatOps com Slack, Microsoft Teams e Mattermost), 3.6.1 (code completion em VS Code e JetBrains), 3.7.1 e 3.7.5 (análise preditiva e detecção de anomalias) e 3.3.5 (vídeo) descrevem capacidades de naturezas distintas. Tais capacidades devem estar disponíveis integralmente na data de assinatura do contrato, ou podem ser disponibilizadas conforme demanda, mediante ordem de serviço, ao longo da vigência? Alguma delas é requisito eliminatório para a aceitação da proposta?

Resposta 14: Todos os recursos descritos como requisitos da solução devem estar disponíveis na plataforma ofertada e aptos para ativação durante a vigência. A implantação ocorrerá conforme plano de trabalho e solicitações da PROCempa; a ativação sob demanda não transforma requisito obrigatório em item opcional nem permite cobrar licença não prevista. A proposta deverá separar claramente funcionalidades incluídas, funcionalidades que dependem de configuração, serviços adicionais do item 2 e prazos de disponibilização. O aceite será realizado por funcionalidade, mediante demonstração e documentação.

Questionamento 15: O item 10.5 estabelece como resultado esperado "≥ 70% das inferências processadas localmente". Trata-se de meta de resultado da PROCempa ou de obrigação contratual da contratada sujeita a penalidade? Em caso de obrigação, como será apurada, considerando que a distribuição entre local e nuvem depende do hardware disponibilizado e das solicitações dos usuários?

Resposta 15: A meta será tratada como indicador de desempenho da solução. Será calculada mensalmente pela razão entre o número de inferências classificadas como sensíveis processadas na infraestrutura local e o total de inferências sensíveis processadas no período, considerando failover e indisponibilidades justificadas. A classificação será definida pela PROCempa no plano de segurança e registrada nos relatórios de auditoria. A contratada deverá fornecer telemetria suficiente para a verificação. O descumprimento injustificado sujeita a contratada a plano de correção e às sanções contratuais aplicáveis, assegurados contraditório e ampla defesa; a meta não cria desconto automático além do que for formalizado no SLA.

Questionamento 16: O Edital veda a subcontratação total ou parcial sem autorização da PROCempa (cláusula I das obrigações e cláusula XVI da minuta de contrato). Admite-se, mediante autorização prévia, a subcontratação da parcela de instalação e operação da infraestrutura local (item 3.1), permanecendo a contratada integralmente responsável pela plataforma, pela API unificada e pelo SLA?

Resposta 16: Sim, exclusivamente para atividades-meio e mediante autorização prévia e expressa da PROCempa. O pedido deverá identificar a subcontratada, escopo, equipe, acessos, localização de dados, controles de segurança e continuidade. É vedada a transferência do núcleo do objeto, da gestão da plataforma ou das obrigações de proteção de dados. A contratada principal permanece integralmente responsável por resultados,

pessoal, sigilo, segurança, incidentes, encargos e cumprimento do contrato. A subcontratação sem autorização caracteriza inadimplemento contratual.

Questionamento 17: O item 3.9.1 exige suporte 24/7 para incidentes críticos e o Anexo VI (SLA) propõe resposta em 15 minutos para críticos. Confirma-se que o regime 24/7 aplica-se exclusivamente a incidentes críticos (indisponibilidade total), sendo o suporte de demais severidades em horário comercial?

Resposta 17: Sim. O atendimento crítico funcionará 24 horas por dia, 7 dias por semana. Consideram-se críticos os incidentes que provoquem indisponibilidade total da plataforma, perda de autenticação geral, incidente de segurança com risco relevante ou interrupção de serviço essencial. Aplicam-se resposta em até 15 minutos e solução ou mitigação em até 4 horas. Incidentes de alta severidade terão resposta em até 1 hora, em horário de atendimento definido no plano operacional, salvo quando escalados para crítico. Os descontos e demais consequências serão os previstos no Anexo VI e no contrato, após validação dos registros de atendimento.

Questionamento 18: O item 5 do Edital estabelece que impugnações e esclarecimentos serão recebidos até 5 dias úteis antes da abertura; o valor estimado da contratação não consta do Edital. O valor estimado é sigiloso nos termos do art. 34 da Lei nº 13.303/2016? Se sim, será divulgado ao final da disputa para fins do item 7.3?

Resposta 18: Sim, o valor estimado é sigiloso, nos termos do art. 34 da Lei nº 13.303/2016. Será divulgado caso seja necessário para fins de negociação.

Questionamento 19: a) Existe contrato semelhante vigente ou recém encerrado?

b) Se sim, qual o número do contrato?

c) Se sim, com qual empresa?

d) Se sim, qual o valor atual do contrato?

Resposta 19: Não existe contrato semelhante vigente ou recém encerrado.