

ERRATA Nº 01 AO TERMO DE REFERÊNCIA

SELEÇÃO PÚBLICA Nº 27/2026 –

Objeto: Contratação de solução integrada de Processamento Computacional de Alto Desempenho (High Performance Computing – HPC) voltada à Inteligência Artificial, pesquisa, desenvolvimento tecnológico e inovação

1. Item 4.5 - versão e escopo geral

ONDE SE LÊ	LEIA-SE
“...observando integralmente as regras oficiais do MLPerf <i>Inference</i> v6.0 e do MLPerf <i>Training</i> v6.0...”	“...observando integralmente as regras oficiais do MLPerf <i>Inference</i> v6.1 e do MLPerf <i>Training</i> v6.1, versões estáveis vigentes na data de publicação desta Errata, com aplicação uniforme das regras, <i>datasets</i> , cenários, divisões, metas de qualidade, auditoria e artefatos oficiais.”

2. Item 4.5(a) - inferência LLM

ONDE SE LÊ	LEIA-SE
“MLPerf <i>Inference</i> v6.0 (Datacenter) para LLMs, incluindo Llama2-70B, Llama3.1-405B, Mixtral-8x7B e DeepSeek-R1...”	“MLPerf <i>Inference</i> v6.1 (Datacenter), Closed Division, para os <i>workloads</i> Llama 3.1 8B, DeepSeek-R1, e GPT-OSS-120B, nos cenários Offline e Server admitidos pela versão, com LoadGen oficial, metas de qualidade, TTFT/TPOT, arquivos <i>summary/detail/accuracy</i> e regras de fair play. O <i>workload Agentic Inference</i> será medido separadamente nos termos do item 4.5.1.”

3. Item 4.5(b) - treinamento

ONDE SE LÊ	LEIA-SE
“MLPerf <i>Training</i> v6.0 para LLMs...”	“MLPerf <i>Training</i> v6.1, empregando exatamente os <i>workloads</i> , <i>datasets</i> , metas de qualidade e regras

ONDE SE LÊ	LEIA-SE
	oficiais da versão. A métrica será time-to-quality e a energia total até a qualidade, sem alteração de pesos, dados ou critério entre proponentes.”

4. Novo item 4.5.1 - IA Agêntica

ONDE SE LÊ	LEIA-SE
Não consta.	4.5.1. <i>Benchmark</i> de IA Agêntica. O proponente deverá executar o <i>benchmark Agentic Inference for MLPerf Inference</i>, conforme a especificação, as regras e a implementação de referência publicadas pelo MLCommons, disponíveis em: https://mlcommons.org/2026/07/agentic-inference-for-mlperf-inference/, na versão oficialmente congelada na data de publicação desta Errata. O <i>benchmark</i> deverá ser executado separadamente para os modelos Kimi K2.6 e Qwen3.6-35B-A3B, ou para os modelos que efetivamente constem da especificação oficial congelada, vedada a combinação dos resultados dos diferentes modelos em uma única pontuação. A execução deverá utilizar as trajetórias, os conjuntos de dados, as configurações dos modelos, os parâmetros de geração, as regras de carga, os critérios de qualidade e os mecanismos de validação definidos na especificação oficial. Deverão ser preservados, em cada trajetória, a ordem dos turnos, as dependências entre as requisições, o histórico acumulado da conversação, o crescimento do contexto e a localidade de sessão necessária à reutilização válida do <i>cache KV</i>. O proponente deverá apresentar a curva de Pareto prevista na metodologia oficial, mediante execuções

ONDE SE LÊ	LEIA-SE
	em níveis fixos de concorrência, reportando, no mínimo: I. tokens de saída por segundo por sistema; II. tokens de saída por segundo por usuário; III. quantidade de trajetórias, turnos e usuários ou conversações concorrentes; IV. tempo total de execução e latência fim a fim das trajetórias, excluídos os atrasos entre turnos quando assim determinado pela metodologia oficial; V. <i>tokens</i> de entrada e de saída processados; VI. configuração completa do sistema submetido ao teste; VII. consumo de energia e eficiência energética, quando a medição estiver prevista ou for compatível com a metodologia oficial; VIII. resultados das verificações de qualidade e exatidão exigidas para cada ponto da curva de Pareto. Cada resultado somente será considerado válido se atender, cumulativamente, aos limites oficiais aplicáveis de comprimento médio das sequências de saída, exatidão durante a execução de desempenho e exatidão autônoma em tarefas, conforme definidos para o modelo e para a versão congelada do benchmark. A execução deverá utilizar o cliente oficial do MLPerf <i>Endpoints</i> e uma interface compatível com a implementação de referência. Serão admitidos servidores, ambientes de execução, bibliotecas e arquiteturas distintos, desde que não sejam alterados o modelo, os pesos, os dados, as

ONDE SE LÊ	LEIA-SE
	trajetórias, os parâmetros, as regras de carga, os limites de qualidade nem os critérios de validação. O proponente deverá entregar os arquivos de configuração, comandos de execução, logs completos, resultados de desempenho e qualidade, versões de software, imagens de contêiner, <i>hashes</i>, variáveis de ambiente, parâmetros de concorrência e demais artefatos necessários à repetição e à reprodução do teste. É vedada a reutilização de cache entre trajetórias quando não autorizada pela especificação oficial, bem como qualquer otimização que utilize conhecimento indevido das respostas ou dos turnos futuros.

5. Substituição integral da Seção 4.6 - Software Stack

Para eliminar inconsistências e versões ultrapassadas, toda a Seção 4.6 deverá ser substituída pelo texto consolidado a seguir.

4.6. Software (SUBSTITUIR PELA REDAÇÃO ABAIXO)

Serão aceitas soluções abertas, proprietárias ou híbridas, desde que atendam integralmente aos requisitos funcionais, de segurança, interoperabilidade, desempenho, reprodutibilidade e portabilidade. O fornecedor deverá entregar, instalar, configurar, documentar, testar e manter uma pilha completa, adequada à arquitetura ofertada e consistente durante homologação e aceitação.

Versões de referência fixadas em 14/09/2026: as versões indicadas a seguir constituem baseline mínimo/fixo desta Errata. Versões posteriores somente serão aceitas pela cláusula de sucessão tecnológica e nunca por escolha unilateral do proponente.

Grupo	Versão fixa de referência	Requisito mínimo
Frameworks principais	PyTorch 2.14.0; TensorFlow 2.21.0; JAX 0.11.1	GPU/CPU, distribuído, mixed precision, profiling e exportação.

Grupo	Versão fixa de referência	Requisito mínimo
Toolkits de aceleradores	CUDA Toolkit 13.3 Update 1 (13.3.1); ROCm 10.0; ou stack equivalente da arquitetura ofertada	Drivers, compiladores, runtimes e bibliotecas coletivas compatíveis.
ML clássico	scikit-learn 1.8.0; XGBoost 3.1.3; LightGBM 4.6.0	CPU obrigatório; aceleração quando disponível.
Dados distribuídos	Apache Spark 4.1.3; Ray 2.54.1; Dask 2026.8.0	Execução distribuída, integração com aceleradores e observabilidade.
LLM e IA generativa	Transformers 4.56.1; Diffusers 0.35.2; vLLM 0.17.0	Treino, fine-tuning e serving de LLMs, MoE e multimodais.

Visão/NLP	OpenCV 4.12.0; spaCy 3.8.16; Ultralytics compatível; Detectron2 estável	CPU e acelerador quando suportado.
HPC	SLURM 25.05.9; OpenMPI 5.0.10; Horovod 0.28.1 ou sucessor mantido; biblioteca coletiva nativa	Scheduling, MPI, jobs distribuídos e afinidade/NUMA.
Contêineres/orquestração	Kubernetes 1.34.11; Podman 5.8.7; Enroot 4.2.1/Pyxis 0.24.0 estáveis; Kubeflow 1.10.0	Imagens OCI, isolamento, GPU device plugins e <i>manifests</i> reproduzíveis.
Inferência/portabilidade	ONNX Runtime 1.23.5; Triton Inference Server 2.72.0 ou equivalente; compilador/otimizador nativo	APIs abertas, métricas p50/p95/p99, batching e <i>cache</i> KV.
MLOps/observabilidade	MLflow 3.16.0; Prometheus 3.14.0; Grafana 12.4.11; OpenTelemetry estável	Rastreabilidade, métricas, logs, auditoria e operação desconectada.
IA simbólica	SWI-Prolog 10.0.2; CLIPS 6.4.2; Clingo 5.8.2; Apache Jena 6.2.0; Drools 10.2.0	Regras, ontologias, grafos e integração neuro-simbólica.

Otimização	Optuna 5.0.0; bibliotecas LAPACK 3.12.1 e implementação BLAS estável e otimizada.	Tuning, álgebra linear e execução CPU.
------------	---	--

4.6.1. Aprendizado de Máquina e distribuído

- PyTorch, TensorFlow e JAX nas versões da matriz, com execução CPU e acelerada, comunicação coletiva e tolerância a falhas.
- scikit-learn, XGBoost e LightGBM com execução CPU; aceleração será admitida sem perda de portabilidade.
- Spark, Ray e Dask para ETL, treinamento, *serving* e workflows distribuídos.
- Optuna ou equivalente para otimização de hiperparâmetros.

4.6.2. Aprendizado Profundo, LLM e IA Agêntica

- Transformers e Diffusers para modelos fundacionais, multimodais e generativos.
- vLLM ou servidor equivalente para *serving*, *continuous batching*, *paged KV cache*, *prefix caching*, quantização e API documentada.
- Suporte a tool *calling*, saídas estruturadas, múltiplos turnos, contexto longo, agentes cooperativos e integração com ferramentas.
- Suporte a treinamento distribuído, *fine-tuning*, LoRA/PEFT, MoE e paralelismos de dados, tensor, pipeline, sequência e especialistas.

4.6.3. HPC, contêineres e orquestração

- SLURM e OpenMPI compatíveis entre si e com a arquitetura ofertada.
- Kubernetes com plugin ou operador oficial da arquitetura, isolamento de recursos, quotas, políticas de rede e alta disponibilidade.
- Contêineres OCI, Podman e Enroot/Pyxis ou equivalentes, com execução sem privilégios quando aplicável.
- Templates de jobs, Helm charts/*manifests* e procedimentos de recuperação documentados.

4.6.4. Portabilidade, independência e interoperabilidade

- Exportação de modelos em *Open Neural Network Exchange (ONNX)* ou formato aberto equivalente, sem dependência exclusiva de runtime do fornecedor.

- APIs abertas e alternativas CUDA, HIP/ROCm, SYCL, OpenCL ou tecnologia equivalente conforme a arquitetura.
- Disponibilização de *build* recipes, imagens, *hashes*, licenças, SBOM e espelhos/repositórios aptos a operação desconectada.
- Nenhuma equivalência poderá reduzir cobertura funcional, qualidade, segurança, desempenho ou reprodutibilidade.

4.6.5. Segurança e cadeia de suprimentos

- Correção de CVEs críticas dentro do SLA e processo documentado de atualização, homologação e rollback.
- Assinatura e verificação de imagens e pacotes; SBOM em formato *Software Package Data Exchange (SPDX)* ou CycloneDX da OWASP.
- Princípio do menor privilégio, segregação administrativa, controle de segredos e trilhas de auditoria.
- Desativação de telemetria externa por padrão e operação integral sem conectividade externa.

4.6.6. Critérios de aceitação e evidências

- Matriz de conformidade com versão completa, *build*, licença, origem, *hash*, dependências e comando de verificação.
- Treinamento distribuído com escalabilidade documentada; inferência com p50/p95/p99, QPS/tokens/s, energia e qualidade.
- Teste de exportação/importação de modelo em ambiente externo à plataforma do fornecedor.
- Teste de IA Agêntica com trajetória, ferramentas, falhas, *retry*, *cache KV* e observabilidade.
- Teste de *scheduler*, MPI, contêineres, Kubernetes, *failover* e retomada de *jobs*.
- Entrega de manuais para CPU, aceleradores, operação, atualização, backup, *disaster recovery* e segurança.

4.6.7. Cláusula de sucessão tecnológica do software

Caso componente, versão, pacote ou projeto nomeado na Seção 4.6 seja oficialmente descontinuado, retirado de suporte, substituído ou declarado incompatível com a arquitetura ofertada antes da data final de apresentação das propostas, poderá ser adotado o sucessor oficial do mesmo mantenedor ou alternativa equivalente, desde que: (i) haja ato ou evidência oficial; (ii) a Comissão Técnica publique decisão ou esclarecimento aplicável indistintamente; (iii) sejam preservadas funcionalidade, interoperabilidade, segurança, licenciamento e

portabilidade; (iv) o desempenho seja igual ou superior; (v) sejam mantidas as versões congeladas dos *benchmarks*, *workloads*, *datasets* e critérios de qualidade; e (vi) a substituição não altere pesos, pontuação nem competitividade. É vedada substituição unilateral pelo proponente.

6. Item 5.1.1.1 - MLPerf Storage

ONDE SE LÊ	LEIA-SE
“Deverá apresentar o resultado da execução do benchmark MLPerf Storage, utilizando workload de checkpointing, tamanho de modelo Llama 3.1-70B, com 512 processos ou aceleradores simulados, modelo de submissão Closed, não mais do que oito processos por nó computacional, escrevendo dez checkpoints, limpando caches e lendo dez checkpoints.”	Deverá apresentar o resultado da execução do MLPerf Storage v3.0 (https://github.com/mlcommons/storage/releases/tag/v3.0), utilizando a release oficial v3.0 do repositório MLCommons Storage, com indicação da versão completa do pacote e do commit correspondente. A execução deverá observar a especificação oficial, o modo de acesso aplicável, POSIX ou S3, as regras de reprodutibilidade, os arquivos de configuração, os logs completos e o mecanismo oficial de validação (https://github.com/mlcommons/storage/blob/main/Rules.md). Para o workload de checkpointing (https://github.com/mlcommons/storage/blob/main/checkpointing) , será utilizado o perfil oficial llama3-405b, com 512 processos, correspondente à escala fixada pela Administração. A execução deverá utilizar a configuração de referência da Closed Division, sendo vedada a combinação entre perfil de modelo e número total de processos incompatível com a versão congelada. A distribuição dos 512 processos entre os nós clientes deverá observar as regras oficiais da versão, inclusive os requisitos de densidade de processos por host e de memória agregada suficiente para acomodar o checkpoint. Não será aplicado o limite autônomo de, no máximo, oito processos por nó computacional. Serão executados dez checkpoints de escrita e dez checkpoints de leitura. Entre as fases de escrita e leitura, deverá ser realizada a limpeza dos caches dos nós clientes quando necessária para impedir que os dados da fase de leitura sejam atendidos por conteúdo mantido em cache, devendo o

ONDE SE LÊ	LEIA-SE
	procedimento e os respectivos logs integrar o relatório. O proponente deverá informar o número de nós clientes, o número de processos por nó, a memória disponível por cliente, o modo de acesso, a configuração do sistema de armazenamento, a versão do código, o commit ou tag utilizado e os resultados de desempenho das operações de escrita e leitura. Deverão ser entregues os arquivos de configuração, os logs completos, a descrição do sistema, o relatório de execução e o resultado produzido pelo mecanismo oficial mlpstorage validate. Não será obrigatória a submissão ou publicação oficial do resultado pela MLCommons, sem prejuízo da observância das regras técnicas e da entrega das evidências exigidas pela Administração. A mesma configuração será aplicada uniformemente a todos os proponentes. ”

7. Item 5.1.1.2 - IO500

ONDE SE LÊ	LEIA-SE
“Deverá ser executado o <i>benchmark</i> IO500, no modo <i>Production</i> .”	“Deverá ser executada a suíte IO500 na categoria <i>Production</i> , utilizando a versão estável vigente na data de publicação desta Errata. O relatório identificará edição/lista de referência, tag ou commit, configuração, dependências, número de clientes/processos, resultados válidos de <i>bandwidth</i> e metadados e questionário de reprodutibilidade. Permanecem as execuções com 50% e 100% dos nós, sem extrapolação não prevista.”

8. Item 5.1.3 - relatório de aceite

ONDE SE LÊ	LEIA-SE
“Relatório completo da execução dos <i>benchmarks</i> MLPerf, IO500 e <i>MLPerf Storage</i> .”	“Relatório completo da execução do MLPerf <i>Inference</i> v6.1, MLPerf <i>Training</i> v6.1, <i>Agentic Inference</i> , MLPerf <i>Storage</i> v3.0 e IO500 <i>Production</i> ,

ONDE SE LÊ	LEIA-SE
	com todos os logs, <i>manifests</i> , imagens, <i>hashes</i> , <i>datasets</i> , configurações e medições de energia aplicáveis.”

9. Item 12 - referência interna

ONDE SE LÊ	LEIA-SE
“...cláusula de equivalência tecnológica prevista no item 5.6.6...”	“...cláusula de equivalência tecnológica prevista no item 4.6.6, complementada pela cláusula de sucessão tecnológica do item 4.6.7...”

10. Item 8.2 - erro material

ONDE SE LÊ	LEIA-SE
“Parcela 3: R\$ 319.680.319,68”	“Parcela 3: R\$ 319.680.319,68”

11. Item 4.10.1 - referência de software

ONDE SE LÊ	LEIA-SE
“Todos os <i>softwares</i> previstos neste documento, item 5.6...”	“Todos os <i>softwares</i> previstos neste documento, especialmente na Seção 4.6, e suas dependências...”

12. Cláusula geral de salvaguarda para benchmarks e modelos

Caso *benchmark*, modelo, *dataset* ou *workload* fixado nesta Errata seja oficialmente descontinuado, removido ou substituído pelo mantenedor antes da data final de apresentação das propostas, a Administração não aceitará substituição automática nem unilateral. A Comissão Técnica avaliará a evidência oficial e, se indispensável, publicará esclarecimento ou nova errata identificando o sucessor, versão, *dataset*, regras, impacto nos pesos e prazo adicional, se necessário, de modo uniforme. Na ausência de manifestação formal, prevalecem as versões e modelos expressamente fixados nesta Errata.

13. Referências técnicas

- [1] MLCommons. MLPerf *Inference* v6.1 Documentation. https://docs.mlcommons.org/Inference/index_gh/
- [2] MLCommons. *Agentic Inference* for MLPerf *Inference*, 8 jul. 2026. <https://mlcommons.org/2026/07/Agentic-Inference-for-mlperf-Inference/>
- [3] MLCommons. MLPerf *Storage* v3.0 Results, 1 set. 2026. <https://mlcommons.org/2026/09/mlperf-Storage-v3-0-results/>
- [4] IO500. *Benchmark, Production* and Research Lists. <https://io500.org/benchmark>
- [5] PyTorch Foundation. PyTorch 2.14 Release, 2 set. 2026. <https://pytorch.org/blog/pytorch-2-14-release-blog/>
- [6] TensorFlow. Releases, TensorFlow 2.21.0. <https://github.com/tensorflow/tensorflow/releases>
- [7] JAX. Change log, JAX 0.11.1. <https://docs.jax.dev/en/latest/changelog.html>
- [8] NVIDIA. CUDA Toolkit 13.3 Update 1 Release Notes. <https://docs.nvidia.com/cuda/cuda-toolkit-release-notes/>
- [9] AMD. ROCm 10 release *information*. <https://rocm.blogs.amd.com/blog/archive.html>
- [10] REDDI, Vijay Janapa et al. MLPerf inference benchmark. CoRR abs/1911.02549 (2019). **arXiv preprint arXiv:1911.02549**, 2019.
- [11] MATTSON, Peter et al. MLPerf training benchmark. **Proceedings of Machine Learning and Systems**, v. 2, p. 336-349, 2020.
- [12] MLCOMMONS. *Inference*: Reference Implementations of MLPerf *Inference Benchmarks*. Repositório oficial. Disponível em: <https://github.com/mlcommons/Inference>. Acesso em: set. 2026.
- [14] MLCOMMONS. Call for Submission: Edge *Agentic Inference Benchmark* for MLPerf *Inference* v6.1. jul. 2026. Disponível em: <https://mlcommons.org/2026/07/mlperf-Inference-v61-edge-Agentic/>. Acesso em: set. 2026.
- [15] WORKATO; MLCOMMONS; AMD; INTEL; NVIDIA. MLPerf Enterprise *Agentic Inference Benchmark*. jul. 2026. Disponível em: <https://www.workato.com/the-connector/mlperf-enterprise-Agentic-Inference-benchmark/>. Acesso em: set. 2026.
- [16] MLCOMMONS. MLPerf Client v2.0 Expands AI PC Benchmarking with Image Generation and *Agentic* AI. ago. 2026. Disponível em: <https://mlcommons.org/2026/08/mlperf-client-v2-0/>. Acesso em: set. 2026.

[17]PATIL, S. G. et al. Berkeley Function Calling Leaderboard (BFCL). Berkeley Sky Computing Lab, University of California, Berkeley. Disponível em: <https://gorilla.cs.berkeley.edu/leaderboard.html>. Acesso em: set. 2026.

[18] WANG, Jiayu et al. Act As a Real Researcher: A Suite of Benchmarks Evaluating Frontier LLMs and Agentic Harnesses in Research Lifecycle. **arXiv preprint arXiv:2606.07462**, 2026.

As referências acima têm caráter público e oficial (organização mantenedora dos *benchmarks* MLPerf, consórcio aberto MLCommons, do qual participam múltiplos fabricantes de hardware) ou acadêmico (artigo em repositório de acesso aberto, com metodologia de avaliação de agentes de IA revisável por pares), adotadas com o propósito de assegurar neutralidade tecnológica e ampla concorrência entre proponentes, em conformidade com a cláusula de equivalência tecnológica prevista no item 4.6.6 deste Termo de Referência.

14. Correções redacionais, remissivas e de coerência

A presente errata consolida correções formais, redacionais, remissivas, técnicas e de coerência interna identificadas no Termo de Referência, preservado o objeto originalmente definido. As alterações abaixo deverão ser incorporadas ao texto oficial e lidas em conjunto com o instrumento convocatório e a minuta contratual.

14.1. Nota de Cautela

As correções de natureza jurídica e técnica que alterem obrigação, critério de julgamento, requisito de habilitação, prazo, garantia, *benchmark* ou condição de pagamento devem ser submetidas à validação das áreas jurídica, técnica, orçamentária e de governança competentes antes da publicação. Caso o certame já tenha sido publicado, deverá ser avaliada a necessidade de reabertura do prazo para apresentação de propostas.

14.2. Sumário

ONDE SE LÊ	LEIA-SE
“Classificação do Critério (9.2.2)”	“Classificação dos Critérios (item 9.2.4.1)”

14.3. Item 3 - descrição da solução

ONDE SE LÊ	LEIA-SE
“conforme requisitos detalhados no item 5”	“conforme requisitos detalhados no item 4”

14.4. Item 3 - descrição da solução

ONDE SE LÊ	LEIA-SE
“comissionamento e testes de aceitação (itens 5.1)”	“comissionamento e testes de aceitação (item 5.1)”

14.5. Item 4.1 - requisitos energéticos

ONDE SE LÊ	LEIA-SE
“limite de energia de 3 MW para os nós de processamento, armazenamento, equipamentos de conectividade, etc.”	“limite de potência elétrica de 3 MW para o conjunto formado pelos nós computacionais, sistemas de armazenamento, equipamentos de interconexão e demais componentes de tecnologia da informação incluídos na solução. A proposta deverá declarar a potência máxima demandada, a potência nominal e o consumo energético estimado, com memória de cálculo, premissas de carga e margem de segurança. O uso de ‘etc.’ fica suprimido por indeterminação técnica.”

14.6. Item 4.2 - aceleradores

ONDE SE LÊ	LEIA-SE
“O suporte a FP4 e FP64 é requisitos desejáveis”	“O suporte a operações em FP4 e FP64 constitui requisito desejável”

14.7. Item 4.2 - originalidade dos componentes

ONDE SE LÊ	LEIA-SE
“O fabricante não poderá ter publicado, até a data da entrega, aviso de <i>End of Life</i> (EOL) ou <i>End of Support</i> (EOS) dos componentes ofertados”	“Até a data de entrega da proposta, os componentes ofertados não poderão estar submetidos a aviso de <i>End of Life</i> (EOL) nem <i>End of Support</i> (EOS) cujo cronograma

ONDE SE LÊ	LEIA-SE
	comprometa a disponibilidade de suporte, peças, <i>firmware</i> e correções de segurança durante todo o prazo de garantia.”

14.8. Item 4.2 - nós de serviço

ONDE SE LÊ	LEIA-SE
“2x unidades de CPU; Mínimo de 128 GB por unidade de CPU”	“2 (dois) processadores físicos por nó; memória RAM mínima de 128 GB por processador físico, totalizando, no mínimo, 256 GB por nó”. A opção adotada deverá ser única para evitar dupla interpretação.

14.9. Item 4.2 – Confidential Computing

ONDE SE LÊ	LEIA-SE
“Confidential Computing: A infraestrutura deve suportar a execução de cargas de trabalho em Trusted Execution Environments (TEEs), garantindo a proteção criptográfica de dados sensíveis e modelos proprietários durante a execução. A conformidade com este requisito deverá ser apresentada durante os testes de aceitação (item 5.1) com demonstração funcional de ao menos um workload executado em ambiente TEE. Na ausência de comprovação, o requisito será tratado como não atendido e o item correspondente receberá pontuação zero na avaliação técnica.”	“Confidential Computing: O suporte à execução de cargas de trabalho em Trusted Execution Environments (TEEs) constitui requisito técnico obrigatório da solução. A infraestrutura ofertada deverá garantir a proteção criptográfica de dados sensíveis e modelos proprietários durante a execução. O proponente deverá declarar expressamente, na proposta técnica, o atendimento ao requisito e identificar a tecnologia, os componentes e a arquitetura de TEE ofertados. A proposta que não contemplar o suporte a Confidential Computing será desclassificada por descumprimento de requisito técnico obrigatório. A conformidade será comprovada durante os testes de aceitação previstos no item 5.1, mediante demonstração funcional de, no mínimo, uma carga de trabalho executada em ambiente TEE.

ONDE SE LÊ	LEIA-SE
	Poderá ser empregada TEE implementada em CPU, em acelerador ou em composição entre CPU e acelerador, desde que a proteção durante o processamento seja efetivamente demonstrada e documentada. Caso o requisito tenha sido declarado na proposta, mas a demonstração funcional não seja aprovada nos testes de aceitação, a solução será considerada não conforme e não será aceita, sem prejuízo das medidas previstas no instrumento convocatório e no contrato.”

14.10. Item 4.3 - rede interna

ONDE SE LÊ	LEIA-SE
“Largura de banda agregada mínima de 1 TB/s por nó para comunicação GPU-GPU”	“Largura de banda agregada bidirecional mínima de 1 TB/s por nó para a malha interna de comunicação entre aceleradores, comprovada por documentação oficial do fabricante e teste reproduzível”. Especificar se o valor se refere à soma das direções, à largura de banda unidirecional ou à largura de banda de bisseção.

14.11. Item 4.3 - latência

ONDE SE LÊ	LEIA-SE
“Latência inferior a 1 microssegundo para mensagens curtas entre GPUs de um mesmo nó”	“Latência inferior a 1 μ s para mensagens curtas entre aceleradores de um mesmo nó, medida segundo metodologia, tamanho de mensagem, ferramenta, número de repetições e percentil previamente definidos no Plano de Testes de Aceitação.”

14.12. Item 4.3.2 - switch de borda

ONDE SE LÊ	LEIA-SE
"4 portas 400/100 Gbps QSFP-DD (...) e 32 portas 25/10 Gbps"	"no mínimo 4 (quatro) portas compatíveis com 400 GbE/100 GbE, em formato QSFP-DD ou equivalente, e 32 (trinta e duas) portas compatíveis com 25 GbE/10 GbE, incluindo todos os módulos ópticos, cabos, licenças e recursos necessários à operação simultânea nas velocidades requeridas."

14.13. Item 4.4 - licenciamento

ONDE SE LÊ	LEIA-SE
"A proposta deverá incluir um modelo de licenciamento que assegure ao comprador: licença perpétua e irrevogável para uso dos <i>softwares</i> embarcados essenciais"	A proposta deverá discriminar, por componente, as licenças perpétuas, as subscrições e os direitos de atualização e suporte. Para os softwares embarcados e essenciais à operação do sistema: (i) licença perpétua, caso a natureza jurídica permita; (ii) subscrição pelo prazo equivalente ao dobro do período de garantia.

14.14. Item 4.5 - benchmarks MLPerf

ONDE SE LÊ	LEIA-SE
"MLPerf <i>Inference</i> v6.0 (...) incluindo (...) Llama2-70B, Llama3.1-405B, Mixtral-8x7B e <i>DeepSeek-R1</i> "	"MLPerf <i>Inference</i> , na versão oficial expressamente indicada no instrumento convocatório, exclusivamente para modelos, cenários, métricas e regras efetivamente integrantes daquela versão. Para a versão 6.1, a relação de <i>workloads</i> deverá ser conferida e

ONDE SE LÊ	LEIA-SE
	ajustada à documentação oficial do MLCommons antes da publicação.”

14.15. Item 4.5 - benchmarks MLPerf Training

ONDE SE LÊ	LEIA-SE
“MLPerf <i>Training</i> v6.0 (...) fine-tuning (Llama 2 70B LoRA / GovReport) e pre <i>Training</i> (Llama 3.1-8B / C4)”	“MLPerf <i>Training</i> , na versão oficial expressamente indicada no instrumento convocatório, com utilização apenas dos <i>benchmarks</i> , modelos, <i>datasets</i> , metas de qualidade e regras integrantes da versão escolhida. A relação indicada no texto deverá ser substituída pela lista oficial da versão vigente (V6.1), anexada ao edital.”

14.16. Item 4.5 - coerência com a tabela técnica

ONDE SE LÊ	LEIA-SE
“O escore técnico será a soma ponderada de 50% para MLPerf <i>Inference</i> (...) e 50% para MLPerf <i>Training</i> ”	“A pontuação dos <i>benchmarks</i> observará exclusivamente os pesos definidos na tabela do item 9.2.3, após sua consolidação. Fica suprimida a distribuição autônoma de 50% para <i>Inference</i> e 50% para <i>Training</i> , salvo se a tabela de julgamento for alterada para refletir exatamente essa mesma distribuição.”

14.17. Item 4.5 - extrapolação de escala

ONDE SE LÊ	LEIA-SE
“se o proponente somente tiver acesso a 10% dos recursos, este será normalizado como o	“Os testes de escalabilidade deverão ser executados nos percentuais de recursos

ONDE SE LÊ	LEIA-SE
valor de <i>Throughput</i>(100%) e os 5% (...) como (...) 50%”	definidos no instrumento convocatório.” “Os testes de desempenho e escalabilidade apresentados na fase de proposta não estarão obrigatoriamente sujeitos à execução em laboratório específico, podendo ser realizados em infraestrutura própria do proponente, do fabricante, de parceiro tecnológico ou em outro ambiente tecnicamente adequado. Será admitida a extrapolação linear dos resultados obtidos com frações dos recursos computacionais da solução, desde que o proponente informe a quantidade e o percentual de recursos efetivamente utilizados, apresente a memória de cálculo da extrapolação e mantenha a mesma configuração de hardware, software, parâmetros, modelos, conjuntos de dados e condições de execução empregadas no teste de origem. A extrapolação deverá ser aplicada de maneira uniforme aos percentuais de recursos definidos no instrumento convocatório e seus resultados deverão ser claramente identificados como valores extrapolados, distinguindo-os dos resultados efetivamente medidos. Não será obrigatória a publicação nem submissão dos resultados ao MLPerf nem ao MLCommons, sendo admitidos relatórios técnicos de execução acompanhados das configurações, parâmetros, registros de execução, logs, resultados medidos, memória de cálculo e demais evidências necessárias à verificação da metodologia e da extrapolação. A Administração poderá solicitar esclarecimentos, realizar diligências e exigir a reprodução dos testes durante a fase de aceitação da solução.”

ONDE SE LÊ	LEIA-SE

14.18. Item 4.5 - terminologia

ONDE SE LÊ	LEIA-SE
“reprodutíveis e reproduzíveis”	“repetíveis no mesmo ambiente e reproduzíveis por equipe independente, mediante disponibilização integral de scripts, parâmetros, imagens de contêiner, versões, logs e configuração do sistema.”

14.19. Item 4.5.2 – Benchmark de CPU

ONDE SE LÊ	LEIA-SE
Não consta	“4.5.2. <i>Benchmark</i> de desempenho de CPU e memória. O desempenho dos processadores de propósito geral e do subsistema de memória da solução será avaliado por meio dos <i>benchmarks</i> HPCG (<i>High Performance Conjugate Gradients</i>) e <i>STREAM</i>, utilizando suas implementações públicas de referência. O <i>benchmark</i> HPCG deverá ser obtido a partir do repositório oficial do projeto, disponível em https://github.com/hpcg-benchmark/hpcg, utilizando versão estável do código-fonte congelada na data de publicação desta Errata. O <i>benchmark</i> deverá ser executado exclusivamente nas CPUs, sem utilização de GPUs, aceleradores dedicados ou mecanismos de <i>offload</i> computacional. A execução deverá utilizar todos os processadores físicos e núcleos disponíveis no nó avaliado, respeitando a

ONDE SE LÊ	LEIA-SE
	arquitetura <i>NUMA</i> do sistema. O problema deverá ser dimensionado de forma que utilize, no mínimo, 25% da memória principal disponível para a execução, evitando que o conjunto de dados seja integralmente acomodado nos níveis de cache dos processadores. Deverão ser reportados, no mínimo: desempenho <i>HPCG</i> em <i>GFLOP/s</i>; desempenho em <i>GFLOP/s</i> por núcleo físico; número de processadores físicos; número de núcleos utilizados; número de <i>threads</i>; frequência das <i>CPUs</i>; configuração <i>NUMA</i>; capacidade e configuração da memória; número de processos <i>MPI</i>; número de <i>threads OpenMP</i> por processo; compilador; opções de compilação; bibliotecas utilizadas; sistema operacional; tempo de execução; parâmetros de entrada; e demais informações necessárias à reprodução do resultado. Adicionalmente, deverá ser executado o <i>benchmark STREAM</i>, utilizando a implementação pública de referência disponível em https://github.com/jeffhammond/STREAM. A execução deverá utilizar dados em FP64, todos os núcleos físicos disponíveis no nó e exclusivamente as <i>CPUs</i>, sem utilização de aceleradores ou mecanismos de <i>offload</i>. Para o <i>STREAM</i> deverão ser reportados os resultados dos <i>kernels Copy, Scale, Add e Triad</i>, em <i>GB/s</i>. O resultado do <i>STREAM Triad</i> será adotado como métrica principal de largura de banda sustentável da memória. Deverão ser apresentados, adicionalmente, os valores de <i>STREAM Triad</i> em <i>GB/s</i> por processador físico e em <i>GB/s</i> por núcleo físico. Para ambos os <i>benchmarks</i> deverão ser informadas a política de afinidade de <i>CPU</i> e memória, configuração

ONDE SE LÊ	LEIA-SE
	de <i>SMT</i> , configuração <i>NUMA</i> , número de repetições, variáveis de ambiente, versões completas do código-fonte e respectivos <i>commits</i> ou <i>tags</i> utilizados. Não será admitida alteração do código-fonte dos <i>kernels</i> computacionais dos <i>benchmarks</i> com o objetivo de produzir implementação específica para determinado processador ou fabricante.”

14.20. Item 4.6.3 - software simbólico

ONDE SE LÊ	LEIA-SE
“Jess (<i>Java Expert System Shell</i>) 80u2”	“O <i>Rule engine Jess</i> , em versão estável suportada pelo fabricante e compatível com o ambiente ofertado”. A designação “80u2” aparenta erro material e deverá ser tecnicamente validada.

14.21. Item 4.6.3 - grafos de conhecimento

ONDE SE LÊ	LEIA-SE
“Grakn/Vectis (<i>knowledge graphs</i>)”	“solução de grafo de conhecimento compatível com RDF/OWL ou modelo equivalente, em versão estável, suportada e com documentação oficial”. Os nomes comerciais ou de projetos deverão ser validados quanto à existência, manutenção e licenciamento.

14.22. Item 4.6.4 - conceitos de IA

ONDE SE LÊ	LEIA-SE
“Ciclo de vida híbrido: integração ANI/AGI/ASI”; “Teoria da Mente (NLP empático)”; “Autoconsciente (Neuro-Simbólica total)”	“Integração de métodos de aprendizado de máquina, aprendizado profundo e inteligência artificial simbólica, com casos de uso, interfaces, critérios de desempenho e evidências objetivamente verificáveis”. Suprimir referências a AGI, ASI, ‘teoria da mente’ e ‘autoconsciência’, por não constituírem requisitos técnicos mensuráveis e maduros para aceitação contratual.

14.23. Item 4.6.6 - remissão final

ONDE SE LÊ	LEIA-SE
“item 5.6.6”	“item 4.6.6”

14.24. Item 4.6.6 - concordância

ONDE SE LÊ	LEIA-SE
“permitam a execução das <i>Stack de software</i> mínima solicitada”	“permitam a execução da <i>Software Stack</i> mínima solicitada”

14.25. Item 4.7 - introdução

ONDE SE LÊ	LEIA-SE
“capaz de atender a duas cargas de trabalho (...) Ela será formada por, pelo menos, três camadas”	“capaz de atender a cargas de trabalho distintas. O subsistema será formado por, no mínimo, três camadas de armazenamento.”

14.26. Item 4.7.1 - Camada 0

ONDE SE LÊ	LEIA-SE
“para serem utilizados como <i>cache</i> as <i>read-only</i> ou <i>read-write</i> ”	“para utilização como <i>cache</i> em modo somente leitura (<i>read-only</i>) ou leitura e escrita (<i>read-write</i>)”

14.27. Item 4.7.1 - requisitos de latência

ONDE SE LÊ	LEIA-SE
“latência abaixo de 100 μ s” e, adiante, “latência (...) igual ou inferior a 80 μ s”	“latência de leitura e escrita igual ou inferior a 80 μ s, no percentil P99, nas condições de teste definidas neste item”. Suprimir o limite concorrente de 100 μ s ou qualificá-lo como requisito geral, prevalecendo expressamente o limite de 80 μ s para aceitação.

14.28. Item 4.7.1 - formatação e unidade

ONDE SE LÊ	LEIA-SE
Trecho corrido iniciado por “O sistema de armazenamento proposto deverá...” sem separação entre frases e requisitos, além de “4KB”, “128KB” e “400,000 IOPS”.	Reestruturar o trecho em alíneas independentes e padronizar as unidades para “4 KiB”, “128 KiB” e “400.000 IOPS”, esclarecendo-se se adotam prefixos binários (KiB, MiB, GiB) ou decimais (kB, MB, GB) em todo o documento.

14.29. Item 4.7.2.1 – Requisitos Gerais de Camada 1 formatação e unidade

ONDE SE LÊ	LEIA-SE
Abaixo estão as características e funcionalidades que o sistema de armazenamento da Camada 1 deve suportar e implementar.	Abaixo estão as características e funcionalidades que o sistema de armazenamento da Camada 1 deve suportar e implementar.

ONDE SE LÊ	LEIA-SE
1. Capacidade de armazenamento a. 20 PB utilizáveis, totalmente flash (NVMe).	1. Capacidade de armazenamento a. 2 PB utilizáveis, totalmente flash (NVMe).

14.30. Item 4.7.2.2 - throughput

ONDE SE LÊ	LEIA-SE
"0,5 GB/s (leitura e escrita) por GPU ativa"	" <i>throughput</i> agregado mínimo equivalente a 0,5 GB/s por acelerador ativo, especificando-se separadamente as metas de leitura e de escrita, a simultaneidade exigida, o número máximo de aceleradores considerados e o método de cálculo da exigência total."

14.31. Item 4.7.3 e 4.7.3.1 – Camada 2

ONDE SE LÊ	LEIA-SE
"30 PB"	"48 PB"

14.32. Item 4.7.3.1 - concordância

ONDE SE LÊ	LEIA-SE
"incluindo, mas não limitado a"	"incluindo, mas não se limitando a"

14.33. Item 4.7.3.1 - crase

ONDE SE LÊ	LEIA-SE
"incluindo, mas não se limitando à: switches de rede"	"incluindo, mas não se limitando a: switches de rede"

14.34. Item 4.8 - nós de serviço

ONDE SE LÊ	LEIA-SE
“A configuração mínima desses equipamentos está descrita no item 4.”	“A configuração mínima desses equipamentos está descrita no item 4.2.”

14.35. Item 4.9 - atualização de segurança

ONDE SE LÊ	LEIA-SE
“prazo máximo de 48 horas para patches críticos de CVE”	“prazo máximo de 48 horas para disponibilização de plano de mitigação ou correção de vulnerabilidades críticas, contado da notificação ou da divulgação oficial; a aplicação em produção dependerá de autorização formal do Contratante.”

14.36. Item 4.10 - engenheiro residente

ONDE SE LÊ	LEIA-SE
“acesso à engenheiros”	“acesso a engenheiros”

14.37. Item 4.10.1 - remissão

ONDE SE LÊ	LEIA-SE
“Todos os <i>softwares</i> previstos neste documento, item 5.6”	“Todos os <i>softwares</i> previstos no item 4.6 deste Termo de Referência”

14.38. Item 4.10.2 - classificação

ONDE SE LÊ	LEIA-SE
Uso simultâneo de “Crítica” para requisição e incidente, com níveis distintos, e de “Urgente” apenas para incidentes/manutenção.	“Adotar nomenclatura uniforme de severidade, preferencialmente S1 - crítica, S2 - alta, S3 - média e S4 - baixa, com critérios objetivos de impacto, abrangência, indisponibilidade, tempo de resposta, tempo de restauração e tempo de solução.”

14.39. Item 4.12.1 - capacitação

ONDE SE LÊ	LEIA-SE
“O programa deverá comprovar que a equipe técnica terá adquirido conhecimento todas em as etapas”	“O programa deverá demonstrar que a equipe técnica adquiriu conhecimentos em todas as etapas de implementação, operação e suporte das tecnologias fornecidas.”

14.40. Item 4.12.1 - cronologia

ONDE SE LÊ	LEIA-SE
“A primeira turma deverá ser iniciada em até 180 dias após o aceite da proposta e concluída até a data da conclusão do processo de comissionamento”	“A primeira turma deverá ser iniciada em até 180 dias contados da assinatura do contrato e concluída até o encerramento do comissionamento.”

14.41. Item 4.12.2 - remissão

ONDE SE LÊ	LEIA-SE
“capacitação de operação prevista no item 4.12”	“capacitação operacional prevista no item 4.12.1”

14.42. Item 4.12.3 – Early Access

ONDE SE LÊ	LEIA-SE
Em até 90 dias após a assinatura do contrato de fornecimento, o fornecedor deverá disponibilizar acesso a uma infraestrutura computacional de tecnologia equivalente à solução ofertada com, pelo menos, 5% da capacidade total ofertada na proposta. Essa infraestrutura deverá estar disponível 24 horas por dia, 7 dias por semana, até a conclusão da entrega e comissionamento da solução.	Em até 180 dias após a assinatura do contrato de fornecimento, o fornecedor deverá disponibilizar acesso a uma infraestrutura computacional (<i>hardware</i> e <i>software</i>) de mesma família ou geração tecnológica da solução ofertada com, pelo menos, 5% do quantitativo de aceleradores e 5% do total de espaço para armazenamento ofertados na proposta. Essa infraestrutura computacional poderá ser disponibilizada por meio de acesso remoto a um ambiente equivalente, tais como: capacidade em nuvem, ambiente do fabricante, parceiro ou cluster dedicado, instalado em localidade diversa daquela da entrega definitiva. Essa infraestrutura deverá estar disponível 24 horas por dia, 7 dias por semana, até a conclusão da entrega e comissionamento da solução.

14.43. Item 4.13 - local

ONDE SE LÊ	LEIA-SE
Repetição integral do local de instalação, também constante do item 5.2.	Manter a identificação completa do local no item 5.2 e, no item 4.13, utilizar: “A infraestrutura deverá ser dimensionada para o local de instalação indicado no item 5.2.”

14.44. Item 4.13 – Infraestrutura

ONDE SE LÊ	LEIA-SE
“A medição de energia será obrigatória e parte integrante da avaliação. Para inferência, a energia deverá ser medida ‘at the wall’ durante a fase de performance, utilizando o workflow MLPerf Power, com correlação temporal aos logs do LoadGen”	“A medição de energia durante os benchmarks será realizada nos pontos de entrada elétrica do <i>System Under Test SUT</i>, conforme a metodologia oficial aplicável. A fronteira do <i>SUT</i> compreenderá os nós computacionais utilizados no teste, aceleradores, processadores, memória, dispositivos locais de armazenamento, equipamentos de interconexão e demais componentes necessários à execução do benchmark e alimentados pelos circuitos elétricos submetidos à medição. Serão incluídas as fontes de alimentação, os ventiladores internos e as bombas ou dispositivos de refrigeração incorporados aos equipamentos abrangidos pela medição. <i>Coolant Distribution Unit</i> (Unidade de Distribuição de Fluido de Resfriamento) <i>CDUs</i>, bombas e unidades de ventilação externas, quando dedicadas exclusivamente à solução e fornecidas pela Contratada, serão medidas e reportadas separadamente, permitindo o cálculo complementar da eficiência energética da solução integrada. Chillers, torres de resfriamento, bombas centrais, iluminação e demais cargas gerais da instalação não integrarão a métrica primária do <i>SUT</i>, sem prejuízo de sua consideração no cálculo do <i>PUE</i> da instalação.”

14.45. Item 4.13 - Infraestrutura

ONDE SE LÊ	LEIA-SE
“PUE = Energia Total da Instalação / Energia Utilizada pelos Equipamentos de T.I.”	“PUE = Energia Total da Instalação / Energia Utilizada pelos Equipamentos de T.I.” “O PUE será determinado conforme a ISO/IEC 30134-2:2026, mediante divisão da energia total do centro de dados pela energia consumida pelos equipamentos de tecnologia da informação, observada a categoria de medição declarada. A energia total do centro de dados compreenderá, dentro da fronteira de medição adotada, a energia dos equipamentos de TI e as cargas de infraestrutura associadas, incluindo sistemas de refrigeração, <i>CDUs</i>, bombas, ventiladores externos, <i>chillers</i>, torres de resfriamento, perdas de transformação e distribuição elétrica e demais cargas auxiliares aplicáveis. A energia dos equipamentos de TI compreenderá servidores, aceleradores, armazenamento, equipamentos de rede e componentes de controle diretamente associados à operação computacional. Deverão ser informados os pontos de medição, o período de apuração, a categoria de <i>PUE</i>, as cargas incluídas, excluídas, compartilhadas, estimadas ou não medidas e o método de rateio adotado.” Para a métrica <i>tokens/s/W</i>, inclua os componentes efetivamente pertencentes e eletricamente conectados ao <i>System under test SUT</i>. Meça separadamente <i>Coolant Distribution Unit CDUs</i> e bombas dedicadas. Não inclua <i>chillers</i> e bombas prediais na métrica primária do benchmark. Para o <i>PUE</i>, inclua todas as cargas de refrigeração associadas à instalação, inclusive

ONDE SE LÊ	LEIA-SE
	CDUs, bombas, ventiladores externos e <i>chillers</i> , respeitando a fronteira e a categoria de medição declaradas.

14.46. Item 4.13 - Infraestrutura

ONDE SE LÊ	LEIA-SE
Não Consta	Todos os nós computacionais devem, obrigatoriamente, ter a refrigeração DLC (<i>Direct Liquid Cooling</i>)

14.47. Item 4.15.1 - numeral

ONDE SE LÊ	LEIA-SE
“mil (1000) especialistas brasileiros”	“1.000 (mil) especialistas brasileiros”

14.48. Item 4.15.1 - Requisitos para Parcerias com Institutos de Pesquisa e Centros de P&D

ONDE SE LÊ	LEIA-SE
A proponente deverá apresentar um Programa de Parceria com Institutos de Pesquisa e Centros de P&D descrevendo as estratégias...	"A proponente deverá apresentar um Programa de Parceria com Institutos de Pesquisa e Centros de P&D públicos descrevendo as estratégias..."

14.49. Item 4.15.1 - Programa de Formação de Recursos Humanos

ONDE SE LÊ	LEIA-SE
Identificação e descrição dos Institutos de Pesquisa e Centros de P&D parceiros envolvidos na execução do programa,...	"Identificação e descrição dos Institutos de Pesquisa e Centros de P&D públicos parceiros envolvidos na execução do programa, ..."

14.50. Item 4.15.1 - forma de comprovação e efeitos

ONDE SE LÊ	LEIA-SE
Não Consta	“Será aceito documento assinado pela autoridade competente reconhecida pela ICT pública”

14.51. Item 4.15.1 - comprovação das parcerias com ICTs

ONDE SE LÊ	LEIA-SE
Não consta	“Para fins de comprovação das parcerias previstas no Plano de Colaboração Estratégica, o proponente deverá apresentar, juntamente com a proposta técnica, documento assinado por autoridade competente. Essa autoridade deverá ser reconhecida pela respectiva ICT pública, instituto de pesquisa ou centro de pesquisa e desenvolvimento públicos. O documento deverá identificar a instituição e o proponente, demonstrar a concordância da instituição em participar do Plano de Colaboração Estratégica e indicar, ainda que de forma preliminar, o objeto, as áreas ou as linhas de colaboração pretendidas. Serão aceitos, entre outros documentos de conteúdo equivalente, carta de intenção, declaração de interesse, termo de compromisso, protocolo de intenções ou instrumento de parceria. Não será exigida, na fase de apresentação da proposta, a celebração definitiva de convênio, contrato ou acordo de parceria, desde que o documento apresentado demonstre de maneira inequívoca a anuência institucional e esteja assinado pela autoridade competente reconhecida pela ICT pública. A Comissão de Seleção poderá realizar diligência para verificar a autenticidade do documento, a competência do signatário e a

ONDE SE LÊ	LEIA-SE
	manifestação institucional nele registrada, vedada a apresentação posterior de documento destinado a constituir parceria inexistente na data de apresentação da proposta.”

14.52. Item 5.1.1.1 - início da garantia

ONDE SE LÊ	LEIA-SE
“ Llama 3.1-70B [512 processos/aceleradores simulados] ou”	“ Llama 3.1-70B [512 processos / aceleradores simulados]”

14.53. Item 5.1.3 - início da garantia

ONDE SE LÊ	LEIA-SE
“A emissão do Aceite Provisório marca o início do período de garantia.”	“A emissão do Aceite Definitivo marca o início do período de garantia técnica.”

14.54. Item 5.1.3 - Aceite definitivo

ONDE SE LÊ	LEIA-SE
Aceite Definitivo: emitido 30 dias após o Aceite Provisório, desde que não tenham sido identificadas não conformidades relevantes durante o período de operação assistida. A emissão do Aceite Definitivo está condicionada adicionalmente à comprovação do início das	Aceite Definitivo: A emissão do aceite definitivo fica condicionada à verificação e à validação de todas as entregas previstas para o aceite provisório. Após a comunicação formal da conclusão da entrega, a Contratante terá o prazo de até 30 (trinta) dias para emitir o aceite definitivo.

ações apresentadas na proposta de colaboração estratégica (item 4.15).	Caso sejam identificadas não conformidades técnicas ou documentais, serão solicitadas ao fornecedor as correções necessárias, iniciando-se, após a conclusão dessas correções, um novo ciclo de verificação e validação das entregas, com novo prazo de até 30 (trinta) dias para a emissão do aceite definitivo.
--	---

14.55. Item 6 - valor estimado

ONDE SE LÊ	LEIA-SE
“conforme detalhado no item 5.1”	“conforme detalhado no item 8.2”

14.56. Item 6 - adequação orçamentária

ONDE SE LÊ	LEIA-SE
“recursos orçamentários”	“recursos orçamentários”

14.57. Item 6 - pagamento

ONDE SE LÊ	LEIA-SE
“O prazo para o pagamento da(s) Nota(s) Fiscal(is)/Fatura(s), será”	“O prazo para pagamento da(s) nota(s) fiscal(is) ou fatura(s) será”

14.58. Item 6 - rescisão

ONDE SE LÊ	LEIA-SE
“Será rescindido o contrato em execução com a Contratada inadimplente, salvo por motivo de economicidade...”	“O inadimplemento contratual poderá ensejar a extinção do contrato, assegurados o contraditório e a ampla defesa, observadas as

ONDE SE LÊ	LEIA-SE
	hipóteses, os procedimentos e as consequências previstos no instrumento contratual e na legislação aplicável.”

14.59. Item 7 - fiscalização

ONDE SE LÊ	LEIA-SE
“fiscal técnico formalmente designado pela Contratante, definida no momento da assinatura”	“fiscal técnico formalmente designado pela Contratante no momento da assinatura do contrato, ao qual competirá”

14.60. Item 8.2 - Parcela 3

ONDE SE LÊ	LEIA-SE
“Parcela 3: R R\$ 319.680.319,68”	“Parcela 3: R\$ 319.680.319,68”

14.61. Item 9.2.2 – Critérios, Pesos e Regras submissão

ONDE SE LÊ	LEIA-SE
"Resultado de benchmark de storage exigido no edital (ex.: IO500 / MLPerf Storage, conforme seção de testes)."	"Resultado de benchmark de storage exigido no edital (MLPerf Storage, conforme seção de testes)." A execução do IO500 não será utilizada como critério para pontuação, tendo um caráter apenas avaliativo do desempenho do sistema de armazenamento. Porém, ele permanece como um item obrigatório do teste de aceitação."

51. Item 9.2.2 – Nova tabela de Pesos e Regras por Critério

14.62. Item 9.9.2 - submissão

ONDE SE LÊ	LEIA-SE
“9.2.2 Critérios, Pesos e Regras por Critério” Em todo o Termo de Referência.	“Será substituído pela nova tabela de Critérios Pesos e Regras por Critério”

Item	Critério	Peso (Wi)	Métrica / Regra	Pontuação
1	Treino LLM – Pré-treino (MLPerf Training)	20%	<i>Time-to-quality</i> (tempo para atingir qualidade alvo). Menor tempo é melhor" Evidência Obrigatória: Logs oficiais do MLPerf Training.	0% até 20%
2	Treino LLM – <i>Fine-tuning</i> (MLPerf Training)	15%	<i>Time-to-quality</i> (tempo até qualidade alvo). Menor tempo é melhor — normalização tratada como benefício por inversão, via regra do MLPerf prevista no edital. Evidência Obrigatória: Logs oficiais do MLPerf Training.	0% até 15%
3	Inferência LLM – Offline (MLPerf Inference)	15%	<i>Throughput</i> (ex.: tokens/s) no cenário Offline. Evidência Obrigatória: Logs do MLPerf Inference gerados pelo <i>software stack</i> oficial.	0% até 15%
4	Inferência LLM – <i>Server/Interactive</i> (MLPerf Inference)	10%	Desempenho no cenário <i>Server/Interactive</i> , respeitando restrições de latência (ex.: TTFT/TPOT) conforme regras do <i>benchmark</i> . Medição e carga via LoadGen, para garantir comparabilidade e justiça.	0% até 10%
5	Eficiência energética em IA ("na parede")	8%	Tokens/W (ou tokens/J) medidos no sistema inteiro (" <i>at the wall</i> ") durante o run de performance. Evidência: Logs de potência sincronizados ao	0% até 8%

Item	Critério	Peso (Wi)	Métrica / Regra	Pontuação
			<i>benchmark</i> (<i>workflow</i> de medição de energia).	
6	Armazenamento para IA	7%	Resultado de <i>benchmark</i> de storage exigido no edital (ex.: IO500 / MLPerf Storage, conforme seção de testes). Evidência: Relatório + logs do <i>benchmark</i> executado conforme plano de testes.	0% até 7%
7	Benchmark de IA Agêntica	5%	Benchmark de IA Agêntica (Agentic Inference for MLPerf Inference), na versão oficialmente congelada na data de publicação da Errata. Os resultados deverão ser apresentados por modelo avaliado (Kimi K2.6 e Qwen3.6-35B-A3B, ou os modelos que constem da especificação oficial congelada), observando a metodologia oficial do MLCommons. Para cada métrica de desempenho reportada, maior resultado é melhor. (Item 8 desta errata)	0% até 5%
8	Capacitação	5%	Número de horas de mentoria (Item 4.12.2) comprovadas (ex.: turmas, ementa, instrutores, execução). Evidência: comprovação documental e plano executável.	0% até 5%
9	Prazo de entrega/comissionamento	15%	Dias corridos até entrega + comissionamento. Prazo máximo desejável: 360 dias (acima disso, 0% no item). Prazo máximo aceitável: 520 dias.	0% até 15%
TOTAL:		100%	TOTAL:	
Benchmarks = 80% / Demais critérios = 20%				

14.63. Item 9.2.3 - totais da tabela

ONDE SE LÊ	LEIA-SE
"Benchmarks = 80% / Demais critérios = 20%"	"Benchmarks e desempenho técnico = 80%; demais critérios = 20%", somente se assim for definido pela autoridade competente. Pela tabela atual, os itens 1 a 6 totalizam 80%, enquanto capacitação e prazo totalizam 20%; a redação deve refletir exatamente essa classificação.

14.64. Item 9.2.4 - normalização

ONDE SE LÊ	LEIA-SE
"Benefício (maior é melhor) - Itens 1 a 7"	"Aplicar a regra 'maior é melhor' apenas aos indicadores cujo aumento represente melhoria. Para os itens de tempo até a qualidade, utilizar $Si(p) = \text{menor tempo válido} / \text{tempo do proponente}$. Para o prazo de entrega, utilizar a mesma lógica de custo, observados os limites e faixas definidos no edital."

14.65. Item 9.2.4 - prazo

ONDE SE LÊ	LEIA-SE
"Prazo máximo desejável: 360 dias (acima disso, 0% no item)" e "Prazo máximo aceitável: 520 dias".	"Prazo mínimo de entrega deverá ser de 360 dias corridos. Propostas com prazo superior a 520 dias corridos serão desclassificadas. Para prazos entre 361 e 520 dias corridos, deverá ser definida fórmula objetiva de pontuação ou faixa de redução. Expressamente mantida e compatibilizada com a fórmula de normalização descrita no item 9.2.1.4."

14.66. Item 9.3 - denominação

ONDE SE LÊ	LEIA-SE
Uso da sigla “RFP” sem definição uniforme no corpo do Termo de Referência.	“Termo de Referência e respectivos anexos”, ou definir previamente “Requisição de Propostas (Request for Proposal - RFP)” e utilizar a sigla de modo uniforme.

14.67. Item 9.4 – Nota final de Técnica e Preço

ONDE SE LÊ	LEIA-SE
9.4. Nota Final de Técnica e Preço A Nota Final - NF será calculada mediante a ponderação da Nota Técnica - NT e da Nota de Preço - NP, ambas expressas na escala de 0 a 100 pontos, atribuindo-se peso de 70% à Técnica e de 30% ao Preço, conforme a fórmula $NF_{(p)} = 0,70 \times NT_{(p)} + 0,30 \times NP_{(p)}$. Será classificado em primeiro lugar o proponente que obtiver a maior Nota Final, observados os critérios de desempate e as demais condições previstas neste Termo de Referência.	9.4. Processo de Produção Nacional 9.4.1. Para fins de pontuação neste edital, será considerada a produção nacional do servidor de muito grande capacidade, classificado na NCM 8471.50.40, mediante o cumprimento do respectivo Processo Produtivo Básico (PPB), estabelecido nos termos das Leis nº 8.248, de 23 de outubro de 1991, e nº 8.387, de 30 de dezembro de 1991, e definido pelas Portarias Interministeriais MDIC/MCTI nº 177 e nº 178, ambas de 16 de junho de 2026. 9.4.2. A comprovação do cumprimento do PPB deverá ser realizada mediante a apresentação de, pelo menos, um dos seguintes documentos oficiais, conforme aplicável: a) Portaria de habilitação emitida pelo Ministério da Ciência, Tecnologia e Inovação (MCTI), que contemple o produto servidor de muito grande capacidade; b) Resolução do Conselho de Administração da SUFRAMA (CAS), que contemple o referido produto; ou c) Publicação do produto ou modelo no sítio eletrônico oficial do MCTI, na relação de

ONDE SE LÊ	LEIA-SE
	empresas habilitadas no âmbito da Lei nº 8.248, de 1991, disponível em: https://inovacaodigital.mcti.gov.br/leiDeInforatica/empresasHabilitadas 9.4.3. O produto ou modelo de servidor de muito grande capacidade ofertado pela empresa proponente deverá estar expressamente identificado em um dos documentos comprobatórios previstos no item 9.4.2, ou ser modelo equivalente ao produto nele identificado. 9.4.4. Para fins de comprovação da produção nacional, o documento apresentado deverá permitir a identificação da empresa responsável pela produção nacional do equipamento ofertado, bem como a vinculação entre o produto ou modelo comprovado e o equipamento objeto da proposta. 9.4.5. Caso o documento comprobatório esteja em nome de empresa diversa da proponente, esta deverá apresentar, adicionalmente, documentação que comprove a vinculação entre as empresas e a legitimidade da proponente para ofertar o equipamento como produto de produção nacional. 9.4.6. A documentação apresentada deverá ser suficiente para demonstrar que o equipamento ofertado atende aos requisitos de produção nacional estabelecidos neste edital, observadas as disposições legais e regulamentares aplicáveis ao PPB. 9.4.7. Será assegurado o direito de preferência previsto no art. 3º da Lei nº 8.248, de 23 de outubro de 1991, observado o procedimento

ONDE SE LÊ	LEIA-SE
	estabelecido nos arts. 5º e 8º do Decreto nº 7.174, de 12 de maio de 2010, quando aplicável. 9.4.8. Para fins deste edital, não serão aplicadas as disposições da Lei nº 8.010, de 29 de março de 1990.

14.68. Item 10.2 - escopo

ONDE SE LÊ	LEIA-SE
“os <i>softwares</i> e <i>firmwares</i> listados no item 4.6 e 4.7”	“os <i>softwares</i> , <i>firmwares</i> e componentes de armazenamento especificados nos itens 4.6 e 4.7”

14.69. Item 10.2 - escopo

ONDE SE LÊ	LEIA-SE
Não consta	Para 9.5. Nota Final de Técnica e Preço A Nota Final - NF será calculada mediante a ponderação da Nota Técnica - NT, Nota do Processo Produtivo Básico - NPPB e da Nota de Preço - NP, todas expressas na escala de 0 a 100 pontos, atribuindo-se peso de 70% à Técnica, 20% ao Processo Produtivo e de 10% ao Preço, conforme a fórmula $NF(p) = 0,70 \times NT(p) + 0,20 \times NPPB(p) + 0,10 \times NP(p)$. Será classificado em primeiro lugar o proponente que obtiver a maior Nota Final, observados os critérios de desempate e as demais condições previstas neste Termo de Referência.

ONDE SE LÊ	LEIA-SE
	Será utilizado como critério de desempate os valores declarados na proposta técnica na seguinte ordem: 1. Treino LLM – Pré-treino 2. Treino LLM – Fine-tuning 3. Inferência LLM – Offline 4. Inferência LLM – Server e Interactive 5. Eficiência energética em IA 6. Armazenamento para IA 7. Prazo de entrega e comissionamento 8. Capacitação Prevalecendo o empate, será utilizado o critério do menor preço total da proposta.

14.70. Item 10.2 – Garantia

ONDE SE LÊ	LEIA-SE
Componentes substituídos: peças e componentes substituídos durante o período de garantia estarão cobertos pelo prazo remanescente da garantia original ou por 12 (doze) meses a partir da substituição, valendo o prazo que for maior. O fornecedor deverá registrar todas as substituições em sistema de gestão de ativos acessível ao comprador.	Componentes substituídos: peças e componentes substituídos durante o período de garantia estarão cobertos pelo prazo remanescente da garantia original. O fornecedor deverá registrar todas as substituições em sistema de gestão de ativos acessível ao comprador.

14.71. Item 10.2 – Garantia EOL e EOS

ONDE SE LÊ	LEIA-SE
Notificação de descontinuação: caso qualquer componente crítico do sistema venha a ter EOL ou EOS anunciado durante o período de	Notificação de descontinuação: caso qualquer componente do sistema venha a ter EOL ou EOS anunciado durante o período de garantia, o fornecedor deverá notificar o comprador em

ONDE SE LÊ	LEIA-SE
garantia, o fornecedor deverá notificar o comprador com antecedência mínima de 18 (dezoito) meses, apresentando plano de substituição ou migração sem custo adicional ao comprador.	até 30 dias após o anúncio do fabricante. Deverá também garantir a continuidade do suporte a esses componentes durante todo o período de garantia.

14.72. Itens 10.3 e 10.4 - percentuais de garantia

ONDE SE LÊ	LEIA-SE
No item 10.4 do TR: “Como condição para formalização contratual, a empresa contratada deverá prestar garantia financeira de execução do objeto, no percentual de 5% (cinco por cento) do valor total do contrato, tendo como beneficiário a Contratante.”	No item 10.4 do TR: “Como condição para formalização contratual, a empresa contratada deverá prestar garantia financeira de execução do objeto, no percentual de 10% (dez por cento) do valor total do contrato, tendo como beneficiário a Contratante.”

14.73. Item 10.4 - concordância

ONDE SE LÊ	LEIA-SE
“O detalhamento das demais condições (...) será apresentados”	“O detalhamento das demais condições (...) será apresentado”

14.74. Item 10.5 - valor de grande vulto

ONDE SE LÊ	LEIA-SE
“patamar atualizado para R\$ 250.902.323,87 (...) pelo Decreto nº 12.343, de 19 de dezembro de 2024, com efeitos a partir de 2025”	“patamar atualizado para R\$ 261.968.421,04 (...) pelo Decreto nº 12.807, de 29 de dezembro de 2025, com vigência a partir de 1º de janeiro de 2026”. Adequar novamente caso a publicação ocorra em exercício posterior.

14.75. Item 11.1 - remissão

ONDE SE LÊ	LEIA-SE
“eventual aplicação de sanções (item 11)”	“eventual aplicação de sanções (itens 11.2 e 11.3)”

14.76. Item 12 - equivalência tecnológica

ONDE SE LÊ	LEIA-SE
“cláusula de equivalência tecnológica prevista no item 5.6.6”	“cláusula de equivalência tecnológica prevista no item 4.6.6”

14.77. Padronização geral

ONDE SE LÊ	LEIA-SE
Ocorrências de estrangeirismos, capitalização e variantes como “ <i>Software stack</i> ”, “ <i>stack</i> ”, “ <i>stacks</i> ”, “ <i>Storages</i> ”, “ <i>workload</i> ”, “ <i>benchmarking</i> ”, “ <i>datacenter</i> ”, “ <i>on-site</i> ”, “ <i>deploy</i> ”, “ <i>hardening</i> ” e “ <i>roadmap</i> ”.	Na primeira ocorrência, empregar o termo em português seguido do equivalente em inglês entre parênteses; nas ocorrências seguintes, utilizar uma forma única. Exemplos: “pilha de software (<i>software stack</i>)”, “sistema de armazenamento”, “carga de trabalho”, “teste de desempenho”, “centro de dados”, “presencial”, “implantação”, “endurecimento de segurança” e “plano de evolução tecnológica”. Manter termos ingleses apenas quando forem denominações oficiais de tecnologias, produtos, protocolos ou <i>benchmarks</i> .

14.78. Padronização geral

ONDE SE LÊ	LEIA-SE
Alternância entre “comprador”, “cliente”, “Contratante”, “beneficiário”, “proponente”, “fornecedor” e “Contratada”, sem definição completa.	Inserir seção de definições e utilizar: “Contratante” para a FACC; “Beneficiário” para o LNCC; “Proponente” durante a seleção; “Contratada” após a celebração; e “Fabricante” quando a obrigação for especificamente atribuída ao fabricante. Evitar “comprador” e “cliente”, salvo definição expressa.

14.79. Data da Sessão divulgado no DOU

ONDE SE LÊ	LEIA-SE
DATA DA SESSÃO PÚBLICA - Dia 08/10/2026 às 10:00h (horário de Brasília)	DATA DA SESSÃO PÚBLICA - Dia 04/11/2026 às 10:00h (horário de Brasília)

Petrópolis, data da assinatura digital.

FUNDAÇÃO DE APOIO AO DESENVOLVIMENTO DA COMPUTAÇÃO CIENTÍFICA – FACC

Danielle Ferrão

Gerente Administrativa

ERRATA_01_SP_27_2026.pdf

Documento número #c770a67b-e4b3-4a36-97bf-59f2d5709d01

Hash do documento original (SHA256): 856dbfaaf310a28a619411e7f5aa29ac4a723b5a0004c733ba199d0ec8b07d30

Assinaturas**Danielle Lafontaine Castedo Ferrão Leal**

CPF: 128.776.127-59

Assinou como parte em 21 set 2026 às 18:12:15

Log

- 21 set 2026, 17:05:25 Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 criou este documento número c770a67b-e4b3-4a36-97bf-59f2d5709d01. Data limite para assinatura do documento: 14 de novembro de 2026 (16:21). Finalização automática após a última assinatura: habilitada. Idioma: Português brasileiro.
- 21 set 2026, 17:07:33 Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 alterou o processo de assinatura. Finalização automática após a última assinatura: não habilitada.
- 21 set 2026, 17:07:33 Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 adicionou à Lista de Assinatura: dani@facc10.org.br para assinar como parte, via E-mail.
- Pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo Danielle Lafontaine Castedo Ferrão Leal e CPF 128.776.127-59.
- 21 set 2026, 18:12:15 Danielle Lafontaine Castedo Ferrão Leal assinou como parte. Pontos de autenticação: Token via E-mail dani@facc10.org.br. CPF informado: 128.776.127-59. IP: 200.20.95.58. Componente de assinatura versão 1.1517.0 disponibilizado em <https://app.clicksign.com>.
- 21 set 2026, 18:13:03 Operador com email crislayne.santos@facc10.org.br na Conta 0d2fec55-6c6d-46ac-93e0-3e2ba7b6a490 finalizou o processo de assinatura. Processo de assinatura concluído para o documento número c770a67b-e4b3-4a36-97bf-59f2d5709d01.

**Documento assinado com validade jurídica.**Para conferir a validade, acesse <https://www.clicksign.com/validador> e utilize a senha gerada pelos signatários ou envie este arquivo em PDF.

As assinaturas digitais e eletrônicas têm validade jurídica prevista na Medida Provisória nº. 2200-2 / 2001

Este Log é exclusivo e deve ser considerado parte do documento nº c770a67b-e4b3-4a36-97bf-59f2d5709d01, com os efeitos prescritos nos Termos de Uso da Clicksign, disponível em www.clicksign.com.