

Request for Quotation

Order Against Non-Judiciary Contract, Incl. GSA Schedule

Lowest Price, Technically Acceptable

Request for Quotation

RFQ Number: **0205-2026-0003x**

Request Date: 08/31/2026

1. Purpose and Overview

The U.S. Probation & Pretrial Services Office is seeking quotes from qualified vendors to design, supply, and support an on-premises AI inference server. The system will be used to host and serve large language models (LLMs) for internal applications, supporting an estimated **45–65 concurrent users** across the organization/department.

The primary workload is **inference only** — this system will not be used for model training or fine-tuning. Vendors should size and recommend hardware accordingly, prioritizing throughput, concurrency, and cost-efficiency for serving pre-trained models rather than training-oriented specs (e.g., NVLink-heavy interconnects are not required unless the vendor believes it materially improves inference performance/economics at this scale).

We welcome vendor input on right-sizing the configuration below to meet our stated concurrency and performance goals most cost-effectively.

2. Scope of Work

The selected vendor will be responsible for:

1. Supplying hardware per the specifications in Section 3 (or a vendor-recommended equivalent/optimized configuration)
 2. Installing and configuring the baseline software stack per Section 4
 3. Basic burn-in testing and validation prior to delivery
 4. On-site or remote deployment support
 5. Documentation of the final configuration (hardware BOM, network settings, installed software versions, credentials handoff)
 6. Warranty and post-deployment support per Section 6
-

3. Hardware Specifications

Qty	Category	Specification / Description
1	GPU Server Chassis	Enterprise 4U rackmount GPU server with redundant power supplies
4	GPU Accelerators	RTX PRO 6000 96GB, H100/H200 class, or equivalent — vendor to recommend based on inference-only workload and price/performance
2	CPU	AMD EPYC or Intel Xeon enterprise processors
1	System Memory	256–512GB ECC RDIMM RAM (8 DIMMs per CPU)
1	OS Drive(s)	2–4TB enterprise NVMe SSD
1–4	AI/Data Storage	4–16TB enterprise NVMe SSD
1	Networking	Minimum 25GbE NIC; 100GbE preferred if it improves multi-user throughput or cluster scalability — vendor to advise
1	Power/Cooling Validation	Vendor to confirm rack PDU/circuit and cooling requirements for the proposed GPU configuration

Note to vendors: Please state clearly if your proposed GPU selection differs from the reference spec above, and explain the reasoning (e.g., VRAM-per-dollar, power/cooling savings, availability/lead time).

4. Baseline Software and Infrastructure Components

Qty	Software / Service	Description
1	Linux Server OS	Ubuntu LTS, Rocky Linux, RHEL, or organization-approved equivalent
1	Container Runtime	Docker or equivalent container environment
1	GPU Runtime	NVIDIA drivers, CUDA toolkit, NVIDIA Container Toolkit
1	Inference Server	vLLM, SGLang, NVIDIA Enterprise (NIM), or equivalent — vendor to recommend based on target model types and concurrency needs
1	Relational Database	PostgreSQL, or equivalent SQL database
1	Vector Database	FAISS, Milvus, pgvector, or equivalent vector database

Vendors should specify which inference server(s) they recommend and why, given the department-wide concurrency target (45–65 users).

5. Performance and Sizing Requirements

Vendors should size the proposed configuration to support:

- **Concurrent users:** 45-65
- **Model sizes:** Based on the quoted LLM, please recommend a model size (e.g., "Llama 70B" or "a mix of 8B–30B models")

- **Time-to-first-token (TTFT):** under 3 seconds at peak load
 - **Sustained throughput:** at least 15 tokens/second per active user under peak load (45–65 concurrent users)
-

6. Support, Warranty, and Service

Please specify:

- Hardware warranty length and coverage (parts, labor, on-site vs. depot)
 - Support SLA (response time, resolution time, support hours)
 - Availability of extended/next-business-day support options
 - Whether software stack setup/configuration is included or billed separately
 - Post-deployment support options (e.g., ongoing tuning, monitoring, upgrades)
-

7. Proposal Requirements

Vendor proposals should include:

1. **Itemized bill of materials** with unit and total pricing
 2. **Total system price**, including any setup/deployment/labor costs
 3. **Lead time** from PO to delivery, and delivery-to-operational timeline
 4. **Alternate configuration(s)**, if the vendor believes a different GPU or architecture better fits an inference-only, 45-65 user workload
 5. **References** — 1–3 references from similar deployments (size/use case), if available
 6. **Warranty and support terms** as outlined in Section 6
 7. **Assumptions or clarifying questions** the vendor has about the requirements
-

8. Evaluation Criteria and Award Basis

The U.S. Probation & Pretrial Services Office intends to make an award based on the **lowest priced, technically acceptable** quote. All items should be quoted as a **fixed price** and **F.O.B. Destination**. Payment terms will be considered **Net 30** unless more favorable terms are offered.

To be considered technically acceptable, proposals must meet the minimum requirements outlined in Sections 3–6, including hardware specifications (or vendor-justified equivalent), baseline software components, and sizing/performance requirements for the stated 45–65 concurrent user workload.

Among proposals deemed technically acceptable, award will be made to the lowest total price (hardware + setup/deployment + first-year support, as applicable).

9. Timeline

Milestone	Date
RFQ Issued	8/31/2026
Vendor Questions Due	09/14/2026
Proposals Due	09/18/2026
Vendor Selection	09/25/2026

10. Submission Instructions

Please submit proposals to Matthew_Leiter@ctp.uscourts.gov and Jeffrey_Johnson@ctp.uscourts.gov by 09/18/2026. Questions regarding this RFQ should be directed both Matthew and Jeffrey by 09/14/2026.

Quote Sheet for RFQ:

Instructions for Quoter:

Provide the information requested here:

Company Name:	
DUNS number UEI:	
Discount terms, if other than Net 30:	

Instructions for Quoter:

Fill in the unit price and extended price for each item.

Fill in the Grand Total Amount

Hardware Specifications

Item	Category	Specification / Description	Qty	Part #	Unit Price	Extended Price
1	GPU Server Chassis	Enterprise 4U rackmount GPU server with redundant power supplies				
2	GPU Accelerators	RTX PRO 6000 96GB, H100/H200 class, or equivalent — vendor to recommend based on inference-only workload and price/performance				
3	CPU	AMD EPYC or Intel Xeon enterprise processors				
4	System Memory	256–512GB ECC RDIMM RAM (8 DIMMs per CPU)				
5	OS Drive(s)	2–4TB enterprise NVMe SSD				
6	AI/Data Storage	4–16TB enterprise NVMe SSD				
7	Networking	Minimum 25GbE NIC; 100GbE preferred if it improves multi-user throughput or cluster scalability — vendor to advise				
8	Power/Cooling Validation	Vendor to confirm rack PDU/circuit and cooling requirements for the proposed GPU configuration				

Baseline Software and Infrastructure Components

Item	Software / Service	Description	Qty	Part #	Unit Price	Extended Price
9	Linux Server OS	Ubuntu LTS, Rocky Linux, RHEL, or organization-approved equivalent				
10	Container Runtime	Docker or equivalent container environment				
11	GPU Runtime	NVIDIA drivers, CUDA toolkit, NVIDIA Container Toolkit				
12	Inference Server	vLLM, SGLang, NVIDIA Enterprise (NIM), or equivalent — vendor to recommend based on target model types and concurrency needs				
13	Relational Database	PostgreSQL, or equivalent SQL database				
14	Vector Database	FAISS, Milvus, pgvector, or equivalent vector database				

Grand Total: \$_____

TERMS AND CONDITIONS

The following judiciary terms and conditions are incorporated into this request and will be included in the resulting order.

SOLICITATION PROVISIONS

Provision 3-5, Taxpayer Identification and Other Offeror Information (APR 2011)

(a) Definitions. "Taxpayer Identification (TIN)," as used in this provision, means the number required by the Internal Revenue Service (IRS) to be used by the offeror in reporting income tax and other returns. The TIN may be either a social security number or an employer identification number.

(b) All offerors shall submit the information required in paragraphs (d) and (e) of this provision to comply with debt collection requirements of 31 U.S.C. §§ 7701(c) and 3325(d), reporting requirements of 26 U.S.C. §§ 6041, 6041A, and implementing regulations issued by the IRS. If the resulting contract is subject to the payment reporting requirements, the failure or refusal by the offeror to furnish the information may result in a 31 percent reduction of payments otherwise due under the contract.

(c) The TIN may be used by the government to collect and report on any delinquent amounts arising out of the offeror's relationship with the government (31 U.S.C. § 7701(c)(3)). If the resulting contract is subject to payment recording requirements, the TIN provided hereunder may be matched with IRS records to verify the accuracy of the offeror's TIN.

(d) Taxpayer Identification Number (TIN): _____

☐ TIN has been applied for.

☐ TIN is not required, because:

☐ Offeror is a nonresident alien, foreign corporation or foreign partnership that does not have income effectively connected with the conduct of a trade or business in the United States and does not have an office or place of business or a fiscal paying agent in the United States;

☐ Offeror is an agency or instrumentality of a foreign government;

☐ Offeror is an agency or instrumentality of the federal government.

(e) Type of Organization:

☐ sole proprietorship;

☐ partnership;

☐ corporate entity (not tax-exempt);

☐ corporate entity (tax-exempt);

☐ government entity (federal, state or local);

☐ foreign government;

☐ international organization per 26 CFR 1.6049-4;

☐ other

(f) Contractor representations. The offeror represents as part of its offer that it is ☐], is not ☐] 51% owned and the management and daily operations are controlled by one or more members of the selected socio-economic group(s) below:

☐ Women Owned Business

☐ Minority Owned Business (if selected then one sub-type is required)

☐ Black American Owned

☐ Hispanic American Owned

☐ Native American Owned (American Indians, Eskimos, Aleuts, or Native Hawaiians)

☐ Asian-Pacific American Owned (persons with origins from Burma, Thailand, Malaysia, Indonesia, Singapore, Brunei, Japan, China, Taiwan, Laos, Cambodia (Kampuchea), Vietnam, Korea, The Philippines, U.S. Trust Territory of the Pacific Islands (Republic of Palau), Republic of the Marshall Islands, Federated States of Micronesia, the Commonwealth of the Northern Mariana Islands, Guam, Samoa, Macao, Hong Kong, Fiji, Tonga, Kiribati, Tuvalu, or Nauru)

☐ Subcontinent Asian (Asian-Indian) American Owned (persons with origins from India, Pakistan, Bangladesh, Sri Lanka, Bhutan, the Maldives Islands, or Nepal)

☐ Individual/concern, other than one of the preceding.

Provision B-1, Solicitation Provisions Incorporated by Reference (SEP 2010)

This solicitation incorporates one or more solicitation provisions by reference, with the same force and effect as if they were given in full text. Upon request, the contracting officer will make their full text available. The offeror is cautioned that the listed provisions may include blocks that must be completed by the offeror and submitted with its quotation or offer. In lieu of submitting the full text of those provisions, the offeror may identify the provision by paragraph identifier and provide the appropriate information with its quotation or offer. Also, the full text of a solicitation provision may be accessed electronically at this address: <http://www.uscourts.gov/procurement.aspx>.

The following provisions marked with an 'X' are incorporated by reference:

X	2-15	Warranty Information (JAN 2003)
---	------	---------------------------------

	2-85A	Evaluation Inclusive of Options (JAN 2003)
	2-85B	Evaluation Inclusive of Options (JAN 2003)
	2-85C	Evaluation of Options Exercised at Time of Contract Award (JAN 2003)
X	2-100	Brand Name or Equal (APR 2013)
	4-155	Alternate Awards (JUN 2014)
	4-165	Price Proposal Instruction – Multi-Year Contract (JUN 2014)

CONTRACT CLAUSES

Applicable to both the solicitation and contract

Clause 7-10, Contractor Representative (JAN 2003)

(a) The contractor's representative to be contacted for all contract administration matters is as follows (contractor complete the information):

Name:

Address:

Telephone:

E-mail:

Fax:

(b) The contractor's representative shall act as the central point of contact with the judiciary, shall be responsible for all contract administration issues relative to this contract, and shall have full authority to act for and legally bind the contractor on all such issues.

Clause B-5, Clauses Incorporated by Reference (SEP 2010)

This contract incorporates one or more clauses by reference, with the same force and effect as if they were given in full text. Upon request, the contracting officer will make their full text available. Also, the full text of a clause may be accessed electronically at this address:

<http://www.uscourts.gov/procurement.aspx>.

The following clauses marked with an 'X' are incorporated by reference:

	2-20A	Incorporation of Warranty (JAN 2003)
	2-20B	Contractor Warranty (Products) (JAN 2010)
	2-35	F.O.B. Destination, Within Judiciary's Premises (JAN 2003)
	2-40B	Delivery of Excess Quantities (JAN 2003)
	2-55	Privacy or Security of Safeguards (JAN 2003)
	2-90A	Option for Increased Quantity (APR 2013)
	2-90B	Option for Increased Quantity – Separately Priced Line Item (APR 2013)
	2-90D	Option to Extend the Term of the Contract (APR 2013)
	2-95	Material Requirements (JAN 2003)
	2-110	Option to Purchase Equipment (JAN 2003)
	2-125	Security for Advance Payment (APR 2013)
	2-130	Energy Efficiency in Energy-Consuming Products (APR 2013)
	2-135	Acquisition of EPEAT®-Registered Personal Computer Products (MAR 2019)
	2-140	Judiciary IT Security Standards (APR 2013)
	3-1	Contractor Use of Mandatory Sources of Products or Services (JUN 2012)
X	3-3	Provisions, Clauses, Terms and Conditions - Small Purchases (APR 2026)
	4-150	Cancellation Under Multi-Year Contracts (JUN 2014)
	5-30	Authorization and Consent (JAN 2003)
	6-10	Deposit of Assets Requirements (APR 2013)
	6-15	Deposit of Assets Instead of Surety Bonds (JAN 2003)
	6-65	Rights in Data – Special Works (JAN 2010)
	6-75	Rights to Data in an Offer (APR 2013)
	6-80	Rights in Data – Existing Works (JAN 2010)
	6-85	Commercial Computer Software License (APR 2013)
	6-90	Notice and Assistance Regarding Patent and Copyright Infringement (APR 2010)
	6-110	Deferred Ordering of Technical Data or Computer Software (JUN 2014)

X	7-1	Contract Administration (JAN 2003)
X	7-5	Contracting Officer's Representative (APR 2013)
X	7-15	Observance of Regulations/Standards of Conduct (JAN 2003)
X	7-25	Indemnification (AUG 2004)
	7-70	Judiciary Property Furnished "As Is" (APR 2013)
	7-95	Contractor Inspection Requirements (JAN 2003)
	7-115	Availability of Funds (JAN 2003)
	7-160	Limitation on Withholding of Payments (APR 2013)
	7-170	Notice of Intent to Disallow Costs (JAN 2003)
	7-180	Prohibition of Assignment of Claims (JUN 2012)
X	7-215	Notification of Ownership Changes (JAN 2003)