Lenovo Validated Design: AI for Social and Human Services

Last update: **10 February 2026**

Version 1.0

Built and validated on Lenovo infrastructure for secure, scalable AI in social services

Flexible CPU & GPU acceleration for cost efficient, scalable AI performance

Accelerates case reviews by turning complex records into searchable AI insights.

Turn complex safeguarding data into clear, actionable outcomes.

Vanita Meyer
Eric Page
Sinan Atan
Abed Islam

LENOVO PRESS

Table of Contents

Introduction

Executive Summary

This Lenovo Validated Design (LVD) describes an on-premises, secure, and responsible AI solution that helps social and human services organizations accelerate case review, improve safeguarding decisions, and reduce administrative burden. The solution is built around Bikal Technology product, Dialog XR, a Retrieval-Augmented Generation (RAG) system that turns complex multi-source records into searchable, explainable insights while preserving data sovereignty and privacy.

This global edition is written for government agencies and multi-sector partners across regions, recognizing that terminology, statutory timelines, and oversight requirements vary by jurisdiction. The platform is designed to map onto local policies and regulated reporting formats without hard-coding any single national process.

The Lenovo Validated Design (LVD) for Social Services AI with BIKAL Dialog XR delivers an on-premises, secure, and responsible AI solution that supports safeguarding and protection programs for children, adults, and other vulnerable populations. Dialog XR is an AI system developed from research and refined with senior domain experts in safeguarding, public safety, and human services practice. It enables practitioners, reviewers, supervisors, and multi-agency teams to analyze case data, detect emerging risks, and generate higher-quality insights, saving time, improving consistency, and strengthening outcomes.

The solution supports three core workflows commonly found across safeguarding systems globally:

- Historical Case Review – Using AI inference over past serious-incident reviews and critical case investigations to inform new cases and policy decisions.
- Expedited Review Generation – Accelerating draft report creation using audio ingestion, transcription, RAG analysis, and explainable AI to support mandated or policy-driven timelines.
- Live Case Review – Enabling secure, conversational access to active case information (office, field, or mobile) to support supervision and frontline decision-making.

Built on Lenovo ThinkSystem infrastructure and powered by Intel Xeon processors, the design targets deterministic performance, secure data handling, and scalable deployment. Intel® Advanced Matrix Extensions (AMX) provides CPU-based acceleration that can enable cost-effective inference where GPUs are constrained, while optional GPUs can be used for high-throughput transcription or large-scale concurrent workloads.

Intended Audience

This document is intended for public-sector and human services leaders, safeguarding program managers, policy architects, and technology decision-makers seeking to use AI to improve case handling, compliance, and risk management for vulnerable populations. It applies to government and multi-agency organizations, including protection services, public safety, health, education, justice, and community support organizations responsible for incident review and case oversight.

IT decision-makers and solution architects will find guidance for deploying secure, on-premises AI platforms

for sensitive workloads. Lenovo partners, system integrators, and resellers can use this document when designing and implementing AI solutions for social and human services across local, regional, and national public-sector environments.

Social Services Challenges and Opportunity

Challenges to Solve

Social care and protective services agencies worldwide face growing pressure to improve outcomes while dealing with rising caseloads, complex family and community dynamics, and strict policy or statutory timelines. The following challenges are consistently observed across jurisdictions:

- High volumes of unstructured information: long reports, case notes, multi-agency records, and audio from meetings or interviews.
- Fragmented multi-agency collaboration: relevant information is distributed across health, education, law enforcement, justice, and social services systems with different access controls and formats.
- Operational pressure and deadlines: regulated or policy-driven review timelines require rapid synthesis of large amounts of evidence, increasing risk of inconsistency or oversight.
- Workforce burnout and retention: practitioners operate in emotionally demanding environments and need tools that reduce administrative load and improve the speed to insight.
- Legal, privacy, and compliance demands: sensitive personal data requires robust controls, transparent processes, and defensible decision-making with strong audit trails.
- Inequity and variation in outcomes: differences in resourcing, policy, and staff experience can lead to inconsistent review quality and missed early-warning indicators.

Opportunity

Lenovo and BIKAL together enable a step-change in how social and human services organizations manage safeguarding activities. By applying Retrieval-Augmented Generation (RAG), semantic search, and responsible guardrails to trusted local data, agencies can improve both efficiency and quality without compromising confidentiality.

Transform Case handling with AI-assisted Decision Support

Dialog XR provides a secure conversational interface that professionals can use to:

- Query historical incident-review and investigation documents to find comparable patterns and risk indicators.
- Compare current cases against relevant past scenarios to support supervision and escalation decisions.
- Generate explainable, evidence-linked summaries that cite the underlying source material.
- Support structured review formats aligned with regional oversight expectations.

Accelerate Expedited Reviews

Using Lenovo servers, agencies can upload audio recordings and documents, transcribe and normalize content, and generate first-draft reports that remain grounded exclusively in retrieved evidence. This can compress review timelines from weeks to days while improving completeness and consistency. The system supports iterative refinement and human sign-off prior to submission.

.

Improve Field Operations and Practitioner Mobility

With Lenovo AI-enabled laptops and servers, practitioners can securely access approved inference services at the point of care, supporting more informed decisions during home visits, community interactions, or inter-agency meetings. Where connectivity is limited, deployments can be designed with offline-first patterns and controlled synchronization.

Scale across public sector organizations

The architecture supports global scale patterns such as:

- Many agencies with decentralized inference nodes and centralized governance.
- Regional optimization clusters for shared model management and observability.
- Cross-border expansion through multinational partners while maintaining local data residency.

Reduce Harm, Recidivism, and Cost

By detecting risks earlier and improving the quality of case reviews, agencies can reduce:

- The number of serious incidents
- Long-term societal costs associated with victim support
- Legal exposure and reputational risk

Operational Savings

Dialog XR delivers substantial operational savings by replacing a slow, manual, multi-step research process with an AI-assisted workflow that completes in minutes instead of days. In a traditional model, identifying relevant historical case reviews can require hours of intermediary staff effort plus additional time for the practitioner to manually read and interpret documents, resulting in significant labor cost, delays, and backlog risk. With Dialog XR, practitioners can directly query the full body of historical cases, receive targeted results, and explore supporting evidence in a guided conversation in about 30 minutes. This reduces effort from many hours or days to a fraction of the time, often representing up to a 90% improvement in efficiency, while freeing specialist staff capacity, accelerating decision-making, and making evidence-based case review practical and scalable across agencies.

Provide a Secure, Audited, Responsible AI Solution

Lenovo ThinkShield security, Lenovo XClarity management, and Bikal Responsible AI (RAI) guardrails combine to provide end-to-end encryption, policy-based redaction, immutable audit logs, and explainability artifacts. These capabilities help agencies demonstrate compliance with local privacy law, records management requirements, and oversight expectations.

Technical Overview

Dialog XR requirements are divided into Functional Requirements and Non-Functional Requirements. Functional requirements describe the capabilities required to ingest data, transcribe audio, process documents, perform semantic retrieval, orchestrate multi-agent workflows, and generate reports. Non-functional requirements describe quality attributes and operational constraints such as performance, scalability, security, reliability, explainability, and maintainability.

Safeguarding analysis is both compute- and data-intensive. Dialog XR processes long-form audio recordings (often 2–3 hours), multi-agency documents, and large archives of serious-incident reviews. Performance depends on orchestration across speech-to-text, document normalization, vector search, Retrieval-Augmented Generation (RAG), and Responsible AI guardrails. To meet these demands, Dialog XR is designed for sovereign, on-premises deployments with optional air-gapping to satisfy regional data residency and jurisdictional privacy requirements.

Functional Requirements

Data Ingestion & Processing

- Ingest long-form audio recordings (2–3 hours) from review meetings, interviews, or case conferences.
- Upload and process unstructured documents (PDF, DOCX, scanned images, meeting transcripts).
- Normalize ingested content into structured JSON chunks with layout and metadata preserved.
- Support anonymization/redaction workflows for names, partial dates of birth, and street-level address information prior to downstream processing, aligned with local policy.

Speech-to-Text (STT)

- Convert uploaded audio recordings into accurate, time-aligned transcripts using GPU-accelerated ASR.
- Support speaker identification and word-error-rate (WER) statistics for quality assessment.
- Produce transcripts suitable for direct downstream semantic indexing and analysis.

Semantic Search & Retrieval

- Embed processed content into a vector database to enable semantic and hybrid (vector + keyword) search.
- Retrieve relevant historical reviews to compare patterns against live cases.
- Filter and rank results by metadata (agency, timeframe, case attributes, jurisdiction, program type).

Retrieval-Augmented Generation (RAG)

- Generate Rapid Review draft content grounded exclusively in retrieved, case-specific context.
- Support structured prompts aligned with regulated and mandated incident review expectations and best practices.
- Enable section-level summarization for executive summaries and condensed narratives.

Multi-Agent Orchestration

- Orchestrate specialized agents for transcription, sanitization, document processing, retrieval, generation, summarization, and report assembly.

- Support asynchronous task execution and queue-based workflows to handle variable workloads.
- Maintain clear separation of responsibilities across agents to improve transparency and maintainability.

Report Review & Editing

- Generate draft reports in editable formats (DOCX/PDF) aligned to local templates.
- Enable reviewers and supervisors to annotate, request targeted regeneration, and iterate prior to sign-off.
- Support role-based workflows for independent review, legal review, and executive approval.

Audit & Archival

- Archive prompts, embeddings, AI outputs, guardrail decisions, and final reports in an immutable store.
- Maintain complete audit trails suitable for oversight, discovery, or regulatory inspection.
- Support retention policies consistent with records-management requirements.

Non-Functional Requirements

Deployment & Integration

- Deploy on-premises within an air-gapped environment to guarantee data sovereignty where required.
- Support containerized deployment using Docker and optional orchestration frameworks (e.g., Kubernetes/OpenShift).
- Integrate with secure object storage and enterprise identity systems.
- Expose APIs for integration with case-management, records-management, and reporting systems.

Performance & Scalability

- Transcribe multi-hour audio within operationally acceptable timeframes using GPU acceleration where available.
- Achieve low-latency semantic retrieval across large embedding stores.
- Scale to support concurrent reviews across multiple agencies and programs.

Reliability & Availability

- Use stateless microservices and restart-safe processing pipelines.
- Preserve processing state and recover gracefully from service restarts or node failures.
- Support capacity planning and high availability patterns as needed.

Security & Privacy

- Encrypt data at rest (AES-256-GCM) and in transit (TLS 1.3).
- Enforce role-based access control across user interfaces and APIs.
- Detect and redact personally identifiable information (PII) during ingestion and generation in alignment with policy.
- Maintain immutable audit logs for compliance with regional regulatory, audit, and oversight frameworks.

Responsible AI & Explainability

- Embed responsible-AI guardrails to detect bias, harmful content, hallucinations, and privacy leaks.
- Provide natural-language explanations and rationale metadata alongside generated content.
- Store explainability artifacts with each generated report section for post-hoc review.

Architectural Overview

Dialog XR Social Services AI is deployed on dedicated hardware using a layered architecture that is scalable, robust, and operationally maintainable. The design separates responsibilities across infrastructure, runtime, inference, data storage, applications, and security to reduce coupling and avoid lock-in. Components interact via open APIs and can be adapted to local integration patterns.

High-level functional layers include:

- Infrastructure: dedicated compute, networking, and storage for maximum efficiency.
- Runtime: OS and container runtime providing controlled access to hardware.
- Inferencing: RAG pipeline execution and results generation.
- Training: optional regional fine-tuning and evaluation under governance controls.
- Data Storage: secure storage for structured and unstructured data at rest.
- Insights: dashboards and operational monitoring.
- Application: web front-end, speech processing, and orchestration services.
- Security: end-to-end security including guardrails and audit trails.

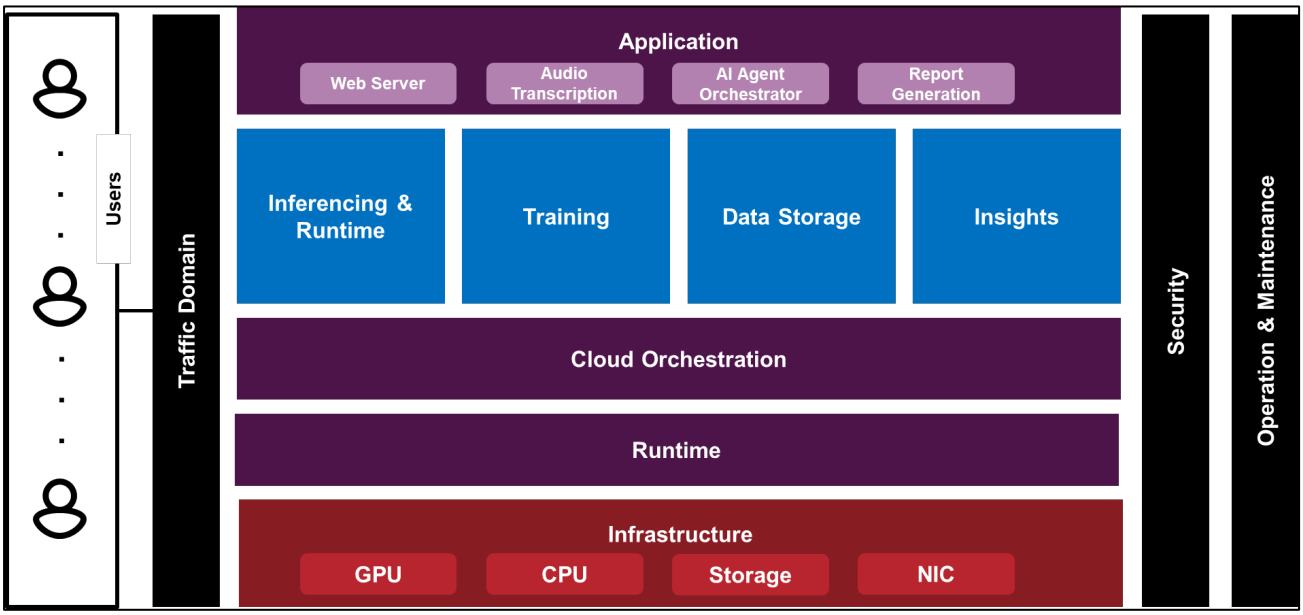

Figure 1 - High-Level Functional Architecture

Solution is designed to be future proof achieving clear roles and responsibilities between functional components. Those components are decoupled to avoid any lock-ins via clear component interactions over open API integrations. As well data pipelines are outlined from the initial capture of data in high quality resulting better social service reports. Last but not least, infrastructure layout provides an overview of all necessary infrastructure as CPU/GPU compute resources, storage and network configurations. Further details of those can be found in BoM under Appendix Section.

Table 1 shows the detailed view of High-Level Functional Architecture with functions breakdown. It also provides insight of technologies and version information used as part of this validation.

Table 1 - LVD Technology Mapping

Layer	Function	Technology	Version	Role
Infrastructure	GPU/CPU/ Storage/NIC	ThinkSystem	SR630 V4	GenAI inference and platform services; supports CPU-only and GPU-accelerated profiles.
Runtime	Operating System	Ubuntu	24.04	Long-term support base OS for secure operations.
Runtime	Hypervisor	KVM, OpenStack		Optional virtualization layer for multi-tenant or segmented environments.
Cloud Orchestration	CaaS	Docker	28.0.1	Container runtime; Kubernetes/OpenShift supported where required.
Data Storage	Document Ingestion	Unstructured Library		Converts PDFs, DOCX, images, and transcripts into JSON chunks with layout metadata.
Data Storage	Vector Store	Milvus		Supports HNSW & scalar filters; integrates easily with Kubernetes.
Data Storage	Data Processing	nomic-embed-text, Milvus		Chunking, embedding, metadata enrichment
Inferencing	Retrieval Agent	Milvus		Hybrid search and ranking for relevant evidence retrieval.
Inferencing	Generation Agent	Llama	3.1	Compose grounded responses with retrieved context
Training	Model Training	Llama	3.1	Pretrained 8B model; can be fine-tuned on jurisdiction-specific, approved safeguarding/incident. Quantised 4-bit GGUF for fast inference; gradient-checkpointing during finetune.
Application	Sanitization Agent	RAIT classifiers		Redaction & compliance check on all inbound and generated text
Application	Audio Transcription Agent	NVIDIA NeMo ASR		Speech-to-text for long-form recordings.
Application	Summarization Agent	Llama	3.1	TL;DR and section stitching for reports.

Application	Report Assembler Agent	python-docx, Pandoc		Populate DOCX/PDF template, route for sign-off
Application	Web Server	Apache		Hosts the web application and supporting services
Security	Guardrail Agent	RAIT Classifiers		Enforce policy on LLM outputs, add explanations; ; adds rationale metadata.
Operation & Maintenance	Experiment Tracking & Observability	Comet ML/Opik + Grafana/Prometheus		Dashboards for model metrics, utilization, and vector-dbDi latency.

Inference services and the RAG pipeline are deployed on Lenovo ThinkSystem SR630 V4. For an 8B class model, high-performing RAG can be deployed on CPU or GPU. For CPU acceleration, Intel AMX is recommended to reduce time-to-first-token and improve throughput to deliver a responsive user experience. GPU acceleration remains valuable for transcription and sustained high concurrency.

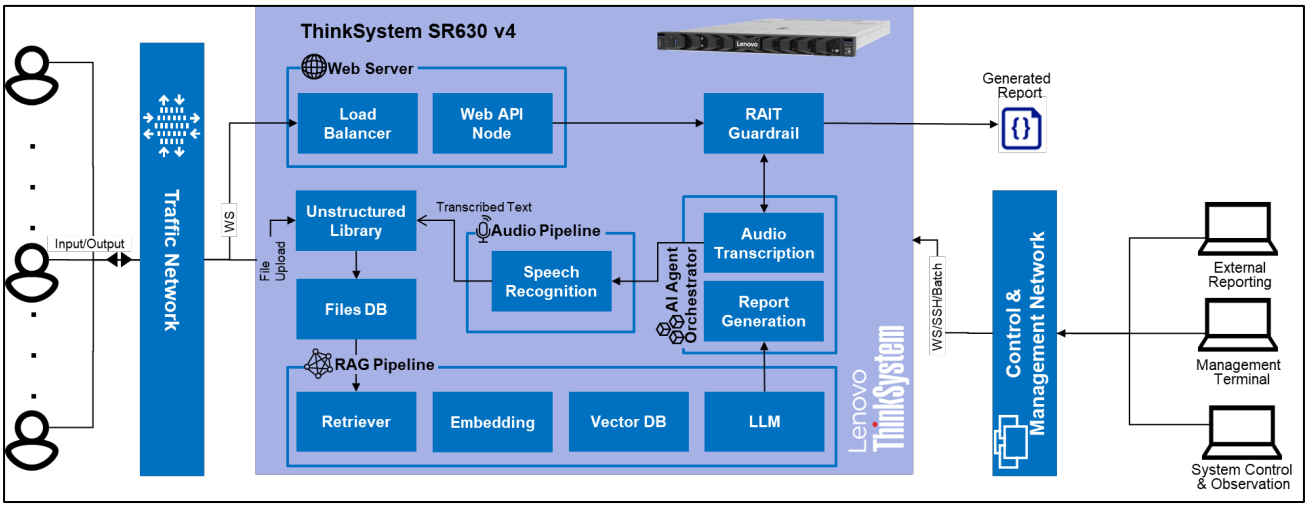

Figure 2 - Deployment Architecture Overview

End-to-end workflow summary:

- Audio Ingestion – recordings uploaded to a secure object store.
- Speech-to-Text (STT) – ASR converts audio to time-aligned transcript chunks.
- Input Guardrail (Scan) – transcript chunks and prompts screened for PII and disallowed content; non-compliant spans redacted or queued.
- Document Pre-Processing – PDFs/DOCX/images/transcripts normalized into structured chunks with metadata.
- Embedding & Vector Storage – chunks embedded and upserted to Milvus for similarity search.
- Semantic Retrieval – retrieval agent performs hybrid vector+keyword search to gather relevant context.

- Generation – context and instructions sent to the LLM endpoint to draft grounded responses.
- Output Guardrail (Shield + Steer) – output filtered; policy explanations and rationale metadata appended.
- Post-Processing & Summarization – section summaries and assembly into DOCX/PDF templates.
- Audit & Storage – prompts, embeddings, classifications, and outputs archived to an immutable store.

Dialog XR Operational Model

Dialog XR is an AI-powered assistant that supports safeguarding professionals and social and human services organizations. It is designed to help teams synthesize complex evidence, identify systemic patterns, and draft compliant reports with transparency and strong governance.

Core Capabilities

- Systemic Failure Analysis – identifies recurring patterns and contributing factors across harm cases to surface systemic issues.
- Successful Intervention Mapping – highlights interventions associated with improved outcomes and reduced risk.
- Thematic Analysis – detects emerging trends and informs policy and practice improvements.
- Policy Support – compares regional practices and identifies gaps relative to local standards and guidance.

System Architecture

Dialog XR uses a modular, service-oriented architecture with three core subsystems:

Data Ingestion Pipeline (Offline)

- Standalone ETL pipeline to retrieve approved documents from enterprise repositories (e.g., SharePoint, network shares, document management systems).
- Semantic chunking and normalization of documents and transcripts.
- Embedding and indexing into a vector database.
- Asynchronous batch processing decoupled from the live user system.

RAG Inference Engine (Backend)

- Stateless API services that manage session context in a state store (e.g., Redis).
- Retrieves relevant evidence from the vector database.
- Constructs augmented prompts and streams responses.
- Horizontally scalable with externalized state.

User Interface (Frontend)

- Web UI to manage user sessions and submit requests.
- Streams and renders responses with citations and rationale metadata.
- Maintains user history within policy constraints.
- Decoupled from backend logic to allow future UI modernization.

Workflows

Offline ingestion workflow (example):

- Administrator triggers ingestion based on governance approvals.
- Approved folders or repositories are scanned.

- Documents are retrieved to a secure staging area.
- Content is chunked, embedded, and indexed into Milvus with metadata.
- Quality checks and sampling validate ingestion completeness.

Online query workflow (example):

- User submits a query via the UI.
- Request is sent to the API with session context.
- Relevant chunks retrieved from the vector database.
- Context injected into the prompt for grounded generation.
- LLM generates a response, streamed back to the UI.
- Guardrails evaluate and, if needed, redact or steer the output.
- Interactions and outputs are archived for audit.

Table 2 – Dialog XR Core Technologies

Layer	Technology	Purpose
API Framework	FastAPI	High-performance async API with SSE support
RAG Orchestration	LangChain	Chain composition and prompt management
Vector Database	Milvus	Efficient similarity search at scale
Embeddings	OpenAI	text-embedding-3-large (3072 dimensions)
LLM Serving	Ollama	Local model inference (Llama 3.1 8B) on CPU or GPU
State Management	Redis	Distributed conversation and session context.
Frontend	Streamlit	Rapid UI development and deployment
Deployment	SLURM	HPC job scheduling and resource allocation

Deployment Considerations

Server / Compute Nodes

The Lenovo ThinkSystem SR630 V4 is a high-density, 2-socket 1U rack server designed for reliability, manageability, and security while maximizing performance and scalability. Powered by Intel Xeon 6700/6500-series processors, it supports both performance-optimized and efficiency-optimized core designs and is suited to compute-intensive workloads common in GenAI pipelines.

Figure 3 - Lenovo ThinkSystem SR630 V4 with optional security bezel

Intel® Advanced Matrix Extensions (AMX)

Intel AMX is a built-in hardware acceleration capability available on select Intel Xeon processors. It introduces dedicated matrix compute engines optimized for inference, machine learning, and deep learning, delivering substantial performance gains while maintaining power efficiency. AMX helps organizations deploy scalable, cost-effective AI processing on CPU where GPU capacity is limited or reserved for other tasks.

Table 3 - SR630 V4 Features and Benefits

Feature	SR630 V4	Benefits
Processor	• 2x Intel Xeon 6700/6500-series (P-cores) up to 86 cores and 172 threads; TDP up to 350W • 2x Intel Xeon 6700-series (E-cores) up to 144 cores; TDP up to 330W	• Higher core density • Improved performance • Consolidate more workloads per server to reduce cost
Memory	• DDR5 up to 6400 MHz • 8 channels per CPU • 32 DIMMs (16 per processor) • Up to 8TB system memory	• Faster DDR5 memory • Support large in-memory pipeline and vector workloads
Internal storage	• Front: up to 10x 2.5" SAS/SATA/NVMe • Front: up to 16x E3.S 1T NVMe or 8x E3.S 2T NVMe • Rear: up to 2x 2.5" hot-swap	• High-speed storage for embeddings and transcripts • Flexible configurations for performance and capacity

	• Onboard NVMe ports; internal/hot-swap M.2 boot	
RAID	• RAID adapters and SAS HBAs supported • Onboard NVMe focus	• Consistent RAID/HBA support • PCIe Gen 5 allows for greater storage performance
Networking	• 2x OCP slots with PCIe Gen5 x16 • Additional PCIe adapters supported	High throughput for east-west service traffic and north-south user/API flows
PCIe slots	• Up to 3x PCIe Gen 5 slots + 2x OCP slots	Expandability for accelerators, NICs, and storage adapters
Management and security	• XClarity Controller 3, • XClarity toolset • Platform Firmware • Resiliency (PFR) hardware Root of Trust (RoT)	• Enterprise manageability • Silicon-level security controls • Strong auditability and operational governance
Power	• 800W/1300W/2000W hot-plug PSUs • AC and DC options (region dependent)	Flexible power configurations across global data-center standards

Networking

A well-designed network ensures responsiveness and scalability for RAG pipelines. Recommended practices include:

- Bandwidth planning – separate management/O&M traffic from application traffic; use VLANs or physical separation for east–west service traffic and north–south user/API traffic.
- Security best practices – enforce firewalls, encryption, and RBAC to protect data flows and prevent unauthorized access.
- Segmentation – isolate training/optimization environments from production inference where required by policy.

Storage integration

High-performance storage supports concurrent users and fast handling of generated data. Recommendations include:

- Fast storage – NVMe SSDs for embeddings, indices, and active working sets.
- Scalable retention – object storage for long-term archival with immutability features (e.g., object lock).
- Organized data management – indexing and metadata standards for rapid retrieval and defensible audit trails.

Solution Validation

Validation Scope

Validation testing assessed performance, scalability, and operational efficiency of Dialog XR under realistic inference workloads. The evaluation focused on token throughput, latency, and stable resource utilization across single-session and concurrent environments. Both CPU-only and GPU-accelerated execution profiles can be validated depending on deployment requirements.

Test scenarios were designed to mirror real-world use cases, including sequential (single-session) processing and parallel multi-session workloads. Measurements included end-to-end latency, time-to-first token (TTFT), token generation throughput, and resource utilization.

Testing Methodology

Validation was performed across two primary execution modalities to establish baseline performance and evaluate scalability under load.

Sequential Baseline (Isolated Queue)

In this modality, safeguarding requests were handled one at a time, allowing each to complete fully and providing a baseline measure of inference performance without resource contention.

Concurrent Stress Testing (Parallel Load)

Multiple queries are submitted in parallel with increasing concurrency to identify saturation behavior and performance degradation thresholds. Continuous batching and parallel execution are evaluated to characterize throughput scaling.

Results Summary

Performance testing shows that enabling Intel AMX and runtime optimizations delivers meaningful CPU performance gains, improving throughput and reducing latency compared to baseline execution. These AMX-optimized pipelines provide strong, production-ready performance for conversational RAG workloads. GPU-accelerated configurations become advantageous when deployments must process long audio files, handle peak operational demand, or maintain responsiveness under heavy parallel workloads. Evaluating both execution paths demonstrates how organizations can select CPU-only, GPU-only, or hybrid deployment profiles based on cost, data sovereignty, and throughput requirements.

Key takeaways:

- AMX can transform CPU inference from a fallback into a production-capable option for many conversational RAG workloads.
- Optimized CPU deployments can meet interactive experience metrics such as TTFT and stable token throughput.
- CPU-only profiles are particularly relevant for cost-sensitive and sovereign deployments.
- GPU acceleration is especially useful for high-volume transcription and workloads requiring heavy parallel throughput.

Solution Summary

This Lenovo Validated Design (LVD) supports social and human services organizations worldwide, including child protection, adult safeguarding, domestic and gender-based violence prevention, homelessness services, behavioral health, and other vulnerable-population programs. While example use cases reference common safeguarding practices, the architecture is designed to align with local regulatory frameworks and data sovereignty requirements across jurisdictions.

This solution delivers a secure, on-premises, responsible AI platform that helps agencies manage complex case data, accelerate mandated reviews, and improve decision quality. By combining Lenovo ThinkSystem infrastructure with Intel Xeon processors and Intel AMX acceleration, it provides a high-performance, cost-efficient AI foundation capable of operating within sovereign, air-gapped environments. Dialog XR leverages retrieval-augmented generation (RAG), semantic search, and multi-agent orchestration to transform large volumes of unstructured, multi-agency information into actionable, explainable insights, enabling faster, more informed casework while maintaining governance and transparency.

Integrated responsible AI guardrails, immutable audit trails, and explainability frameworks support compliance with legal, ethical, and regulatory expectations. This LVD demonstrates how CPU-accelerated AI, enhanced by Intel AMX and runtime optimizations, can reliably support generative AI workloads in sensitive public-sector environments, reducing operational burden, improving service outcomes, and providing a scalable path to adoption.

In addition to CPU-optimized inferencing, the architecture supports optional GPU acceleration for workloads that benefit from high-parallel compute, including:

- Long-form speech-to-text (STT), where GPU-accelerated Automatic Speech Recognition (ASR) significantly reduces transcription time for multi-hour recordings.
- High-concurrency environments, where GPUs can sustain larger volumes of parallel RAG requests or simultaneous reviews.
- Future model scaling, supporting larger parameter LLMs or multimodal pipelines as adoption matures.

This hybrid CPU+GPU approach ensures agencies can tailor deployments to their regulatory, budget, and performance needs, running fully on CPU where cost efficiency is paramount, or enabling GPU acceleration for throughput-intensive workloads.

Appendix B: Lenovo Bill of materials (BOM)

Note: *CPU, GPU, and memory configurations are subject to adjustment based on specific customer requirements.

ThinkSystem SR 630 V4

Part Number	Product Description	Total Qty
7DG9CTO1WW	ThinkSystem SR630 V4 - 3yr Warranty	1
C1XE	ThinkSystem 1U V4 10x2.5" Chassis	1
C3J9	ThinkSystem General Computing - Max Performance	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
C5QM	Intel Xeon 6787P 86C 350W 2.0GHz Processor	**2** *
C1XJ	ThinkSystem 1U V4 Performance Heatsink	2
C0TQ	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM	**32** *
5977	Select Storage devices - no configured RAID required	1
C0JK	ThinkSystem M.2 B340i-2i NVMe Enablement Adapter	1
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1
BK1J	ThinkSystem Broadcom 57508 100GbE QSFP56 2-Port PCIe 4 Ethernet Adapter	1
C1YH	ThinkSystem SR630 V4 x16/x16 PCIe Gen5 Cable Riser 1	1
C1Z7	ThinkSystem SR630 V4 Full Height+Low Profile Riser1 Cage	1
C0U3	ThinkSystem 2000W 230V Titanium CRPS Premium Hot-Swap Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
C1YT	ThinkSystem 1U V4 Performance Fan Module	4
C1YP	ThinkSystem 1U V4 Standard Media Bay	1
C2DH	ThinkSystem Toolless Slide Rail Kit V4	1
BPKR	TPM 2.0	1
B7XZ	Disable IPMI-over-LAN	1
BK15	High voltage (200V+)	1
BZ7F	ThinkSystem WW Lenovo LPK, Birch Stream	1
C20J	ThinkSystem SR630 V4 Service Label for WW	1
AWF9	ThinkSystem Response time Service Label LI	1
C20U	ThinkSystem 2000W TT power rating label WW	1
C20D	ThinkSystem SR630 V4 model name Label	1
B97B	XCC Label	1
BQPS	ThinkSystem logo Label	1
C1ZN	ThinkSystem SR630 V4 Agency label with ES&CE&UKCA	1

AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	1
C212	ThinkSystem BHS SR630 V4 1U PCIe number 1-2 and OCP number Label (BF+Rear M.2)	1
C1YA	ThinkSystem M.2 Signal&Power Cable, ULP 82P-SLX4/2X10 SB, 540/680mm	1
BE0E	N+N Redundancy With Over-Subscription	1
B0ML	Feature Enable TPM on MB	1
C3NP	ThinkSystem SR630 V4 MI-BF Cbl Riser to P9	1
CAR5	SR630 V4 Laser service indicator	1
C4DV	ThinkSystem SR630 V4 MotherBoard	1
C3K9	XClarity Platinum Upgrade v3	1
CA7N	ThinkSystem SR630 V4 System I/O board v2	1
C26Z	ThinkSystem GNR XCC CPU HS Clip	2
C5XG	ThinkSystem SR630 V4 General Config PKG AC+CL	1
AVEN	ThinkSystem 1x1 2.5" HDD Filler	4
C1YY	ThinkSystem 1U V4 Low Profile Riser Cage Filler	1
BCEB	ThinkSystem 1U V2 2x3 2.5" HDD Dummy	1
BPP5	OCP3.0 Filler with screw	1
B8NK	ThinkSystem 1U Super Cap Holder Dummy	1
BTTY	M.2 NVMe	1
CBDC	ENERGY STAR Certification Country	1
7S0XCTO8WW	XClarity Controller Prem-FOD	1
SCY0	Lenovo XClarity XCC3 premier - FOD	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1
3444	Registration only	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1
QAJQ	SR630 V4 3Yr	1
QAK6	KYD	1
QA0Y	Months	36
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
QA18	Premier	1
QAJQ	SR630 V4 3Yr	1
QA0Y	Months	36
QA12	24x7 4hr Resp	1

Appendix C: Abbreviations

Acronym	Meaning
AI	Artificial Intelligence
API	Application Programming Interface
CaaS	Container as a Service
CPU	Central Processing Unit
GPU	Graphics Processing Unit
RAG	Retrieval-Augmented Generation
RBAC	Role-Based Access Control
STT	Speech to Text
SSE	Server-Sent Events
PII	Personally Identifiable Information
TTFT	Time to First Token

Resources

Resources	Links
Dialog XR	www.Dialog XR.ai
SR630 V4	Lenovo ThinkSystem SR630 V4 Server Product Guide > Lenovo Press
Lenovo XClarity	Systems Management
LOC-A	Lenovo Open Cloud Automation (LOC-A)

Document history

Version 1.0	February 2026	First version includes Lenovo SR630 V4, Intel CPU with AMX, Dialog XR

Trademarks and special notices

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ThinkEdge®
ThinkShield®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

AMD and AMD EPYC™ are trademarks of Advanced Micro Devices, Inc.

Intel® and Intel Core® are trademarks of Intel Corporation or its subsidiaries.

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.
All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.
Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serves as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites are at your own risk