# **Model Selection & Hardware Specification Report**

**Project:** Saudi Arabic Document Assistant in LMS  
**Prepared for:** MoD  
**Date:** 5 Aug 2025

## **1. Model Comparison – Selection Criteria**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Qwen‑2.5‑72B‑Instruct** | **JAIS‑30B** |
| **Arabic Ability** | Strong multilingual Arabic (Modern Standard + Gulf dialects) via large, diverse corpus | Native Arabic focus, optimized for Arabic/English |
| **Grounded Answers** | 128K context allows whole-document grounding (less chunking, fewer hallucinations) | Shorter context (16K), more chunking needed |
| **Hallucination Resistance** | Rigorous alignment (DPO, SFT) – lower hallucination rate | Strong in Arabic but fewer large-scale eval metrics |
| **Conciseness** | Highly instruction-tuned for concise, structured outputs | Concise in Arabic, natural phrasing |
| **Performance in Mixed Domains** | Strong general-purpose, domain-agnostic | Best in Arabic-heavy domains if fine-tuned |
| **Scaling** | Handles multi-document chat with minimal engineering | Works well but needs tighter RAG integration |
| **Recommendation** | Preferred primary model better grounding & context handling | Consider only if Arabic tone/domain nuance outweighs context needs |

### Conclusion:

* **Primary recommendation:** Qwen‑2.5‑72B‑Instruct, due to ultra-long context window and lower hallucination rate, ideal for RAG over Arabic documents without excessive chunking.
* **Secondary option:** JAIS‑30B, only if *native Arabic tone & domain nuance* is prioritized over extreme context length.

## **2. Hardware Specifications**

### Part 1: System Specification Tables with User Capacity

#### A. Minimum Viable Configuration (per system)

|  |  |
| --- | --- |
| **Component** | **Specification** |
| Server Platform | Supermicro SYS-420GP-TNAR+ / ASUS ESC-N8A-E11 |
| GPUs | 4× NVIDIA A100 80GB SXM4 |
| CPU | 2× Intel Xeon Silver 4316 (2×16-core) |
| RAM | 512 GB DDR4 ECC |
| GPU Interconnect | NVLink (4-GPU topology) |
| Storage | 2× 3.84 TB NVMe SSD (Gen4) |
| Networking | Dual 25 GbE NICs |
| Power Draw (per node) | ~3.0–3.5 kW |
| Cooling Requirement (per node) | ~12,000–14,000 BTU/hr |
| Approx. Concurrent Users Supported | ~200–300 real-time LLM sessions |

#### B. High Headroom / Future-Proof Configuration (per system)

|  |  |
| --- | --- |
| **Component** | **Specification** |
| Server Platform | NVIDIA DGX H100 / Dell PowerEdge XE9680 / Supermicro 420GP |
| GPUs | 8× NVIDIA H100 80GB SXM5 (NVSwitch) |
| CPU | 2× Intel Xeon 8592+ (2×64-core) |
| RAM | 1.5–2 TB DDR5 ECC |
| Storage | 4× 7.68 TB NVMe Gen4 SSDs (RAID 10) |
| Networking | Dual 100 GbE or InfiniBand HDR (200 Gbps) |
| Power Draw (per node) | ~6.5–7.0 kW |
| Cooling Requirement (per node) | ~24,000–26,000 BTU/hr |
| Approx. Concurrent Users Supported | ~450–600 real-time LLM sessions |

### Part 2: Rack Planning with User Distribution

#### Assumptions:

* Standard 42U rack
* 4U per high-end server, 4–5 servers max per rack
* 4–6 kW per system, high-density cooling needed
* Cooling via in-rack liquid cooling or rear-door heat exchangers recommended

#### A. Per Rack Capacity Summary

|  |  |  |  |
| --- | --- | --- | --- |
| **Rack Type** | **Server Type** | **Servers per Rack** | **Users per Rack** |
| Minimum Config Rack | 4× A100 80GB nodes | 4 | ~1,000 |
| High Headroom Rack | 8× H100 80GB SXM5 nodes | 4 | ~2,000 |
| Embedding / Vector DB | 2–3× L40S or A100 light nodes | 3 | Supports ~6,000+ |
| Admin/Monitoring Rack | No GPU / Light GPU | 2–3 | N/A |

#### B. Cooling Setup Recommendation

|  |  |
| --- | --- |
| Configuration | Cooling Setup |
| Minimum Config | Hot aisle/cold aisle + 20 kW CRAC units per rack or rear-door heat exchangers |
| High Headroom Config | Direct-to-chip liquid cooling (e.g., CoolIT) or rear-door liquid-cooled units, 30–35 kW per rack |
| All Racks | Raised floor + environmental monitoring (for airflow, pressure, humidity) |

### Part 3: Deployment Planning for 6,000 Concurrent Users

|  |  |  |  |
| --- | --- | --- | --- |
| **Configuration Type** | **Servers Needed** | **Racks Required** | **Total Users Covered** |
| High Headroom | 12–15 servers | 3–4 racks | 6,000–7,500 |
| Minimum Config | 20–24 servers | 5–6 racks | 6,000–7,200 |

### Part 4: Summary Table for IT/Procurement Teams

|  |  |  |  |
| --- | --- | --- | --- |
| **Layer** | **Count** | **Capacity per Unit** | **Total Capacity** |
| Inference Servers | 12–15 (H100) or 20–24 (A100) | 400–600 (H100) / 200–300 (A100) users | 6,000+ users |
| Rack Units | 3–4 (H100) or 5–6 (A100) | 2,000+ users (H100) / 1,000+ (A100) | Scalable as needed |
| Retrieval Nodes | 2–3 | Shared async vector DB (e.g., Faiss, Qdrant) | N/A |
| Admin/Storage | 2–3 | Monitoring, logging, backups | N/A |
| Storage Backend | ~50–100 TB (NVMe + NAS/RAID) | PDF/doc ingestion, logs, finetuned weights | Scalable |

## 3. Basic Architectural Overview

