QSAN Official Document

On-Premises AI Setup Guide

Application Note

July 2026

ANNOUNCEMENT

Copyright

© Copyright 2026 QSAN Technology, Inc. All rights reserved. No part of this document may be reproduced or transmitted without written permission from QSAN Technology, Inc.

QSAN believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

Trademarks

- QSAN, the QSAN logo, and QSAN.com are trademarks or registered trademarks of QSAN Technology, Inc.
- Microsoft, Windows, Windows Server, and Hyper-V are trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries.
- Linux is a trademark of Linus Torvalds in the United States and/or other countries.
- UNIX is a registered trademark of The Open Group in the United States and other countries.
- Mac and OS X are trademarks of Apple Inc., registered in the U.S. and other countries.
- Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
- VMware, ESXi, and vSphere are registered trademarks or trademarks of VMware, Inc. in the United States and/or other countries.
- Citrix and Xen are registered trademarks or trademarks of Citrix Systems, Inc. in the United States and/or other countries.
- Other trademarks and trade names used in this document to refer to either the entities claiming the marks and names or their products are the property of their respective owners.

TABLE OF CONTENTS

Announcement i
Notices v
Preface vi
Technical Support vi
Information, Tip, and Caution vi
1. Introduction to ON-Premises LLM 1
1.1. KSM × LobeChat × OpenAI 20B LLM 1
2. Configuration Settings 2
2.1. Environment and Topology 2
2.2. Application Setup Overview 3
2.3. Vector Database Workflow 4
3. Conclusion 6
4. Appendix 7
4.1. Apply To 7

©2026 QSAN Technology, Inc. All rights reserved.

FIGURES

Figure 1-1 On-Premises AI Topology 1
Figure 2-1 Demonstration Topology 3

Announcement
©2026 QSAN Technology, Inc. All rights reserved.
www.qsan.com

Official Document

TABLES

Table 2-1 Persistent Volume Configuration..... 4
Table 2-2 Environment Variables..... 4

Announcement
©2026 QSAN Technology, Inc. All rights reserved.
www.qsan.com

NOTICES

Information contained in this document has been reviewed for accuracy. But it could include typographical errors or technical inaccuracies. Changes are made to the document periodically. These changes will be incorporated in new editions of the publication. QSAN may make improvements or changes in the products. All features, functionality, and product specifications are subject to change without prior notice or obligation. All statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

©2026 QSAN Technology, Inc. All rights reserved.

PREFACE

Technical Support

Do you have any questions or need help trouble-shooting a problem? Please contact QSAN Support, we will reply to you as soon as possible.

- Via the Web: https://www.qsan.com/technical_support
- Via Telephone: +886-2-77206355
- (Service hours: 09:30 - 18:00, Monday - Friday, UTC+8)
- Via Skype Chat, Skype ID: qsan.support
- (Service hours: 09:30 - 02:00, Monday - Friday, UTC+8, Summer time: 09:30 - 01:00)
- Via Email: support@qsan.com

Information, Tip, and Caution

This document uses the following symbols to draw attention to important safety and operational information.



INFORMATION

INFORMATION provides useful knowledge, definition, or terminology for reference.



TIP

TIP provides helpful suggestions for performing tasks more effectively.

©2026 QSAN Technology, Inc. All rights reserved.



CAUTION

CAUTION indicates that failure to take a specified action could result in damage to the system.

©2026 QSAN Technology, Inc. All rights reserved.

1. INTRODUCTION TO ON-PREMISES LLM

This guide helps KSM users to deploy AI workloads locally with a secure and efficient container orchestration platform. This document demonstrates how QSAN's KS series platform can host open-source LLM (Large Language Model) services, such as LobeChat and the OpenAI 20B model, providing a fully internal, cost-effective AI infrastructure with minimal effort. It will also enable the deployment of AI workloads locally using a secure and efficient container orchestration platform.

1.1. KSM × LobeChat × OpenAI 20B LLM

This use case illustrates how KSM hosts LobeChat, an open-source conversational interface, with an internal inference service (OpenAI 20B) and optional vector database for knowledge retrieval (RAG pipeline). This provides enterprises with secure, AI-enhanced knowledge workflows.

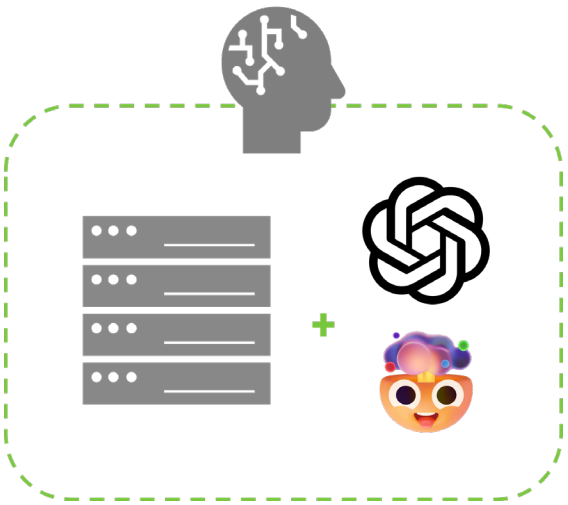


Figure 1-1 On-Premises AI Topology

©2026 QSAN Technology, Inc. All rights reserved.

2. CONFIGURATION SETTINGS

KSM provides a simple way to deploy container-based services, achieving scalability and flexibility for enterprise AI infrastructure. When combined with LobeChat and OpenAI 20B, users can deploy a complete LLM stack without relying on the public cloud.

2.1. Environment and Topology

Demonstration Environment

- Server
  - Model: KS2204-8C x 4 (1 x Master Node + 3 x Worker Nodes)
    CPU: 8 cores
    Memory: 16 GB
    Local Capacity: 1 TB NVMe SSD
    Network Connectivity: 10 GbE SFP+
    OS: KSM 1.0.0

Demonstration Topology

The connections are very simple, here are the connections in brief and any topology will work if the logical connection is pingable.

©2026 QSAN Technology, Inc. All rights reserved.

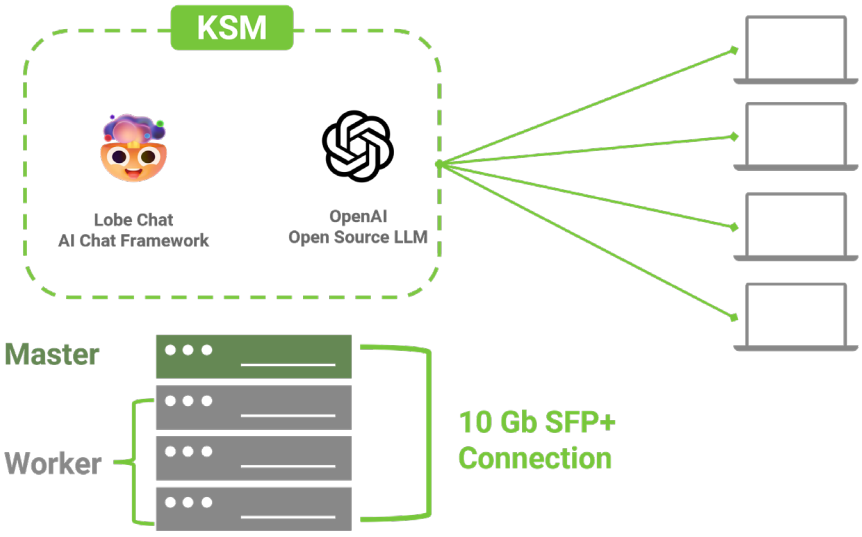


Figure 2-1 Demonstration Topology

2.2. Application Setup Overview

Step 1. KSM Cluster Setup

1. Initialize KSM node with static IP (e.g., 192.168.200.100).
2. Create a storage pool for persistent volumes.
3. Start Kubernetes services and verify cluster health.
4. Check node connectivity over 1/10GbE.

Step 2. Deploy Container Images

1. ollama:latest — LLM inference service
2. lobehub/lobechat:latest — Chat frontend
3. qdrant/qdrant:latest — Vector database (optional)

Step 3. Persistent Volume Configuration

©2026 QSAN Technology, Inc. All rights reserved.

Table 2-1 Persistent Volume Configuration

VOLUME NAME	DESCRIPTION	CAPACITY
llm-model	Model storage	300 GB
lobechat-data	User settings	5 GB
vector-store	Knowledge base	20 GB (optional)

Step 4. Environment Variables

Table 2-2 Environment Variables

VARIABLE	DESCRIPTION
LLM_API_BASE	Backend service endpoint
DEFAULT_MODEL	Default model name
QDRANT_ENDPOINT	Vector DB URL
ACCESS_CODE	Authentication key

Step 5. Verification

1. Web access: http://:31300
2. Check pod status: ksmcli get pods -l app=lobechat
3. Health check: curl http://<node-ip>:31300

2.3. Vector Database Workflow

Document Processing

1. Content Management – Upload enterprise documents.

©2026 QSAN Technology, Inc. All rights reserved.

2. Chunking – Split into semantic sections.
3. Embedding Generation – Convert text into vectors.
4. Indexing – Store vectors in the database.

Query Processing

1. Query – User submits prompt.
2. Embedding – Encode query.
3. Retrieval – Find relevant chunks.
4. Augment – Add context to the prompt.

©2026 QSAN Technology, Inc. All rights reserved.

3. CONCLUSION

Deploying the KSM On-Premises LLM deployment allows enterprises to build internal, secure, and high-performance AI environments. This integration eliminates dependency on cloud-based LLMs and protects sensitive data while maintaining scalability and cost control.

By combining QSAN's KSM orchestration with open-source AI tools, users can streamline knowledge processing workflows, develop context-aware assistants, and accelerate internal AI adoption with confidence and compliance.

©2026 QSAN Technology, Inc. All rights reserved.

4. APPENDIX

4.1. Apply To

- KSM firmware 1.0.0 and later

©2026 QSAN Technology, Inc. All rights reserved.