PrivateLLM構築士

3 級 公式テキスト

Private LLM Engineer Certification Level 3

発行	プライベートAI推進協会（PAPA）
バージョン	ver.1.0 ／ 2026年2月
対象資格	PrivateLLM構築士 3級
配布	無料（privatellm.jp）

本テキストは PrivateLLM構築士 3級 試験の全出題範囲をカバーした公式学習資料です。
ITの専門知識は不要です。独学1〜2週間で合格を目指せます。

目次

はじめに ― 本テキストの使い方

本テキストは、PrivateLLM構築士3級の試験合格を目指す方のための公式学習資料です。プログラミングやIT専門知識は一切必要ありません。

■ 推奨学習の流れ

STEP 1 本テキストを通読する（3〜4時間）

STEP 2 各章末の重要ポイントを確認する

STEP 3 巻末の練習問題を解く

STEP 4 間違えた箇所を重点的に復習する

STEP 5 privatellm.jpの模擬試験で仕上げる

■ 試験概要（3級）

項目	内容
出題数	40問（4択）
試験時間	60分
合格基準	70%以上（28問正解）
受験形式	オンライン
受験料	8,000円（税込）

第1章 プライベートLLMの基礎概念

出題比率：25%（約10問）

1-1 LLMとは何か

LLM（Large Language Model：大規模言語モデル）とは、膨大なテキストデータを学習した人工知能モデルです。文章の生成・要約・翻訳・質問応答など、あらゆる言語処理タスクをこなせます。

■ LLMの仕組み（概要）

LLMはトランスフォーマーという構造を持ち、単語の並びのパターンを学習しています。入力された文章の「次に来る言葉」を予測することで、自然な文章を生成します。

主な特徴として以下が挙げられます：

- 膨大なデータ（数千億〜数兆文字）を学習している
- 指示（プロンプト）に従って回答を生成する
- 文脈を理解して会話を継続できる
- 専門知識から日常会話まで幅広く対応できる

1-2 クラウドLLMとプライベートLLMの違い

LLMの運用方式は大きく2種類に分かれます。

比較項目	クラウドLLM	プライベートLLM
代表例	ChatGPT・Claude・Gemini	Llama・Mistral・Gemma
データの扱い	外部サーバーに送信	社内で完結
情報漏洩リスク	高い	ほぼゼロ
月額コスト	従量課金（青天井）	固定費（初期投資後）
カスタマイズ	制限あり	自由にカスタマイズ可能
導入の手軽さ	即日利用可能	構築・設定が必要
インターネット	必要	不要（オフライン動作可）

重要　プライベートLLMの最大の利点は「データが外部に出ない」ことです。医療・法律・金融など機密情報を扱う業種で特に注目されています。

1-3 主要なオープンソースモデル

プライベートLLMで使われる代表的なオープンソースモデルを覚えておきましょう。

Llama（ラマ）

Meta（旧Facebook）が開発。世界で最も広く使われているオープンソースLLM。

Mistral（ミストラル）

フランスのMistral AI開発。軽量で高性能。Mixtralはその大規模版。

Gemma（ジェンマ）

Googleが開発した軽量モデル。小型GPUでも動作可能。

Qwen（チュエン）

Alibaba（中国）開発。日本語・中国語に強い。

Phi（ファイ）

Microsoftが開発。超軽量で低スペック端末でも動作。

■ 第1章 重要ポイントまとめ

- ☒ LLM＝膨大なデータで学習した言語生成AI
- ☒ クラウドLLMはデータを外部送信する→情報漏洩リスクあり
- ☒ プライベートLLM＝社内で動かすAI→データが外に出ない
- ☒ 代表モデル：Llama・Mistral・Gemma・Qwen・Phi

第2章 セキュリティ・コンプライアンス

出題比率：20%（約8問）

2-1 情報漏洩リスク

クラウドAIを利用する際、入力したデータは外部サーバーに送信されます。以下のようなリスクがあります。

- 社内の機密情報・顧客情報が外部に送信される
- AIの学習データとして利用される可能性がある
- 競合他社と同じサーバーにデータが保存される場合がある
- セキュリティインシデント時にデータが漏洩する恐れがある

重要　実際に2023年、某大手企業の社員がChatGPTに社内機密を入力し、情報が漏洩した事例が報告されています。プライベートLLMはこのリスクを原理的にゼロにできます。

2-2 個人情報保護法・GDPR・EU AI Act

■ 個人情報保護法（日本）

AIに個人情報を入力・処理させる場合、個人情報保護法の規定に従う必要があります。第三者提供の制限、利用目的の特定など、クラウドAI利用時は注意が必要です。

■ GDPR（EU一般データ保護規則）

EUの個人データ保護規則。EU域内の個人データを扱う場合、日本企業も対象になります。違反した場合は高額の制裁金（最大売上高の4%）が課されます。

■ EU AI Act（EU AI規制法）

2024年に施行されたAIに関する世界初の包括的規制。AIをリスクレベルで分類し、高リスクAIには厳格な要件を課します。プライベートLLMは自社管理のため規制対応がしやすいメリットがあります。

2-3 業種別規制と社内AIガバナンス

業種	AI利用の主な規制・注意点
医療	患者情報の外部送信禁止。電子カルテデータの取扱いに法規制あり
金融	顧客情報・取引データの外部送信を厳しく規制
法律	依頼人の機密情報（守秘義務）の外部送信リスク

行政	個人情報・機密情報の外部サービス利用を原則禁止
教育	生徒の個人情報・成績データの管理に厳格な規定

社内AIガバナンスとは

AI利用のルールや管理体制のこと。「どの情報をAIに入力していいか」「誰がAIを管理するか」「問題が起きたときの責任はどこか」などを明確にした社内規定です。

■ 第2章 重要ポイントまとめ

- ☒ クラウドAIへのデータ送信→情報漏洩リスクあり
- ☒ 個人情報保護法・GDPRに基づく適切なデータ管理が必要
- ☒ EU AI Actは世界初の包括的AI規制法
- ☒ 医療・金融・法律・行政は特に厳格な規制あり
- ☒ 社内AIガバナンス＝AIの利用ルール・管理体制

第3章 導入判断と費用対効果

出題比率：20%（約8問）

3-1 導入検討のフレームワーク

プライベートLLMが向いているケースと向いていないケースを整理しましょう。

プライベートLLMが向いているケース	クラウドLLMが向いているケース
機密情報・個人情報を扱う業務	すぐに使い始めたい
法規制でデータ外部送信が禁止	小規模・短期間の利用
大量のデータを継続的に処理	GPUなどの設備投資が難しい
インターネット接続できない環境	常に最新モデルを使いたい
長期的なコスト削減を重視	技術担当者がいない

3-2 コスト比較とROI

■ コスト構造の違い

	クラウドLLM（API）	プライベートLLM
初期費用	ほぼゼロ	GPU購入費（数十〜数百万円）
月額費用	従量課金（使うほど増加）	サーバー電気代・保守費のみ
大量利用時	月数十〜数百万円に膨張	コスト変わらず
損益分岐	—	一般的に1〜2年で回収

重要 ROI（投資対効果）の観点では、月のAPI利用料が30万円を超える規模の組織では、プライベートLLMへの移行で1〜2年以内にコスト回収できるケースが多いです。

3-3 ハードウェアと導入ステップ

■ 必要なハードウェアの基礎知識

GPU（グラフィックスカード）

LLMの計算処理に必須。NVIDIA製が主流。VRAMが多いほど大きなモデルを動かせる。

VRAM（ビデオメモリ）

モデルサイズに応じて必要量が変わる。7Bモデル→約8GB、13Bモデル→約16GB程度

RAM（メインメモリ）

VRAMに加えてシステムメモリも必要。32GB以上推奨

ストレージ

モデルファイルの保存に使用。SSDが高速で推奨。1モデルあたり4〜30GB程度

■ 導入ステップの概要

STEP 1 目的・要件の定義（何に使うか、どのデータを使うか）

STEP 2 PoC（概念実証）— 小規模テストで効果を確認

STEP 3 検証フェーズ — 実際の業務データで精度確認

STEP 4 本番導入 — セキュリティ設計・運用体制の整備

STEP 5 継続改善 — モデル更新・精度向上・コスト最適化

■ 第3章 重要ポイントまとめ

☒ 機密データ扱う・法規制あり・大量処理→プライベートLLMが有利

☒ クラウドAPIは従量課金→大量利用でコスト膨張

☒ GPU（特にVRAM容量）がパフォーマンスの鍵

☒ 導入は目的定義→PoC→検証→本番の順序で進める

第4章 主要ツール・技術の概要

出題比率：20%（約8問）

4-1 Ollama・vLLM

Ollama（オラマ）はローカル環境でLLMを簡単に動かすためのオープンソースツールです。コマンド一つでモデルをダウンロードして実行できます。

主な特徴：

- コマンドラインで簡単にモデルを管理できる
- Windows・Mac・Linuxに対応
- 多数のオープンソースモデルに対応（Llama・Mistral・Gemma等）
- 個人利用から小規模企業まで幅広く使われている

vLLM（ブイエルエルエム）は大規模な同時接続に対応した高速LLM推論フレームワークです。

多数のユーザーが同時にアクセスする企業環境での本番運用に適しています。Ollamaより設定は複雑ですが、処理速度と同時接続数で優れています。

4-2 RAG（検索拡張生成）

RAG（Retrieval-Augmented Generation）は、社内文書などの独自データをAIに参照させる技術です。

■ RAGの仕組み

→ 社内文書・マニュアル・データベースをベクトル化して保存

→ ユーザーが質問を入力

→ 関連する文書をベクトル検索で取得

→ 取得した文書をLLMに渡して回答を生成

重要　RAGを使うことで、LLMを再学習させずに社内固有の知識（製品マニュアル・社内規定・顧客情報など）をAIに活用させることができます。コストと時間を大幅に節約できます。

4-3 量子化・Dify・LibreChat

■ 量子化（Quantization）

モデルの数値精度を下げてファイルサイズと必要VRAMを削減する技術。たとえばQ4（4bit量子化）にすると元の1/4程度のVRAMで動作可能になります。精度は多少下がりますが、実用上問題ないケースがほとんどです。

■ Dify（ディファイ）

LLMを活用したアプリケーションをノーコード・ローコードで構築できるオープンソースプラットフォーム。RAG構築・ワークフロー設計・エージェント作成などをGUI操作で実現できます。

■ LibreChat（リブレチャット）

ChatGPTのような使い勝手を社内のプライベートLLMで実現するオープンソースのチャットインターフェース。複数モデルの切り替えや会話履歴の管理が可能です。

■ 第4章 重要ポイントまとめ

- ☒ Ollama＝ローカルLLMを簡単に動かすツール
- ☒ vLLM＝企業の本番環境向け高速推論フレームワーク
- ☒ RAG＝社内文書を再学習なしにAIに活用させる技術
- ☒ 量子化＝モデルを軽量化してVRAM消費を削減
- ☒ Dify＝ノーコードでAIアプリを構築するプラットフォーム

第5章 ビジネス活用事例

出題比率：15%（約6問）

5-1 業種別活用事例

製造業

製品マニュアル・設計図をRAG化。現場作業員が音声で質問→即座に回答。トラブルシューティング時間を大幅短縮。

医療機関

電子カルテの要約・退院サマリー作成を自動化。患者情報は外部送信なしで処理。医師の記録業務を50%削減した事例あり。

法律事務所

契約書のリスク箇所の自動抽出・法律文書の要約。依頼人の機密情報を外部送信せずにAI活用が可能。

金融機関

社内規定・コンプライアンス文書をRAG化。社員からの質問に自動回答。内部情報の外部漏洩リスクゼロ。

地方自治体

住民からの問い合わせ対応を自動化。個人情報を含む問い合わせもオンプレミスで安全に処理。

5-2 社内FAQ・文書要約・マルチモーダル

■ 社内FAQシステム

社内規定・人事制度・業務マニュアルをRAG化して、社員からの質問に24時間自動回答。特に新入社員・異動者の問い合わせ対応コスト削減に効果的です。

■ 文書要約・レポート生成

長い会議議事録・報告書・メールを自動要約。週次レポートの自動生成。社内文書の検索・分類の自動化。これらをクラウド送信なしで実現できます。

■ マルチモーダルAI

テキストだけでなく画像・音声・動画も処理できるAIのこと。たとえば製品の傷を写真で撮影してAIが検品、設計図の画像から部品情報を抽出、などの活用例があります。

5-3 失敗事例と対策

失敗パターン	原因	対策
ハルシネーション（幻覚）	LLMが事実でない内容を自信満々に生成する現象	RAGで正確な情報源を参照させる。重要情報は必ず人間が確認する。
精度不足	モデルが小さすぎて実務に使えない	用途に合ったモデルサイズを選択。ファインチューニングを検討。
レスポンス低下	GPUスペック不足で応答が遅い	必要VRAMを事前に試算。量子化でモデルを軽量化。
セキュリティ設定漏れ	ローカルLLMのAPIを社外に公開してしまう	ファイアウォール・アクセス制御を適切に設定。

■ 第5章 重要ポイントまとめ

- ☒ RAGで社内文書をAI化→情報漏洩ゼロで社内知識を活用
- ☒ ハルシネーション対策→RAG活用＋人間による確認
- ☒ マルチモーダル＝テキスト以外も処理できるAI
- ☒ セキュリティ設定（アクセス制御）は必須

練習問題（20問）

以下の問題に答えてください。解答は次ページにあります。

Q1 LLMの正式名称として正しいものはどれか。

A. Large Learning Machine

B. Large Language Model

C. Local Learning Module

D. Linear Language Method

Q2 プライベートLLMの最大のメリットはどれか。

A. 無料で使える

B. データが外部に出ない

C. 常に最新モデルが使える

D. 設定が不要

Q3 Ollamaの説明として正しいものはどれか。

A. クラウドAIサービス

B. ローカルLLMを動かすツール

C. データベース管理システム

D. 画像生成AI

Q4 RAGの目的として正しいものはどれか。

A. モデルを軽量化する

B. 社内文書をLLMに参照させる

C. GPUの速度を上げる

D. モデルを再学習させる

Q5 量子化の主な目的はどれか。

A. モデルの精度を上げる

B. 学習速度を速める

C. 必要VRAMを削減する

D. インターネット速度を上げる

Q6 EU AI Actの説明として正しいものはどれか。

A. 米国のAI規制法

B. AIに関する世界初の包括的規制法（EU）

C. 日本のAI基本法

D. AIの特許に関する規制

Q7 ハルシネーションの説明として正しいものはどれか。

A. モデルが学習を拒否すること

B. AIが事実でない内容を生成すること

C. GPUが過熱すること

D. データが消えること

Q8 次のうち、プライベートLLMの導入に最も向いているケースはどれか。

A. 個人ブログを書く

B. 患者情報を使った医療文書作成

C. 旅行の観光情報を調べる

D. 1日だけAIを試してみる

Q9 vLLMが特に向いている用途はどれか。

A. 個人のPC学習環境

B. 企業での大規模・多並列処理

C. スマートフォンでの利用

D. オフィスの文書作成

Q10 Difyの説明として正しいものはどれか。

A. GPU製造メーカー

B. ノーコードでAIアプリを構築するプラットフォーム

C. オープンソースのLLMモデル

D. クラウドストレージサービス

Q11 Meta社が開発したオープンソースLLMはどれか。

A. GPT-4

B. Gemini

C. Llama

D. Claude

Q12 GDPRの違反時の最大制裁金はどれか。

A. 売上高の1%

B. 売上高の4%

C. 固定で100万ユーロ

D. 制裁金はない

Q13 マルチモーダルAIの説明として正しいものはどれか。

A. 複数の言語に対応したAI

B. テキスト以外も処理できるAI

C. 複数のGPUを使うAI

D. 複数の企業が共同開発したAI

Q14 RAGとファインチューニングの違いとして正しいものはどれか。

A. 違いはない

B. RAGはモデルを再学習する、ファインチューニングはしない

C. RAGはモデルを再学習しない、ファインチューニングはする

D. どちらもモデルを再学習する

Q15 プライベートLLM導入の損益分岐点として一般的に言われる期間はどれか。

A. 1週間

B. 1ヶ月

C. 1〜2年

D. 10年以上

Q16 LibreChatの説明として正しいものはどれか。

A. LLMのモデルファイル

B. ChatGPT風の社内チャットインターフェース

C. GPU管理ツール

D. 翻訳専用AI

Q17 7BモデルをOllamaで動かす際に必要なVRAMの目安はどれか。

A. 約1GB

B. 約8GB

C. 約100GB

D. VRAMは不要

Q18 社内AIガバナンスに含まれるものとして最も適切なものはどれか。

A. AIの価格設定

B. AIの利用ルールと管理体制

C. AIのデザイン規定

D. AIの広告規制

Q19 プライベートLLM導入のステップとして正しい順序はどれか。

A. 本番導入→PoC→目的定義→検証

B. 目的定義→PoC→検証→本番導入

C. PoC→目的定義→本番導入→検証

D. 検証→PoC→目的定義→本番導入

Q20 プライベートLLMのセキュリティ設定で最も重要なものはどれか。

A. モデルの見た目を整える

B. アクセス制御とファイアウォールの設定

C. キーボードの設定

D. 画面の明るさ調整

練習問題 解答

問題	正答	問題	正答	問題	正答	問題	正答
Q1	B	Q2	B	Q3	B	Q4	B
Q5	C	Q6	B	Q7	B	Q8	B
Q9	B	Q10	B	Q11	C	Q12	B
Q13	B	Q14	C	Q15	C	Q16	B
Q17	B	Q18	B	Q19	B	Q20	B

重要用語集

LLM

Large Language Model。大規模言語モデル。膨大なテキストデータで学習したAIモデル。

プライベートLLM

社内・オンプレミス環境で動かすLLM。データが外部に出ない。

オープンソースモデル

無料で利用・改変できるAIモデル。Llama・Mistral・Gemmaなど。

Ollama

ローカル環境でLLMを簡単に動かすオープンソースツール。

vLLM

高速・大規模並列処理に対応したLLM推論フレームワーク。本番環境向け。

RAG

Retrieval-Augmented Generation。社内文書をLLMに参照させる技術。再学習不要。

量子化

モデルの数値精度を落としてファイルサイズ・VRAM使用量を削減する技術。

ファインチューニング

既存モデルを特定のデータで追加学習させること。

ハルシネーション

AIが事実でない内容を自信満々に生成する現象。幻覚とも呼ばれる。

VRAM

GPUに搭載されたメモリ。LLMの動作に必要。

Dify

ノーコード・ローコードでAIアプリを構築できるオープンソースプラットフォーム。

LibreChat

ChatGPT風UIをプライベートLLMで実現するオープンソースチャットツール。

EU AI Act

EUが2024年に施行した世界初のAI包括規制法。AIをリスクで分類。

GDPR

EU一般データ保護規則。個人データの保護・管理を規定。違反時は高額制裁金。

オンプレミス

サーバーを自社施設に設置して運用する形態。クラウドの対義語。

AIガバナンス

AI利用のルール・管理体制・責任体制のこと。

PoC

Proof of Concept。概念実証。本番導入前の小規模テスト。

マルチモーダル

テキスト・画像・音声など複数の形式を処理できるAI。

トランスフォーマー

LLMの基盤となるニューラルネットワークの構造。

API

Application Programming Interface。システム間の接続インターフェース。