ChatGPT

みずほFGの自社LLMが「GPT-5.2と同精度」でオンプレ運用可能とする報道の検証

エグゼクティブサマリ

- みずほFGの一次情報（プレスリリース相当）では、銀行の**多肢選択式**の実務テストにおいて、金融特化LLM（Qwen3-32Bベース）が「推論なし＋RAG等のコンテキスト付与あり」の条件で**正答率89.0%**、業務実装を想定した評価で**平均回答時間1秒未満**と記載されている。 [1]
- 同じ一次情報の表では、汎用LLM（GPT-5.2）も「推論なし＋コンテキスト付与あり」で**正答率89.0%／1秒未満**とされており、見出しの「同精度」は**“特定の評価セット・特定条件”で成立**する主張である（一般能力の同等性を意味しない）。 [1]
- 一方で、GPT-5.2の「推論あり（reasoning_effort=xhigh）＋コンテキストなし」の比較（正答率89.7%だが平均67.4秒）を併記しており、**条件が揃っていない比較**（RAG有無・推論設定・ネットワーク遅延の混在）として解釈上の注意が要る。 [2]
- Qwen3-32Bは公開情報として、**32.8Bパラメータのdense Transformer**（64層、GQA等）で、学習データは**36兆トークン／119言語**規模とされ、公開ベンチマーク結果も提示されている。 [3]
- 金融機関としての運用は、性能だけでなく、**モデル・リスク管理（ガバナンス、インベントリ、承認、継続モニタリング、独立検証）**や、生成AI特有の**データ保護・サイバーセキュリティ・第三者依存**への統制が要点になる。 [4]
- 推奨としては、(1)「同精度」主張を支える評価の再設計（公開ベンチマーク併用・RAG条件統一・p95遅延公開等）、(2) オンプレ推論基盤の冗長化と監視、(3) プロンプトインジェクション／モデル盗用対策を含むセキュリティ設計、(4) 法務・監査を巻き込んだ運用規程整備を、短期〜中期ロードマップで進めるべきである。 [5]

事実確認

出典URL一覧

以下は、調査で参照した主要URL（一次情報優先）である。なお、Yahoo!ニュース本文ページは技術的制約（取得制限）により本文確認を完遂できていないものがあるため、「未確認」を明記する。

```
【一次情報（みずほFG側）】
- PR TIMES（みずほFG発表）: https://prtimes.jp/main/html/rd/p/000000002.000177612.html
- みずほFG（MIZUHO DX記事・2025/12/23）: https://www.mizuho-fg.co.jp/dx/articles/ai-finance-specialized-llm/index.html
- みずほFG（PDF・2025/12/11）: https://www.mizuho-fg.co.jp/dx/assets/pdf/ai-finance-specialized-llm-2512.pdf

【報道・二次情報（Yahoo!含む）】
- ビジネス+IT / FinTech Journal（報道）: https://www.sbbit.jp/article/fj/182069
- Yahoo!ニュース（配信元: ITmedia NEWS）: https://news.yahoo.co.jp/articles/cc08e1bcefe02033a7469dc7680d021900c3590a　※本文未確認
- Yahoo!フォロー（関連情報一覧ページ）: https://follow.yahoo.co.jp/themes/0e69ae97bed17f501661/
```

```
【比較対象・周辺一次情報（モデル/規制/運用）】
- OpenAI（GPT-5.2モデル説明）: https://developers.openai.com/api/docs/models/gpt-5.2
- OpenAI（GPT-5.2紹介・ベンチ詳細）: https://openai.com/index/introducing-gpt-5-2/
- OpenAI（reasoning.effort / reasoning_effort）: https://developers.openai.com/api/reference/resources/responses/methods/create/
- Qwen3 Technical Report（arXiv）: https://arxiv.org/html/2505.09388v1
- Hugging Face（Qwen3-32B-FP8 model card）: https://huggingface.co/Qwen/Qwen3-32B-FP8
- Qwen公式GitHub（ライセンス等）: https://github.com/QwenLM/Qwen3
- vLLM（OpenAI互換サーバ）: https://docs.vllm.ai/en/latest/serving/openai_compatible_server/
- vLLM（Kubernetesデプロイ）: https://docs.vllm.ai/en/stable/deployment/k8s/
- AWS（p5.4xlarge 料金の参照例：Capacity Blocks）: https://aws.amazon.com/ec2/capacityblocks/pricing/
- AWS（p5.4xlarge スペック）: https://docs.aws.amazon.com/ec2/latest/instancetypes/ac.html
- NVIDIA（H100仕様）: https://www.nvidia.com/en-us/data-center/h100/
- 金融庁（AIディスカッションペーパー第1.1版・2026/3）: https://www.fsa.go.jp/news/r7/sonota/20260303/aidp_version1.1.pdf
- 金融庁（モデル・リスク管理に関する原則・2021/11/12）: https://www.fsa.go.jp/common/law/ginkou/pdf_02.pdf
- 個人情報保護委員会（生成AI利用に関する注意喚起）: https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/
```

公式発表の要点

みずほFGの2026年3月5日付の発表（PR TIMES掲載）で、事実として確認できる記載は次の通りである。 [1]

- 「銀行の実務テスト（預金・融資・外国為替等、多肢選択式）」で、金融特化LLMは**推論なし条件**で**正答率89.0%**。 [1]
- 「業務実装を想定した評価」で**平均回答時間1秒未満**。 [1]
- 比較表では、GPT-5.2も「推論なし＋コンテキスト付与あり」で**正答率89.0%／1秒未満**、一方で「推論あり＋コンテキストなし」だと**正答率89.7%／平均67.4秒**とされる。 [1]
- 金融特化LLMは「Qwen3-32Bベース」。また応答時間の注記として、「単一H100 GPUインスタンス（p5.4xlarge）で計測」「GPT-5.2側はOpenAI APIで計測し、推論ありはreasoning_effort=xhigh、推論なしはnone」といった条件が明示されている。 [2]
- 「オンプレミス（銀行内のセキュアな環境）で運用できる点」が最大の特長として述べられ、外部へのデータ送信なしで処理を完結できる旨が強調される。 [1]
- ただし同リリース自身が「本結果は特定の評価セットや条件に基づく」「評価セット拡充や継続的検証」「応答時間も中央値や分布を確認し枠組み整備」といった**限定・継続検証**も明記している。 [1]

報道の主張と一次情報との差分

報道（例：FinTech Journal）は、上記一次情報のポイント（正答率89.0%、1秒未満、GPT-5.2比較、オンプレ）を概ね踏襲している。 [6]
一方、見出しや短い要約では「GPT-5.2と同精度」「オンプレ運用可能」が前面に出やすく、一次情報が付した**条件（推論なし／コンテキスト付与／多肢選択式／評価セット限定／APIのネットワーク遅延混在）**が省略されると、受け手が「総合的な知能や広範なタスクでGPT-5.2と同等」と誤解し得る。 [7]

下表は、一次情報と報道（確認可能範囲）の差分を、論点ごとに整理したものである。

論点	一次情報（みずほFG発表）	報道（確認可能範囲）	差分の意味合い
「同精度」の定義	**銀行実務テスト（多肢選択式）**で、金融特化LLM（推論なし＋コンテキスト付与あり）とGPT-5.2（推論なし＋コンテキスト付与あり）が**同じ89.0%**。 [1]	見出し等で「GPT-5.2と同精度」を強調しやすい。 [6]	“条件付き同等”を“包括的同等”と誤解させるリスク。
レイテンシ比較	GPT-5.2の「推論あり＋コンテキストなし」は**67.4秒**、一方で「推論なし＋コンテキストあり」は1秒未満と注記。 [2]	「推論あり設定は遅い」点が強調される傾向。 [6]	**推論設定とRAG条件がばらつく**ため、純粋なモデル性能差ではなく“設定・構成差”を含む。
オンプレ運用可能性	オープンウェイトモデル（Qwen3-32B）ベースで、銀行内閉域で完結可能と主張。 [1]	同趣旨を踏襲。 [6]	実際のオンプレ要件（冗長化、監査、運用体制）は別途設計が必要（未開示）。
限界の明示	評価セット限定、今後の指標提示方法整理、分布評価などの記載あり。 [1]	省略されやすい（記事短縮の宿命）。	妥当性評価では一次情報の限定記載を重視すべき。

モデル情報

Qwen3-32Bの公開仕様

Qwen3は、オープンウェイトのLLMシリーズとして、denseモデルとMoEモデル双方を含むファミリーとされる。 [8]
Qwen3のdenseアーキテクチャは、Qwen2.5と同系統で、GQA、SwiGLU、RoPE、RMSNorm（pre-norm）などを用い、学習安定化のためにQK-Normの導入等が記載されている。 [9]

また学習データについて、Qwen3は**119言語・方言**をカバーし、総トークン数は**36兆**規模とされる。 [8]

Qwen3-32B（FP8版のモデルカード）に基づく、「Qwen3-32B」の公表仕様（抜粋）は以下である。 [10]

項目	公開値（モデルカード）
パラメータ数	32.8B（非埋め込み31.2B） [10]
レイヤ数	64 [10]
Attention	GQA（Q=64 heads / KV=8 heads） [10]
コンテキスト長	32,768（ネイティブ）、YaRNで131,072検証と記載 [10]
推論サーバ候補	vLLM / SGLang等でOpenAI互換API提供が可能と記載 [10]
ライセンス	Apache 2.0（公式GitHubで全open-weightがApache 2.0と明記） [11]

加えてQwen3は「Thinking Mode / Non-thinking Mode」のハイブリッドを特徴として掲げ、タスクに応じて“考える量”を切り替える思想が説明されている。 [12]

GPT-5.2の公開仕様

GPT-5.2は、OpenAIの公開情報上、APIモデルとして `gpt-5.2` 等の名称で提供され、**context window 400,000**、**knowledge cutoff 2025-08-31**、推論負荷を調整する `reasoning.effort`（none/low/medium/high/xhigh）のサポートが明記されている。[13]

またOpenAIの「Introducing GPT-5.2」では、GPQA Diamond、SWE-Bench Pro、ARC-AGI等のベンチマークでのスコアが提示され、APIモデル名の対応（ChatGPT側の表記とAPI名）や、価格も記載されている。[14]

ベンチマーク比較の注意点と「比較可能指標」

みずほFGの主張「GPT-5.2と同精度」は、一次情報の表が示す通り、**自社評価セット（銀行実務テスト）**かつ**RAG等の“コンテキスト付与”と“推論設定”を含む条件**の中で成立している。[1]

一方、公開ベンチマークで比較する場合、少なくとも次の前提差があるため、スコアの単純比較は危険である。

- Qwen3側：技術報告の表は主に **Baseモデル（事前学習モデル）**のスコアが中心（用途によりInstruct/RL版の方が適切な場合がある）。[9]
- GPT-5.2側：OpenAI掲載のスコアは、`GPT-5.2 Thinking` 等の条件（ツール有無、推論設定）を伴うものが混在し得る。[15]

それでも「公開情報だけで比較可能な指標」を提示する、という要請に合わせ、**“同名ベンチマークでスコアが公開されているもの”**を中心に、代表例を整理する。

指標（公開ベンチ）	Qwen3-32B（技術報告の表）	GPT-5.2（OpenAI掲載）	注記
MMMLU	83.83 [9]	89.6（GPT-5.2 Thinking、Appendix）[14]	評価設定差（ショット数、ツール、推論設定、汚染耐性等）が不明確なため“参考比較”。
GPQA	49.49 [9]	GPQA Diamond 92.4（no tools、推論最大等の注記あり）[14]	**GPQA と GPQA Diamond は同一ではない**可能性が高く、直接比較は不可（名称は近いが粒度が違う）。
SWE系	（Qwen3-32Bの該当値は、当該表には明示なし）[9]	SWE-Bench Pro (public) 55.6 [14]	Qwen側の同一指標が同ページに無いため未比較。

結論として、みずほFGの「同精度」主張は**公開ベンチマークで裏付けられた“汎用知能の同等性”を意味しない**。一次情報の範囲では、あくまで「銀行実務テスト（多肢選択式）における、推論なし＋コンテキスト付与あり条件での精度同等」を指す。[1]

オンプレ運用要件

一次情報が示す「最低限の実証条件」

みずほFGの発表では、応答時間の計測条件として「単一H100 GPUインスタンス（p5.4xlarge）」が明示され、H100 1枚での推論が前提に置かれている。[16]

この `p5.4xlarge` は 1 × NVIDIA H100（80GiB）を搭載するインスタンスとしてAWS公式のインスタンスタイプ仕様に記載されている。 [17]

したがって、少なくとも「同等レベルのオンプレ推論」を狙うなら、**H100　80GB級（ないし同等のVRAM容量・帯域のGPU）を1枚以上**確保する設計が素直である（ただし、実運用は冗長化が必要）。 [16]

推論基盤のハードウェア要件

Qwen3-32Bのパラメータは32.8Bであるため、重みをFP16/BF16相当で保持するだけでも、概算で「パラメータ数×2バイト」程度のVRAMが必要になる（32.8Bなら約65.6GB相当。ここにKVキャッシュ等の運用メモリが上乗せされる）。パラメータ数の根拠はモデルカード通りである。 [10]

オンプレ運用要件は、要件定義（同時接続数、最大トークン長、RAG有無、思考ON/OFF）で大きく変わるため、ここでは「金融機関の社内業務用途（高機密・低遅延・監査対応）」を想定した推奨レンジを示す。

レイヤ	推奨（目安）	根拠・注記
GPU	最小：80GB VRAM級×1（検証用）／本番：×2以上（冗長化）	p5.4xlargeがH100 80GiBであること、一次情報が単一H100で測定と明記。 [16]
CPU/メモリ	CPUは推論前後処理・RAG・暗号処理を考慮し 16〜64 vCPU、RAM 128〜512GB	モデルサービングに加え、RAG（検索、埋め込み、前処理）と監査ログが走る前提。一次情報では“コンテキスト付与”を用いる。 [1]
ストレージ	NVMe SSD（少なくとも 1〜2TB）、モデル重みとベクタDB、ログの分離	vLLMのK8s例でもモデル重み保存用PVCを使う例が提示される。 [18]
ネットワーク	閉域網＋APIゲートウェイ、監査ログ転送経路	金融機関はアクセス制御と監査が必須（後述のモデル・リスク管理原則とも整合）。 [19]

GPUの電力・冷却については、H100の最大TDP（SXMで最大700W、PCIeで350-400W等）が公式仕様として提示されている。 [20]
データセンター側の間接電力（冷却等）はPUEに依存し、経済産業省資料ではデータセンター業の目標として「事業者平均PUEを1.4以下」といった言及がある。 [21]

推論レイテンシとスケーリング設計

一次情報が公表しているレイテンシは「平均回答時間1秒未満」であるが、これは「銀行実務テスト」「推論なし」「コンテキスト付与あり」等の条件が紐づく。さらにGPT-5.2側はAPI経由のためネットワーク遅延が含まれる旨も明記されている。 [1]

実運用では、平均値よりも **p95/p99遅延**、最大トークン長、同時リクエスト時の劣化（キューイング）が支配的になりやすい。一次情報も中央値・分布を将来整備する旨を述べているため、オンプレ化の意思決定では、少なくとも以下をKPIとして固定するのが実務的である。 [1]

- p50/p95/p99のレイテンシ（入力長別、出力長別）
- スループット（req/s、token/s）と落ち込み（ピーク時）
- RAG有無と“コンテキスト長”の感度
- 推論モード（思考ON/OFF、温度等）別の品質×コスト

運用ソフト構成の標準形

Qwen側のモデルカードは、デプロイ手段として vLLM や SGLang 等を挙げ、OpenAI互換API提供の例を示す。[10]
vLLMはOpenAI互換サーバを提供し、HTTP経由でOpenAIクライアントから呼び出せる説明がある。[22]
またKubernetes上でのデプロイ手順（PVC、Secret、Deployment/Service等）も公式ドキュメントとして提示されている。[18]

このため、金融機関のオンプレでは、一般に次のようなレイヤ分割が設計しやすい。

```
flowchart LR
  U[社内ユーザー/業務アプリ] --> G[API Gateway / WAF]
  G --> A[認証・認可 / RBAC]
  A --> P[ポリシー/プロンプトガード]
  P --> R[RAG: 検索 + コンテキスト生成]
  R --> L[LLM推論サーバ vLLM/SGLang]
  L --> O[後処理: 形式チェック/マスキング]
  O --> U

  subgraph Data[データ層]
    V[ベクタDB]
    D[参照文書ストア]
    M[モデルレジストリ/署名付き重み]
    LOG[監査ログ/メトリクス]
  end
  R --> V
  R --> D
  L --> LOG
  P --> LOG
  A --> LOG
  M --> L
```

コスト概算

前提

- クラウド単価は、AWSのCapacity Blocks価格ページに掲載の **p5.4xlarge = $3.933/時（H100×1）** を参照値として用いる（リージョン差はある）。[23]
- p5.4xlargeのGPUがH100 80GiBであることはAWS公式仕様表に基づく。[17]
- オンプレのハードウェア調達価格はベンダ見積依存で公表情報が乏しく、本報告書では**未確認**とする。その代替として、クラウド24/7実行コストから“費用感”を逆算する（あくまで参考）。[23]
- 電力コストは、H100のTDP（最大700W等）と、PUE目標1.4の資料を参照し、電力単価は契約依存のため例示（未確認）とする。[24]

試算例

シナリオ	想定	初期投資（CAPEX）	運用費（OPEX）概算	コメント
検証（最小）	GPU 80GB級×1、単一ノード	**未確認**（要見積）	クラウドなら $3.933/時 → 年$34.5k（24/7の場合）[23]	一次情報の計測条件（単一H100）に近い。[1]
本番（可用性重視）	GPU 80GB級×2（Active/Standby or Active/Active）	**未確認**（要見積）	クラウド換算で概ね2倍（単純化）[23]	金融用途では業務継続（BCP）と障害時切替が必須になりやすい。[25]
本番（スケール重視）	GPU 80GB級×4〜8、K8sで水平スケール	**未確認**（要見積）	リクエスト量に応じて増分	vLLMのK8s運用は公式ガイドがあり、複数レプリカ化が現実的。[26]

ポイントは、オンプレ運用では「GPU台数」だけでなく、**監査ログ保管、セキュリティ運用、MLOps（評価・再学習・承認）**まで含めた人的コストが支配的になり得ることである（後述のモデル・リスク管理原則が実務要件として効いてくる）。[27]

セキュリティ・コンプライアンス

金融庁の観点

金融庁[28]のAIディスカッションペーパー（第1.1版）は、生成AIを含むAI活用に関する課題を整理しており、図表では「データセキュリティ」「個人情報保護」「サイバーセキュリティ」「AIガバナンス」等が検討・導入・利用時の課題として挙がる構成になっている。[29]
また同資料は、AIシステムが金融安定に与え得る影響として、サードパーティ依存やサイバーセキュリティ、モデル・リスク管理、市場の群衆効果、フェイクニュース拡散等の論点に触れている。[30]

オンプレ運用（閉域で完結）自体は「外部送信を減らす」方向に寄与するが、**閉域＝安全**ではない。特に、RAGやマルチエージェント構成は攻撃面（データ供給面）が増えるため、設計段階から脅威分析を内蔵する必要がある。[31]

モデル・リスク管理

金融庁[28]の「モデル・リスク管理に関する原則」では、取締役会等・上級管理職による態勢整備、方針・規程と文書化、役割責任の明確化、モデルの特定とインベントリ管理、リスク格付、モデル開発プロセス（モデル記述書、モデルテスト）、承認、継続モニタリング、独立検証といった枠組みが示されている。[32]

「社内LLM」は、クレジット審査・市場分析・顧客対応などに入り込むほど、実質的に“意思決定補助モデル”となるため、従来の統計モデル同様に、モデル台帳（インベントリ）・承認・監査証跡・ドリフト監視が必須と解釈するのが安全側である。[33]

個人情報保護

個人情報保護委員会[34]は、生成AIサービス利用時に、入力情報が提供者側で学習データに利用される場合、利用者が個人データを入力すると「第三者提供」に該当し得るなど、個人情報保護法上の注意点を注意喚起している。[35]

オンプレ運用は「外部提供者が学習に使う」局面を減らせるが、（1）社内の目的外利用、（2）ログへの個人情報混入、（3）権限の弱い利用者への過剰開示、（4）RAG文書のアクセス制御漏れ、といった内部統制リスクが残るため、DLPや権限管理が要件になる。[36]

モデル盗用・逆コンパイル・プロンプト攻撃

オンプレであっても、APIを社内公開する以上、**モデル盗用（model extraction）**や**学習データ抽出**、**プロンプトインジェクション／エージェント乗っ取り**のリスクは残る。

- 予測APIへのクエリだけでモデルの機能を高精度に複製できる可能性（モデル抽出攻撃）が古典的に示されている。[37]
- 大規模言語モデルに対して、クエリを通じて学習データの一部が抽出され得ることも報告されている。[38]
- OWASP[39] は LLMアプリの主要リスクとして Prompt Injection 等をTop 10形式で整理している。[40]
- NIST[41] も、間接的プロンプトインジェクション（agent hijacking）の評価・強化の必要性を指摘している。[42]

したがって、オンプレ運用のセキュリティ要件としては、少なくとも以下が必要になる（未実装なら“未確認”として監査指摘対象になり得る）。

- 認証・認可（RBAC/ABAC）、レート制限、監査ログ（誰が何を聞いたか、どの文書を参照したか）[19]
- RAGの文書アクセス制御（検索結果に“権限外”が混ざらない保証）[29]
- プロンプト／出力のセキュリティレビュー（Prompt Injection対策、出力の危険操作禁止）[43]
- モデルと重みのサプライチェーン対策（署名、ハッシュ固定、SBOM、脆弱性管理）[44]

実務的影響

みずほFG内で想定される利用ケース

みずほFG側が示すロードマップ型の利用イメージは、2025年末時点のDX記事でも「金融特化LLM → 特定業務LLM → 協調型エキスパートLLM」という三層構造として説明されている。[45]
2026年3月の発表でも同趣旨の段階的整備（照会応答・資料作成 → 部署特化 → 複数モデル連携）が記載されている。[1]

これを実務に落とすと、代表的には以下の領域が対象になりやすい（一次情報の記載に沿う）。

- 社内照会・ヘルプデスク（規程・手続の検索と回答、資料作成支援）[46]
- 融資・与信判断支援、稟議書の素案作成（第2段階の例示）[46]
- 市場分析・トレーディング支援（応答遅延が機会損失になり得るという文脈で言及）[1]
- 複数領域（法務×市場×与信等）の“協調”による部門横断判断支援（第3段階の構想）[46]

他行・海外事例との比較

国内外で見ると、「オンプレ／自社運用」一択ではなく、「統制された社内環境で商用LLMを使う」アプローチも広がっている。

- 三菱UFJ銀行[47] は ChatGPT Enterprise[48] を全行員規模で段階展開する旨のプレスリリース（PDF）を公表している。[49]

- SMBCグループ [50] は、生成AI活用の推進体制（LLM-CoE等）を紹介するDX記事を公開している。 [51]
- JPMorgan Chase [52] は社内向け生成AIプラットフォーム「LLM Suite」を自社のテクノロジー発信として紹介している。 [53]
- Morgan Stanley [54] は OpenAI [55] と連携した業務ツール（ミーティング要約等）を公式発表している。 [56]
- Goldman Sachs [57] の全社向けAIアシスタント展開も報道されている。 [58]

これらを「データ統制」「社内展開速度」「運用負荷」で見ると、みずほFGの“オープンウェイトを社内最適化してオンプレ運用”は、機密データ処理や低遅延を重視する文脈に合う一方、MLOpsとガバナンス負担は重くなりやすい（モデル台帳、承認、監視、セキュリティ運用が常時必要）。 [59]

技術的妥当性評価と推奨事項

「GPT-5.2と同精度」主張の妥当性

一次情報に照らす限り、「同精度」という表現は、**限定条件下では形式的に成立**する。具体的には「銀行実務テスト（多肢選択式）」「推論なし」「コンテキスト付与あり」で、金融特化LLMとGPT-5.2がともに正答率89.0%と記載されているためである。 [1]

ただし、技術的に“強い主張”として受け取るには根拠が弱い点が残る。

- **評価セットの実体が非公開**：問題数、難易度分布、更新頻度、汚染（学習データへの混入）対策が不明。一次情報も「評価セット拡充・継続検証」を示唆しており、現時点で一般化はできない。 [1]
- **比較条件が揃っていない**：表の3行目（GPT-5.2 推論あり）はコンテキストなしで、他行はコンテキストあり。RAGの有無は性能と遅延の両方に影響するため、厳密な同列比較ではない。 [1]
- **推論設定の解釈**：OpenAI側の `reasoning.effort` は品質と速度・コストをトレードオフする設計であるため、xhighで遅くなるのは仕様に整合するが、比較としては「同じRAG条件」で両設定を提示しないと読み手は誤解し得る。 [60]

より頑健な主張にするには、例えば日本語金融ドメインに特化した公開ベンチマーク（例：Ebisu、EDINET-Bench等）を併用し、RAG条件と推論条件を統一した上で、再現可能な評価プロトコルを提示することが望ましい。 [61]

導入・運用チェックリスト

金融機関の実装では、性能の前に「統制・監査」がボトルネックになりやすい。 [27]
以下は、監査対応まで含めたチェック項目の最小セットである。

領域	チェック項目（抜粋）	根拠
技術	RAGの権限分離（検索結果の“越権混入”ゼロ）、p95遅延、障害時フェイルオーバ	生成AI活用の課題としてデータセキュリティ等が挙がる。 [29]
セキュリティ	Prompt Injection対策、機微情報マスキング、レート制限・監査ログ	OWASP Top 10がPrompt Injection等を主要リスクに整理。 [40]
モデル保護	社内APIでもモデル抽出・データ抽出に備えた制御（出力制限、監視）	モデル抽出攻撃・学習データ抽出攻撃が知られる。 [62]
組織	モデルオーナー、独立した統制（第2線）、モデル台帳、リスク格付	金融庁のモデル・リスク管理原則。 [63]

領域	チェック項目（抜粋）	根拠
法務・個情	個人情報の入力・ログ混入ルール、目的外利用防止、同意取得要否整理	個人情報保護委員会の注意喚起。[64]
監査	モデル記述書、テスト結果、変更管理、継続モニタリングの証跡	原則で文書化・テスト・継続モニタリングが示される。[65]

短期・中期の実行プラン案

```
gantt
  title オンプレLLM導入タイムライン案
  dateFormat  YYYY-MM-DD
  axisFormat  %m/%d

  section 企画・統制
  ユースケース定義/リスク格付      :a1, 2026-03-10, 21d
  モデル台帳/承認フロー整備         :a2, after a1, 21d

  section 技術PoC
  RAG権限分離設計・検証             :b1, 2026-04-01, 30d
  速度/精度評価（条件統一・p95）     :b2, after b1, 30d

  section セキュリティ
  Prompt Injection/データ漏えい対策  :c1, 2026-05-15, 45d
  監査ログ/監視/運用手順書          :c2, after c1, 30d

  section 本番展開
  段階導入（社内照会→特定業務）      :d1, 2026-07-01, 120d
  定期再評価・独立検証（四半期）      :d2, 2026-07-01, 180d
```

短期（〜3か月）で最も重要なのは、一次情報にある「条件付きの性能」を、**運用条件（RAG、推論設定、入力長、同時接続）に落としたSLO**として固定し、モデル・リスク管理原則に沿って「承認・変更・監視」を制度化することである。[66]
中期（3〜12か月）では、部署特化LLMやマルチエージェント化を進めるほど、間接プロンプトインジェクション等の攻撃面が広がるため、NISTやOWASPが挙げるクラスの脅威を前提に、レッドチーミングと継続監視を“運用の一部”として組み込む必要がある。[67]

---

[1] [2] [7] [16] [28] [31] [34] [47] [48] [50] [59] [66] https://prtimes.jp/main/html/rd/p/000000002.000177612.html
https://prtimes.jp/main/html/rd/p/000000002.000177612.html

[3] [10] https://huggingface.co/Qwen/Qwen3-32B-FP8
https://huggingface.co/Qwen/Qwen3-32B-FP8

[4] [19] [25] [27] [32] [33] [41] [55] [63] [65] https://www.fsa.go.jp/common/law/ginkou/pdf_02.pdf
https://www.fsa.go.jp/common/law/ginkou/pdf_02.pdf

[5] [40] [43] [44] https://owasp.org/www-project-top-10-for-large-language-model-applications/
https://owasp.org/www-project-top-10-for-large-language-model-applications/

6 54 https://www.sbbit.jp/article/fj/182069
https://www.sbbit.jp/article/fj/182069

8 9 52 https://arxiv.org/html/2505.09388v1
https://arxiv.org/html/2505.09388v1

11 https://github.com/QwenLM/Qwen3
https://github.com/QwenLM/Qwen3

12 https://qwenlm.github.io/blog/qwen3/
https://qwenlm.github.io/blog/qwen3/

13 39 60 https://developers.openai.com/api/docs/models/gpt-5.2
https://developers.openai.com/api/docs/models/gpt-5.2

14 15 https://openai.com/index/introducing-gpt-5-2/
https://openai.com/index/introducing-gpt-5-2/

17 https://docs.aws.amazon.com/ec2/latest/instancetypes/ac.html
https://docs.aws.amazon.com/ec2/latest/instancetypes/ac.html

18 26 https://docs.vllm.ai/en/stable/deployment/k8s/
https://docs.vllm.ai/en/stable/deployment/k8s/

20 24 https://www.nvidia.com/en-us/data-center/h100/
https://www.nvidia.com/en-us/data-center/h100/

21 https://www.meti.go.jp/shingikai/enecho/shoene_shinene/sho_energy/pdf/047_00_04.pdf
https://www.meti.go.jp/shingikai/enecho/shoene_shinene/sho_energy/pdf/047_00_04.pdf

22 https://docs.vllm.ai/en/latest/serving/openai_compatible_server/
https://docs.vllm.ai/en/latest/serving/openai_compatible_server/

23 https://aws.amazon.com/ec2/capacityblocks/pricing/
https://aws.amazon.com/ec2/capacityblocks/pricing/

29 30 https://www.fsa.go.jp/news/r7/sonota/20260303/aidp_version1.1.pdf
https://www.fsa.go.jp/news/r7/sonota/20260303/aidp_version1.1.pdf

35 https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert
https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert

36 64 https://www.ppc.go.jp/files/pdf/generativeAI_notice_leaflet2023.pdf
https://www.ppc.go.jp/files/pdf/generativeAI_notice_leaflet2023.pdf

37 62 https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf
https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf

38 https://www.usenix.org/system/files/sec21-carlini-extracting.pdf
https://www.usenix.org/system/files/sec21-carlini-extracting.pdf

42 67 https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations
https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

45 https://www.mizuho-fg.co.jp/dx/articles/ai-finance-specialized-llm/index.html
https://www.mizuho-fg.co.jp/dx/articles/ai-finance-specialized-llm/index.html

46 https://www.mizuho-fg.co.jp/dx/assets/pdf/ai-finance-specialized-llm-2512.pdf
https://www.mizuho-fg.co.jp/dx/assets/pdf/ai-finance-specialized-llm-2512.pdf

[49] [57] https://www.mufg.jp/dam/pressrelease/2025/pdf/news-20251112-001_ja.pdf
https://www.mufg.jp/dam/pressrelease/2025/pdf/news-20251112-001_ja.pdf

[51] https://www.smfg.co.jp/dx_link/article/0219.html
https://www.smfg.co.jp/dx_link/article/0219.html

[53] https://www.jpmorganchase.com/about/technology/news/llmsuite-ab-award
https://www.jpmorganchase.com/about/technology/news/llmsuite-ab-award

[56] https://www.morganstanley.com/press-releases/ai-at-morgan-stanley-debrief-launch
https://www.morganstanley.com/press-releases/ai-at-morgan-stanley-debrief-launch

[58] https://www.reuters.com/business/goldman-sachs-launches-ai-assistant-firmwide-memo-shows-2025-06-23/
https://www.reuters.com/business/goldman-sachs-launches-ai-assistant-firmwide-memo-shows-2025-06-23/

[61] https://arxiv.org/abs/2602.01479
https://arxiv.org/abs/2602.01479