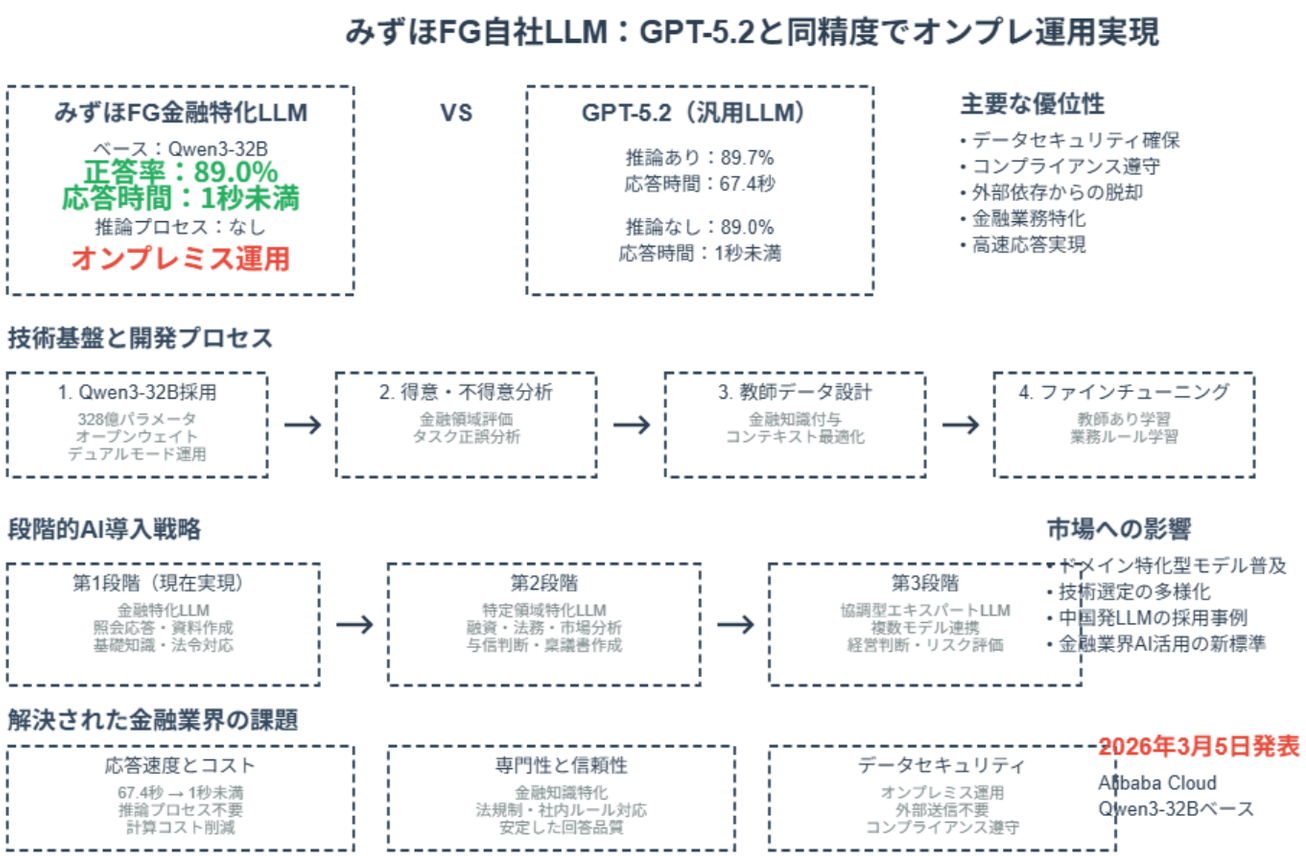
# みずほ FG の自社 LLM、「GPT-5.2 と同精度」でオンプレ運用可能

**Felo AI**




## 概要

みずほフィナンシャルグループ（みずほ FG）は 2026 年 3 月 5 日、独自に開発した金融業務特化型の大規模言語モデル（LLM）が、銀行の実務テストにおいて主要な汎用 LLM である「GPT-5.2」と同等の精度を達成したと発表した [4] [5] [11]。この LLM は、Alibaba Cloud が開発したオープンウェイトモデル「Qwen3-32B」を基盤としており、推論プロセス

なしで 89.0%の正答率を記録し、平均応答時間を 1 秒未満に抑えることに成功した [4] [11] [25]。
この成果の最大の戦略的意義は、高性能な AI を銀行内のセキュアなオンプレミス環境で運用できる点にある [4] [5] [13]。これにより、機密性の高い顧客情報や内部データを外部に送信することなく、高度な AI 処理を安全に適用することが可能となる [11] [12]。応答速度の遅延やコンプライアンス遵守といった金融業界特有の課題を解決し、データセキュリティと業務効率化を両立させる先進的な取り組みとして、今後の金融機関における AI 活用の新たな標準を示すものと評価される。

# 詳細報告

## 開発背景：金融業界における生成 AI の課題

近年、金融業界では生成 AI の活用が急速に進んでいるが、その導入には特有の課題が存在する [11] [25]。

- **応答速度とコスト**: 複雑な金融商品や社内規定に関する照会において、汎用 LLM は詳細な推論プロセスを要するため、回答生成に時間がかかり、計算コストも増大する傾向があった [11] [25]。特に、市場機会が秒単位で変動するトレーディングや法人向け金融業務では、応答の遅延が直接的な機会損失につながるリスクがあった [11] [25]。
- **専門性と信頼性**: 金融業務では、法規制、社内ルール、リスク許容度といった厳格な基準に基づいた回答が求められる [11] [25]。汎用モデルでは、これらの専門的な前提条件の解釈が不安定であったり、必要な観点を十分に反映できず、期待される品質の回答を安定して得ることが難しい場合があった [11] [25]。
- **データセキュリティとコンプライアンス**: 金融機関が取り扱うデータは機密性が極めて高く、外部の API を利用するクラウドベースの LLM では、情報漏洩のリスクやデータガバナンスの観点から導入の障壁となっていた [13] [25]。

これらの課題を解決するため、みずほ FG は金融知識と内部ルールに特化し、かつオンプレミスで安全に運用可能な LLM の開発に着手した [11]。

## 技術的基盤：Alibaba Cloud「Qwen3-32B」の採用

みずほ FG の金融特化 LLM は、Alibaba Cloud が開発したオープンウェイトモデル「Qwen3-32B」をベースに構築されている [4] [5]。Qwen3-32B は、Qwen シリーズの最新世代モデルであり、その高性能と柔軟性が採用の決め手となった [2] [9]。

**Qwen3-32B の主な特徴**:

- **モデル規模**: 328 億パラメータを持つ高密度（Dense）モデルであり、複雑なタスクに対応可能な能力を持つ [9] [19]。
- **デュアルモード運用**: 最大の特徴は、複雑な論理的推論を行う「思考モード」と、効率的な対話を行う「非思考モード」をシームレスに切り替えられる点にある [2] [9] [23]。これにより、タスクに応じて最適なパフォーマンスと効率性を両立できる。
- **高度な推論能力**: 数学、コード生成、常識的推論などのベンチマークで、旧世代のモデルを大幅に上回る性能を実証している [2] [9] [31]。
- **多言語対応**: 100 以上の言語と方言をサポートしており、グローバルな金融業務への応用ポテンシャルも秘めている [2] [9]。

オープンウェイトモデルである Qwen3-32B を活用することで、みずほ FG は最先端の基盤モデルの能力を享受しつつ、自社の要件に合わせて自由にカスタマイズする道を選んだ [4] [11]。

## みずほ FG による独自カスタマイズと性能評価

みずほ FG は、Qwen3-32B を単に導入するのではなく、金融業務に最適化するための独自のファインチューニングを施した [4] [5]。

**カスタマイズプロセス**:

1. **得意・不得意領域の特定**: まず、ベースモデルが生成した回答やタスクの正誤を分析し、金融領域におけるモデルの能力を詳細に評価した [4] [5]。
2. **高品質な教師データの設計**: 不得意とされた領域に対し、正答を導き出すために必要な金融基礎知識、業務手続き、社内ルール、コンプライアンス上の注意点などをコンテキストとして教師データに付与した [4] [5]。
3. **教師ありファインチューニング**: 回答とその根拠となるコンテキストの対応関係をモデルが深く学習するようデータ設計を最適化し、教師ありファインチューニング（Supervised Fine-Tuning）を実施した [4] [5]。

このプロセスにより、モデルは汎用的な知識だけでなく、みずほ FG 固有の業務知識とルールを深く理解することが可能となった。

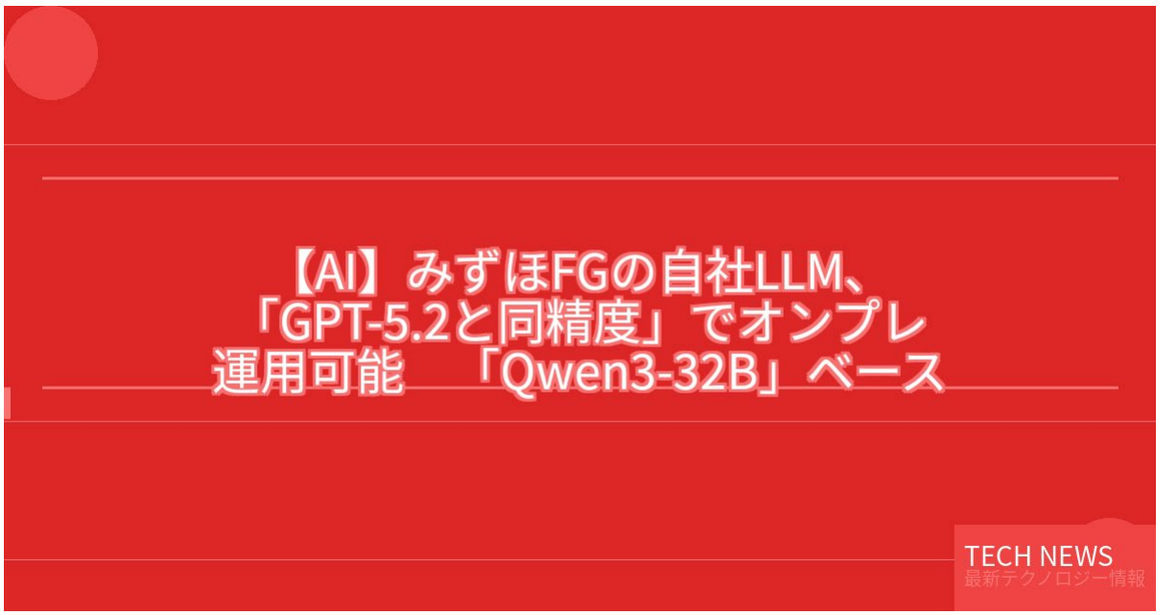
**性能ベンチマーク**: 銀行の実務を想定したテストでは、汎用 LLM である GPT-5.2 と比較して、その有効性が実証された [25] 。

| LLM の種類 | 推論 | コンテキスト付与 | 正答率 | 平均回答時間 |
|---|---|---|---|---|
| **金融特化 LLM (Qwen3-32B ベース)** | **なし** | **あり** | **89.0%** | **1 秒未満** |
| 汎用 LLM (GPT-5.2) | なし | あり | 89.0% | 1 秒未満 |
| 汎用 LLM (GPT-5.2) | あり | なし | 89.7% | 67.4 秒 |

この結果は、みずほ FG の金融特化 LLM が、時間のかかる推論プロセスを経ずに、GPT-5.2 が推論を行った場合とほぼ同等の精度（正答率差はわずか 0.7%）を達成しつつ、応答時間を 67 分の 1 以下に劇的に短縮したことを示している [4] [11] [16] 。

## 戦略的インプリケーション：オンプレミス運用の重要性

この取り組みの核心は、高性能 LLM をオンプレミス環境で運用可能にした点にある [4] [13] 。




- **データ主権の確保**: 顧客情報、取引データ、リスク分析、コンプライアンス関連情報といった機密性の高いデータを、外部のクラウドサービスに送信することなく、完全に管理された行内ネットワークで処理できる [11] [13]。これにより、データセキュリティとプライバシー保護を最高レベルで維持できる。
- **規制遵守とガバナンス**: 金融監督当局が求める厳格なデータ管理要件やトレーサビリティの要求に対応しやすくなる [13]。AI の意思決定プロセスに対する説明責任を果たし、リスク管理を徹底することが可能となる。
- **外部依存からの脱却**: 外部 API の利用に伴うコスト変動、サービス仕様の変更、あるいはサービス停止といったリスクから解放される [13]。自社でモデルを管理することで、安定的かつ持続可能な AI 活用基盤を構築できる。

この「性能」と「セキュリティ」の両立は、これまで大手金融機関が AI 導入を進める上での大きなジレンマであったが、みずほ FG のモデルはこの課題に対する一つの解を示したと言える [13]。

## みずほ FG の段階的 AI 導入戦略

今回の金融特化 LLM は、みずほ FG が描くより大きな AI 戦略の第一段階に位置づけられている [11] [25]。

- **第 1 段階：金融特化 LLM（今回実現）**：金融の基礎知識、法令、社内手続きなどを幅広く学習し、一般的な照会応答や資料作成を支援する [11] [25]。
- **第 2 段階：特定領域特化 LLM**: 融資、法務、市場分析など、各部署の専門領域に特化したデータを追加学習させ、手続き案内、与信判断支援、稟議書作成といったより高度な実務を支援するモデルを開発する [4] [11] [25]。
- **第 3 段階：協調型エキスパート LLM**: 複数の特定領域特化モデルを連携させ、部門を横断するような複雑な経営判断やリスク評価を支援する統合的な AI システムを構築する [4] [11] [25]。

このロードマップは、AI の適用範囲を単純作業の自動化から、専門的な意思決定支援へと段階的に拡大していく明確な

ビジョンを示している。

## 市場への影響と今後の展望

みずほ FG の発表は、日本の金融業界における AI 活用の潮流に大きな影響を与える可能性がある。

- **ドメイン特化型モデルの普及**: これまで主流であった汎用 LLM と RAG（Retrieval-Augmented Generation）を組み合わせる手法に加え、特定の業務領域に深く特化した「ドメイン特化型モデル」の内製化という選択肢が現実的であることを示した 15 。
- **技術選定の多様化**: 米国製 LLM が市場を席巻する中、中国発の高性能なオープンウェイトモデルを基盤に採用した点も注目される 10 。これにより、企業は自社のニーズや戦略に応じて、より多様な選択肢から最適な基盤モデルを検討するようになる可能性がある。
- **今後の展開**: みずほ FG は今後、よりパラメータサイズの大きいモデルでの学習を通じてさらなる精度向上を目指すとともに、融資、外国為替、法務といった専門分野への適用を拡大していく計画である 4 。継続的な評価と改善を通じて、AI を真の業務パートナーへと進化させていくことが期待される。


1. MIZUHO FINANCIAL GROUP （MFG） H2 2023 Earnings ...
2. Qwen/Qwen3－32B
3. Qwen3 32B － Intelligence, Performance & Price Analysis
4. みずほ FG の自社 LLM、「GPT－5.2 と同精度」でオンプレ運用可能 ...
5. みずほ FG の自社 LLM、「GPT－5.2 と同精度」でオンプレ運用可能 ...
6. みずほ FG の自社 LLM、「GPT－5.2 と同精度」でオンプレ運用可能 ...
7. Blog － Markets
8. In your experience and opinion, is Qwen3 32B better than ...
9. Qwen3－32B from Alibaba – Run On－Device with Mirai.
10. みずほが自社 LLM のベースモデルに中国 AI の Qwen3－32B を採用 ...
11. みずほ FG 開発「金融特化 LLM」、銀行実務テストで正答率 89
12. 【AI】みずほ FG の自社 LLM、「GPT－5.2 と同精度」でオンプレ運用 ...
13. Mizuho FG déploie un LLM interne au niveau de « GPT－5.2
14. Qwen3－Next－80B－A3B
15. “使えない AI”から脱却？ 2026 年注目の「ドメイン特化型モデル ...
16. GPT－5.2 と同精度」でオンプレ運用可能 「Qwen3－32B」ベース － MSN
17. posi__posi （@posi__posi8） / Posts / X
18. Qwen3 Usage Guide － vLLM Recipes

19. Qwen3－32B ｜ NVIDIA NGC
20. みずほ FG の自社 LLM、「GPT－5.2 と同精度」でオンプレ運用可能 ...
21. みずほフィナンシャルグループの関連情報 － フォロー － Yahoo！ JAPAN
22. 蒸留と追加学習するだけで自社 LLM になるのか… 今時どの企業でも ...
23. Qwen 3: Alibaba's Dual－Mode LLM That Changes How We ...
24. So ... a new qwen 3 32b dense models is even a bit better than 30b ...
25. みずほフィナンシャルグループが実現した金融特化 LLM の高精度と ...
26. 『みずほ FG の自社 LLM、「GPT－5.2 と同精度」でオンプレ運用可能 ...
27. My simple test: Qwen3－32b ＞ Qwen3－14B ≈ DS Qwen3－8 ...
28. Qwen3－32B: 32B Autoregressive Transformer － Emergent Mind
29. AI 治療チャットの意外な落とし穴：LLM が妄想協力者に変貌する ...
30. Qwen/Qwen3－32B－GGUF
31. Qwen3 32B is available on Lambda's Inference API
32. みずほ FG の自社 LLM、「GPT－5.2 と同精度」でオンプレ運用可能 ...
33. two models big difference in how it converses/answers. ie ...
34. The Qwen3 32B dense model just fails for me due to a template ...
35. にっくす （@_____nix_____） / Posts / X
36. Qwen 3 32b vs QwQ 32b : r/LocalLLaMA
37. デジタル庁と OpenAI が連携 職員用 AI プラットフォームに AI モデル ...