NTT版LLM tsuzumi

tsuzumi 2 進化のポイント

NTT株式会社
執行役員 研究企画部門長
木下 真吾

tsuzumi 2 進化のポイント

① 日本語性能のさらなる向上

② 特化型モデル開発効率の向上

③ 低コスト・高セキュアの維持、国産AI

© NTT, Inc. 2025

tsuzumi 2 進化のポイント

① 日本語性能のさらなる向上

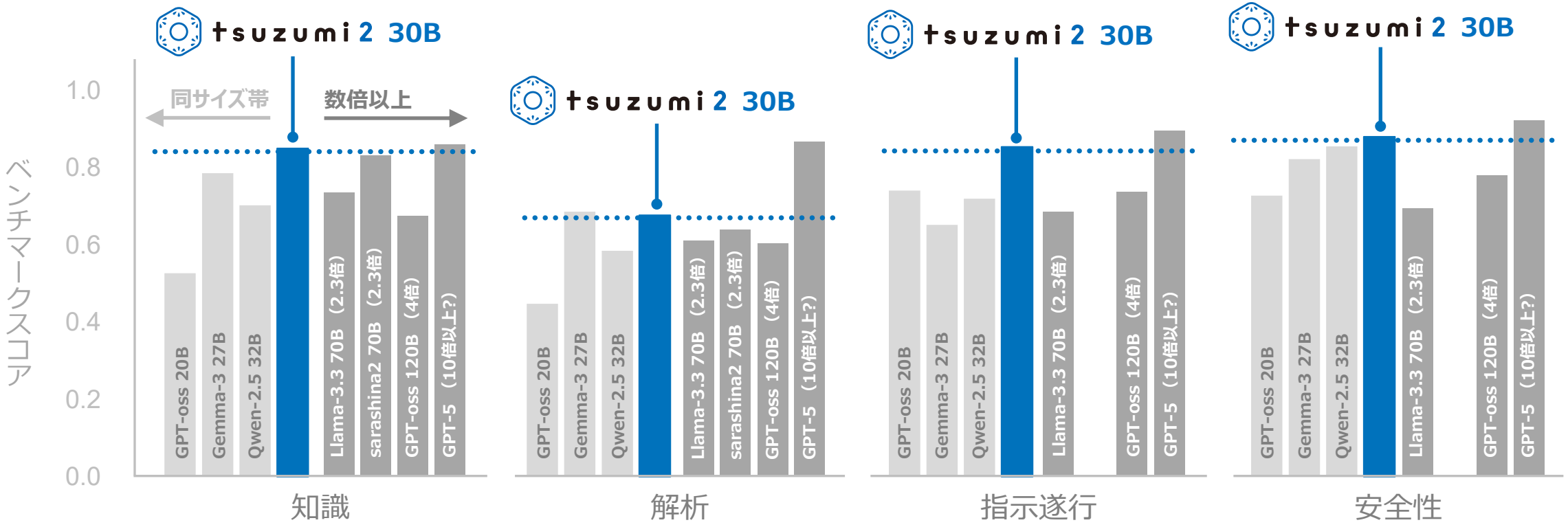
② 特化型モデル開発効率の向上

③ 低コスト・高セキュアの維持、国産AI

© NTT, Inc. 2025

特長①　日本語性能のさらなる向上

- 同サイズ帯のモデル（Gemma-3等）では世界トップクラスの日本語性能
- 数倍以上大きなモデル（GPT-oss 120B, GPT-5等）にも引けを取らない性能でコストパフォーマンスに優れる

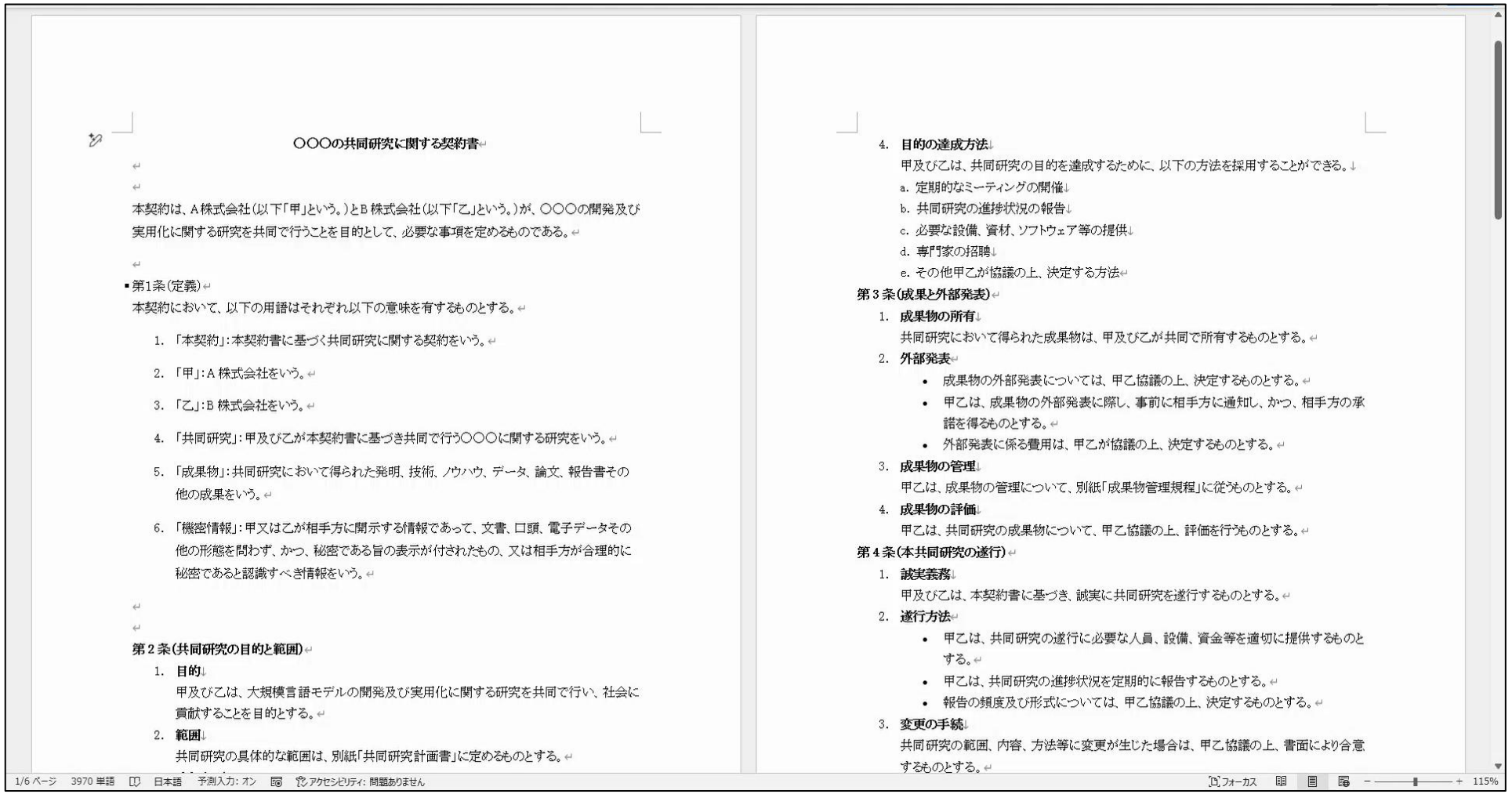


※評価条件
- 知識：llm-jp-eval（JAQKET (日本語QAタスク), NIILC（質問応答）, JCommonsenseQA（常識推論能力））, pfgen-bench（PFN日本知識ベンチ）の平均値
- 解析：llm-jp-eval（Wikipedia読み推定, Wikipedia係り受け解析）の平均値
- 指示遂行: M-IFEval_Ja
- 安全性: AnswerCarefully（x0.2）
- 知識と解析においては、事前学習モデルを使用。ただし、GPT-oss 20B/120B, GPT-5は事前学習モデルが公開されていないため事後学習モデル（Reasoningモデル）を使用
- 指示遂行と安全性において、全て事後学習モデルを使用。sarashina2 70Bは事後学習モデルが公開されてないため対象外

© NTT, Inc. 2025

特長①　日本語性能のさらなる向上

契約書案を社内チェックリストを使って確認。条項の不足や問題点を的確に指摘



○○○の共同研究に関する契約書

本契約は、A株式会社（以下「甲」という。）とB株式会社（以下「乙」という。）が、○○○の開発及び実用化に関する研究を共同で行うことを目的として、必要な事項を定めるものである。

第1条（定義）
本契約において、以下の用語はそれぞれ以下の意味を有するものとする。

1. 「本契約」：本契約書に基づく共同研究に関する契約をいう。
2. 「甲」：A株式会社をいう。
3. 「乙」：B株式会社をいう。
4. 「共同研究」：甲及び乙が本契約書に基づき共同で行う○○○に関する研究をいう。
5. 「成果物」：共同研究において得られた発明、技術、ノウハウ、データ、論文、報告書その他の成果をいう。
6. 「機密情報」：甲又は乙が相手方に開示する情報であって、文書、口頭、電子データその他の形態を問わず、かつ、秘密である旨の表示が付されたもの、又は相手方が合理的に秘密であると認識すべき情報をいう。

第2条（共同研究の目的と範囲）

1. **目的**
   甲及び乙は、大規模言語モデルの開発及び実用化に関する研究を共同で行い、社会に貢献することを目的とする。
2. **範囲**
   共同研究の具体的な範囲は、別紙「共同研究計画書」に定めるものとする。
4. **目的の達成方法**
   甲及び乙は、共同研究の目的を達成するために、以下の方法を採用することができる。
   a. 定期的なミーティングの開催
   b. 共同研究の進捗状況の報告
   c. 必要な設備、資材、ソフトウェア等の提供
   d. 専門家の招聘
   e. その他甲乙が協議の上、決定する方法

第3条（成果と外部発表）

1. **成果物の所有**
   共同研究において得られた成果物は、甲及び乙が共同で所有するものとする。
2. **外部発表**
   - 成果物の外部発表については、甲乙協議の上、決定するものとする。
   - 甲乙は、成果物の外部発表に際し、事前に相手方に通知し、かつ、相手方の承諾を得るものとする。
   - 外部発表に係る費用は、甲乙が協議の上、決定するものとする。
3. **成果物の管理**
   甲乙は、成果物の管理について、別紙「成果物管理規程」に従うものとする。
4. **成果物の評価**
   甲乙は、共同研究の成果物について、甲乙協議の上、評価を行うものとする。

第4条（本共同研究の遂行）

1. **誠実義務**
   甲及び乙は、本契約書に基づき、誠実に共同研究を遂行するものとする。
2. **遂行方法**
   - 甲乙は、共同研究の遂行に必要な人員、設備、資金等を適切に提供するものとする。
   - 甲乙は、共同研究の進捗状況を定期的に報告するものとする。
   - 報告の頻度及び形式については、甲乙協議の上、決定するものとする。
3. **変更の手続**
   共同研究の範囲、内容、方法等に変更が生じた場合は、甲乙協議の上、書面により合意するものとする。

1/6 ページ　3970 単語　日本語　予測入力：オン　アクセシビリティ：問題ありません　フォーカス　115%

© NTT, Inc. 2025

特長①　日本語性能のさらなる向上

ニュースリリースの改善提案。重要な改善点（簡潔な表現や全角半角の混在）を指摘

© NTT, Inc. 2025

tsuzumi 2 進化のポイント

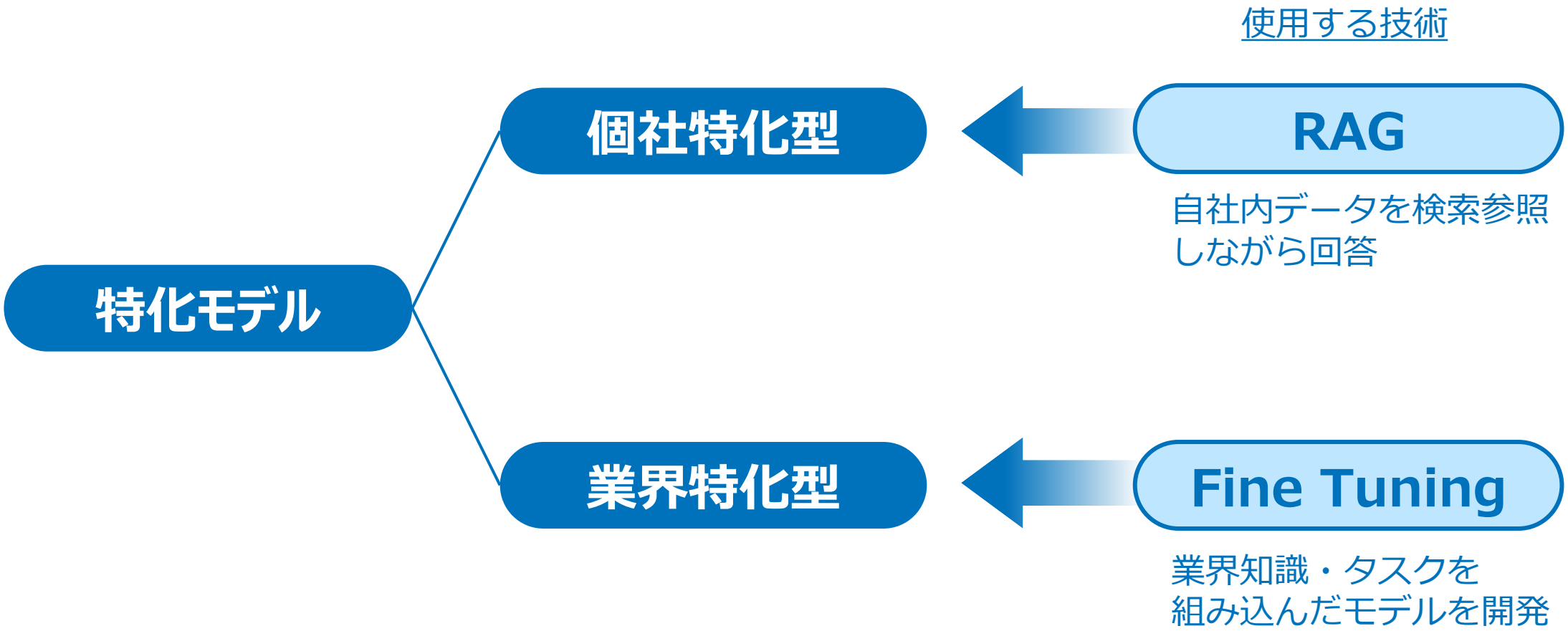
① 日本語性能のさらなる向上

② 特化型モデル開発効率の向上

③ 低コスト・高セキュアの維持、国産AI

© NTT, Inc. 2025

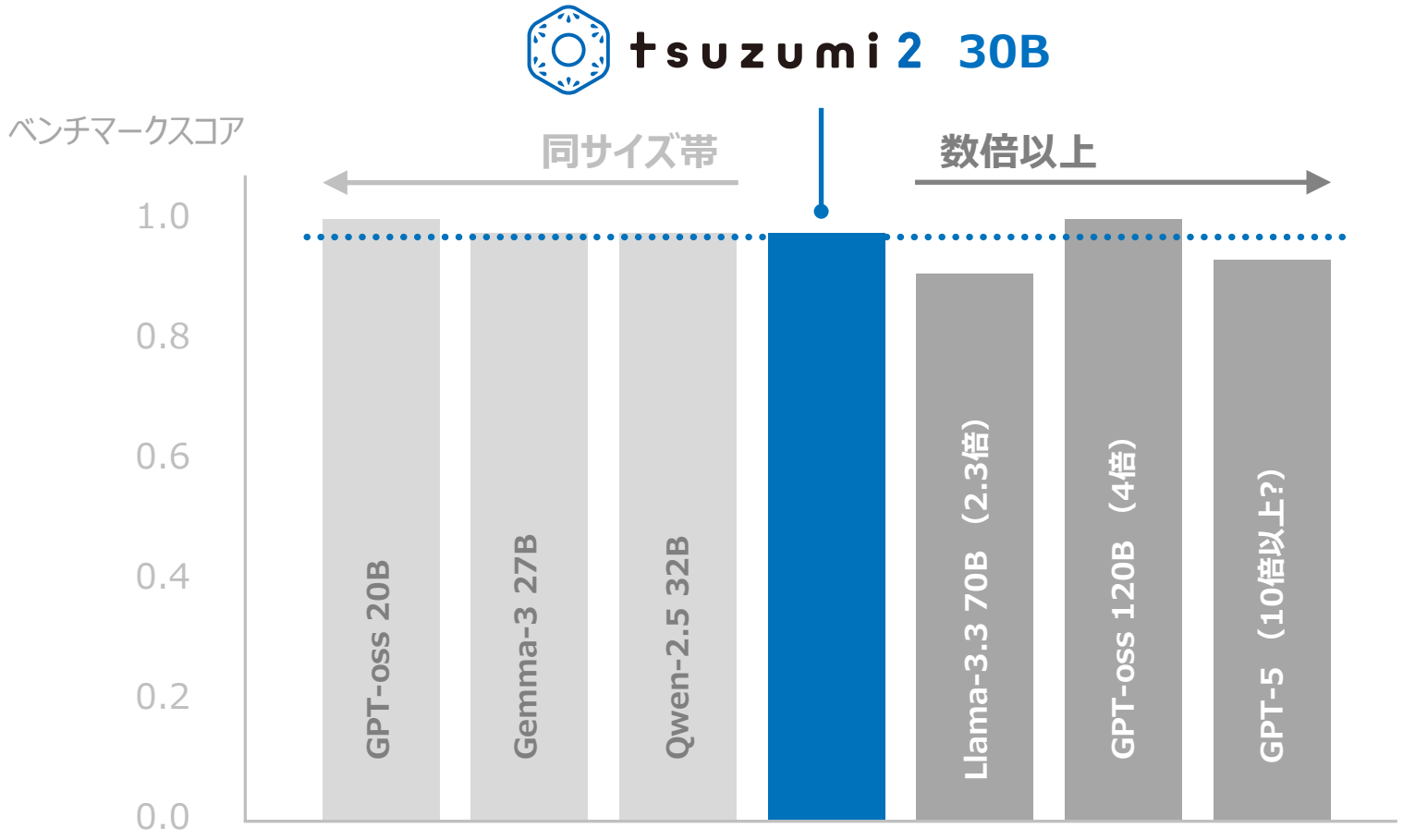
特長②　特化モデル開発効率の向上



© NTT, Inc. 2025

特長②　特化モデル開発効率の向上（RAG）

実システムへの適用評価において世界トップクラスのRAG性能を実現（財務システムに関する社内ヘルプデスク）

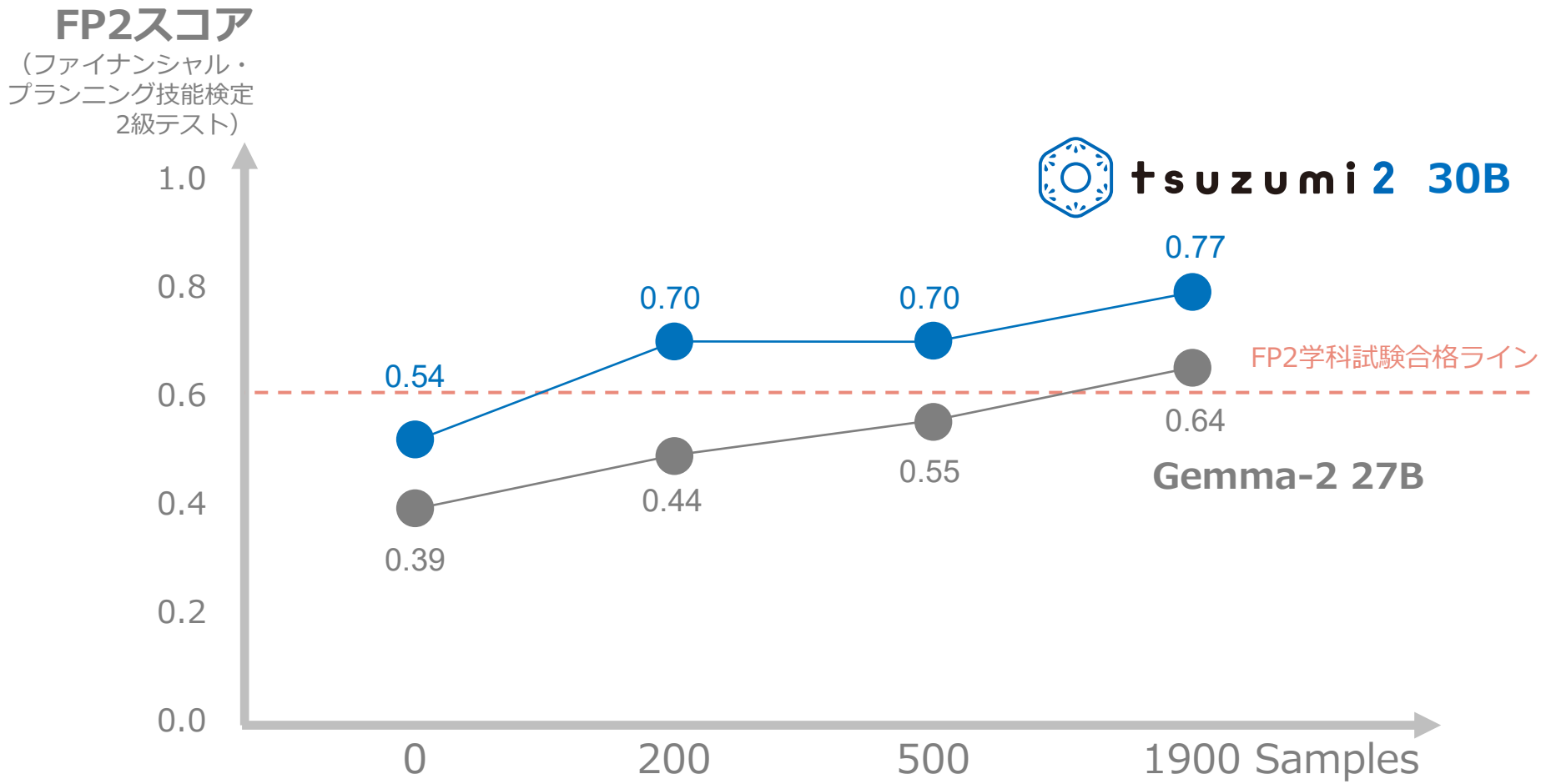


※評価条件　NTT社内業務（財務システムに関する社内ヘルプデスク）における
トライアル案件において、RAGによる問い合わせ回答正答率を独自に評価。

© NTT, Inc. 2025

特長②　特化モデル開発効率の向上（F.T）

- 金融業界の追加知識＋タスク特化チューニングによって、世界トップクラスの金融タスク性能を実現
- 他モデルに比べて少ないチューニングデータで高性能を実現



© NTT, Inc. 2025

tsuzumi 2 進化のポイント

① 日本語性能のさらなる向上

② チューニング（個社・業界特化）性能の向上

③ 低コスト・高セキュアの維持、国産AI

© NTT, Inc. 2025

特長③　低コスト・高セキュアの維持

tsuzumi 2

30B

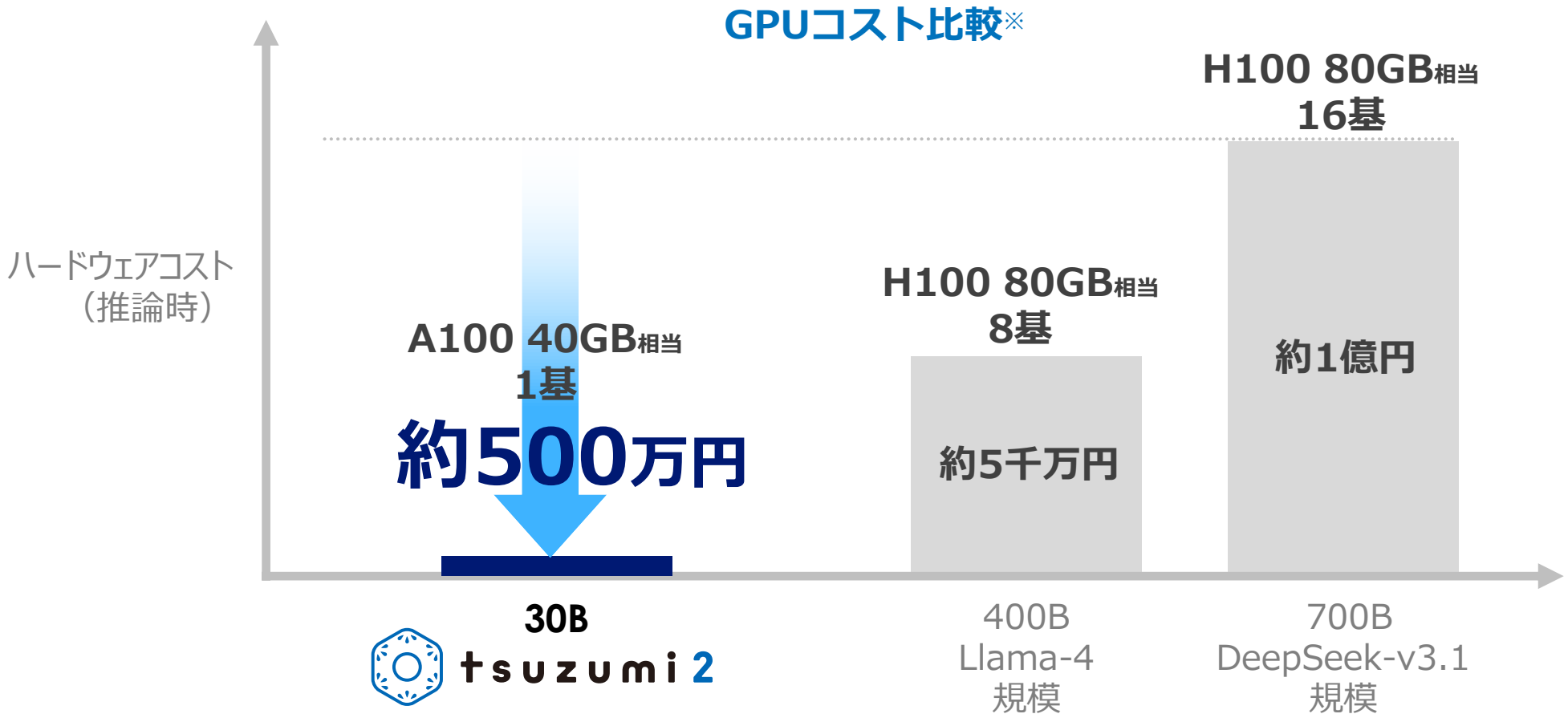
Why 30B? → **Only 1 GPU** → **On-premise**

→ **低コスト&高セキュア AI**

© NTT, Inc. 2025

特長③　低コストの維持

大規模クラスと比べて、推論コストを約10～20分の1に低減可能



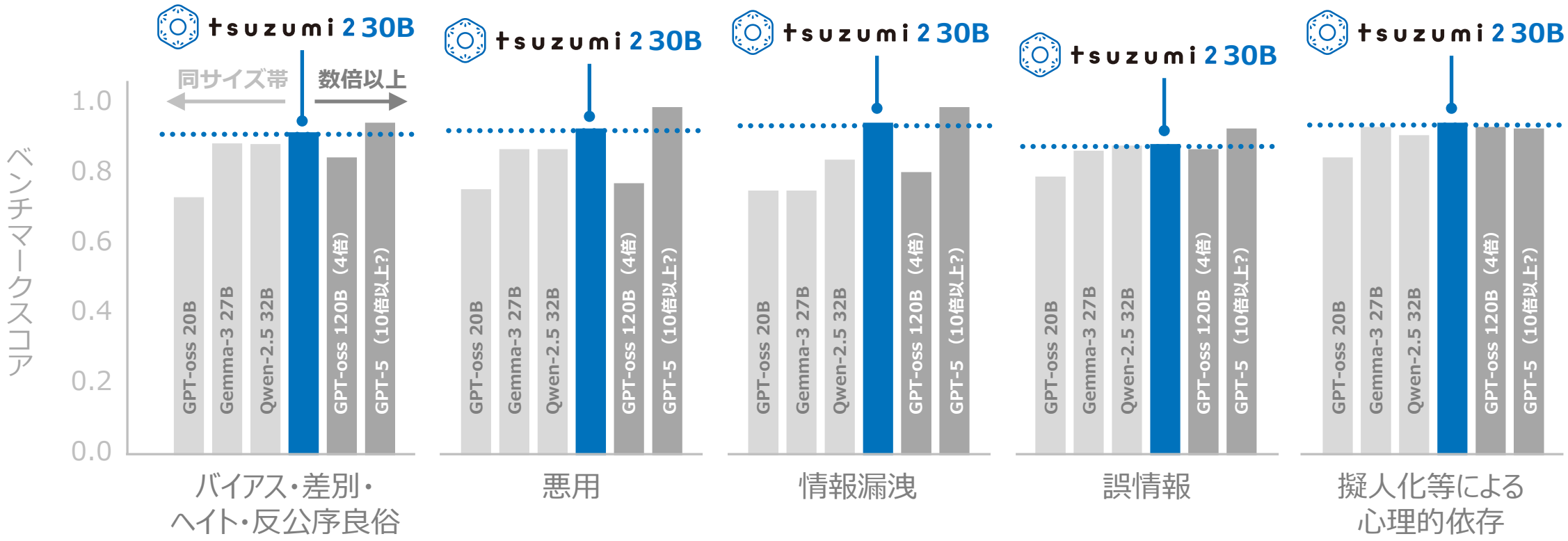
※試算条件
- 量子化: 8ビット
- 必要GPUメモリサイズ：パラメタ数 x 量子化サイズ/8bit（30Bは30GB、400Bは400GB、700Bは700GB）
- Llama-4(Llama 4 Maverick)は、MoE, H100 x 8 (1ノード)動作、DeepSeek-v3.1は、MoE, H100 x 16 (2ノード)動作を前提
- ハードウェアコストは、上位GPU H100 80GB: 1,000万円／台, 下位GPU A100 40GB: 500万円／台として換算、その他の運用などの費用は含まず

© NTT, Inc. 2025

特長③　高セキュアの維持

モデル自身の安全性も主要モデルと比較して高いスコアをマーク

日本語安全性比較※



※評価条件　AnswerCarefully（x0.2）

© NTT, Inc. 2025

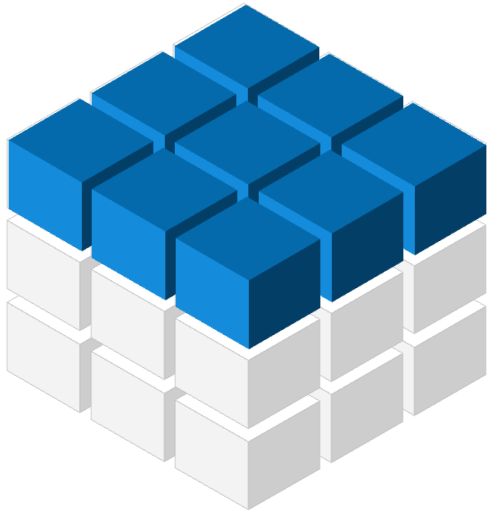
特長③　国産AI

海外製オープンAIに頼ることなく、スクラッチで開発

LLMの開発アプローチ

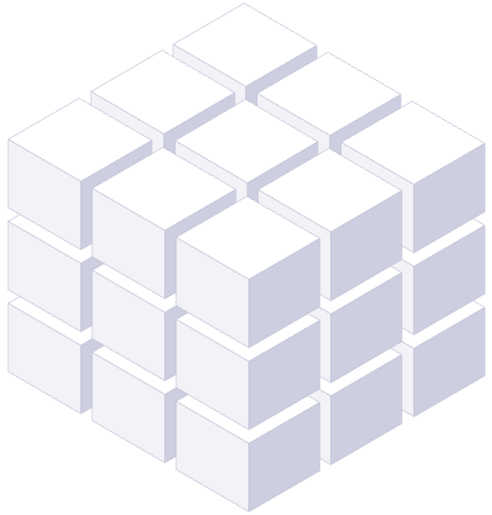
海外製オープンAIをベースに日本語データで追加学習

- ELYZA, RICOHなどはLlamaベース



基盤モデルを一から作成

- NTT tsuzumi



© NTT, Inc. 2025