日本語マルチターンデータセットの作成と効果検証

RICOH
imagine. change.

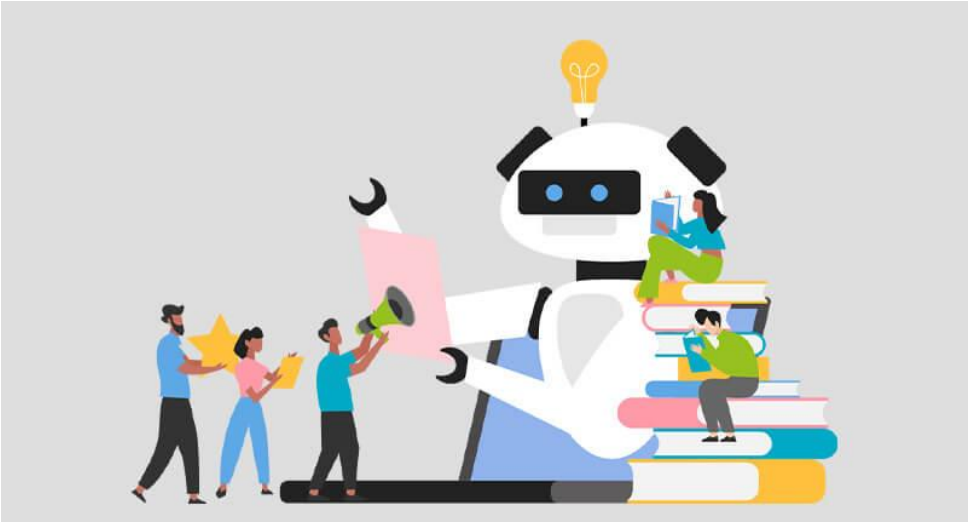
株式会社リコー　デジタル戦略部
言語AI開発室
佐藤奈穂子　伊藤真也
株式会社リコー商会
栗川朋子

本発表の概要

RICOH
imagine. change.

- 背景と目的
- 課題
- データセット構築
  - データデザイン
  - 構築体制
- データセットの効果検証
  - 評価指標
  - 人手による評価結果
- まとめと今後の課題

背景



2021年 「仕事のAI」シリーズ

2023年 新サービス「ノーコード開発ツール」
（テキスト分類AI）

2024年 生成AI、リコーデジタルバディ

大規模言語モデル（以下LLM）

「お試し」から「業務利用」へ
⇒お客様の業務に寄り添うAIソリューション
⇒専門領域特化LLM / プライベートLLM

独自LLM
オープンLLM
← **学習**

ドメイン適応 特定ドメインの語彙＆表現

タスク適応 特定タスクの指示応答ロジック

2023～ 理研プロジェクト参画
独自インストラクションデータ開発

目的

2025年はAIエージェント元年（と言われている）
生成AIの本格利用が進む中、弊社もAIエージェントの開発に取り組んでいる※

AIエージェントとは？　特定された目標に対し、必要な情報を収集し、その情報に基づいて必要なアクションを自律的に決定、実行しながら目標を達成するシステム

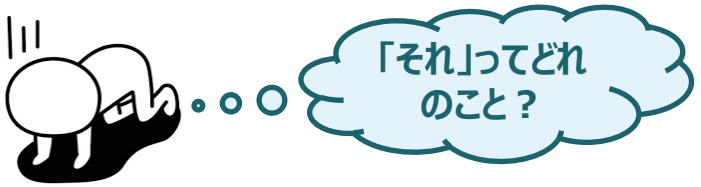
期待シーン

・カスタマーサポート
・営業/マーケティング支援
・パーソナルアシスタント

主にユーザーとの対話を通じて必要情報の取得、提供を行なうことで目標に至るタスクを特定し実行するために、**文脈を維持した自然な対話**が求められる

しかし！ **ロングターンの対話出力がイマイチなモデルが多い**

既存LLM
ロングターンの自然な対話が苦手 💦
シングルターンデータだけでは不足?!



一問一答を繋げるだけでは不自然、対話の訓練が要る

実際の顧客の運用シーンを想定した対話を訓練する
日本語のマルチターンデータセットが欲しい

※「AIエージェントを使い倒す未来」 https://jp.ricoh.com/about/ricoh-bil/topics/202501-01

課題1

RICOH
imagine. change.

データデザイン　どんなデータであれば良いか？

■ 求めるモデルの挙動

自然なマルチターン対話ができること
⇒**対話履歴の文脈情報も考慮に入れて**指示に対して適切な応答をする

■ モデル運用上の条件

- ターゲットは製造業、金融/保険業
- シーンは現実的なビジネス上の対話が生じる場
- 現行ビジネスマナーに則った自然な日本語対話
- 話者はhuman（ユーザー）とAI（アシスタント）
- 人権侵害や名誉棄損になりえる内容は避ける

課題2

RICOH
imagine. change.

データ効果検証　**効果をどうやって確認すれば良いか？**

そういえば！

対話の成功指標とは？

マルチターン対話出力の評価方法って？

対話内でシナリオから逸脱していないか？
ユーザの課題が解決したかどうか？

流暢さや正確さ、直前の発話に追従するだけでなく、
対話が成功していることを評価する指標が必要

実際の運用シーンを想定した対話が成立しているかどうかを評価する指標と評価手法を決める

データセット構築

RICOH
imagine. change.

データデザイン　**どんなデータであれば良いか？**

- 言語　日本語、機種依存文字は使用しない
- ターン数　4ターン中心
- 文字数　制限無し
- 対象ドメイン　製造業、金融・保険業
- 対話シナリオ　現実的なビジネスのシーン（複数）を想定
- 文体　human（ユーザー）の発話は任意
  AI（アシスタント）の発話は敬体とする

データセット構築

データセット構築

RICOH
imagine. change.

データデザイン　どんなデータであれば良いか？

対話シナリオ

シナリオ	説明
挨拶	「こんにちは」や「ありがとう」などの挨拶や会話の終話のみの対話
状況確認	顧客の状況を確認する対話
プランニング	製品の利用プランや、ツールの導入案などの計画を立てる対話
サポートサービス	顧客に対する、契約内容や利用方法の説明を含む基本的なサポートや保証・修理に関する対応をカバーする対話
商品特性/技術/システム	商品の主な機能や新機能の紹介、他社製品や商品との違いなどを知る目的の対話 ユーザーが製品・商品やシステムの特性について理解を深めるための対話
料金プラン	特定の商品やサービスの料金プランの紹介や、支払いオプションの説明、プランごとの機能比較を目的とする対話
セキュリティポリシー	データ保護とプライバシーポリシーやデータのバックアップ方法などを知る目的の対話
顧客の声	クレームや商品使用後の感想、お褒めの言葉などの対話
一般的なFAQ	業種や特定のサービスに関係なく、一般的な質問とそれに対する標準的な応答の対話 一般的な情報提供や問い合わせに対応する内容

データセット構築

RICOH
imagine. change.

データデザイン　どんなデータであれば良いか？

項目No.	記載項目		例
1	ID		1
2	ターン数		4
3	シナリオ		プランニング
4	業種		製造業
5	ターン1	human_1	グループホームに介護ロボットを導入しようか迷っています。
		発話のタスク	コミュニケーション
		AI_1	お問合せいただきありがとうございます。 弊社は介護ロボットを中心に様々なロボットの開発、製造、販売をおこなっております。 どのような目的の介護ロボットをお探しですか。
		発話のタスク	情報提供/情報要求
6	ターン2	human_2	コミュニケーションです。人手不足で、1人の利用者さんにかける時間がどうしても短くなってしまって。
		発話のタスク	回答
		AI_2	コミュニケーションを目的とした介護ロボットをお探しでしたら、お話ロボットはなちゃんはいかかでしょうか。 お話ロボットはなちゃんは、人と豊かなコミュニケーションをとることができるロボットです。 お話ロボットはなちゃんが利用者の相手をしてくれるので人手不足の介護施設などでの導入実績も数多く、人気がある商品です。
		発話のタスク	情報提供
7	ターン３	同上　_3	＜上記同様ターン毎に発話とタスクを記載＞
8	ターン４	同上　_4	
9	参考URL		https://www.xxx.co.jp/yyyy/

データセット構築

RICOH
imagine. change.

構築体制

データデザイナー　　1名
アノテーター　　10名
基本は個人作業、班体制で各アノテーターが作成したデータを確認、合議により修正（3名一致）
human （ユーザー）の発話とAI(アシスタント)の発話は別のアノテーターが担当

作業期間

期間：2024年6月～11月末

構築結果

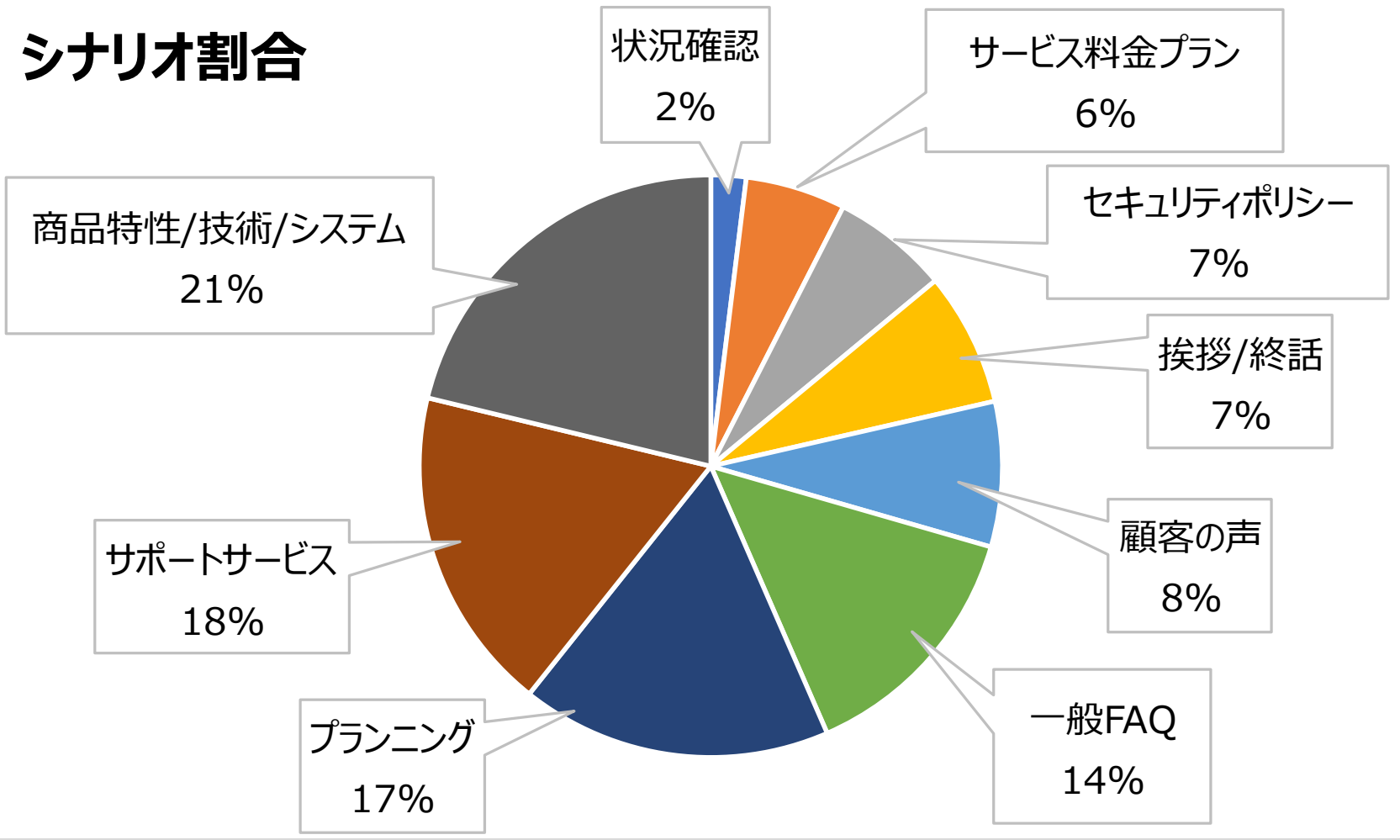
ドメイン：　製造業、金融保険業
シナリオ：　挨拶含む9カテゴリ
データ数：　1,079
ターン数：　4ターン（３～6もあり）
発話総数：8,610

データセット構築

RICOH
imagine. change.

構築結果

シナリオ割合



データセットの効果検証

データセットの効果検証

RICOH
imagine. change.

データ効果検証　効果をどうやって確認すれば良いか？

評価指標

- 1ターン毎に、流暢性・関連性・正確性について評価
- 対話の成功指標は、対話全体のシナリオ合致性・文脈追従性・課題解決度について評価
- これらの指標を用いて、最高点３を出発点として、減点法で評価をスコア化

評価指標	評点	判断ポイント
流暢性	3	正しい日本語、文法的誤りがない 他
	2	常態と敬体が混在、もしくは常態で応答
	1	誤字脱字、文法的誤り、文字化け
	0	文章が成立していない、文やフレーズの重複
関連性	3	指示への適切な応答
	2	部分的応答
	1	関連話題だが、応答になっていない
	0	無関係、無意味な回答
正確性	3	事実を的確に回答
	2	真偽不明
	1	虚偽が1つ含まれている
	0	虚偽が2つ含まれている
	-1	ほとんどが虚偽、もしくは虚偽が3つ以上含まれている

データセットの効果検証

RICOH
imagine. change.

データ効果検証　効果をどうやって確認すれば良いか？

評価指標	評点	判断ポイント
シナリオ合致性	3	対話全体がシナリオの定義から逸脱していない
	2	1つのターンだけシナリオの定義、話題から逸脱している
	1	2つのターンがシナリオの定義、話題から逸脱している
	0	全て、もしくは3ターン以上、シナリオの定義、話題から逸脱している
文脈追従性	3	前の文脈を読んで応答できている
	2	1ターンだけ前の文脈を無視した応答がある 2ターン目以降で文脈に沿った回答に戻っている
	1	前の文脈を無視して応答しているターンが２つある 2ターン目以降で文脈が戻らずに会話が終わっている
	0	全て、もしくは3ターン以上、文脈を無視した応答である
課題解決度	3	ユーザーの課題が最後のターンまでに解決している
	2	ユーザーの課題が部分的に解決している
	1	1つのターンだけ解決へ導く応答があるが全体の解決には至っていない
	0	どのターンにおいても、解決へ導かれていない

データセットの効果検証

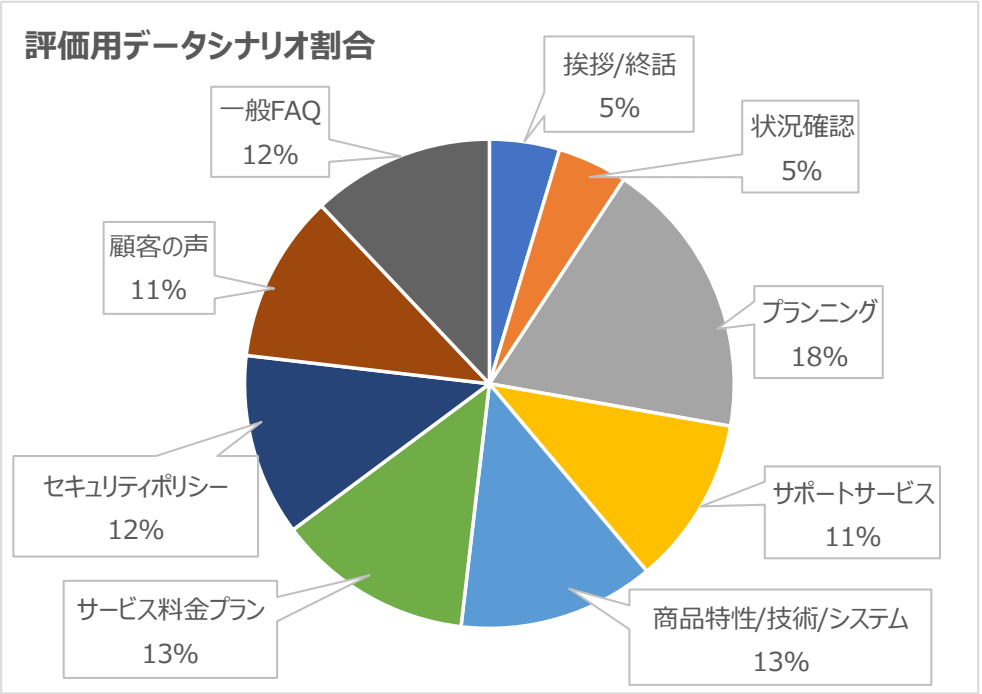
RICOH
imagine. change.

データ効果検証　**効果をどうやって確認すれば良いか？**

評価用データ

インストラクション用データと同ドメイン、同シナリオの評価用データを準備

- ✓ 同ドメイン、同シナリオを網羅した108対話
- ✓ 模範的なシナリオ会話と見做す



評価用モデル

130億パラメータの独自日本語 LLM※を用いて下記①②のモデルで評価用データの対話応答文を推論、出力

①Base:インストラクション無しモデル

②Inst済：作成したマルチターンデータによるインストラクションチューニングモデル

※ 2024年6月3日ニュースリリース済

データセットの効果検証

RICOH
imagine. change.

データ効果検証　**効果をどうやって確認すれば良いか？**

人手による評価　　アノテーター　　10名

- ✓ 流暢性・関連性・正確性・シナリオ合致性・文脈追従性・課題解決度の６つの指標の各判断ポイントに基づき、下記①②のモデルが出力した対話文にスコアを付与

  ①Base:インストラクション無しモデル
  ②Inst済：作成したマルチターンデータによるインストラクションチューニングモデル

- ✓ 基本は個人作業、班体制をとり班毎にシナリオを担当し、各アノテーターが評価した結果を3名で合議し、最終評価スコアを決定

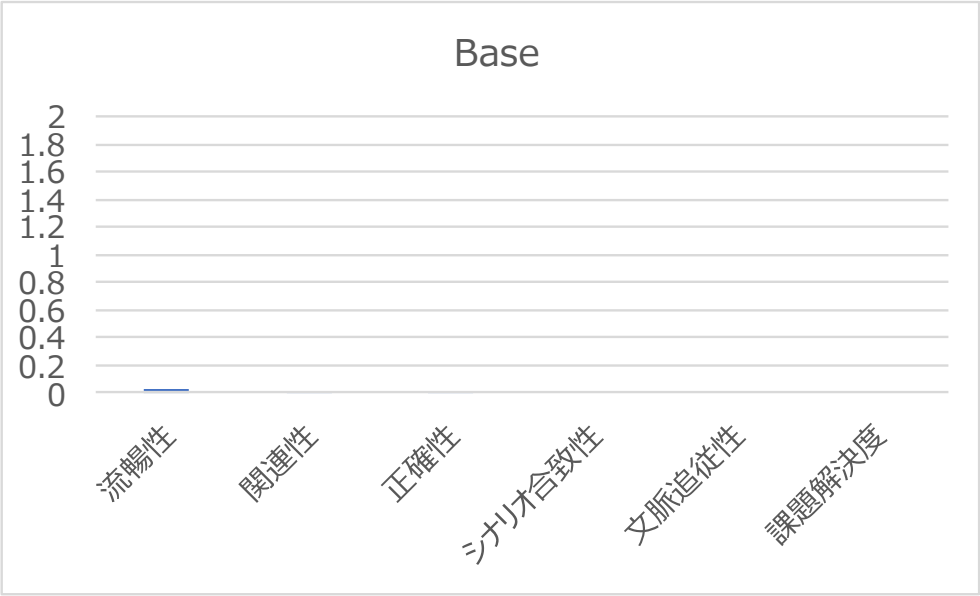
データセットの効果検証

RICOH
imagine. change.

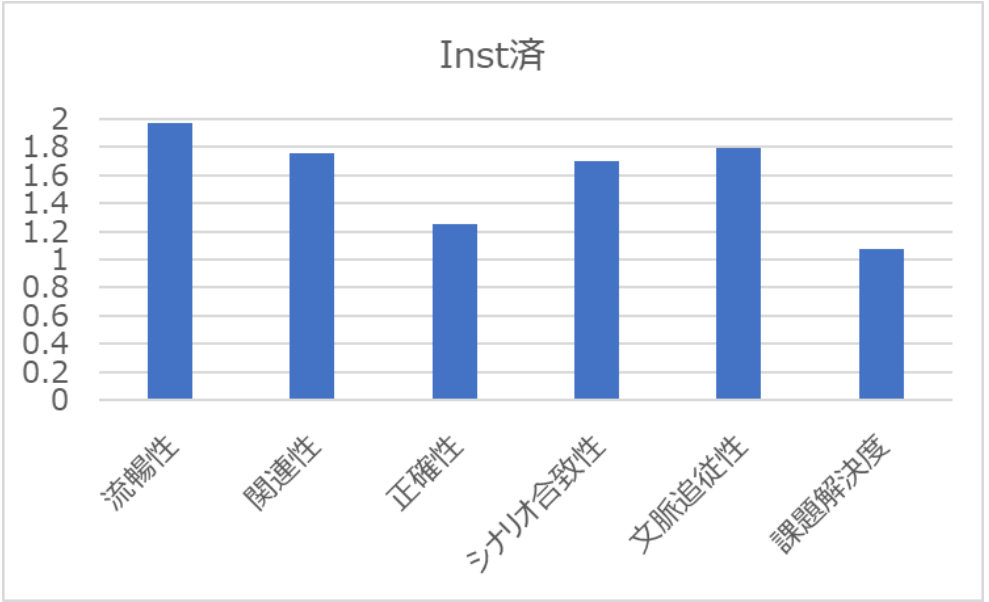
データ効果検証　**効果をどうやって確認すれば良いか？**

結果　その１

①Baseモデルに対し、②Inst済モデルは、全指標平均1.58ポイント向上、
マルチターンデータによるインストラクションの効果が明らかに見られた



①Base:インストラクション無しモデル



②Inst済：作成したマルチターンデータによる
インストラクションチューニングモデル

データセットの効果検証

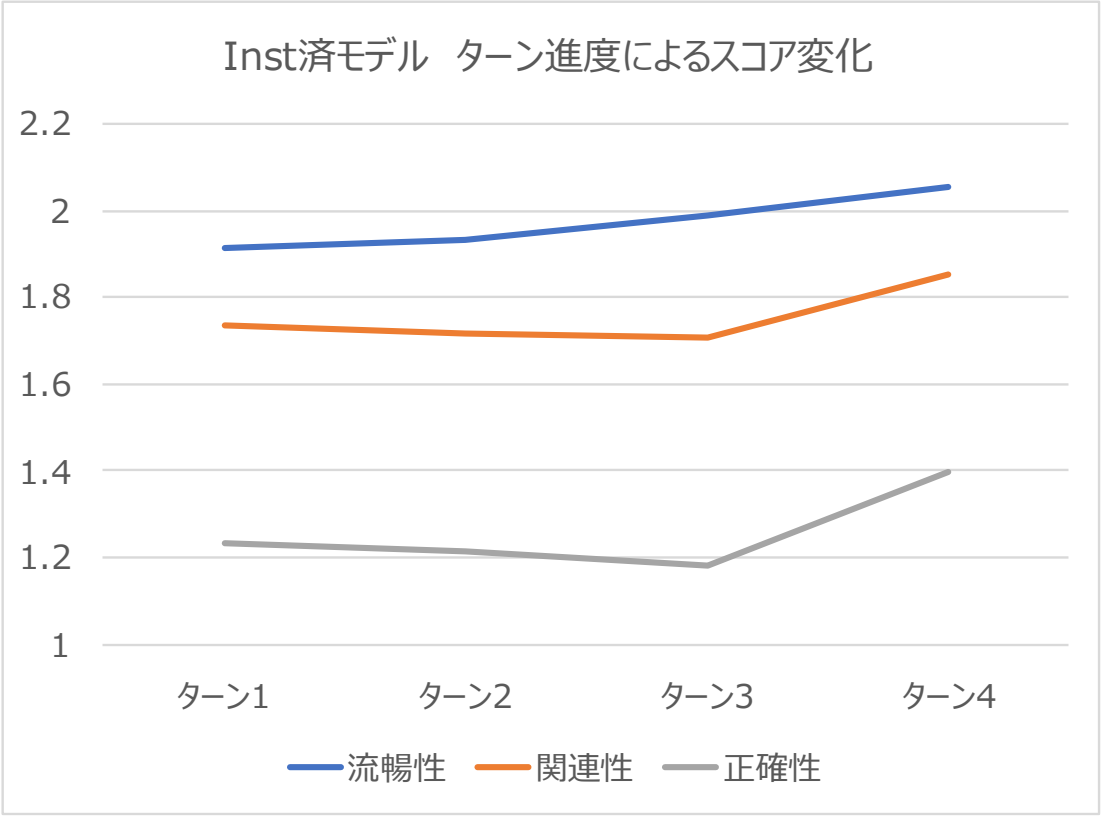
RICOH
imagine. change.

データ効果検証　**効果をどうやって確認すれば良いか？**

結果　その２

流暢性・関連性・正確性の平均スコアは、対話ターンが進むにつれ向上している

マルチターンデータによるインストラクションの効果と考えられる



データセットの効果検証

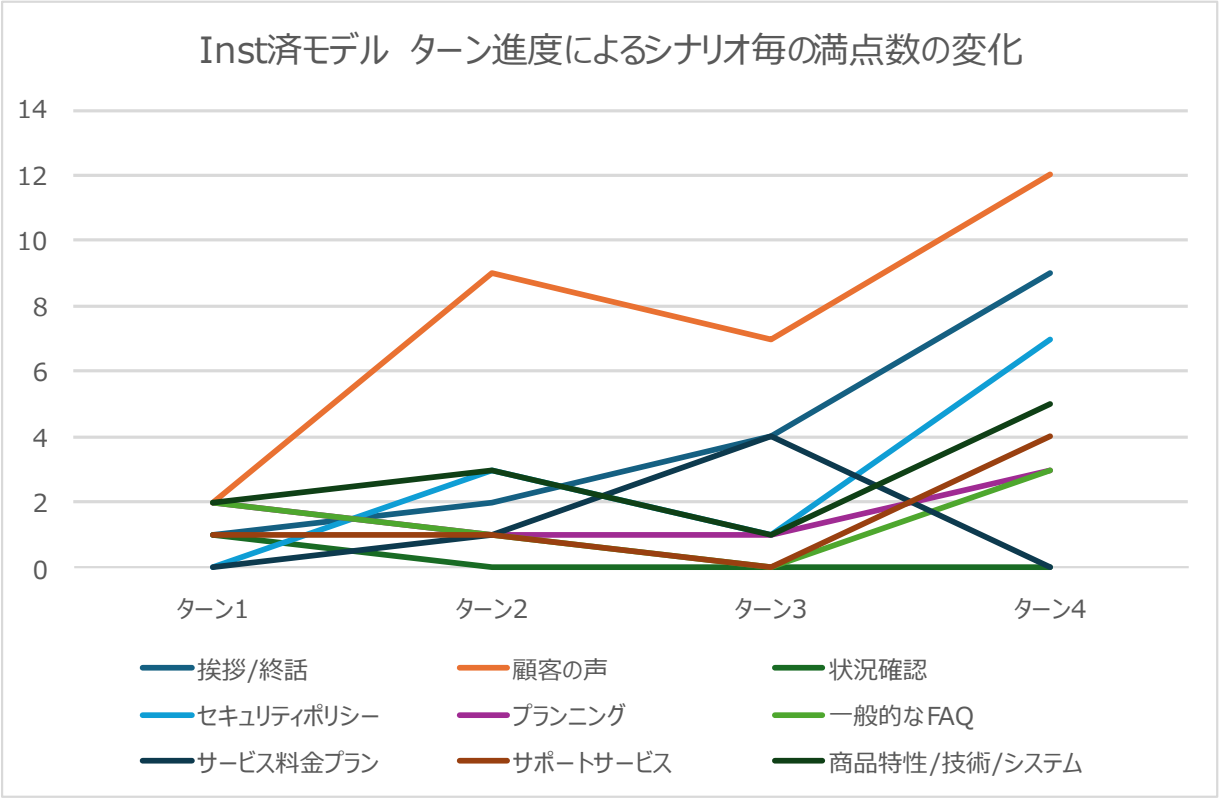
RICOH
imagine. change.

データ効果検証　**効果をどうやって確認すれば良いか？**

結果　その3

流暢性・関連性・正確性の満点数も、対話ターンが進むにつれ高くなっているシナリオが多い

マルチターンデータによるインストラクションの効果と考えられる



データセットの効果検証

RICOH
imagine. change.

データ効果検証　**効果をどうやって確認すれば良いか？**

定性的な分析

Inst済モデルの応答発話には、質問に追従はしているが内容が曖昧でスコアを落としている事例が多く見られた

あんしんプランの方が手厚いんですね。標準プランとあんしんプランの保険料を教えてください。

Bebeの標準プランの保険料は、「年齢」と「補償内容」の組み合わせで決まります。「保険料と補償内容一覧」お渡ししますね。

★標準プランは～、あんしんプランは～といった金額をずばり答えて欲しいところ、保険料の決め方を答えたり、一覧表を渡す、などの応答になってしまっている。

データセットの効果検証

RICOH
imagine. change.

データ効果検証　効果をどうやって確認すれば良いか？

対策案：インストラクションデータの応答文を、聞かれたことにまず答えるという構成にする

human	AI
ご紹介いただいた運搬ロボットは、どのような仕組みで走行するのか簡単に教えてください。	ご紹介いたしました「オートノマスマシーン」は「自律走行搬送ロボット」と呼ばれるロボットで、磁気テープなどのガイドを必要とせず、搭載されたカメラやセンサーなどで自己位置の推定を行い、ルートを自動算出して走行する仕組みです。
走行ルート上に障害物があった場合はどうなりますか。	走行ルート上に障害物があった場合でも、「オートノマスマシーン」は搭載カメラやセンサーで検知し自動で回避します。

AI（更新バージョン）
磁気テープなどのガイドを必要とせず、搭載されたカメラやセンサーなどで自己位置の推定を行い、ルートを自動算出して走行する仕組みで、「自律走行搬送ロボット」と呼ばれます。
搭載カメラやセンサーで検知し自動で回避します。

まとめと今後の課題

RICOH
imagine. change.

まとめと残課題

RICOH
imagine. change.

顧客の運用シーンを想定した対話を訓練する日本語のマルチターンデータセットを構築し、評価した

- ✓ 実用を想定し、データをデザイン
- ✓ 構築したデータの効果を検証するための指標を策定
- ✓ 策定した指標を用いて人手による検証を実施
- ✓ マルチターンデータの効果を確認

残課題

- ✓ 応答発話の文章構成見直し
- ✓ ドメイン、シナリオの拡張
- ✓ 評価指標をLLMによる自動評価に展開

RICOH

imagine. change.