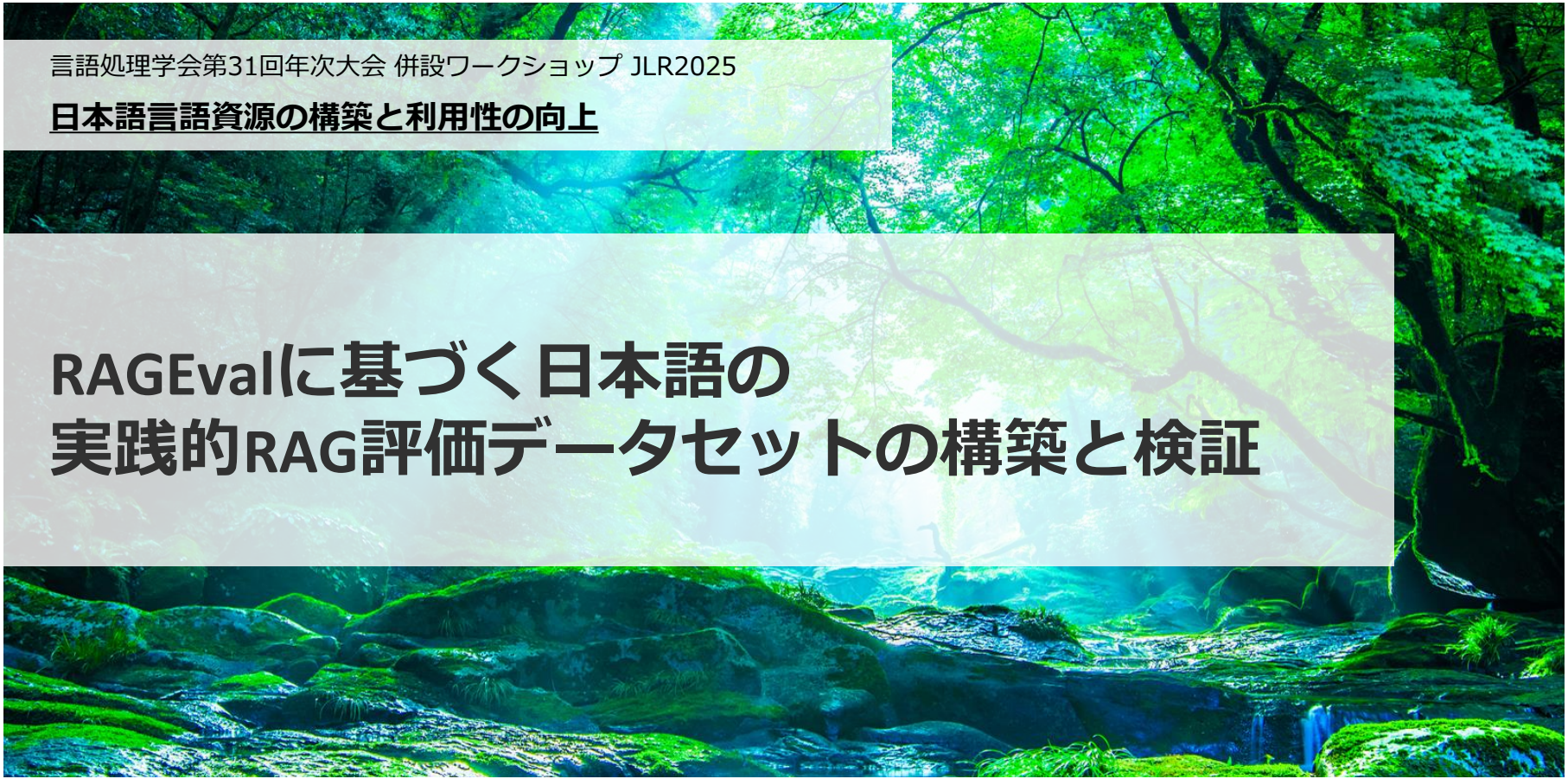


言語処理学会第31回年次大会 併設ワークショップ JLR2025

日本語言語資源の構築と利用性の向上

RAGEvalに基づく日本語の実践的RAG評価データセットの構築と検証

白邦裕千, 石井愛（BIPROGY株式会社）

2025/03/14

BIPROGY | Foresight in sight

Take Home Message

目的

顧客の個別ドメイン・想定される利用シーンに即したRAGシステムを
事前検証できるようにするための**RAG評価データセット**の作成

本発表の内容

RAG評価データセット**SynRAG**の構築・検証

- 構築：RAGEvalフレームワークを修正・拡張
- 検証：
  - ドメイン知識を反映した文書および質問をLLMとの協働で作成できることがわかった
  - RAG評価指標による検証により通常のRAGでは難易度が高い問題を示した

→ 従来のと比較し、実用面でより有用性の高いデータセットである可能性

アジェンダ

1 背景・目的

2 先行研究

3 手法

4 実験結果・考察

5 まとめ

アジェンダ

1 背景・目的

2 先行研究

3 手法

4 実験結果・考察

5 まとめ

背景

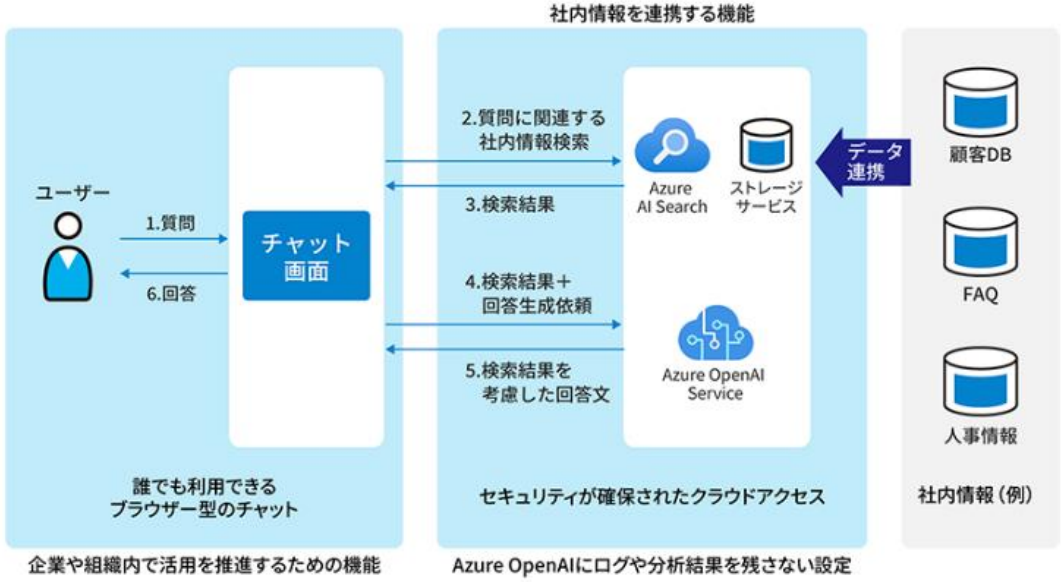
弊社では**Azure Open AI Serviceスターターキット Plus**を提供しており、RAGによる様々なドメインでの社内ナレッジの有効活用をサポートしております。

提供サービス例

- ChatGPT利用環境
- チャットインターフェース
- 活用シナリオ検討

AI適用事例

- ChatGPTによる会話型アプリ
- CoTアプリ
- **RAGアプリ**



課題と目的

顧客の社内へのRAGシステム適用における課題

1. 顧客の社内情報を用いたRAG構築は精度検証のハードルが高い
2. 顧客のドメインに合ったデータセットや、日本語のデータセットがない
3. 既存のデータセットはWikipediaベースのものが多く、LLMが事前学習に用いていない文書群が検索対象の場合に情報を適切に活用できるかを検証できない

本研究の目的

顧客ごとの個別ドメイン・想定される利用シーンに即したRAGシステムの事前検証を可能にする

- **ドメイン**の情報を踏まえた**架空**の業務文書を作成
- それに基づく**利用シーンに即した**RAG用データセットを作成

アジェンダ

1 背景・目的

2 先行研究

3 手法

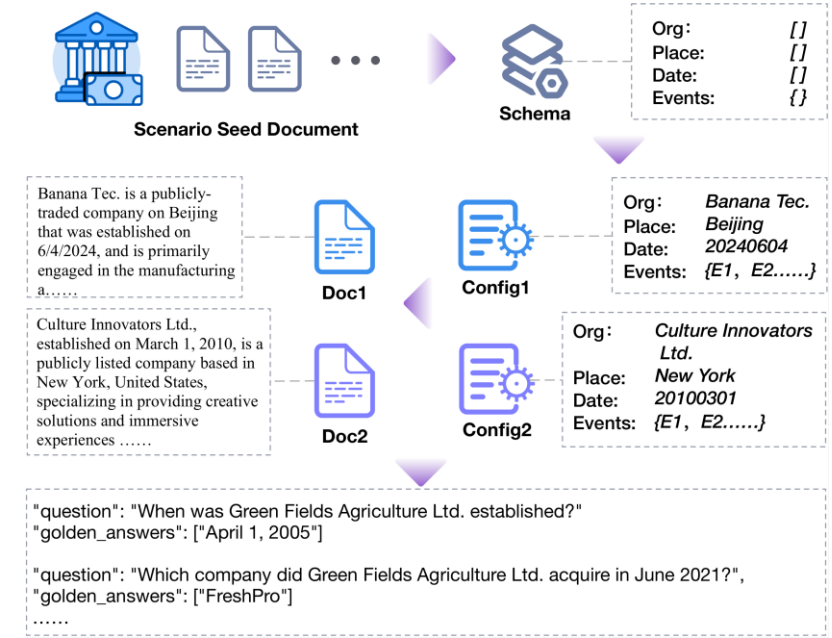
4 実験結果・考察

5 まとめ

RAGEval：RAG評価データセット生成プロセスと課題

既存研究：**ドメイン特化・架空**なデータセットの作成フレームワークを提案

- 少数のシード文書をもとに、RAGの検証用データを自動生成するプロセスが体系化



RAGEvalのデータセット生成プロセス

1. シード文書をもとにLLMでスキーマを抽出
2. スキーマの値を埋めたConfigを作成
   - 値埋めにはルールベース手法およびLLMを使用
3. Configに沿った文書生成
4. Configと文書を基に質問-参照-回答（QRA）生成

RAGEvalデータセットの課題

スキーマ・Configで文書同士の関係性がない

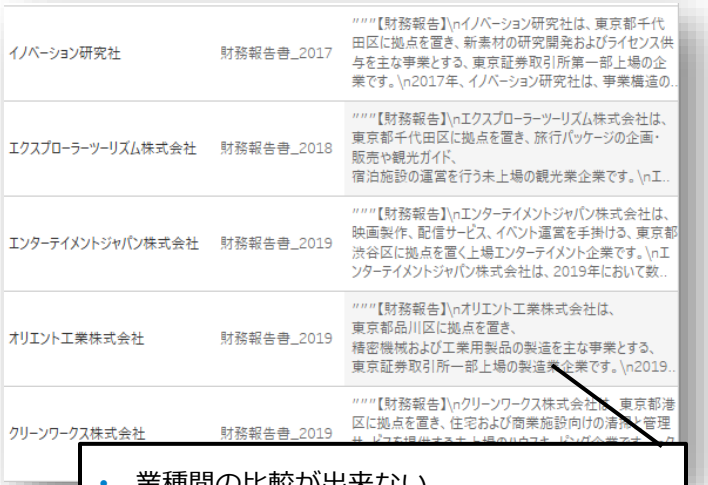
→ 実践的な複雑な質問作成が難しい

RAGEval：RAGEvalデータセットの課題

用意されたスキーマを日本語化して作成した文書とQRAデータセットの例

文書

- 文書同士の関係性や整合性が無い



イノベーション研究社	財務報告書_2017	"""【財務報告】\nイノベーション研究社は、東京都千代田区に拠点を置き、新素材の研究開発およびライセンス供与を主な事業とする、東京証券取引所第一部上場の企業です。\n2017年、イノベーション研究社は、事業構造の..
エクスプローラーツーリズム株式会社	財務報告書_2018	"""【財務報告】\nエクスプローラーツーリズム株式会社は、東京都千代田区に拠点を置き、旅行パッケージの企画・販売や観光ガイド、宿泊施設の運営を行う未上場の観光業企業です。\nエ..
エンターテイメントジャパン株式会社	財務報告書_2019	"""【財務報告】\nエンターテイメントジャパン株式会社は、映画製作、配信サービス、イベント運営を手掛ける、東京都渋谷区に拠点を置く上場エンターテイメント企業です。\nエンターテイメントジャパン株式会社は、2019年において数..
オリエント工業株式会社	財務報告書_2019	"""【財務報告】\nオリエント工業株式会社は、東京都品川区に拠点を置き、精密機械および工業用製品の製造を主な事業とする、東京証券取引所一部上場の製造業企業です。\n2019..
クリーンワークス株式会社	財務報告書_2019	"""【財務報告】\nクリーンワークス株式会社は、東京都港区に拠点を置き、住宅および商業施設向けの清掃と管理サービスを提供する未上場のリクエ[illegible]

- 業種間の比較が出来ない
- 同一企業の時系列による比較が出来ない

QRAデータセット

- 不自然な質問

西区中央病院の入院記録によると、河野学の出生地はどこですか？
逗子市立病院の入院記録によると、林美羽の母の健康状態は何ですか？
平塚市総合病院と豊島総合病院の入院記録によると、吉田詩織と石井大樹の手術内容は何ですか？
北区総合病院と千葉市立病院の入院記録によると、菊地守と中野心春のどちらが先に入院しましたか？
2020年にメディアインク株式会社が改定したポリシーは何ですか？
2021年に未来学園株式会社が締結したライセンス契約の相手はどのような機関ですか？

- 表面的な質問
- 実用上起こりえないであろう質問

- 不自然なデータ構造

成田市立総合病院の入院記録により、遠藤聡の治療計画を要約してください。
成田市立総合病院の入院記録によると、遠藤聡の2型糖尿病の診断根拠はいくつありますか？
⋮
成田市立総合病院の入院記録により、遠藤聡の入院後の経過[illegible]

- 1病院1患者（1業種1企業）

アジェンダ

1 背景・目的

2 先行研究

3 手法

4 実験結果・考察

5 まとめ

RAGEvalを改善したデータセット生成プロセス

実際に起こり得る質問を生成するため、既存の生成プロセスを改善

①**想定質問**起点の文書設計, ②構造化データ**生成ロジック**の精緻化 でデータセットを高度化

RAGEvalを改善したデータセット生成プロセス

新たに追加したプロセス

1. 少数のシード文書をもとにLLMでスキーマを作成
   - **1-1: LLM・専門家との反復的対話プロセスに基づき、想定質問リストを作成**
   - **1-2: 想定質問から、文書の依存関係を定義・Configを更新**

   ①
2. ルールベース手法およびLLMでスキーマの値を埋めたConfigを作成
   - **2-1: ドメイン固有の制約と文書同士の依存関係を考慮したルールベースのプログラム構築**
3. Configに沿った文書生成
   - **3-1: LLMによる文書レビュ、修正点の反映（必要に応じて2-1～3-1を反復）**

   ②
4. Configと文書を基に質問-参照-回答（QRA）生成

改善した生成プロセス詳細

データセット作成

①想定質問起点の文書設計

1-1: LLM・専門家との反復的対話プロセスに基づき、想定質問リストを作成

架空の入院記録を参照しながら答える質問応答データセットを作成します。
入院記録を業務で用いる人が実際に調べたいような問題を作成します。

その入院記録を参照しなければ答えられない問題、もしくは、その入院記録と一般的な知識を掛け合わせて答える問題の例を挙げてください。

入院記録は、添付のような形式で同じ患者の単一の記録が存在する場合と、複数の記録が存在する場合があります。
…
問題をカテゴリごとに分類して質問と回答生成の例を示してください。

入院記録を用いた質問応答データセットのカテゴリと例を示します。
1. 単一記録からの事実確認
質問：患者P502596の入院時の主訴は何でしたか？また、発症からどのくらいの期間で入院に至りましたか？
回答：入院時の主訴は頭痛、下痢、筋肉痛でした。4月28日に症状が発現し、5月4日に入院したため、発症から6日後に入院に至りました。
2. 複数記録の比較分析
質問：患者の2回の入院における肝機能検査（AST/ALT）の値を比較し、どのように変化したか説明してください。
回答：1回目の入院時（2024年1月）はAST 19U/L、ALT 27U/L と正常範囲内でしたが、2回目の入院時（2024年5月）…

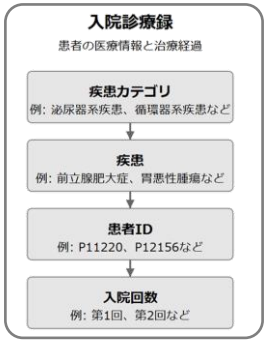
②構造化データ生成ロジックの精緻化

2-1: ドメイン固有の制約と文書同士の依存関係を考慮したルールベースのプログラム構築

- ドメイン固有の制約（疾患ごとの制約条件）

カテゴリ	性別	年齢範囲	疾患数
女性特有疾患 (female_specific)	女性	20～50歳	13種 乳房悪性腫瘍、子宮頸部悪性腫瘍、妊娠管理など
男性特有疾患 (male_specific)	男性	40～90歳	1種: 前立腺肥大症または炎症
高齢者疾患 (elderly)	制約なし	65～95歳	3種 老年期精神病、パーキンソン病、白内障
新生児疾患 (neonatal)	制約なし	0歳	4種 早産児および未熟児、分娩損傷など

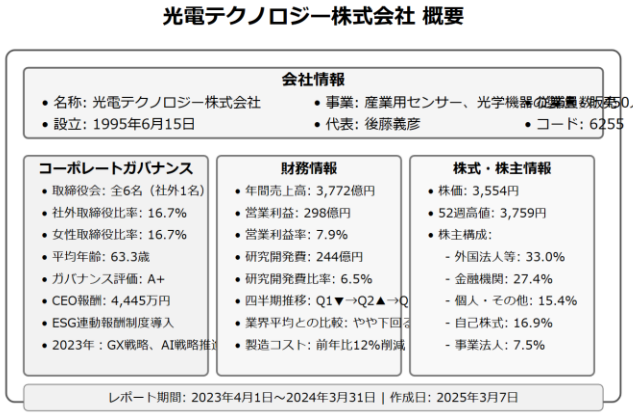
- 文書同士の依存関係



Configの例：財務ドメイン（統合報告書）

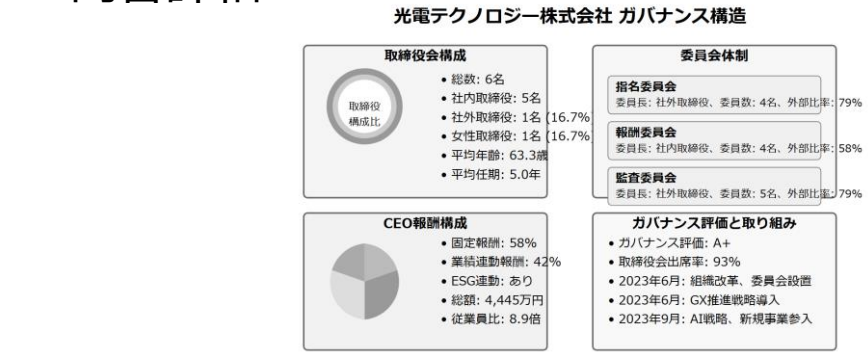
データセット作成

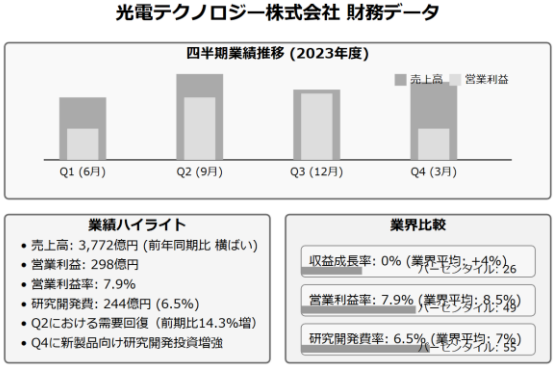
全体

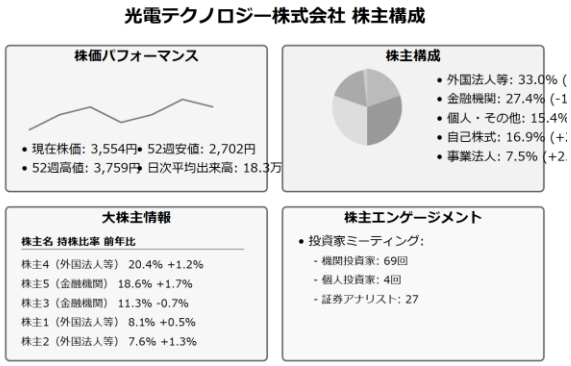


内容詳細

Claude 3.7 SonnetでConfigを可視化







Configの例：医療ドメイン（入院診療録）

データセット作成

■ 全体

入院記録データ構造

基本情報
- 氏名: 平野絵美
- 患者ID: P11220
- 性別: 女性 / 年齢: 17歳
- 診療科: 内分泌・代謝内科

治療計画
- 食事療法（炭水化物調整）
- 運動療法（週150分以上）
- メトホルミン 500mg 1日2回 食後
- バイタルサイン測定（8時間ごと）

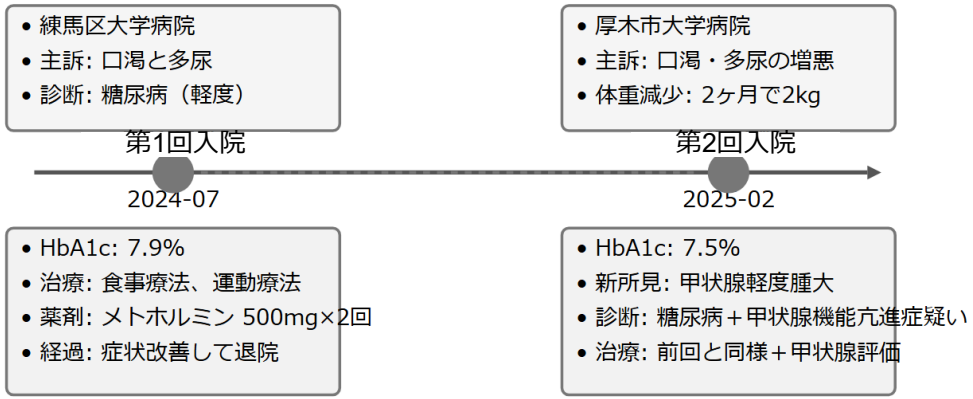
診断情報
- 主訴: 口渇、多飲および多尿の増悪
- 暫定診断: 糖尿病（軽度）
- 合併症: 甲状腺機能亢進症の疑い
- 検査値: HbA1c 7.5%, 血糖値 139mg/d
- 家族歴: 父親が2型糖尿病

入院経過
- 第1回: 2024年7月26日（練馬区大学病院）
  - 初診時 HbA1c 7.9%
  - 食事・運動療法、薬物療法開始
  - 症状改善して退院
- 第2回: 2025年2月19日（厚木市大学病院）
  - 症状再燃、甲状腺腫大を確認
  - 治療継続、甲状腺機能評価を追加

※ データは入院記録JSONファイルより抽出

■ 時系列複数データ（複数回入院）

平野絵美さん(P11220)の入院記録時系列



Claude 3.7 SonnetでConfigを可視化

実験設定

データセットの質・検索／回答生成性能の評価・検証を実施

■ データセット生成

- ドメイン／業務文書：
  1. 財務ドメイン／統合報告書
  2. 医療ドメイン／入院診療録
- LLM：

モデル	用途
Claude 3.5, 3.7	• 想定質問リスト作成 • スキーマ更新 • 構造データ生成プログラム生成
ChatGPT Deep Research / o3-mini-high	生成された文書のレビュー
Azure OpenAI Service API GPT-4o	• Config生成 • 文書生成 • QRA生成

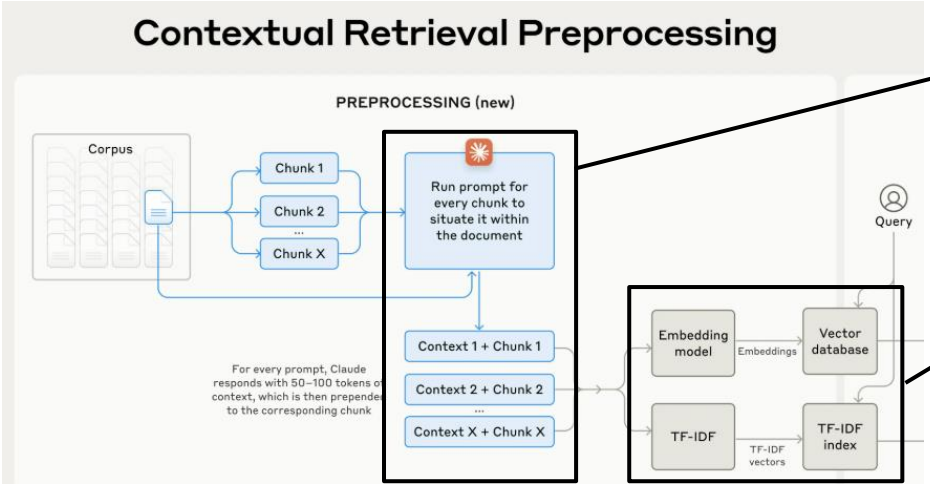
■ 評価用RAGシステム

- 検索処理：
  - 密ベクトル検索：
    - 検索エンジン：Numpyによるコサイン類似度検索
    - 埋め込みモデル：OpenAI API text-embedding-3-small
  - 疎ベクトル検索：
    - 検索エンジン： Elastic Search
    - 検索アルゴリズム： BM25
    - 日本語トークナイザ：analysis-kuromoji
- 回答生成
  - Azure OpenAI Service API GPT-4o

評価用RAGシステム

データセットの評価のため、以下の２種類のRAGシステムを実装

1. 一般的な構成のRAG：密ベクトル検索＋LLM
2. Contextual Retrieval ：

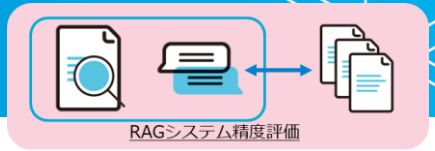


- ドキュメントとチャンクをLLMに渡して、チャンクの文脈（Context）を生成する
- 作成した文脈をチャンクの先頭に付与する

- 密ベクトル検索を実行する
- 疎ベクトル検索を実行する
- 上記検索結果をマージする

参考：Introducing Contextual Retrieval

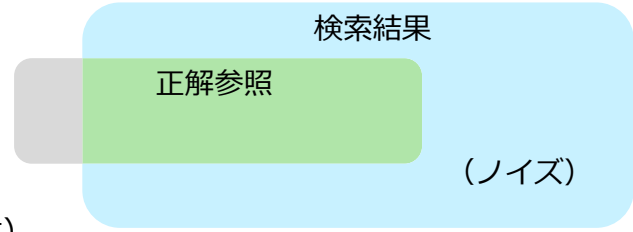
3. RAG評価指標



データセットの評価には先行研究と同じ評価指標を使用

検索

- Recall
  - （**検索結果**に占める**正解参照**の数）/（**正解参照**の数）
    → 検索の正確性の指標

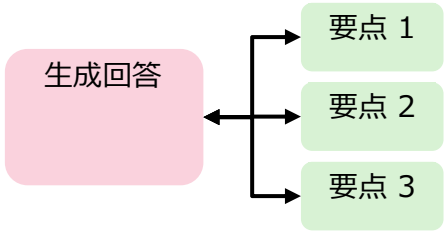


- EIR（Effective Information Rate）
  - （**検索結果**に占める**正解参照**と一致する単語数）/（**検索結果**の単語数）
    → 検索の効率性の指標

回答生成

生成回答と**正解要点**を比較し、生成した回答が以下のいずれかに該当するかを評価する※（keypoint metrics）

- Completeness（完全性）：正解要点を含んでいる
- Hallucination（幻覚）：正解要点と矛盾しているか
- Irrelevance（無関係性）：完全性、幻覚どちらにも該当しない場合
  → 生成回答の質・信頼性の指標



※ 評価にはLLM（本研究ではGPT-4o）を使用

アジェンダ

1 背景・目的

2 先行研究

3 手法

4 実験結果・考察

5 まとめ

SynRAG※データセットの特徴

■ 財務ドメイン（統合報告書）

- ◆ 構造的特徴
  - ➢ 文章構造：叙述形式
  - ➢ 平均文書長：12,000字程度
- ◆ 情報的特徴
  - ➢ 時系列の数値データ比較
  - ➢ 表データ（Markdown）を含む
  - ➢ 企業の多様性の表現
    - ◆ 業種、規模、成長段階
  - ➢ 財務指標間の数学的整合性
  - ➢ ガバナンス構造と業績の関連性

■ 医療ドメイン（入院診療録）

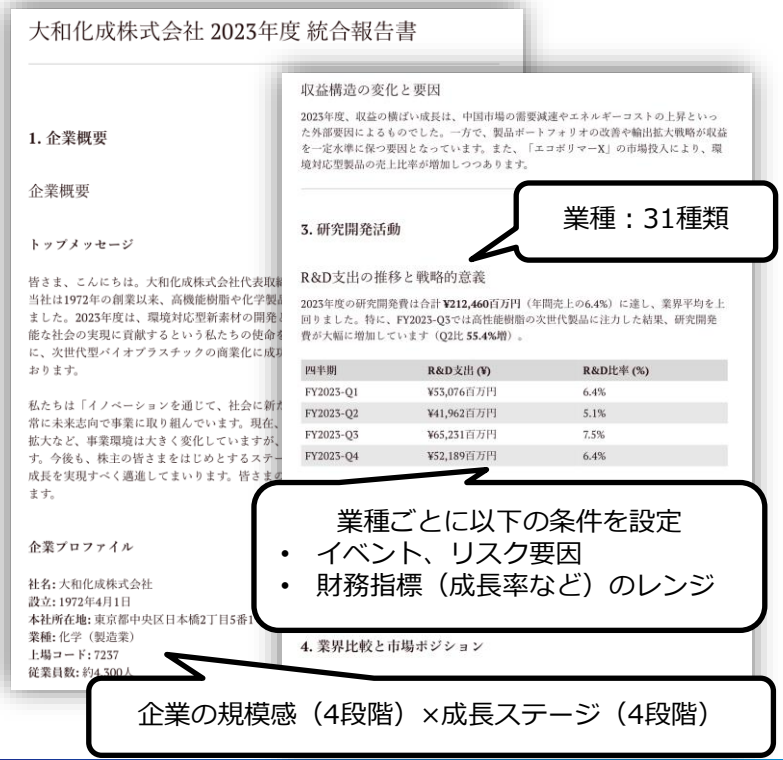
- ◆ 構造的特徴
  - ➢ 文章構造：箇条書き形式
  - ➢ 平均文書長：2,000字程度
- ◆ 情報的特徴
  - ➢ 時系列の数値データ比較
  - ➢ コーパス内の分布の計算
  - ➢ 疾患進行の時系列的一貫性
  - ➢ 医学的整合性
    - ◆ 臨床検査値・バイタルサイン
  - ➢ 治療効果と症状変化の因果関係

※ RAGEvalの生成プロセス改善前後を明確にするため、改善後のデータセットを”SynRAG”と表記

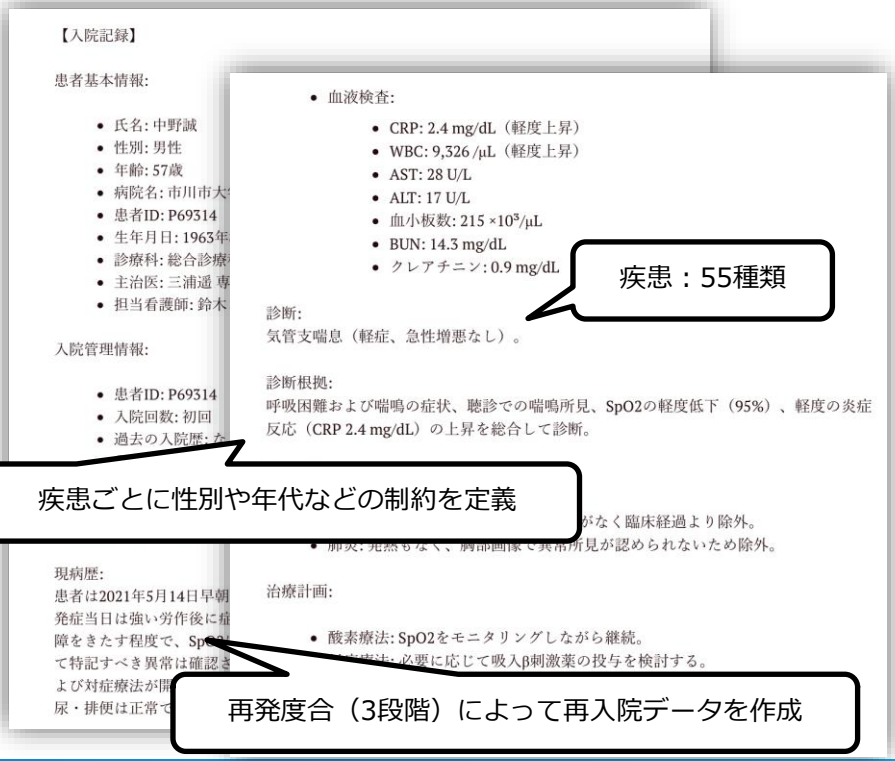
SynRAGデータセット（文書）：文書の特徴

データセット作成

財務ドメイン:



医療ドメイン:

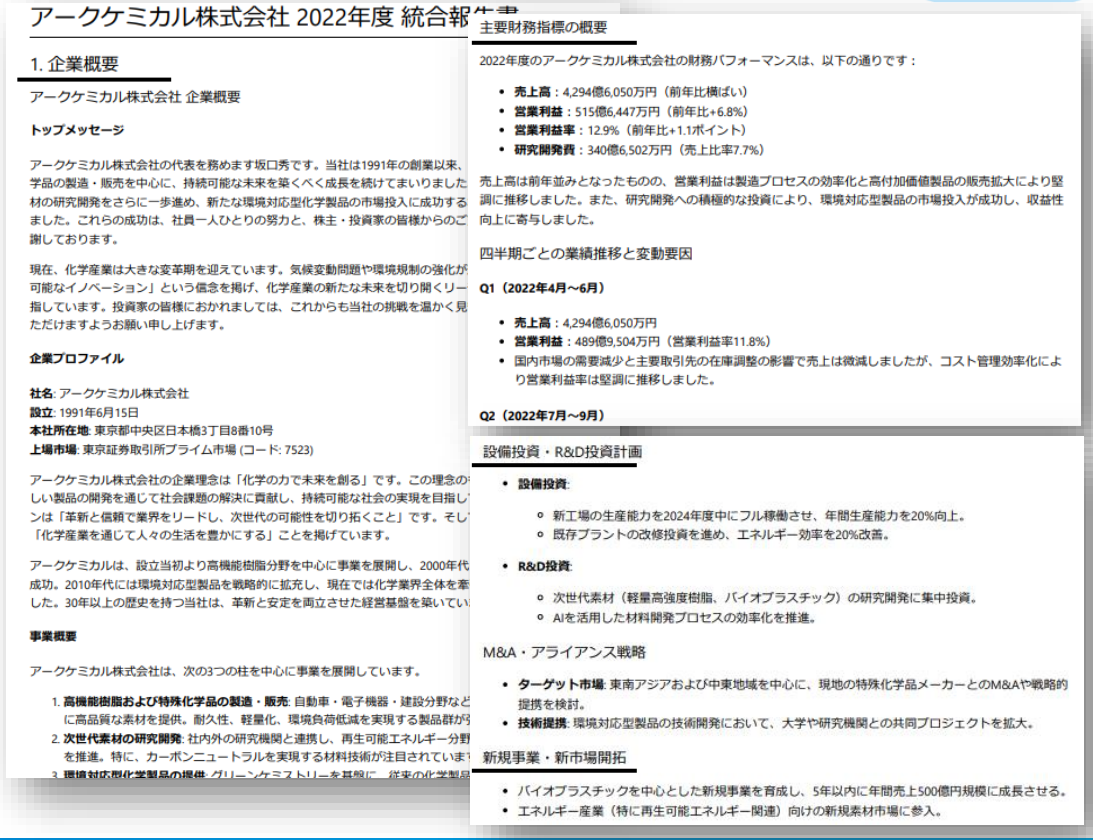


SynRAGデータセット（文書）：文書の例（財務ドメイン）

データセット作成

Case：化学メーカの統合報告書

- 情報密度
  - 企業概要, 事業概要, 経営戦略
  - 財務分析レポート
  - コーポレートガバナンス
  - 将来の見通し
    - ESG戦略, 投資戦略



アークケミカル株式会社 2022年度 統合報告書

1. 企業概要

アークケミカル株式会社 企業概要

トップメッセージ

アークケミカル株式会社の代表を務めます坂口秀です。当社は1991年の創業以来、
学品の製造・販売を中心に、持続可能な未来を築くべく成長を続けてまいりました
材の研究開発をさらに一歩進め、新たな環境対応型化学製品の市場投入に成功する
ました。これらの成功は、社員一人ひとりの努力と、株主・投資家の皆様からのご
謝しております。

現在、化学産業は大きな変革期を迎えています。気候変動問題や環境規制の強化が
可能なイノベーション」という信念を掲げ、化学産業の新たな未来を切り開くリー
指しています。投資家の皆様におかれましては、これからも当社の挑戦を温かく見
ただけますようお願い申し上げます。

企業プロファイル

社名: アークケミカル株式会社
設立: 1991年6月15日
本社所在地: 東京都中央区日本橋3丁目8番10号
上場市場: 東京証券取引所プライム市場 (コード: 7523)

アークケミカル株式会社の企業理念は「化学の力で未来を創る」です。この理念の
しい製品の開発を通じて社会課題の解決に貢献し、持続可能な社会の実現を目指し
ンは「革新と信頼で業界をリードし、次世代の可能性を切り拓くこと」です。そし
「化学産業を通じて人々の生活を豊かにする」ことを掲げています。

アークケミカルは、設立当初より高機能樹脂分野を中心に事業を展開し、2000年代
成功。2010年代には環境対応型製品を戦略的に拡充し、現在では化学業界全体を牽
した。30年以上の歴史を持つ当社は、革新と安定を両立させた経営基盤を築いてい

事業概要

アークケミカル株式会社は、次の3つの柱を中心に事業を展開しています。

1. **高機能樹脂および特殊化学品の製造・販売**: 自動車・電子機器・建設分野など
に高品質な素材を提供。耐久性、軽量化、環境負荷低減を実現する製品群が
2. **次世代素材の研究開発**: 社内外の研究機関と連携し、再生可能エネルギー分野
を推進。特に、カーボンニュートラルを実現する材料技術が注目されてい
3. **環境対応型化学製品の提供**: グリーンケミストリーを基盤に、従来の化学製品

主要財務指標の概要

2022年度のアークケミカル株式会社の財務パフォーマンスは、以下の通りです：

- **売上高**：4,294億6,050万円（前年比横ばい）
- **営業利益**：515億6,447万円（前年比+6.8%）
- **営業利益率**：12.9%（前年比+1.1ポイント）
- **研究開発費**：340億6,502万円（売上比率7.7%）

売上高は前年並みとなったものの、営業利益は製造プロセスの効率化と高付加価値製品の販売拡大により堅調に推移しました。また、研究開発への積極的な投資により、環境対応型製品の市場投入が成功し、収益性向上に寄与しました。

四半期ごとの業績推移と変動要因

Q1（2022年4月～6月）

- **売上高**：4,294億6,050万円
- **営業利益**：489億9,504万円（営業利益率11.8%）
- 国内市場の需要減少と主要取引先の在庫調整の影響で売上は微減しましたが、コスト管理効率化により営業利益率は堅調に推移しました。

Q2（2022年7月～9月）

設備投資・R&D投資計画

- **設備投資**:
  - 新工場の生産能力を2024年度中にフル稼働させ、年間生産能力を20%向上。
  - 既存プラントの改修投資を進め、エネルギー効率を20%改善。
- **R&D投資**:
  - 次世代素材（軽量高強度樹脂、バイオプラスチック）の研究開発に集中投資。
  - AIを活用した材料開発プロセスの効率化を推進。

M&A・アライアンス戦略

- **ターゲット市場**: 東南アジアおよび中東地域を中心に、現地の特殊化学品メーカーとのM&Aや戦略的提携を検討。
- **技術提携**: 環境対応型製品の技術開発において、大学や研究機関との共同プロジェクトを拡大。

新規事業・新市場開拓

- バイオプラスチックを中心とした新規事業を育成し、5年以内に年間売上500億円規模に成長させる。
- エネルギー産業（特に再生可能エネルギー関連）向けの新規素材市場に参入。

SynRAGデータセット（文書）：文書の例（医療ドメイン）

データセット作成

Case：新生児の破傷風発症

- 時系列一貫性
  - 再感染による再発
  - 前回入院記録を保持
- 医学的整合性
  - 疾患と症候との対応

基本情報

- **患者名**: 藤田里奈
- **性別**: 女性
- **年齢**: 生後12日
- **記録者**: 菊地彩香 研修医
- **入院施設**: 八千代市大学病院
- **患者ID**: P79541
- **生年月日**: 2023年10月22日
- **診療科**: 新生児科
- **担当看護師**: 伊藤悠人 看護師
- **主治医資格**: 周産期専門医

入院管理情報

- **患者ID**: P79541
- **入院回数**: 初回入院
- **前回入院日**: 記録なし
- **入院日時**: 2023年11月3日 15:00

主訴

哺乳力低下、開口障害（軽度）、軽度の筋強直

診断

- **暫定診断**: 新生児破傷風
- **診断根拠**:
  臍部感染所見（発赤、浸出液）、痙攣、筋強直、哺
  血液検査結果に基づく。
- **鑑別診断**:
  - 新生児敗血症: 血液検査で白血球数が上昇し
    外可能。
  - 髄膜炎: 髄膜刺激徴候を現時点で認めず除外。

基本情報

- 患者氏名: 藤田里奈
- 性別: 女性
- 年齢: 8か月
- 記録者: 松本五郎 医局長
- 病院名: 足立区大学病院
- 患者ID: P79541
- 生年月日: 2023年10月22日
- 診療科: 新生児科
- 主治医: 松本五郎 医局長 (周産期専門医)
- 担当看護師: 池田竜也 看護師
- 入院回数: 2回目

入院管理情報

- 前回の入院日: 2023年11月3日 15:00
- 今回の入院日: 2024年6月19日 11:45

既往歴

- 発育・発達: 年齢相応
- 疾患歴: 前回、新生児破傷風の診断で2023年11月3日に入院（軽症で改善）
- 感染症歴: 特記事項なし
- 予防接種歴: 破傷風トキソイド未接種
- 手術・外傷歴: 出生時、臍帯切断後の特別な処置なし
- 輸血歴: 特記事項なし
- アレルギー歴: 薬物アレルギーなし

診断

- 新生児破傷風（再発）

診断根拠

- 臍部の発赤と浸出液、Clostridium tetani陽性の培養結果
- 筋強直、哺乳力低下の臨床症状
- 前回の疾患歴と一致した所見

破傷風（Tetanus）とは

厚生労働省 検疫所 FORTH

症状

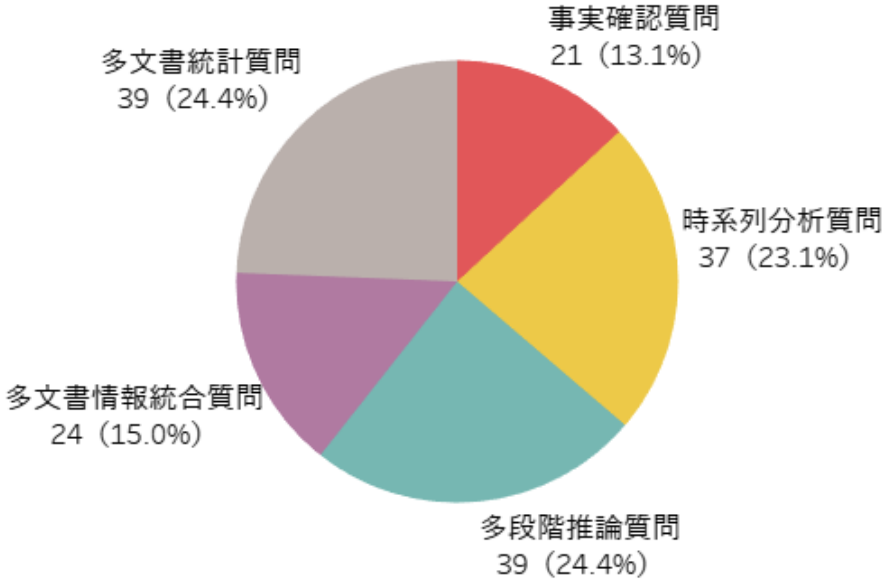
感染して3日から3週間からの症状のない期間があった後、口を開けにくい、首筋が張る、体が痛いなどの症状があらわれます。その後、体のしびれや痛みが体全体に広がり、全身を弓なりに反らせる姿勢や呼吸困難が現れたのちに死亡します。

参考：厚生労働省検疫所 FORTH：破傷風

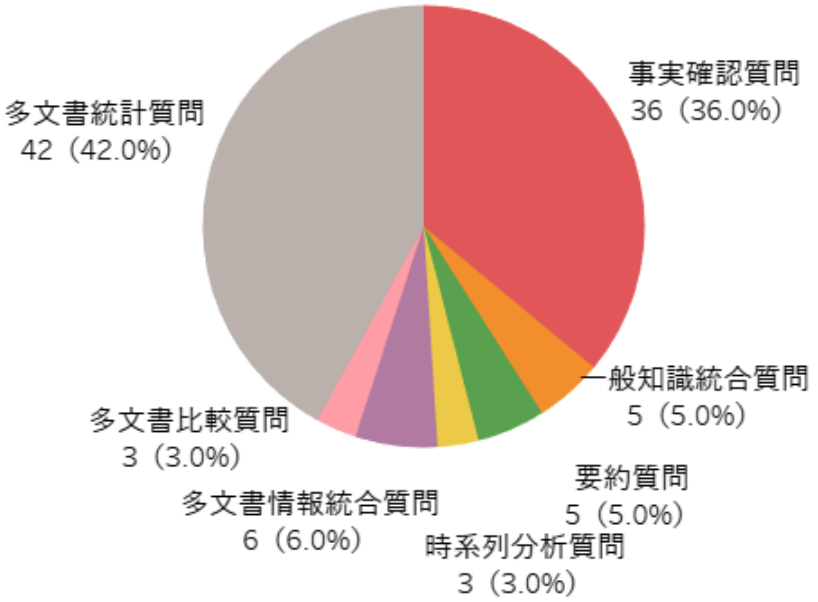
SynRAGデータセット（QRAデータセット）：質問タイプの分布

データセット作成

■ 財務ドメイン:



■ 医療ドメイン:

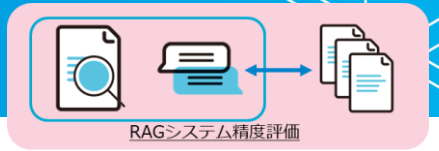


SynRAGデータセット（QRAデータセット）：質問タイプごとの例

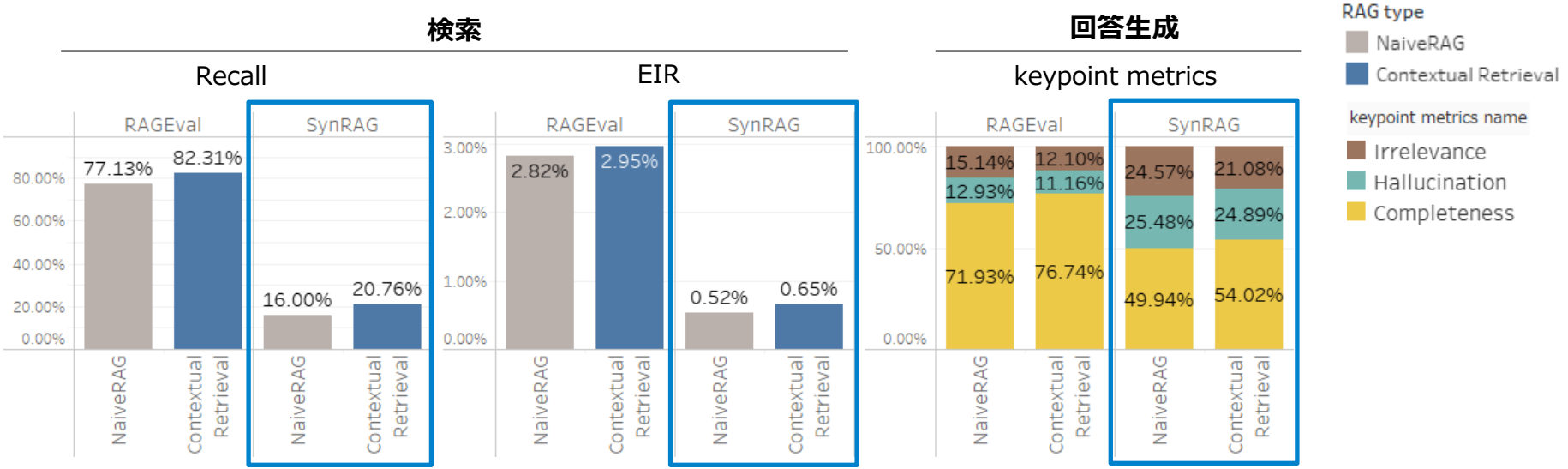
予め想定質問を作成しておくことで、実践的かつ網羅的な質問を作成

事実確認質問	日本陸運ネットワーク株式会社の設立年は何年ですか？
一般知識統合質問	患者ID P73022の痙攣発作及び開口障害の症状は、**破傷風の診断基準**と比較してどのように評価されますか？
時系列分析質問	• クローバー・デジタル株式会社の**2020年から2023年**における研究開発費（R&D費用）の変化はどのような**傾向**を示し、それが営業利益率や新製品開発にどのような影響を与えたのかを分析してください。 • 平野健二の重要な検査値の**推移パターン**を教えてください。
多段階推論質問	エコトランスポート株式会社の取締役会の構成とガバナンス体制の改善が、業績および株主還元に**どのように影響した**と考えられますか？
多文書比較質問	柴田佳子の各入院時の治療反応を**比較**するとどのような違いがあるか？
多文書統計質問	**住血吸虫症患者群**における肝脾腫の**出現頻度**はどのくらいですか？

精度評価結果：財務ドメイン

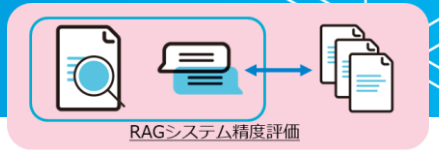


SynRAGでは文書長が大きく、検索難易度が高いことが結果に表れたと考えられる

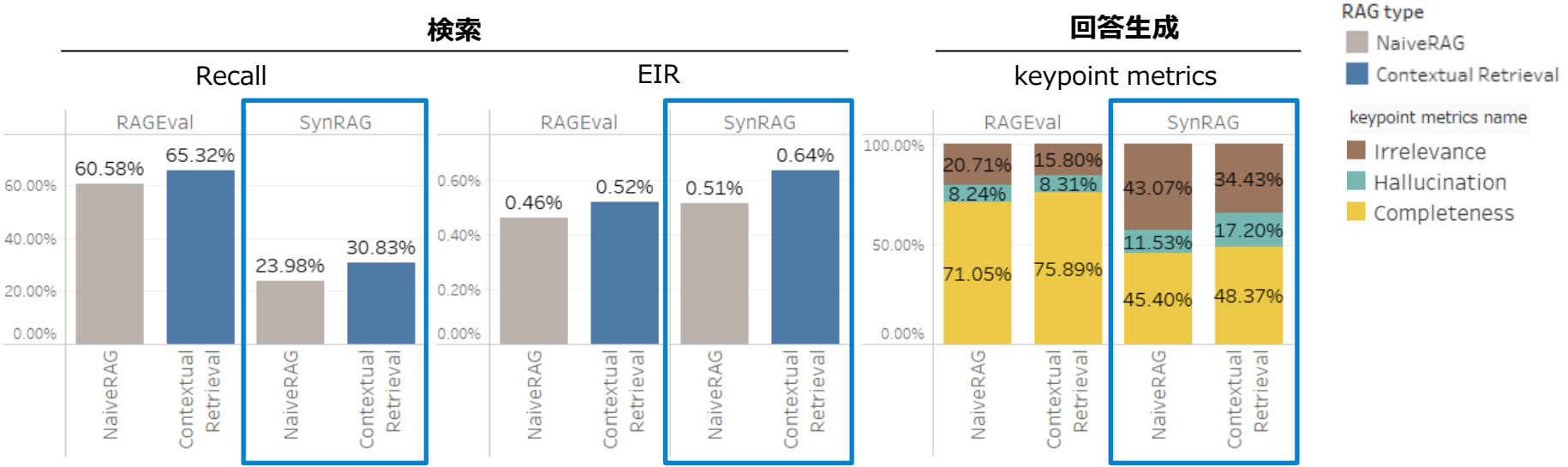


※ チャンクサイズは1,000、チャンクオーバーラップは0で固定。
チャンク数はそれぞれRAGEvalデータセット: 289, SynRAGデータセット：1,664
検索上位5件を取得

精度評価結果：医療ドメイン

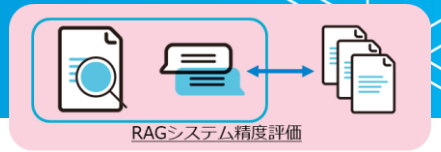


回答のIrrelevanceの割合が高い要因として、関連性が無く、内容が似たドキュメントが誤って検索されていることが考えられる



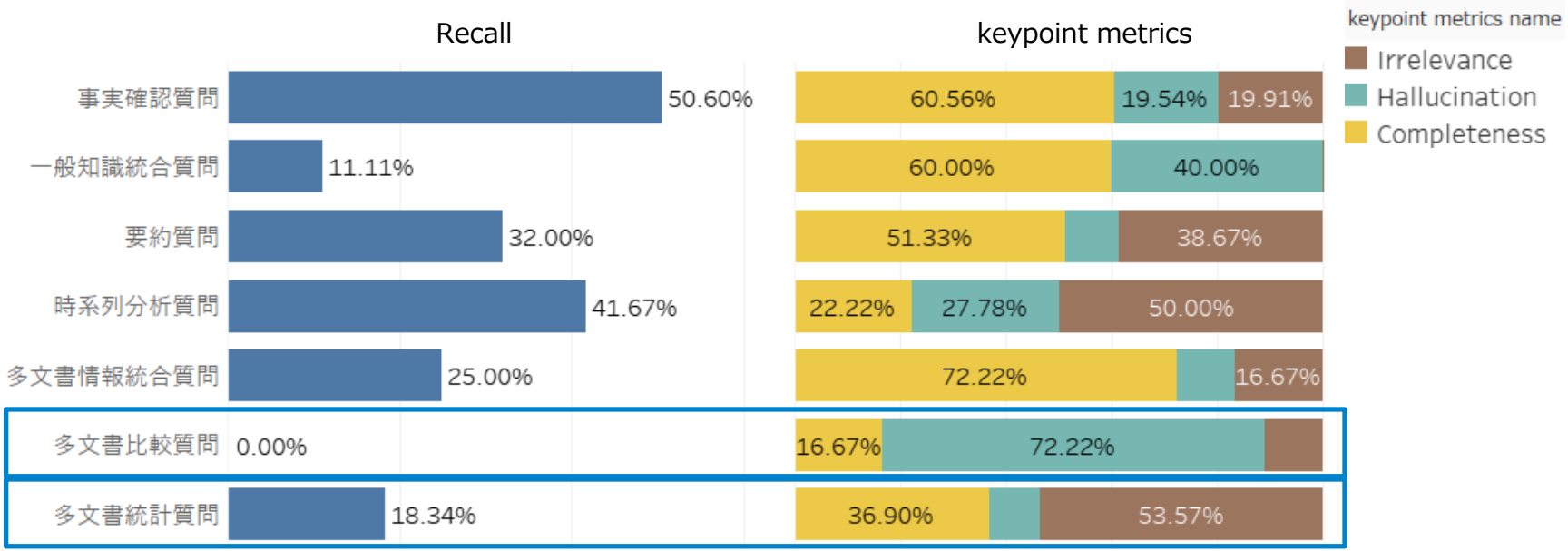
※ チャンクサイズは1,000、チャンクオーバーラップは0で固定。
チャンク数はそれぞれRAGEvalデータセット: 289, SynRAGデータセット：1,664
検索上位5件を取得

質問タイプ別精度評価結果：医療ドメイン

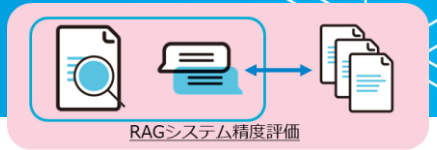


Cotextual Retrievalによる各質問タイプごとの精度指標は以下の通り

- 精度が特に低い質問タイプを調査し、質問の難易度を評価



データセットの難易度：多文書比較質問



専門知識の理解及び意味的解釈に基づいて、マルチホップな検索が要求される

- 高い検索精度が求められる質問

質問	高橋香織さん（患者ID: P69511）の各入院時の治療反応を比較してください。
参照文書数	3
正解参照	1. 適切な抗菌薬治療により7～10日以内の改善を期待。再発リスクは中程度と評価。 2. 14日程度での症状改善が期待される。 3. 患者はICUで管理され、酸素療法および抗菌薬治療を継続予定。
検索結果	対象の計3回の入院記録のチャンクは取得出来ているものの、正解参照となるチャンクを全て取得することは出来ていない。 → 治療反応の記述部分の検索には専門知識（「予後」）の理解が必要

データセットの難易度：多文書統計質問

生成モデルの制御能力、統合・集計性能が要求される

- RAGシステムの適切な判断能力が問われる質問

質問	尿路結石症の患者群における男性と女性の性別分布はどのようになっていますか？
参照文書数	7
正解参照	1. ‘性別: 男性’, 2. ‘性別: 男性’, … , 7. '性別: 男性'
正解要点	1. 尿路結石症の患者群では男性が6名。　2. 尿路結石症の患者群では女性が1名。
検索結果	尿路結石症の患者に関する検索はできているものの、対象者全てを検索できていない。
回答生成結果	「男性は女性に比べて尿路結石を発症するリスクが高いとされています。 具体的な性別分布は地域や研究によって異なることがありますが、一般的には...」 → ・検索結果ではなく、事前学習された一般知識に基づく回答がされる ・類似問題においては文脈から回答根拠を得ようとする

アジェンダ

1 背景・目的

2 先行研究

3 手法

4 実験結果・考察

5 まとめ

まとめと今後の展望

ドメインに特化・利用シーンに即したRAG評価データセットを構築・検証

- RAGEvalフレームワークを改善し、新たに**SynRAGデータセット**を作成
  - 医療ドメイン, 財務ドメイン
- データセットの質・検索／回答生成性能の評価・検証を実施
  - 文書：多様性・情報密度・整合性が改善されたことを確認
  - QRAデータセット：RAGシステムの性能を評価するのに十分な質問の種類・難易度

→ **特定のドメイン・利用シーンに即しただけでなく、**
RAGシステム性能評価用としても実用的なデータセット

今後の展望

- データセット・フレームワーク修正版の公開
- 他ドメイン（P&P等の社内文書, 製造, 小売,...）への適用
- データセットの追加評価：より高度なRAGシステム, 一般的なRAG評価指標

BIPROGY | Foresight in sight