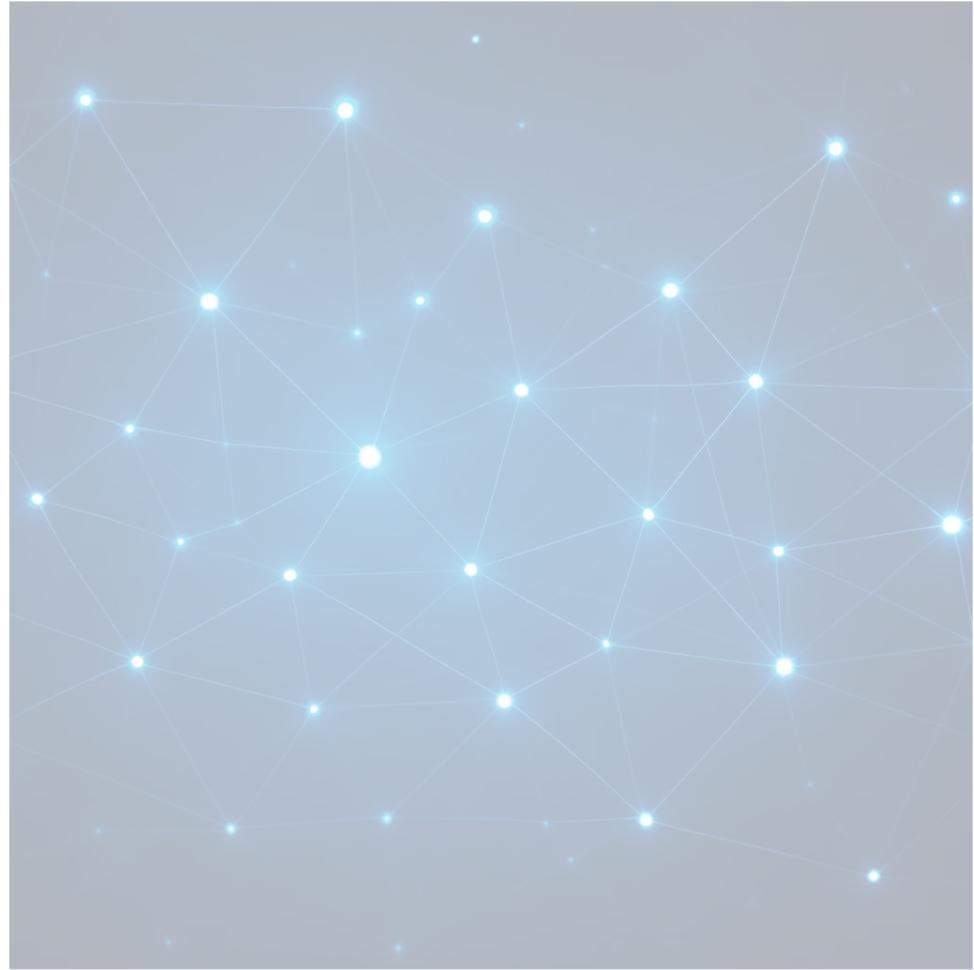


RAG として AIChatBot に 与えるデータベースの 構造に関する一考察

*On Database Structures for
Retrieval-Augmented Generation
in AI Chatbots*

神奈川工科大学
川越 航太



概要 (Overview)

- RAG とベクトル DB の概要
- データセット & ファイル形式
- 比較実験: 設定と結果
- 考察・結論と今後の課題

研究背景と目的

1 目的 (Goal)

- RAGベースのAIチャットボットを構築し、顧客の質問に自動回答してサポート業務を効率化する。
- 応答ログを活用した、新人オペレーター等の社内教育への活用。

2 背景 (Background)

- オペレーターと顧客のやり取り、ユーザマニュアルなどから自動生成したFAQを、LLMが正確にリトリーブするためにどうすればよいか。

3 データ提供元

- 日本のEdTech企業で職能団体・学会・自治体などの研修をオンライン/オフライン問わず一元管理できる「研修DXプラットフォーム」を提供。

RAGシステムのフロー



大規模言語モデルは外部知識ベースから関連情報を検索し、回答生成に活用する。

RAG (Retrieval-Augmented Generation) は、クエリに対してベクトルデータベースから類似文書を検索し、LLM へ渡して応答を生成する仕組み。

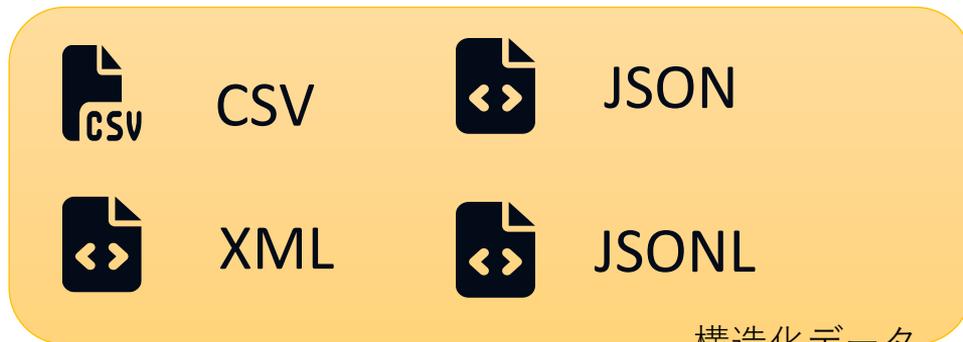
本研究では FAQ データセット (100件/1000件) を複数形式に変換し、VDB 化して検索精度を評価した。

データセットとファイル形式



DOCX
PDF
TXT

非構造化データ

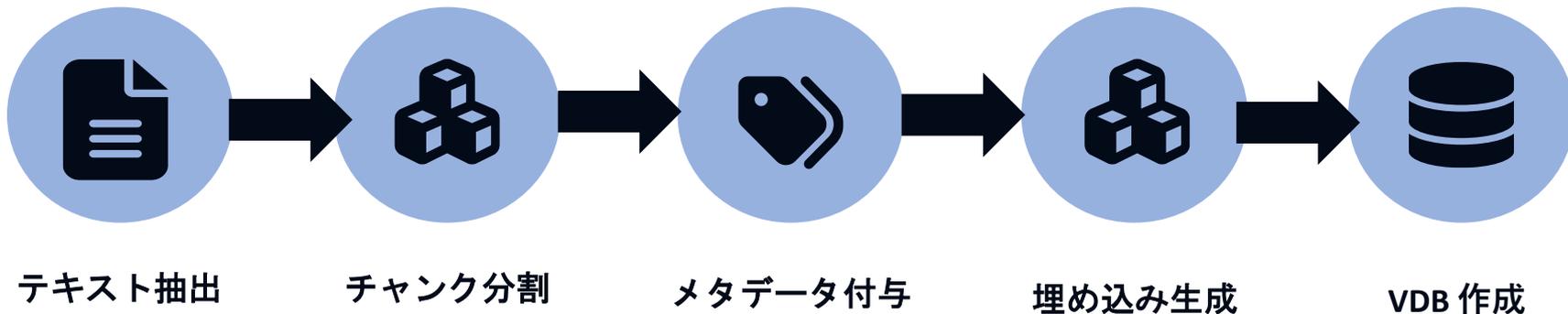


CSV
XML
JSON
JSONL

構造化データ

形式	抽出ライブラリ	分割単位	メタデータ
DOCX	python-docx	段落	ファイル名+段落番号
PDF	PyPDF2	ページ	ファイル名+ページ
TXT	open().read	全文	ファイル名
CSV	pandas.read_csv	行	ファイル名+行番号
XML	xml.etree	エントリ要素	ファイル名+エントリ番号
JSON	json.load	オブジェクト	ファイル名+エントリ番号
JSONL	json.loads	行	ファイル名+行番号

前処理パイプライン



- チャンク長: 1000 文字, オーバーラップ: 200 文字
- 埋め込みモデル: intfloat/multilingual-e5-large
- 使用ライブラリ:  LlamaIndex
- 言語モデル:  Gemini 2.0-flash

非構造化データの VDB化の例

● チャンク

● オーバーラップ

Question:無料トライアル期間が終了した後、自動的に有料プランへ移行され、課金が開始されますか？

Answer:いいえ、自動で有料プランへ移行されることはありません。トライアル期間終了後に引き続きご利用いただく時は、管理者設定画面から明示的にクレジットカード情報を登録し、プランのアップグレード手続きを行っていただく必要があります。

Question:経費精算の申請が承認者から差し戻された場合、元のデータを修正してそのまま再提出することはできますか？

Answer:はい、可能です。「申請一覧」の「差し戻し・却下」タブから該当の申請データを開き、指摘された箇所（金額や領収書の添付など）を修正した上で、再度「申請する」ボタンを押下してください。

Question:受講者がログインパスワードを忘れてしまったら、管理者がシステム側から代わりに再設定することは可能ですか？

Answer:はい、可能です。管理者画面の「ユーザー管理」から対象の受講者を選択し、「パスワードのリセット」を実行してください。受講者の登録メールアドレス宛に、パスワード再設定用の専用URLが自動送信されます。

構造化データの VDB化の例

● 質問

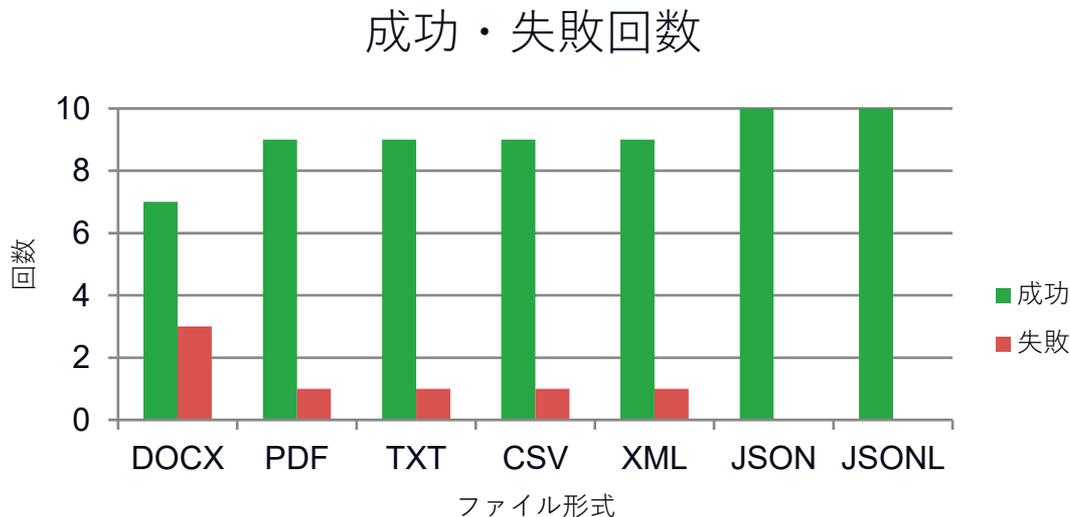
● 回答

```
[
  {
    "question": "無料トライアル期間が終了した後、自動的に有料プランへ移行され、課金が始まりますか？",
    "answer": "いいえ、自動で有料プランへ移行されることはありません。トライアル期間終了後に引き続きご利用いただく時は、管理者設定画面から明示的にクレジットカード情報を登録し、プランのアップグレード手続きを行っていただく必要があります。"
  },
  {
    "question": "経費精算の申請が承認者から差し戻されたら、元のデータを修正してそのまま再提出することはできますか？",
    "answer": "はい、可能です。「申請一覧」の「差し戻し・却下」タブから該当の申請データを開き、指摘された箇所（金額や領収書の添付など）を修正した上で、再度「申請する」ボタンを押下してください。"
  }
]
```


実験設定① 100 問に対する10問の質問

1	管理者側から受講者を申し込んだ場合、受講者には申込み完了のメールが送信されますか？
2	クレジットカード払いでキャンセルした場合、返金はいつ頃になりますか？
3	代表者アカウントのメンバー画面で「メンバー情報変更」と「アカウントの削除」を制限するにはどうすれば良いですか？
4	修了証はどのようにダウンロードできますか？
5	コースの言語設定を変更できますか？
6	受講者の進捗状況を一括で更新できますか？
7	システムメンテナンス情報はどこで確認できますか？
8	講師の都合で講義の日程や順序を変更する方法を教えてください。
9	受講料の一部を振り込んだ後、残りの金額を支払う方法を教えてください。
10	振込コードを入力しなかった場合のリスクは何ですか？

100 問実験: 成功 / 失敗回数



- JSON / JSONL では、全問正解 (10/10)。
- DOCX は 70% (7/10) と低調。
- その他の形式は 90% (9/10) の精度を示した。

考察 (100問)

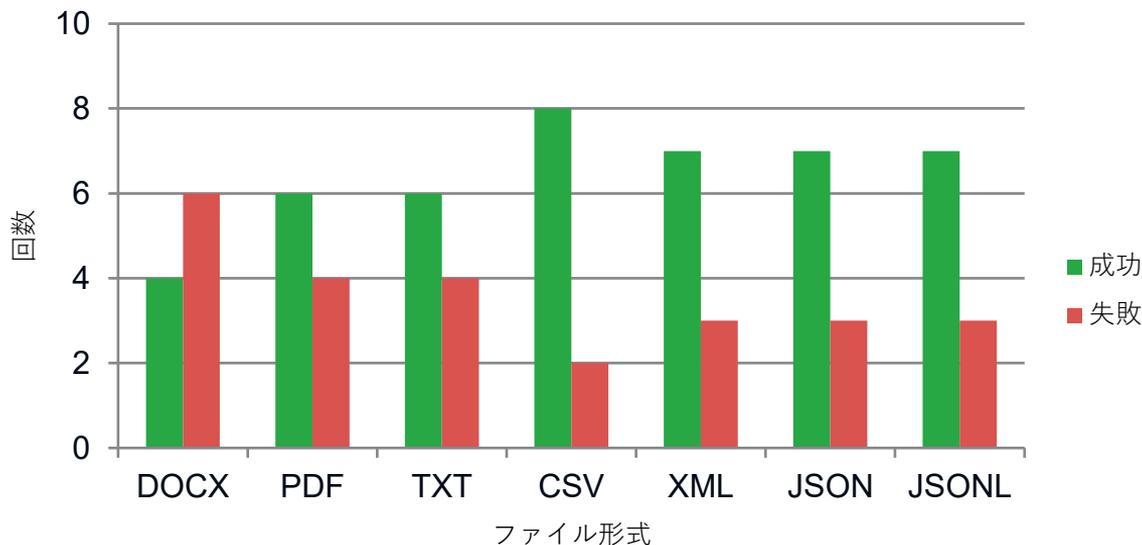
- DOCX 形式は段落単位の分割が適切に働かず、レイアウトやネスト構造の影響で正答率が低下した。
- JSON / JSONL はオブジェクト/行単位に分割されたので、構造化度の高さが有利に働いた。
- PDF ・ TXT ・ CSV ・ XML は 90% と高精度。シンプルな構造が検索性能を上げた。

実験設定② 1000 問に対する10問の質問

1	レッスンに紐づく課題の回答ボタンが表示されません。どの受講状況で回答可能になりますか？
2	施設側から除名すると、管理画面上ではどうなりますか？
3	ログインセッションのタイムアウト時間を変更できますか？
4	オーダーIDのない支払いカードが生成されることがありますが、どういう場合に起こりますか？
5	二段階認証を有効化する方法は？
6	アカウントを削除（退会）したい場合は？
7	ユーザーのログイン履歴を確認できますか？
8	アップロードファイルのサイズ制限は？
9	イベント発生時のWebhook通知は設定できますか？
10	データ保持ポリシー（GDPR対応）の設定方法は？

1000 問実験: 成功 / 失敗回数

成功・失敗回数 (1000 問)



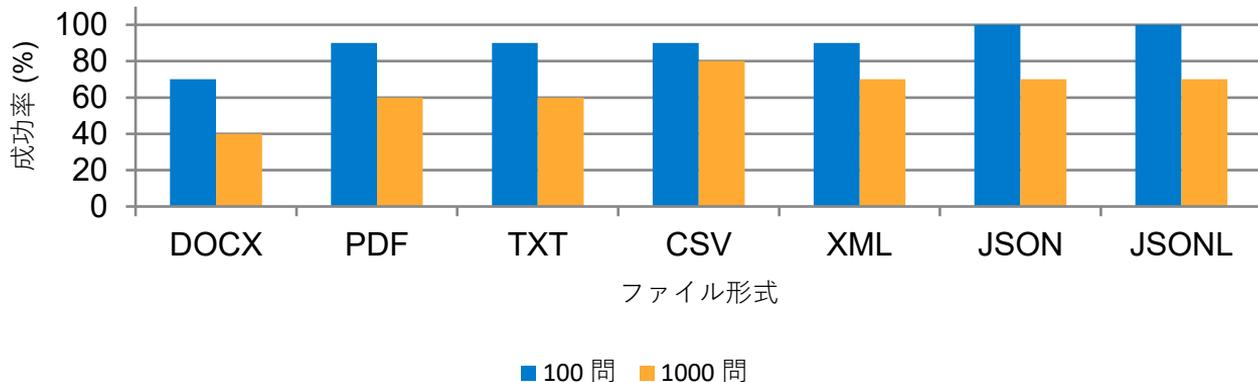
- ・ CSV が 8/10 で最高成績を維持。
- ・ 一方、JSON 系は 7/10、DOCX は 4/10 と大きく低下した。

考察 (1000問)

- データの増加に伴い、検索成功率が低下。
- 多くなったデータを3つのファイルに分割して保存したが、100問での実験の3倍以上の問題数を1つのファイルにまとめた。100問ずつに分割すれば、検索精度は高まった可能性が残る。
- ファイルのカテゴリでの分割手法を用いなくても、CSV,JSONなどの構造化データは一定の精度を維持できたので、**構造化ファイルはQAシステムのRAGにとって相性が良い。**

実験比較

成功率比較



- JSON 系は 100 問では 100% だったが 1000 問では 70% に低下。
- CSV は安定して 80% を維持し、最も堅牢であることが示された
- DOCX は大規模化で 70%→40% と急落し、**レイアウト依存性の脆弱さ**が顕在化した。

形式別の検討



レイアウトや段組み、構造化データではないので、チャンク化が難しく、大規模時に成功率が低下。



全文を単一の文章として扱うため文脈が保持されるが、メタデータが少なくバリエーションに弱い。



PyPDF2による改行検出が不十分で、空白・タブの揺れが検索を阻害する可能性がある。



オブジェクト単位に分割でき、データ構造がしっかりと固定されていてVDB化しやすい。



レコード単位での分割が自然で、構造が単純なため大規模でも高精度を維持。



オブジェクト単位でデータを分割して管理できる。統一されたデータ構造により、安定した解析や変換が可能。

ファイル形式の違いによる失敗例と成功例

ハルシネーションについて：1000問程度では誤った回答を生成することはなく、検索が失敗した場合には、「分からない」と出力した。

課題をダウンロードすると、受講者が作成したファイル名ではなく、別の名前でダウンロードされます。どこで設定を変更すれば、アップロードされたファイル名でダウンロードできますか？

課題のダウンロードに関する設定変更方法についての情報は見つかりませんでした。【使用したファイル】
faq_output1.docx (page: 不明),
faq_output3.docx (page: 不明)

DOCXでの失敗例

課題をダウンロードすると、受講者が作成したファイル名ではなく、別の名前でダウンロードされます。どこで設定を変更すれば、アップロードされたファイル名でダウンロードできますか？

現在の仕様では、ダウンロード時にランダムなファイル名が割り当てられるため、設定を変更してアップロードされたファイル名でダウンロードすることはできません。仕様改善が進められていますので、改善をお待ちください。【使用したファイル】
faq_output1.csv (page: 不明),
faq_output3.csv (page: 不明)

CSVでの成功例

検索精度に差が生じた原因

質問部分は両方とも検索に成功しているが、DOCX形式では**質問と回答の紐付けが出来ないケース**が散見され、これが原因となって回答を探すことに失敗している。

```
qasystem-1 | > [Node f60ca1df-e11f-4f47-acab-aa55e470fdfe] [Similarity score: 0.848401] Answer: ファイルを右クリックして [共有リンクを作成] を選び、リンク設定画面で「パスワード保護」を有効にし、任意のパスワードを入力してください。リンクを受け取った相手は入力が必要になります。
qasystem-1 | > [Node 81b13df9-e0ca-4ea0-b853-0d80526d44b3] [Similarity score: 0.843327] Answer: マイページの [講師一覧] から新しく受講したい講師を選び、希望日時でレッスン予約をしてください。既存の予約がある場合はキャンセル料なしで変更が可能です。
qasystem-1 | DEBUG:llama_index.core.indices.utils:> Top 2 nodes:
qasystem-1 | > [Node 55d8b55b-6840-410a-a563-180b86638810] [Similarity score: 0.963113] Question: 課題をダウンロードすると、受講者が作成したファイル名ではなく、別の名前でダウンロードされます。どこで設定を変更すれば、アップロードされたファイル名でダウンロードできますか？
qasystem-1 | > [Node 55d8b55b-6840-410a-a563-180b86638810] [Similarity score: 0.868878] Answer: 課題は、受講者のステータスが「受講予定」に変更された時点で表示されます。支払いが完了していない場合、受講者のステータスが「受講予定」にならないため、課題は表示されません。
```

DOCXでのログ

```
qasystem-1 | > [Node a73996a9-29f2-447e-8208-5da874161820] [Similarity score: 0.833109] Q: オンラインストレージで共有リンクにパスワードを設定するには？
qasystem-1 | A: ファイルを右クリックして [共有リンクを作成] を選び、リンク設定画面で「パスワード保護」を有効にし、任意のパスワードを入力...
qasystem-1 | > [Node 09502534-2e26-4f1f-8017-bb8c9316cbbb] [Similarity score: 0.82888] Q: オンライン英会話スクールで講師の変更をしたい場合は？
qasystem-1 | A: マイページの [講師一覧] から新しく受講したい講師を選び、希望日時でレッスン予約をしてください。既存の予約がある場合はキャンセル料...
qasystem-1 | > [Node b4de197f-13cb-4c88-9f6d-ef756dfd2b79] [Similarity score: 0.954508] Q: 課題をダウンロードすると、受講者が作成したファイル名ではなく、別の名前でダウンロードされます。どこで設定を変更すれば、アップロードされたファイル名でダウンロードできますか？
qasystem-1 | A: 現在の仕...
qasystem-1 | > [Node b4de197f-13cb-4c88-9f6d-ef756dfd2b79] [Similarity score: 0.871595] Q: 課題はどのタイミングで利用者の提出物に表示されますか？
qasystem-1 | A: 課題は、受講者のステータスが「受講予定」に変更された時点で表示されます。支払いが完了していない場合、受講者のステータスが「受講...
```

CSVでのログ

その他の失敗例

申込者の人数制限を設定できますか？

この情報だけでは、申込者の人数制限を設定できるかどうかは判断できません。【使用したファイル】

faq_output3.json (page: 不明),

faq_output2.json (page: 不明)

失敗例

その他の原因

失敗のパターン：クエリの内容に近い単語が類似度が高くなりピックアップされるが、関係のないデータである場合が多い。

申込者の人数制限を設定できますか？



```
qasystem-1 | > [Node 40950d08-f04a-407f-b438-4198ef79a771] [Similarity: 0.85]
旅行の参加人数を当初登録から変更するにはどうすればよいですか？
qasystem-1 | A: 旅行出発日の30日前まではマイページの「団体管理」からキャンセル料が発生...
qasystem-1 | > [Node 20485a20-f59d-47d5-bd43-6c2cc5375520] [Similarity: 0.85]
・学園向け認証システムでシングルサインオンを設定する方法は？
qasystem-1 | A: 管理コンソールの「認証設定」で「SAML2.0有効化」を
ください。そ...
qasystem-1 | DEBUG:llama_index.core.indices.utils:> Top 2 nodes:
qasystem-1 | > [Node 6813cd0e-9930-4a75-89d0-b9d91c63f13f] [Similarity: 0.85]
ント発生時のWebhook通知は設定できますか？
qasystem-1 | A: 管理画面の「システム設定」→「Webhook管理」で、対象
と通知先URLを指定すると、P...
```

運用ベストプラクティス & 結論

- DOCX, PDFなどから構造化形式（CSV、JSON等）へ統一
- VDB化する前に、元ファイルのクレンジング & スキーマ自動検出
- 構造化データにできない場合には、適切なチャンク長/オーバーラップを設計
- ファイル名などのメタデータの付与で検索文脈保持
- 単純なファイル分割ではなく、QAのカテゴリごとにファイルやディレクトリを分割することで、誤検出を減らす
- 大規模での運用は「自動前処理 + 構造化」で実装

限界と今後の課題

- データセット規模が限定的 (1000 FAQ) であり、一般化にはさらなる検証が必要。
- 埋め込みモデルや LLM の種類を変えた場合の評価は未実施。
- 固定チャンク長のため動的なチャンクサイズ調整の効果を検証できていない。
- 今後はクレンジング・スキーマ検出の自動チューニングやエラー検出機構の導入を検討。

Thank you!

ご清聴ありがとうございました

Questions?

