注意制限事項

本書は本サービスのAPIおよび各種WebUIの注意事項・制限事項を記載します。

注意事項

API

- 履歴付き対話のAPIでは、同じ履歴に対して複数人で同時に会話をした場合、他の人の会話も履歴に含まれないため会話の内容に不整合が生じる可能性があります。履歴付き対話は1つの会話履歴に対して複数人で共有しないことを推奨します。
- ユーザテンプレートのユーザプロンプトは必ず何かの文字列を記載してください。未記載の場合、対話API実行時にテンプレートが見つかりません。
- 拡張対話（Webコンテンツ）機能では使用可能なURLの文字列長上限は4096文字までとします。4096文字を超えるURLの送信はご遠慮ください。
- 拡張対話（Webコンテンツ）機能ではURLがリファレンスマニュアルに記載された条件を満たしているにもかかわらず、エラーが出力される場合、参照先のサーバーにWebスクレイピングを禁止する設定がされている可能性があります。
- 拡張対話（Webコンテンツ）機能ではURLにHTML、テキストファイル、jsonファイルが指定可能ですが、参照先のWebサーバーやサービスによっては処理がうまくいかない可能性があります。読み込み可能なURLでエラーが発生した場合はブラウザで表示可能かどうかを確認してください。HTML、テキストファイル、jsonファイル以外を参照するURLは動作対象外となります。
- インデックスへの文書登録機能では対応しているPDFであっても、PDF内のUnicodeマッピング情報が不足している場合、回答が正常に動作しないことがあります。
- 検索対話機能では一度に大量のリクエストが発生すると、一時的に500エラーを返却することがあります。その場合は時間をおいて再度実行してください。
- 拡張対話で指定できるファイル、Webコンテンツには条件があります。詳細は拡張対話チュートリアルを参照してください。本注意事項はAPIだけでなくチャット画面でも同様になります。
- 回答根拠確認APIの注意事項は「チャット画面およびチャット部品」の注意事項をご確認ください。
- 外部LLMを用いた対話機能の応答性能はサーバのネットワーク環境、外部LLMモデルのクォータ制限などに依存します。また負荷が高い場合は一時的にエラーが発生する可能性があります。

チャット画面およびチャット部品

- 回答根拠の確認機能（推論を用いた回答根拠の確認機能も同様）では、回答文章に対し、「。」などで区切られた一文単位で回答根拠を確認します。そのため、箇条書きなどの記載によっては複数の記載が1つの文として認識される可能性があります。
- 回答根拠の確認機能（推論を用いた回答根拠の確認機能も同様）では、同時に大量のリクエストがあった場合、処理に時間がかかることがあります。
  - 特に、推論を用いた回答根拠の確認は推論時間が長くなるため、多重度に応じて処理時間が大幅に増えることがあります。
- 推論を用いた回答根拠の確認機能の応答性能はサーバのネットワーク環境、Azure OpenAI Serviceのクォータ制限などに依存します。
- 推論を用いた回答根拠の確認機能では回答および参照文書が判定に適さないものであった場合、推論に失敗しエラーになることがあります。
  - 特に、文章にならない文字の羅列、1文が非常に長い文章、不自然に制御文字を含む文章などの場合はエラーの原因となります。
- 推論を用いない回答根拠の確認機能では、回答、参照文書それぞれの1文の長さが512～1024文字を超過した場合、上限超過で処理が失敗します（判定上限の文字数は文章によって異なります）。エラーとなった場合は推論を用いた回答根拠の確認機能を使用することを推奨します。
- Webコンテンツの内容を対象とした会話は拡張対話（Webコンテンツ）機能の仕様に準拠します。詳しくは「拡張対話チュートリアル」の「拡張対話（Webコンテンツ）機能仕様」をご確認ください。
- Web検索した内容を対象とした会話は拡張対話（Web検索）機能の仕様に準拠します。詳しくは「拡張対話チュートリアル」の「拡張対話（Web検索）機能仕様」をご確認ください。

管理ポータル

- ドキュメント管理画面において、合計20MB以上になる複数のファイルの登録を行うと、タイムアウトが発生します。
  このタイムアウトは、管理ポータル上で発生しているものであり、文書登録はバックエンドで引き続き行われています。1MBあたり2分を目安に十分な時間をおいていただき、登録文書一覧に目的の文書が表示されているかご確認ください。
  また、RAG文書登録プログラム(script_addDocuments.py)を用いることでタイムアウトのエラーを回避できます。

RAG文書登録操作コマンド、インデックス管理

- 登録できる文書の種別は「PDF、Excel、PowerPoint、 Word、CSV、Text、Markdown（拡張子：pdf, xls, xlsx, pptx, docx, csv, txt, md)」となります。これ以外の拡張子のファイルは登録することが出来ません。

- 対応文字コードは.txtと.csvファイルについてはShift-JIS,JIS, UTF-8(BOMあり),UTF-8(BOMなし) UTF-16(BE),UTF-16(LE), EUC-JPを対応しており、.mdファイルについてはUTF-8(BOMあり),UTF-8(BOMなし) のみの対応となっております。
- 複数のユーザが同一インデックスかつ同一ファイル名で異なる内容のファイルに対して同時刻に文書登録を実行した場合、文書登録が失敗する場合や上書き処理で削除されるはずの内容が一部登録されたままになってしまう可能性があります。その場合は再度時間を置いて登録していただくようお願いいたします。
- API実行時のタイムアウトは1時間であり、1時間を超える文書登録はタイムアウトエラーで失敗する可能性があります。大容量や複数のファイルが登録できなかった場合は分割し、複数回に分けて登録してください。文書登録時にタイムアウトが発生したかどうかは、登録開始から1時間以上経過したのち、登録文書一覧画面に登録したファイルが表示されているかどうかで確認してください。
- 拡張子がtxtやcsvの文書を登録する際、内部で文字コードの特定ができず登録エラーになる場合があります。その際は文字コードが判別しやすい形式（例：UTF-8 BOM付き）で該当ファイルを保存し直し、改めて登録してください。
- V2.1.2でインデックスに登録したPDFファイルのページ番号が1ずれる不具合を修正しております。ただしV2.1.2以前に登録済みのPDFファイルは、アップデート後もページ番号がずれたままになっています。ページ番号を正常化させる場合は、V2.1.2以降の環境で当該PDFファイルを再登録(上書き登録)してください。また、PDF以外の文書については問題は発生しないため、再登録は不要になります。
- 文書を自動的にチャンク分割したうえで埋め込み処理を行いますが、文書構造によっては、1チャンクあたりのテキスト量が埋め込みモデルのトークン上限を超過し、エラーとなる場合があります。その際は、マニュアル記載のRAG文書登録コマンドの `split_chunk_size` パラメータを小さい値に調整したうえで再度登録をお試しください。
- 文書登録機能では、ファイル内に含まれる「文字情報」のみを読み込み・解析対象とし、埋め込み画像・スキャン画像・図表画像内の文字情報を直接認識することはできません。画像内の文字情報も検索対象としたい場合は、図表文脈理解機能を使用して文字情報に変換してから、ご登録いただくようお願いいたします。
- V2.2.3でmdファイルのパース方式を修正し、Markdown記法を保持するようになりました。そのため、V2.2.3以前に登録済みのmdファイルについては、精度向上の観点から再登録を推奨します。

LLM

- cotomi v3では長文のリクエストを処理している間は他にリクエストの応答が遅くなることがあります。
- LLMは入力文字数が多くなるほど推論時間が長くなるため、応答が返却されるまでの時間が長くなります。

- LLMに対し同じ質問を繰り返した場合、使用するOSSのキャッシュ機能が働き2回目以降の応答が速くなることがあります。キャッシュ機能は基本的に無効にできません。速度検証などで一時的に無効にしたい場合はお問い合わせ窓口まで相談してください。

認証基盤

- カスタム認証のIdP連携(Active Directory)認証を利用する場合、連携するActive Directory 上のユーザには、「ユーザログオン名」と「ユーザログオン名(Windows2000より前)」の両方の名前が設定されている必要があります。
  - 「ユーザログオン名」と「ユーザログオン名(Windows2000より前)」が異なる場合、本サービスのログイン時には「ユーザログオン名(Windows2000より前)」を利用します。
  - 「ユーザログオン名」が設定されていないユーザは本サービスを利用できません。Active Directory で該当ユーザの「ユーザログオン名」を設定してください。「ユーザログオン名」の変更が本サービスに反映され利用可能になるまでに最大24時間かかります。
- カスタム認証のIdP連携(Active Directory)認証を利用する場合、連携できるユーザは Active Directory の 1つのコンテナ、あるいは、1つのOU(組織単位)のユーザです。複数のコンテナや複数のOUのユーザとの連携は出来ません。
- カスタム認証のIdP連携(Active Directory)認証利用時、「ユーザーログオン名」「ユーザーログオン名(Windows 2000より前)」に 半角英数字以外の文字（日本語など）を含む Active Directoryのユーザは Generative AI FW をご利用いただけません。
- カスタム認証のIdP連携(EntraID)認証を利用する場合、ユーザの表示名は128文字より短い文字列を設定してください。長すぎた場合チャット画面、管理ポータル画面にログインできません。
- カスタム認証のIdP連携(Active Directory)認証利用時、「<ユーザーログオン名>@<ドメイン名>」の形式の文字列は255文字より短い文字列を設定してください。長すぎた場合チャット画面、管理ポータル画面にログインできません。

制限事項

API

- 会話で保持できる対話の履歴の最大数は100です。上限を超えた場合は古い対話から回答に含まれなくなるため会話の内容が意図しない内容になる可能性があります。
- 会話履歴は最大数に達しなくてもLLMのトークン数の上限などにより、履歴付き対話を継続できなくなることがあります。その場合は新規に会話を開始してください。
- 会話履歴一覧APIで取得できる会話および対話の最大数はそれぞれ100です。取得数の上限を超えた場合は古いものから取得できなくなります。会話履歴は一覧から取得できないものでも履歴ID(historyId)を指定することで、履歴付き対話APIから利用できます。

チャット画面およびチャット部品

- 特にありません。

管理ポータル

- テンプレート編集およびテンプレート引用の画面において、F5キー押下または当該ページのURLの直接指定を行った場合、「template id is not specified」のエラーが発生し、入力中の情報が消失します。
  再度、テンプレート一覧画面に戻り、テンプレート編集またはテンプレート引用の画面を開いてください。
- カスタム認証利用時は、ログインユーザの削除や役割変更のガードが行われません。

監査ログ

- カスタム認証利用時の監査ログでは、監査ログの種類によってユーザを特定するための情報が異なります。
  - 対話履歴ログ：ユーザを特定するための情報として、ユーザIDとメールアドレスがログに含まれます
  - 管理ポータル画面操作ログ：ユーザを特定するための情報として、メールアドレスがログに含まれます
  - アクセスログ：ユーザを特定するための情報として、ユーザIDがログに含まれます

RAG文書登録操作コマンド、インデックス管理

- 特にありません。

認証基盤

- IdP連携時のユーザ削除の為には、連携先の IdP だけでなく、Generative AI FW からのユーザ削除操作も必要となります。

Generative AI FW

- 削除された会話履歴などMongoDBデータベース内の不要なディスク容量は「会話履歴の最大保存期間を変更したい」に記載する「MONGO_CAPPED_SIZE」の設定だけでは削除されません。削除するための手順については別途お問い合わせください。
- 「MONGO_CAPPED_SIZE」について、古い会話履歴の削除機能はリアルタイムでは行われないため、ディスク容量の空き容量に十分余裕がある数値を設定してください。特に長文での対話はディスク容量を大きく使用する可能性があります。
- 監査ログのバックアップ・リストアについて、監査ログのリストア時に実行日が異なるバックアップデータを使用した場合、リストア日の監査ログに「バックアップ実行日の監査ログ」と「リストア実行日の監査ログ」が混ざって出力されますのでご注意ください。

- ディスク容量に十分な空きがない場合、インデックスに関わる操作(管理ポータル画面、API、バックアップ・リストアなど）が正常に動作しない可能性があります。opt配下のディスク容量の使用率が85%を超えないようにしてください。特にバックアップ・リストアはディスク容量に十分に空きがあることを確認の上実施してください。
- MongoDBの会話履歴は保存容量が大きくなるについて、会話履歴の取得かかる時間が長くなります。そのためチャット画面のログイン時などで会話履歴を取得する際に時間がかかる可能性があります。
- セットアップ完了後にOSの時間は変更しないでください。変更した場合、正常にサービスが起動できなくなる可能性があります。