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

- 本サービスが提供する各種サービス(LLM、履歴付き対話チャット、テンプレート対話チャット)は240秒でタイムアウトとなります。
- チューニング済モデル利用の契約数変更時にはデプロイ済みモデルの数を変更後の契約数以下にする必要があります。モデル管理APIで変更後の契約数を超えるモデルのアンデプロイを行ってください。なお、全テナント共通で提供されるNEC cotomiのモデルはデプロイ済みモデルの数には含まれません。
- チューニング済モデルのアップロード用URLとトークンの有効期限は1時間です。有効期限が切れた場合は、モデルリポジトリを削除した後、アップロード用URLとトークンを取得しなおしてください。
- チューニング済モデルのアップロードで5,000MiBを超えるサイズのファイルをアップロードする場合、ファイルを4,000MiB以下に分割してアップロードする必要があります。
- チューニング済モデルのアップロードのタイムアウトは1MBあたり10分です。1MBあたり10分を超える時間がかかるとタイムアウトします。
- 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。

- 拡張対話（ファイル添付）機能や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。
- 拡張対話（ファイル添付）機能では一度に多数の画像入力リクエストを実行すると、対話が履歴に保存されないことがあります。その場合は時間をおいて再度実行してください。
- 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。

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。
- 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、処理時間が240秒を超えた場合、エラーになります。
  - 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、エラーの原因となります。
- 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。
- 回答根拠の確認機能（推論なし）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。

管理ポータル

- サインアップを行う際のメールアドレスに、管理ポータルから登録したメールアドレスと大文字/小文字のみが異なる値を指定した場合、対象のユーザにおける管理ポータル上の操作でエラーが発生することがあります。
  本エラーが発生した場合は、管理ポータル画面より対象のユーザを全てのテナントから一度削除した上で再登録を行い、登録したメールアドレスと大文字/小文字まで一致するメールアドレスでサインアップするようお願いいたします。

RAG文書登録操作コマンド、インデックス管理

- 登録できる文書の最大ファイルサイズの諸元は16MBとなります。（ファイル内のテキストサイズは1MBを上限とします、ただし xls,xlsx については 500KB を上限とします。）
- 登録時間が230秒を超過するとタイムアウトが発生して登録が失敗する場合があります。大きなサイズのテキストが含まれるファイルは、この制限に抵触する可能性がありますのでご注意ください。このような場合には、文章を分割して登録する、もしくは非同期の文書登録コマンドを使用することを推奨します。詳細はベクトルDB管理APIの非同期実行についての記載をご覧ください。
- 登録できる文書の種別は「PDF、Excel、PowerPoint、 Word、CSV、Text、Markdown（拡張子：pdf, xls, xlsx, ppt, pptx, doc, 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。
- 管理ポータルの利用状況で確認できるインデックス使用量の合計には、インデックスの管理データ(数百byte程度)が含まれています。そのため、文書登録後に全文書を削除しても、使用量が完全に0にならない点にご注意ください。
- 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。
- 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。
- 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。
- インデックス使用量取得API（GET /document/size/{index_name}）が返却する使用量は、内部計算による概算値です。チャンク内容（トークン数）の偏り等により、実際の使用量と乖離する場合があります。容量逼迫の判断は、管理ポータルの「利用状況」に表示される全体の使用容量（正確値）を基準にしてください。
- 非同期文書削除（POST /document/deleteDocumentAsync）の処理中に、リクエスト時のメタデータフィルタ条件に一致する文書を登録すると、削除後の残存件数として扱われ、結果情

報の失敗件数に計上される場合があります。再登録は削除完了後に実施してください。
- 非同期文書削除API（POST /document/deleteDocumentAsync）では、metadataAnyFiltersを指定しない場合、削除対象が絞り込まれず、対象インデックスに登録されている文書がすべて削除されます。全件削除が目的でない場合は、metadataAnyFiltersを必ず指定し、削除対象を明示したうえで実行してください。

AIガードレール

- AI ガードレールはリスクの低減を目的とした機能で、すべての不適切なコンテンツを完全に防止できるものではありません。入力や出力内容は利用者で問題ないか確認が必要です。
- AIガードレールを有効にした場合、入出力のチェックを行うためLLMからの応答が返却されるまでの時間が長くなります。
- AIガードレール機能を有効にした場合、対話機能でのstream応答には対応しておりません。そのためチャット画面や推論APIでLLMからの回答がまとまって表示されるようになります。推論APIのstreamオプションを指定しても無効になります。
- AIガードレールの検出には誤検出が発生する可能性があります。誤検出をなくすことはできませんのでご留意ください。
- AIガードレールにブロックされた場合には、当該の会話は会話履歴に保存されないのでご留意ください。

図表文脈理解

- 結果テキストに同じ文字が繰り返し出力される事象が発生する場合があります。
- 結果テキストに記載と関係のない内容が出力される事象が発生する場合があります。

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

- 管理ポータルのAPIコール数画面においては、「multilingual-e5-large」モデルの利用に関するコール数が集計対象となりません。
- ユーザまたはグループが多数のインデックスに紐付いている場合に、ユーザおよびグループの一括削除に失敗することがあります。この際、一括削除対象の一部のユーザ・グループの削除が不能になることがあります。本エラーが発生した場合は、お手数ですがサポート窓口までご連絡ください。
  なお本エラーの抑止のため、一度に削除するユーザおよびグループに紐付くインデックスの合計は1000件程度までを推奨します。
- ユーザおよびグループの一括追加・一括削除の操作を行った場合に進捗バーが表示されますが、途中の進捗状況は表示されません。

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。

Agentic AI for Search

- インデックス名「index-usage」は予約語となります。Agentic AI for Search のベクトル検索Toolを利用する為の特別なインデックスとして利用されるため、RAG用途でのご利用の場合は別名のインデックスで再作成してください。