

高精度大規模**RAG**エンジン ChatBrid の技術的優位性

- LLMは大規模化で精度向上するがRAGは通常精度低下
- 様々な技術的、内容的工夫で遂にナレッジマネジメントの夢が実現！

独自のベクトルストレージとTF*IDFの重み調整

- 独自のベクトルストレージにより、TF*IDFの重みを自在に調整
- 日本語のSTOP WORDS (不要語: の、こと、もの、する 等の重みをゼロに) を数百語登録
- 単語分割を精緻化。複合語全体と構成語と共にベクトルの要素とし検索精度を向上
 - 沖縄基地問題 vs 沖縄米軍問題 が 67%類似
 - 3.3mg vs 3月3日 が類似しない

(text:型)^3.0 text:溶解 (text:倉庫)^2.0 text:取引
text:t (text:除く)^2.0 (text:手形)^13.0 (text:日)^15.0
text:ラード text:鉛 (text:買)^3.0 text:ビートグフ text:
段ボール (text:名古屋)^2.0 text:雑貨 (text:鋼
板)^2.0 text:油脂 text:営業 text:テレフタレート (text:
製品)^2.0 (text:米)^2.0 (text:卸)^3.0 text:叩 text:天
(text:ポリエチレン)^2.0 (text:纖維)^2.0 ...

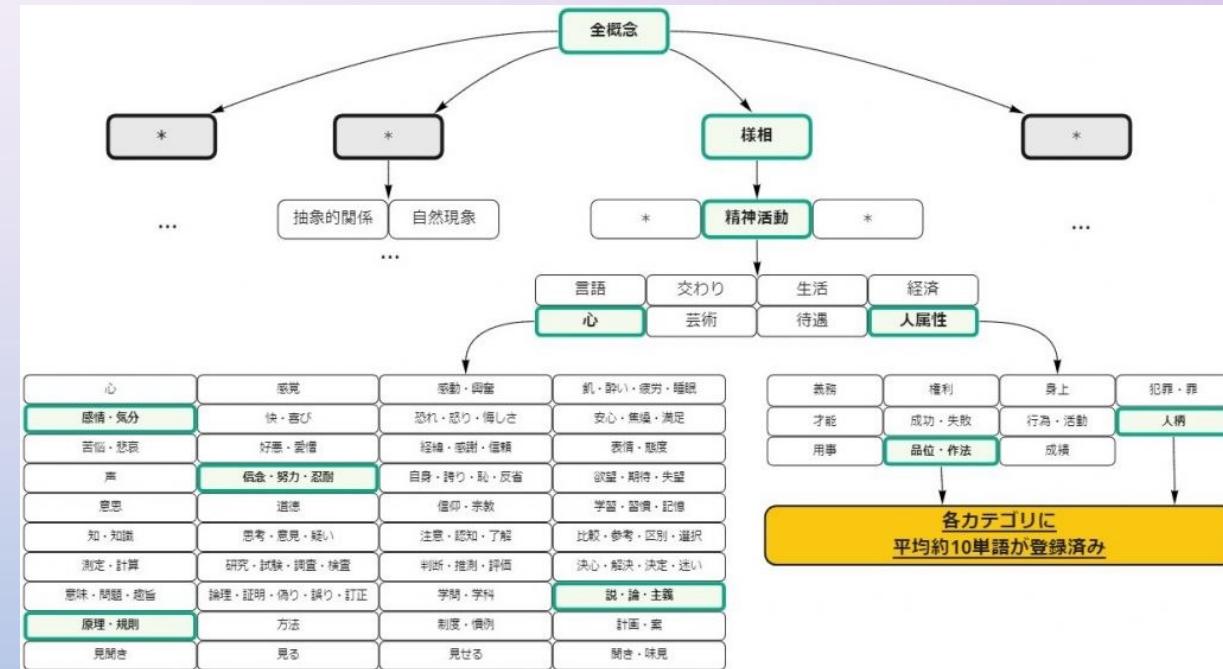


経済特集: 日本市場における多様な素材と製品の最新取引動向
【名古屋発】近年、名古屋を中心とした日本の工業市場では、多岐にわたる素材と製品の取引が活発化しています。特にポリエチレンテレフタレートやポリエチレンなどのプラスチック素材、鉛、油脂といった化学製品の需要が増加しており、これに伴い、営業戦略も急速に変化しています。

最新の市場データによると、名古屋の倉庫ではこれらの素材が大量に取り扱われ、特にポリエチレン製の段ボールやビートグフと呼ばれる特殊な纖維製品の在庫管理が重要視されています。これらの製品は、国内外の雑貨市場での需要が高く、日々の取引量も増加傾向にあります。...

5階層シソーラスの活用

- 30数万語の5階層シソーラスを活用し、単語の意味カテゴリをベクトルの要素に追加
 - 意味カテゴリをベクトルに追加
 - 質問文中に、マニュアルに無い類義語のみが使われている場合でも正しい知識がヒット
 - 一般の質問者は様々な言葉、言い方で表現
 - CHATGPTはそれらの意味を解釈してくれるが、GPT側に送付前の普通のRAGは表現の違いに対応できなかった
 - **CHATBRID**なら、程よく類義語に展開可能！
 - 対義語(反対語)も含まれる



ベクトル検索結果のビジュアル化 ～知識デバッグのための視覚フィードバック

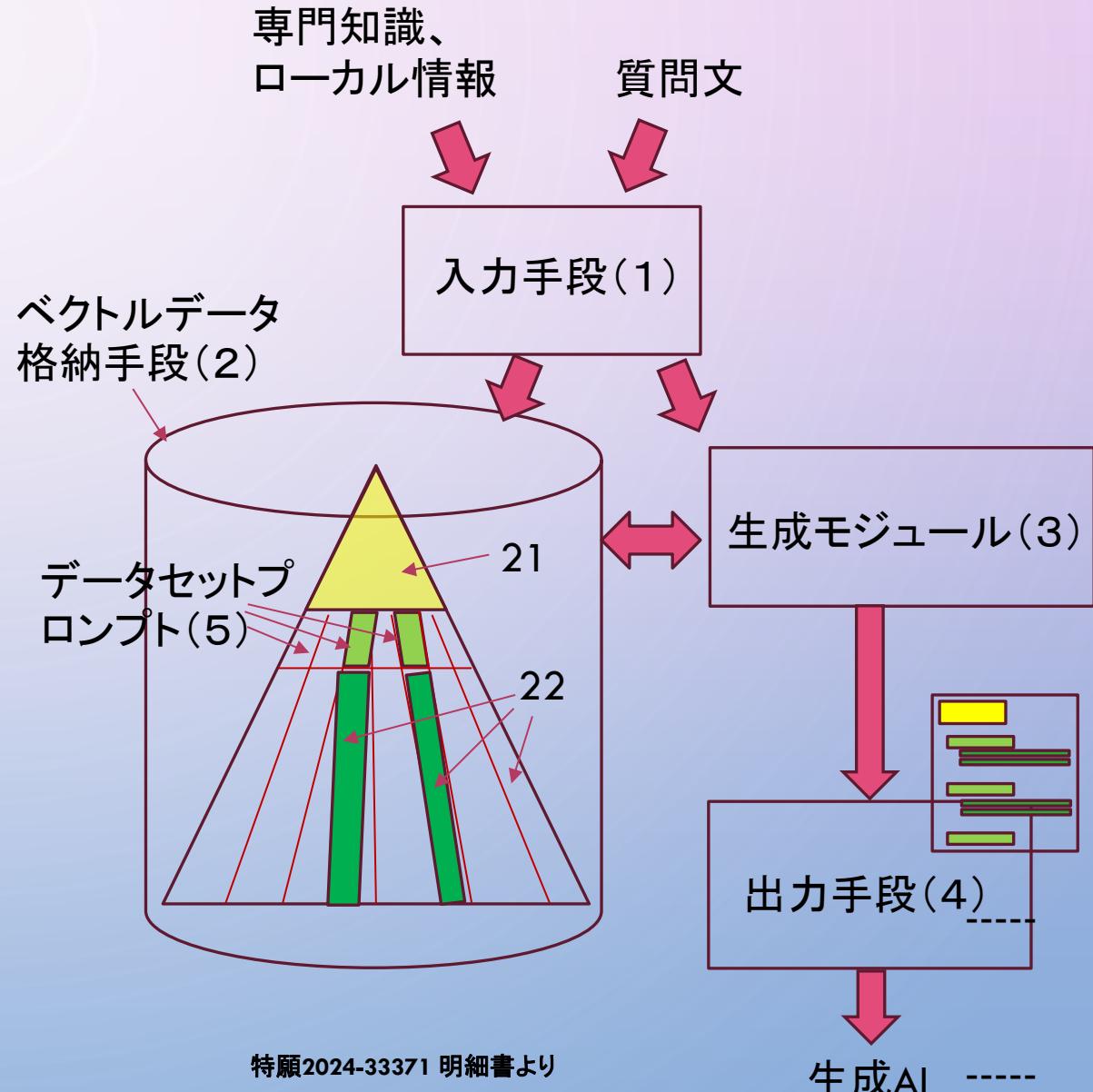
- ・自前のベクトルストレージにより、回答不能や不適切な回答のベクトル検索結果を色付け表示
- ・ビジュアル類似検索でAIの回答を秒速で評価
→ ナレッジ自体の有無、品質を評価、改良
- ・マニュアルの記述が不適切だった場合の即座のビジュアル確認と修正が容易。
- ・マッチした単語と意味コードを強調表示
- ・不適切な回答に対する即時の視覚フィードバック
- ・MARKDOWNをその場で修正 → 即検索に反映

The screenshot shows a search interface with two main panels. The left panel is a search form with dropdowns for '文字列' (Text), 'ID (行番号)', 'ネガポジ', '意味カテゴリ', and '呼び出し日時'. It also includes a '検索対象' section with checkboxes for '本文' (Text), '重要語', 'メモ', '質問テキスト', and '備考', and a '類似度・閲連度 正規化' section with a slider from 0.0~1.0. Below these are buttons for '検索条件のクリア' (Clear search conditions) and '検索・絞込' (Search/filter). The right panel displays search results for 'ファイル名 = 458+707-[q2..._] 202306190723'. The results table has columns for '順位' (Rank), 'ID', '本文' (Text), '類似度・閲連度' (Similarity), 'グラフ' (Graph), 'エージェント名' (Agent name), '質問ファイル' (Question file), and '行' (Line). The first result is highlighted with a red border and the text 'xAI : explainable AI !' overlaid. The results show various text snippets with color-coded highlights for matching words and meaning codes.

順位	ID	本文	類似度・閲連度	グラフ	エージェント名	質問ファイル	行
1	1103	これまで non-FTE のセキュリティールルは、常駐か非常駐かを輪で整理してい た日々の在勤務体制が長期化しており、これまでの考え方でセキュリティ ールルを整備することは不適切委託業務内容をベースに、non-FTE のセ キュリティールルを定義し直す対象：社員以外全て (amigo、派遣、業務 委託、代理店、インター)	248.35896	-	nfte	465	
2	1117	Q&A（随時追加）これまで non-FTE のセキュリティールルは、常駐か非常 駐かを輪で整理して日々の在勤務体制が長期化しており、これまで の考え方でセキュリティールルを整備することは不適切委託業務内容をバ ースに、non-FTE のセキュリティールルを定義し直す対象：社員以外全て (amigo、派遣、業務委託、代理店、インター)	243.80168	-	nfte	479	
3	609	non-FTE のセキュリティールル定義は、従来のセキュリティールルと異なる点 がいくつかあります。まず、non-FTE のセキュリティールル定義は、FTE (Full- TimeEquivalent) に比べて、アクセス権限やセキュリティ対策の面で厳しい 要件があります。これは、non-FTE が不在の期間が長いため、セキュリティ上 の脅威に対してより高いレベルの防御が必要とされるためです。また、non- FTE のセキュリティールル定義では、データの取り扱いに関するルールが強化さ れています。具体的には、非公開情報や個人情報をどの程度の高いデータ の取り扱いに関する要件が設けられています。さらに、non-FTE のセキュリティ ールル定義では、社内規定や法令に基づくセキュリティ対策が求められま す。これにより、企業や組織が定めるセキュリティポリシーや、国が定める法令 に適合する必要があります。以上のように、non-FTE のセキュリティールル 定義は、従来のセキュリティールルと比べて、より厳しい要件が求められる 点がいくつかあります	148.68065	-	latest	q2-180	151
4	610	non-FTE のセキュリティールルは、主に以下の要素を基準に定義されています。 1. アクセス権限non-FTE がアクセス可能なデータシステムについて、どの ような権限を有するかが重要な要素の一つです。例えば、非公開情報や個人 情報へのアクセス権限は、より厳しい権限制御が求められます。2. セキュ リティ対策non-FTE が利用する環境や通信環境、セキュリティ対策などが、セ キュリティールルの決定要素の一つです。例えば、セキュリティソフトウェアの導 入や、パスワードの強度によって、セキュリティールルが異なる場合があります。 3. 所属する組織や業務内容non-FTE が所属する組織や、業務内容が、セ キュリティールルの決定要素の一つとなります。例えば、金融機関や医療機 関で働くnon-FTE は、より高いセキュリティールルが求められる場合がありま す。4. 業務の重要度non-FTE が担当する業務の重要度が、セキュリティール ルの決定要素となることがあります。例えば、重要業務に從事するnon-FTE は、より高いセキュリティールルが求められる場合があります。以上のようによ り、non-FTE のセキュリティールルは、アクセス権限やセキュリティ対策、所属組 織や業務内容など複数の要素を考慮して定められます。	137.80426	-	latest	q2-180	152

特許技術による影プロンプト送出量の削減&高精度化

- ナレッジをジャンル分けし各データセットに分割登録することにより影プロンプトの平均量を削減。
 - RAG内の検索でヒットしたデータセットに付随する影プロンプトのみをLLMに送信。
 - これによりLLMのレスポンスが「速い」「安い」「高い(高精度)」
 - 高速な応答と低成本のLLM利用
 - 大規模化しても高精度を維持
 - 通常1専門分野に特化するRAGをマルチ専門家に
 - 質問文に関連の深い上位ナレッジに付随のデータセットプロンプトのみをLLMに送出
 - コンシェルジェ、総務的な機能が可能に
 - 数百本、数千本のマニュアル、数万ページ、数10万件のナレッジを取り入れられる。



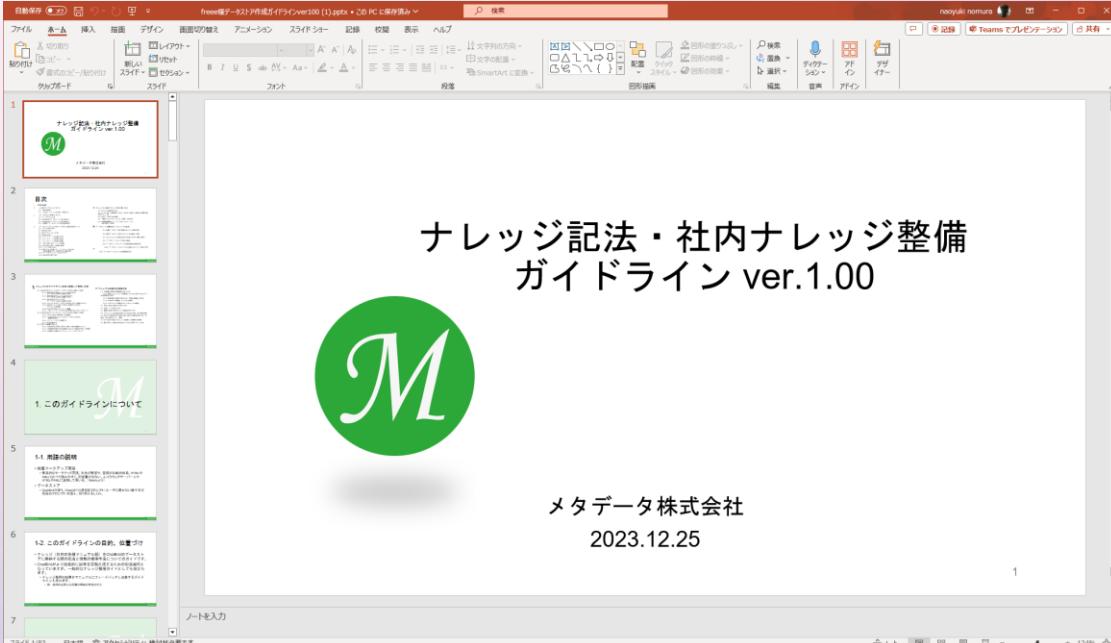
データセットや各種絞込み条件下の単語分布を目視

- ・高頻度の未知語から、用語説明を付すべき社内専門用語などを素早く発見



知識管理のための包括的なマニュアル100数10ページ

- DOCX, XLSX, PPTXをLLMが理解しやすいマークダウンに変換し、知識管理を効率化。
- 100ページ超の効率的なMARKDOWN変換マニュアル
- DOCX、XLSX、PPTXフォーマットに対応



ナレッジ記法・社内ナレッジ整備
ガイドライン ver.1.00

目次

用語の説明	4. マニュアルをMrテキスト分析に取り込み
1. このガイドラインについて	4-1. マニュアル類の取り込み
1-1. 用語の説明	4-2. データ一覧・分類画面、および、ID=2の「本文」に含まれる単語の意味カテゴリ一覧
1-2. このガイドラインの目的、位置づけ	4-3. CSVデータ取り込み画面
2. マークダウン形式について	4-4. 「単語ランキング」デフォルト画面（上位100）
2-1. マークダウンとは	4-5. 表示語数調整スライドバーを右へスライドして上位201件、400件
2-2. CHATBRIDのデータストアで使う記法1	5. データセットの構成法とプロンプトの拡充
2-3. CHATBRIDのデータストアで使う記法2	5-1. 目標データセット数と内部のマニュアル数の目安
2-4. 【参考】データストアでは不要な記法	5-2. 新規データセット作成とマニュアルの追加、削除
3. データストアとしてのマークダウン記法のポイント	5-3. マニュアルごとに異なる区切り文字とそのメモ欄への記入
3-1. リストの取り扱い	5-4. データセットプロンプト記入の基本
3-2. 表の取り扱い	5-5. データセットプロンプトへの用語説明の効率的方法
3-3. 表をテキストにした例	5-5-1. データセットプロンプトから全体プロンプトへの格上げ条件
3-4. 図の取り扱い	5-5-2. データセットプロンプトの効果的活用法
3-5. フローチャートの文章化方法	6. マニュアルをマークダウン形式に変換した事例と注意点
3-6. フローチャートの文章化の例	6-1. PPTX形式のマニュアルをマークダウン形式に変換した事例
3-7. フローチャートの文章化結果	6-1-1. 請求書作成依頼マニュアル(PPTX形式)をマークダウン形式に変換した例
3-8. URLの取り扱い(リンク表記)	6-1-2. 領未審査申請マニュアル(PPTX形式)をマークダウン形式に変換した例
3-9. URLの取り扱い(画像の表示)	6-1-3. PPTX形式のマニュアルをマークダウン形式に変換した流れ
3-10. テストの取り扱い	6-1-4. マニュアルをマークダウン形式に上手く変換するコツ
3-11. 試行により判明したデータストア内の用語の調整TIPS: 発見方法を追記予定	6-1-5. 効率良くマークダウン形式に変換するための、マニュアル改善案
3-12. マークダウンファイル名について	6-1-6. POWERPOINTのアウトライン編集
3-13. Q&A形式の取り扱い	6-1-7. 「今ここ」マーク付きで目次が繰り返し出てくるタイプ
	6-2. DOCX形式のマニュアルをマークダウン形式に変換した事例
	6-2-1.『人事制度』マニュアルの変換例
	6-2-2. DOCX形式のマニュアルをマークダウン形式に変換する流れ
	6-2-3. ハイバーリンクを変換する
	6-2-4. 表を変換する
	6-3. 見出しの階層を整える
	6-3-1. CHATBRIDが分割する見出し単位、概ね7種類について
	6-3-2. 分割結果が概ね1800字未満になるよう最深の見出しきを調整
	6-3-3. 分割部分を極力コンテキストフリーにすべきこと
	7. マニュアル改善点の指摘方法
	7-1. 用語解説対象の候補単語の見つけ方
	7-1-1. 単語ランキングで「定義文」を入れたほうがよさそうな高頻度語を探す
	7-1-2. 複数候補で単語の使われ方、文脈の難解さを評価
	7-1-3. ChatGPTが理解しているかを確認
	7-1-4. いがどこにも定義されていないことの確認
	7-2. 見出し語の不適切さの見つけ方
	7-3. 欠落ページの見つけ方
	7-4. 複数の意味で使われている言葉の見つけ方
	7-5. 元マニュアルの体裁を整えるための空白類、改行類の削除
	7-6. 明らかな漢字脱字や文脈に継ぎ過ぎた省略表現などは、対象を一意に明示すべく、補筆
	7-7. GPT内部で英訳されることを考慮した曖昧性の排除
	7-8. 最小見出しを概ね1800字以下にする分割テクニック例
	8. データセット、プロンプト類のバックアップ、移設
	8-1. CSV出力とインポート
	8-1-1. CSV出力対象フィールドについて
	8-1-2. 取り込み時の注意～タイムスタンプ、データ形式等
	8-1-3. CSVをExcel等で編集する際の注意
	8-1-4. CSVインポート時のエラー、修正方法等
	8-2. プロンプトの手動バックアップとその対象
	8-3. パラメータ設定のバックアップ
	8-4. ユーザ辞書のバックアップ
	9. マークダウン化3rd partyツール、Office365からの変換
	9-1. オンラインのマークダウン編集ツールNotePM
	9-2. MS Office文書からの変換
	9-3. xdoc2txtの活用
	9-4. ChatBrid上で編集。マスター管理する際の留意点～メモ欄の活用

自動レコード分割

～マークダウンの見出し階層指定に基づきチャンキング

職務発明規定
《社外秘》メタデータ株式会社

第1章 総 則
第1条（目的）
- この職務発明規程において、次の各号に掲げる用語の意味は、当該各号に定めるところによる

第2条（用語の定義）
- この職務発明規程において、次の各号に掲げる用語の意味は、当該各号に定めるところによる。

- 職務発明 その性質上会社の業務範囲に属し、かつ、その発明をするに至った行為が会社における従業者等の現在または過去の職務に属する発明として第5条に基いて会社が認定したもの。
- 発明者 発明をした従業者等。
- 従業者等 期間の定めの有無を問わず会社が雇用する者および会社の役員。

The screenshot shows a user interface for managing data sets. At the top, there are several icons: 'データセット一覧' (Data Set List), 'マニュアル' (Manual), 'プロンプト変更' (Prompt Change), '新規データアップロード' (New Data Upload) which is highlighted in green, 'データ一覧' (Data List), and '階級-順位ランキング' (Ranking). On the far right is a 'ログアウト' (Logout) button.

In the center, there's a large input field for '区切り文字 (スペース区切りで複数の区切り文字を指定できます)' (Separator character (Specify multiple separator characters separated by spaces)). The placeholder text 'ex. # ## ###' is shown, and the user has entered '##'. This input field is highlighted with a red rectangle.

Below this input field, a preview window shows the processed data. It contains the heading '# 職務発明規定' and '《社外秘》メタデータ株式会社', followed by the content of the first two sections from the left panel. A blue button at the bottom right of the preview window says '取り込み開始' (Import Start).

知識粒度を小さめにして高精度化しつつコンテキスト情報(上位見出し)を自動付加しLLMの理解を促進

- 知識粒度(レコードサイズ=文字量)がバラつき過ぎると精度低下
- 特に巨大レコードには複数トピックが存在し質問文に無関係なのに関係ありげな記述が混入し精度低下
- 個々のレコードは1パラグラフ程度が好みがコンテキスト不明になってLLMが理解失敗することも



- 小粒度でのコンテキスト維持のために上位見出しを自動付加！

検索条件のクリア

検索・絞込

回答

見出し

データ形式

マニュアル名

責任部門

カテゴリ 2

カテゴリ 3

マニュアル名

データ形式

日付時刻

メモ 削除

?

マニュアル

データセット一覧

データセットプロンプト変更

条件:
ファイル名 = 社内規約類抄コンテキストフリー化

折りたたみ

3件 / 全3件中

クリックで表示/非表示を切替

ID	回答	見出し	責任部門
1		第1章 総則	法務部
2	- この就業規則（以下「規則」という。）は、メタデータ株式会社（以下「会社」という）の秩序を維持し、業務の円滑な運営を期すため、従業員の労働条件、服務規律その他の就業に関する事項を定めるものである。 - この規則に定めのない事項については、労働基準法その他の法令の定めるところによる。	第1条 (目的) << 第1章 総則	法務部
3	- この就業規則（以下「規則」という。）は、メタデータ株式会社（以下「会社」という）の秩序を維持し、業務の円滑な運営を期すため、従業員の労働条件、服務規律その他の就業に関する事項を定めるものである。 - この規則に定めのない事項については、労働基準法その他の法令の定めるところによる。	第2条 (***) << 第1章 総則	法務部

厳密一致遵守機能と属性による知識絞り込み

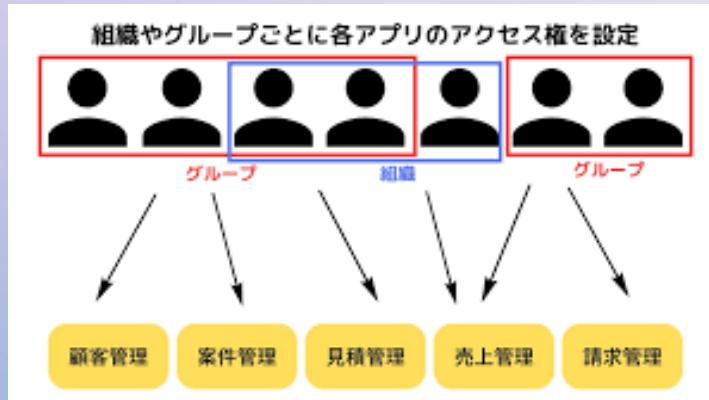
- 指定した文字列との厳密一致遵守(EXACT MATCH)を条件にすることで材料名の取り違えなどの致命的な誤りを防止。
- 属性、法律・条令の有効期間、適用地域等による知識の厳密な絞り込み。

材料ID	材料名	純度	供給元	在庫量 (kg)	備考
Cu-101	電気銅 (C1100)	99.99%	A社	1000	高純度銅
Cu-102	りん脱酸銅 (C1220)	99.90%	B社	500	少量のりんを含む
Cu-103	無酸素銅 (C1020)	99.96%	C社	750	酸素含有量が極めて低い
Cu-104	タフピッチ銅 (C1100)	99.90%	D社	1200	一般的な純銅
Zn-201	高純度亜鉛 (Z1100)	99.99%	E社	600	高純度亜鉛
Sn-301	高純度スズ (Sn99.9)	99.90%	F社	400	高純度スズ
Ni-401	電解ニッケル (N02200)	99.90%	G社	300	高純度ニッケル
Al-501	工業用純アルミ (A1050)	99.50%	H社	800	一般的な工業用純アルミ

Cu-101とCu-102やZn-201が類似検索されてはならない。厳密一致遵守(Exact match)で絞り込み、該当レコードのデータのみを回答に反映させるべし。

知識アクセス権制御

- 質問者の所属や資格を確認し、アクセス権限内のナレッジのみ参照し回答生成へ。
 - セキュアで関連性の高い情報の取得
 - 企業のセキュリティとコンプライアンスに重要



- ・ジャンル別マニュアル群
- ・データセット単位でも、その中のマニュアル単位でも、知識レコード単位でもアクセス権設定可能

The screenshot displays two windows of a knowledge management application. The top window shows a list of access logs for various datasets, with arrows indicating the flow of access rights from organizational levels down to individual users. The bottom window shows a detailed view of a specific document's access rights, with arrows highlighting the granular level of control (down to individual clauses) and the scope of those rights (including specific roles like 'CEO' and 'Manager').

APIモードでの統合運用 ～企業システムとの相互通接続性

- CHATBRID全体がAPIモードで動作
- 各種企業情報システムと一体化して運用可能。
- アプリ構築環境MIIBOやDIFYと統合済。
 - シナリオベース、ワークフローベース

webhookによる連携 (ChatBridのAPI)

