ChatGPT

RAGはなぜ幻滅期に？〜現状、課題、将来展望と企業への示唆〜

要約 (Summary)

- **Gartner社のハイプサイクル2025**によると、生成AIの一形態である**RAG（検索拡張生成）**は現在「幻滅期（Trough of Disillusionment）」に位置づけられています [1] [2]。これは多くの企業でRAG導入への**過剰な期待**が現実の課題に直面し、失望感が広がっている状況を反映しています。初期には「ハルシネーション（誤情報生成）の万能な解決策」として過度に期待されたRAGですが、**精度や安定性が期待値に届かない**ことが明らかになりつつあります [3]。
- 本レポートでは、**RAGの技術的仕組みと利点**から始め、**ガートナーのハイプサイクル**と「幻滅期」の意味を整理します。そして、**RAGが幻滅期に陥った具体的な理由**（検索精度や運用コスト・複雑さ、評価の難しさ等）を掘り下げ、**初期の期待と実際の導入効果のギャップ**を分析します。さらに、**幻滅期を乗り越え次の「啓発期」へ進むための技術的アプローチ**（例：検索手法の改善や評価フレームワーク整備など）や、**国内外の導入事例**から見る成功要因と課題を比較検討し、RAG実用化のハードルを浮き彫りにします。最後に、これらの情報を統合して**RAGの現状と将来展望**を多角的に解説し、企業がRAG導入を検討する際に留意すべきポイントを提言します。

---

1. GartnerがRAGを幻滅期に位置づけた理由と背景

Gartner社が2025年のクラウド＆AIハイプサイクルにおいて、**RAG（Retrieval-Augmented Generation、検索拡張生成）**を「幻滅期」に入った技術として位置づけたのは、現場で直面している課題の深刻さを反映しています [1]。同ハイプサイクルでは、生成AI分野で**AIエージェントが「過度な期待」のピーク**にある一方、**多くの企業が取り組むRAGが幻滅期に入った**と報じられています [4]。その背景には、**RAGの精度向上に企業が苦心している現実**があります [3]。RAGは本来、生成AIを実用化する上で重要な技術として大きな期待を集めていました。しかし**実際の運用では、期待したほどの回答精度やシステム安定性を実現できずにいる**ことが明らかになってきたのです [3]。この結果、RAGへの期待値が低下し、ひいては生成AI全体への期待の落ち込みにつながる可能性があるとGartnerは警告しています [3]。

特にGartnerが指摘するのは、**RAG導入企業でROI（投資対効果）の低さや経営層の不満**が出始めている点や、当初「脱ハルシネーション」の万能薬と目されたRAGが**万能ではなかった現実**です [5] [6]。例えば一部の専門家は、RAGプロジェクトにおいて**期待通りの成果が得られず**経営側の失望を招いたケースを挙げています（ROIが見えにくい、回答精度が低迷等）。Gartnerのレポートでも、多くの企業で**RAGの精度や安定性への不満**が高まり、当初の熱狂的な期待が急速にしぼんでいる現状が示されています [2] [5]。

要するに、**「RAGはすごい」「完璧に近い回答精度が得られる」といった初期のブーム的期待**が、実際の導入フェーズで裏切られつつあるのが幻滅期の実態です。その理由を次章以降で詳しく見ていきますが、端的に言えば**技術自体の限界というより、実装・運用上の課題**に起因する部分が大きいと分析されています [7]。このギャップが**幻滅（Disillusionment）**を生み、RAGが一時的な停滞局面に入ったと位置付けられた背景と言えるでしょう。

2. 基本概念の整理

2(a) RAG（Retrieval-Augmented Generation）とは：仕組み・目的と従来型生成AIとの違い

RAG（検索拡張生成）とは、大規模言語モデル（LLM）による生成AIに**外部の検索機能を組み合わせ**、回答の正確性や最新性を高める技術です [8] [9]。通常の生成AI（LLM）は事前に学習したデータに基づき回答を生成しますが、その知識は訓練データ時点の情報に限られ、**最新の事実や訓練データに含まれない個社固有の情報には弱い**という課題があります [9] [10]。また、LLMは学習データにない質問に対して**「ハルシネーション」**と呼ばれる誤情報の創作を行ってしまうこともあります [11]。RAGはこうした限界を克服する目的で開発されました [12]。

RAGの仕組みは大きく3つのステップから成ります [13]（**図表**）: 1. **ドキュメント検索**: ユーザ質問に関連する外部データ（社内文書やウェブ情報など）を検索し、関連性の高い文章片（**チャンク**）を取得する [14]。 2. **生成AIによる回答生成**: 検索で得た関連文書の内容を、大規模言語モデルに入力プロンプトとして組み込み、それらの**根拠情報に基づいて回答を生成**する [15]。 3. **回答の出力・整形**: 必要に応じて回答から所定の形式の情報を抽出・整形し、ユーザに提示する（例えば引用文の抽出やリスト化など） [16]。

ポイントは、LLMが**回答を生成する際にリアルタイム検索で得た検証済みデータを参照**することで、**知識不足を補完**し [5]、回答の正確性・信頼性を高めることです [17]。たとえば**医療・法律・製造業**など専門知識が要求される分野や、企業内の非公開情報・最新ニュースに関する質問では、RAGが特に有効だと報告されています [18]。LLM単体では答えられない最新の社内ナレッジを検索で取得し、それをLLMが自然な文章にまとめて回答するため、**信頼性が求められるビジネス現場でも正確な回答を得やすくなる**点がRAGの大きな特徴です [19]。また、この**外部データの参照**によって、ChatGPTのような汎用LLMでも**企業固有の最新ナレッジに対応**できるため、**「生成AIの進化形」**として注目されています [20]。

従来の生成AIに対するRAGの利点としては、以下が挙げられます： - **正確性・信頼性の向上**: 外部情報に基づいて回答を補強するため、**ハルシネーション（幻覚回答）の発生リスクを軽減**できます [21] [22]。根拠となる文書を提示することで、回答の裏付けを示すことも可能です（例えば出典の提示など）。 - **最新情報や自社データへの対応**: 事前学習モデルだけでは難しい**最新の事象や個社データ**も、検索によって組み込めるため、**陳腐化しない知識**をAIに反映できます [9] [20]。これにより、常にアップデートされた回答が期待できます。 - **追加学習コストの低減**: モデルを一から追加訓練せずとも、外部データを渡すだけで応答精度を上げられるため、**ファインチューニング（追加学習）のコストが抑えられる**という報告があります [23]。大量の社内文書をモデルに取り込む場合でも、検索インデックス化さえしておけばLLM側の再学習は不要です。 - **スケーラビリティと安全性**: RAGは**LLMを都度最新データで強化**するアプローチのため、静的モデルに比べ柔軟性があります。また必要に応じて**機密データのみを社内検索させる**ことで、**データ漏洩リスクやコンプライアンス面**の懸念にも対応しやすいとされます [24] [25]。外部に出せない情報も、自社内でRAGを構築すればモデルに直接学習させることなく活用できます。

こうした理由から、RAGは**「LLMの強力な相棒」**として2023年前後に急速に脚光を浴び、多くの企業が**「自社データにチャットGPTで答えさせる」**用途で導入検討を始めました [26]。**専門知識や最新情報を組み込んだ高精度なAI回答**を実現するアーキテクチャとして、RAGはまさに**生成AIブームの立役者**の一つだったのです。

2(b) ガートナーのハイプサイクルモデルと「幻滅期」の意味

ガートナーのハイプサイクル（Hype　Cycle）は、新技術が普及するまでの市場・心理的な変遷を**5つの段階**で表した曲線モデルです [27]。技術や概念の**認知度と期待値の推移**を視覚化したもので、各テクノロジーが現在どの段階にあるかをプロットします [27]。5段階のフェーズは一般に以下のように定義されます： 1. **黎明期（Innovation Trigger）**：技術の萌芽期。画期的なアイデアやブレイクスルーが登場し注目され始める段階。

2. **過度な期待のピーク期（Peak of Inflated Expectations）**：メディアや業界の熱狂により、技術への期待値が現実以上に膨らむ段階。成功事例が喧伝される一方で、多くの未成熟な試みも乱立します。　3. **幻滅期（Trough of Disillusionment）**：実装や運用上の課題が顕在化し、**期待が大きくしぼむ段階**です。プロジェクトの失敗や遅延が相次ぎ、**関心が急速に低下**します [28]。多くの企業が投資を見直し、中には撤退するところも出てきます [29] [6]。 4. **啓発期（Slope of Enlightenment）**：一部の残ったプレイヤーが課題克服の経験知を蓄積し始め、**現実的な解決策や新たな応用方法が見えてくる段階**。技術への理解が深まり、改善された第2世代・第3世代の製品やベストプラクティスが登場します。期待も現実に即した形で再び高まります。 5. **生産性の安定期（Plateau of Productivity）**：技術が成熟し、市場に広く受け入れられる段階。**実用価値が実証**され、標準化やコモディティ化が進みます。技術自体は目新しさを失いますが、インフラの一部として当たり前に使われるようになります。

このモデルにおいて「**幻滅期**」は、**熱狂から冷却への転換点**とも言える重要な局面です。過剰な期待がしぼむことで、「失望感からその技術への投資や関心が一時的に落ち込む」状態が特徴です [28]。幻滅期では、**短期的成果が見えにくく**なるため多くの企業がプロジェクトを中止・縮小しがちですが、一方で**生き残った取り組み**は粛々と課題解決と改善を積み重ねています [29]。Gartnerの分析では、**幻滅期に入った技術は往々にして「実装面での困難さ」に直面しており**、それを乗り越える組織的能力が問われる段階だとされます [29] [30]。

実際、「幻滅期に入ると技術への関心が急速に低下し、継続的な学習や改善へのモチベーションが失われるリスクがある」と指摘されています [28]。しかしここで踏みとどまり、地道な改善を続けた企業は、**啓発期に向けた知見を先行して蓄積**できます [6]。幻滅期を経て残った技術は、より現実に即した形で再評価され、次の啓発期へ進むことで初めて真の価値を発揮し始めます [29] [6]。

今回、**RAGが幻滅期に入った**というのは、**技術そのものより運用上の問題で期待を下回った**ことが主因と考えられます [7]。次章以降で詳述するように、RAGには実装・運用面の課題が多く、このフェーズで多くの組織がつまずいているのです。ハイプサイクルの概念を踏まえれば、**現在はRAGに対する期待と現実のギャップが表面化し、マーケット全体が教訓を得ている段階**と言えます。むしろ、この幻滅期の学びを糧にして適切な改善策を講じれば、やがてRAGも啓発期へ移行し**生産性の安定期に至る可能性**があります [29] [6]。そのためには何が必要か――これを理解するためにも、まずは**RAGが直面する具体的課題**を整理してみましょう。

3. RAGが幻滅期にある具体的課題：精度・運用・コスト・評価の視点

RAGが幻滅期に陥った背景には、**技術的課題と運用上の難しさ**が複合的に存在します。ここでは主な課題を(i)検索精度・回答品質、(ii)導入と運用の複雑さ、(iii)コスト要因、(iv)性能評価の難しさ、の観点から解説します。

- **(i) 検索精度の問題と回答の正確性**: RAGの**回答品質は検索された文書の質と関連性に大きく依存**します [31]。しかし実際には、**「検索結果にばらつきがある」「回答が不正確」**といった声が多くの導入企業・自治体で上がっています [32]。これは、インデックス化したデータにノイズが多かったり、適切な文書がヒットしなかったりすることに起因します。**データ量や構造が統一されていない**と、同じ質問でも異なる回答が出たり、誤った・不十分な回答が生成されてしまい、実務利用に支障をきたすのです [32]。たとえば、**検索対象ドキュメントに答えが無い場合（欠落したコンテンツ）**はRAGが無理にそれらしい回答を生成してしまう可能性が指摘されています [33]。また、関連する文書が検索上位に来ない（トップランク文書の見逃し）と正しい回答が生成されませんし [34]、検索で見つかった文書が質問の文脈に合っていない場合（文脈不整合）も不適切な回答につながります [35]。こうした**検索精度不足**やマッチングの問題が、RAGの回答精度を著しく妨げる主要因となっています [36] [37]。実際、BCGの分析でも**「RAG実践には検索精度の向上という古典的課題がある」**と明言されています [38]。要するに、**情報検索という昔からの難題（ゴミデータを入れればゴミが出る問題）**に

RAGも直面しており、そこをクリアしない限りLLM側でどんな高度生成をしても正確な回答は得られません。

- **(ii) 導入・運用の複雑さ（データ準備とシステム構築）**：RAGを効果的に機能させるには、**裏で支えるデータ基盤の整備**が不可欠です [39] [40]。しかし、多くの企業は最初この**手間と難しさ**を甘く見積もっていました。実際に導入してみると「**データの量や形式がバラバラで、クレンジングや構造化が追いつかない**」という状況に陥りがちです [32] [41]。例えば社内文書には**誤字脱字・表記揺れ・古い情報・重複**などのノイズが多く、特にPDFなどから起こしたテキストには不要な改行や余計な要素が含まれます [42]。これらを**データクレンジング**しないままインデックス化すると、検索がうまく働かず**期待精度が出ない**一因となります [42]。さらに**チャンク分割**の設計も難しく、適切にチャンキングしないと文章の意味が分断され検索ミスにつながります [41]。有効なチャンクサイズはケースバイケースで、ページや段落ではなく**意味単位で分割する**ほうが精度向上に有効だと報告されています（大和総研の例ではNLPを使い意味に基づくチャンク分割で精度改善） [43]。またデータの分類も重要で、ソフトバンクの事例では**ユースケース別にデータを分けて管理**し、それぞれに適切な前処理をすることが成功のポイントだったといいます [44]。このように、**RAG導入の裏側では相当のデータ整備とチューニングが必要**であり、「導入して終わり」ではなく**継続的なチューニングとフィードバックループ**が欠かせません [45]。実際、「RAGは導入自体は比較的簡単でも、**回答精度を出すことが難しい**」と言われます [46]。多くの企業が**十分なデータ前処理をしない**まま試して精度が伸び悩み、そこで挫折するケースが散見されます [47] [48]。またシステム的にも、LLM＋ベクトルDB＋検索APIと複数コンポーネントを統合する必要があり、その**アーキテクチャ管理やスケーリング**も運用担当者にとって新たな負荷となっています [49] [50]。Gartnerのレポートでは、RAGの課題の多くは**技術そのものの限界より「実装と運用」に起因**すると指摘されており、データ品質・クエリ最適化・検索結果とLLMの統合方法などが鍵だとされています [7]。

- **(iii) コスト要因（構築・維持コストとスケーラビリティ）**：RAG導入には**新たなコスト構造**も発生します。従来のLLM　API利用に比べ、**事前のデータ埋め込み（ベクトル化）**やインデックス構築にかかる手間と費用が無視できません [51]。例えば数万〜数十万件の社内文書をすべてベクトル変換する場合、その埋め込みAPI利用料が相当額になります [52]。加えて見落とされがちなのが、**モデルやEmbeddingのアップデートに伴う再埋め込みコスト**です。現在主流のEmbeddings（OpenAI等）には**後方互換性がなく**、モデルバージョンが上がるたびに**全データの再ベクトル化が必要**と指摘されています [52]。クラウドネイティブ社の分析では、「半年に一度は再埋め込み作業が発生する前提で予算を考えるべき」とまで言われています [53]。つまり**RAG構築時に発生した埋め込み費用は、モデル更新のたび繰り返し発生**するのです [52]。この維持費用は、静的に一度学習させれば終わりの従来モデルにはなかったランニングコストです。また、**回答生成そのもののコスト**（トークン消費）も質問内容に比例して発生し続けます [54]。RAGはユーザーの質問ごとに検索と生成を行うため、利用が拡大すると**検索インフラのスケーリングコスト**や**LLM　API利用料**が増大します。さらに、オンプレミスで構築すれば**ベクトルデータベースの運用コスト**（クラスタ管理やストレージ拡張など）もかかります。こうした**費用対効果の不透明さ**は経営層の懸念材料になりやすく、明確なROIを示せないとプロジェクト継続が難しくなる一因です [55]。幻滅期に入った技術には「コスト高騰や価値不明確さから中止される」例が多く、RAGも**ROIの見えづらさ**が投資判断を鈍らせている可能性があります [56] [57]。

- **(iv) 性能評価と継続改善の難しさ**: RAGシステムの**評価指標を定めづらい**ことも、課題改善を遅らせる要因です。一般的なAIシステムであれば精度や再現率など定量評価が可能ですが、RAGの場合「**どんな質問にどれだけ正しく答えられるか**」を測る必要があり、汎用的な指標設定が難しいのです [39]。NTT東日本の指摘によれば、正答率・再現率・ユーザー満足度（CSAT）といった**KPIを設定しないままでは、精度の評価や改善が行えず運用が形骸化する**恐れがあるとされます [39]。しかし実際には、このKPI策定が疎かにされるケースが多々あります [39]。ビジネス部門と現場の要件を詰めずにRAGチャットボットを展開した結果、「何をもって正解とするか」が不明瞭なままユーザーの苦情（誤回答や使いにくい等）だけが蓄積し、プロジェクトチームも改善の糸口を掴めない――という悪循環に

陥りがちです [58] [59]。Dan Giannone氏は「多くの組織が**何を解決したいのか定義しないままRAGに飛びつき**、デフォルト設定でボットを公開してしまう。最初は“自社データと対話できる”魔法にユーザーも驚くが、すぐに**ハルシネーションや間違いに気付いて幻滅**し始める」と指摘しています [60]。結局、**質問と正解のゴールドセットを用意し、それに合うデータソースを揃えて初めて評価と改善が回せる**のですが [61] [62]、そのプロセスを怠ると「**改善しようにも何をどう直すべきか測れない**」状態となります [63]。さらに、経営側にとってはユーザー満足度や業務効率化といった**成果指標が曖昧**なままだと、追加投資に慎重になります。幻滅期にありがちなのは、「**当初の期待に比べて効果が測定できず、周囲の支持が得られなくなる**」ことです [64]。RAGの場合も、**定量的な価値の証明が難しい**ために失望を招いている面があるでしょう。

以上のように、**RAGは単なる技術要素の組合せ以上に、データ品質管理・システム構築・費用管理・効果測定という**広範な課題と向き合う必要があります。**「幻滅期」に入った主因**は、これらの課題によって**初期期待ほどの成果が出せていない**ことにあります [2] [5]。特に日本企業においては、社内データ整備の遅れや人材不足もあって、期待倒れに終わっているケースが散見されます。このままでは「やはり使えない技術だった」と烙印を押されかねませんが、逆に言えば**課題を一つ一つ解決していけば、幻滅期を脱することも可能**です。それにはまず、次章で述べる**過度な期待と現実のギャップ**を正しく認識することが出発点となります。

4. 初期の過度な期待と現実のギャップ：ハルシネーション解決の幻想と導入現場の実態

RAGがこれほど幻滅期の渦中にあるのは、**当初の「万能感」への期待**と**現実の冷厳な結果**との間に大きなギャップが生じたためです。当初、生成AIブームの中でRAGは**「ハルシネーション問題を解消する切り札」**としてもてはやされました [21]。**「社内の膨大なドキュメントさえ食べさせれば、ChatGPTが何でも正確に答えてくれる」**――そのような**夢物語的な宣伝**も一部で語られ、経営層やユーザーの期待値は急上昇していました [64]。具体的には次のような**過度な期待**がありました。

- **ハルシネーションが完全になくなる**: RAGならばAIが必ず根拠のある回答を生成するので、もはやAIの嘘回答を心配しなくて良い、といった期待。【※実際は一定の低減はするものの、根拠データが不十分だったり統合の仕方によっては依然として捏造が発生します。】
- **自社データを入れれば精度100%近くになる**: 社内ナレッジを全部取り込めば、人間以上に正確で詳細な回答がすぐ得られる、と誤解されました。【※実際はデータ量が増えるほどノイズも増え、むしろ**的外れな回答が増えるリスク**があります。きちんと選別・管理しない大量データ投入は「Garbage in, Garbage out」を招くという指摘があります [67] [65]。】
- **導入すれば即業務が劇的効率化**: 特別な調整なしに、現場社員が何でも質問すれば瞬時に最適解が返ってくる、と期待されました。【※現実には、適切に使うためのユーザー教育やプロンプト工夫が要りますし、出てきた回答を**社員が検証・活用するプロセス**も依然必要です。導入直後は「便利だ！」という驚きがあるものの、すぐに**誤回答への不満**に転じるケースも多いです【14†L107-L115】 [66]。】

これらの期待に対し、**実際の導入現場で得られた成果**は往々にして**そこまで理想的ではありません**でした。いくつかギャップを具体的に見てみます。

ギャップ1: ハルシネーションは減ったが消えてはいない – RAGは確かに**AI回答の暴走を抑える一定の効果**を見せました [21]。しかし「完全解決」には程遠く、文書に答えが無い場合や文書が曖昧な場合、AIは依然として**もっともらしい誤情報を作り出す**ことが確認されています [33]。初期には「RAGなら間違いようがない」と信じたユーザーもいましたが、実際には根拠文書が偏っていたり不足していたりすると**誤答が出る**ため、**「なんだ、まだ嘘つくじゃないか」**という失望に変わりました。例えば社内QAボットにRAGを導入したある企業では、想定外の質問に対してやはり事実誤認の回答が出て現場ユーザーからクレームが上がり、**「結局AIは信用できない」と評価が下がった**というケースが報告されています（ユーザー調査で「知りたい情報が

出てこない」「誤った回答が多い」という声が広まった）[32]。**RAGはハルシネーション問題を緩和するが、万能薬ではない**――この現実は多くの企業で思い知らされた点です。

ギャップ2: データを入れただけでは精度向上しない – 初期には「社内文書を全部読み込ませればAIが賢くなる」と考えた向きもありました[59][67]。しかし、Dan Giannone氏が指摘するように「**ビジネスチームには『AIにデータを食べさせれば食べさせるほど賢くなる』という誤解がある**」のが実情で、結果として**無秩序に何千もの文書をインデックスに放り込んでしまう**例が頻発しました[67][68]。これはかえって**検索汚染**を引き起こし、適切な文書に当たりづらくなるだけでした[69][70]。現実には、**少数の厳選した良質文書の方が1000件の玉石混交データより有用**であるにも関わらず[69][70]、初期期待では「量が多いほど良い」と思われていたのです。多くの企業がこの落とし穴に陥り、「せっかく全社の知識を集めたのに正しい答えが出ない」という壁に突き当たりました[68][32]。**データ品質＞データ量**であることを痛感した組織も多いでしょう。

ギャップ3: 現場への定着に時間がかかる – RAG導入当初は、現場社員も「自分の質問にAIが答えてくれる!」と期待して試します。しかし前述のように誤答やばらつきがあると、**ユーザーの信頼が揺らぎ利用率が低下**するケースが報告されています[32]。「**使いにくい**」「**思った情報が出てこない**」という声が広がると、せっかく導入しても社内で定着せず形骸化します[32]。これはまさに幻滅期の兆候で、最初の「ワオ！」から「なんだ期待外れだ」に変わる瞬間です。原因は上述の通りシステム側にもありますが、**ユーザー側のリテラシーや期待値調整不足**も一因でした。当初、技術に疎い決裁者ほど「40%程度の回答精度では低すぎる」と感じてしまい[64]、**現実的な精度水準への理解が不足**していたケースもあります。東京ガスなどはこの対策として**アジャイル開発で現場ユーザーと密にコミュニケーションを取り、段階的に精度を高める**取り組みを行ったといいます[71]。つまり**PoC段階で一気に完璧を目指すのではなく、小さな成功と学習を積み重ねてユーザーの期待値も調整していく**必要があったのですが、そこを怠ったプロジェクトほど「期待外れ」に終わりました。

このように、**初期の過度なブーム（過熱）と、現場で明らかになった地道な課題とのコントラスト**が幻滅を生んでいます。Gartnerレポートも「RAGは期待されたほどの精度・安定性が実現困難であることが明らかになってきた」と指摘しており[3]、**期待と現実の乖離**が技術全体への期待度低下につながりかねないと警鐘を鳴らしています[3]。

しかし裏を返せば、**これらのギャップは今後の改善ポイントを示唆**しています。初期の幻想が消えた今、RAGに求められるのは**冷静な現状認識に基づく技術革新とアプローチの見直し**です。次章では、まさに**幻滅期を脱し啓発期へ進むために必要な技術的・運用的ブレイクスルー**について考察します。

5. 幻滅期から啓発期へ：RAG再浮上に必要な技術革新とアプローチ

RAGが幻滅期を乗り越え、次の**啓発期（Slope of Enlightenment）**へ移行するには、前章で見た課題を克服するための**技術的イノベーション**と**運用面での工夫**が不可欠です。ここでは主要な展望をいくつか挙げます。

- ① **検索技術の高度化とハイブリッドアプローチ**: RAGの成否を握る検索精度向上には、従来のベクトル検索に加えて**新たなアプローチ**が期待されています。その一つが**Graph RAG（グラフRAG）**です[72]。Microsoftが提唱したこの手法は、**知識グラフ**の考え方を取り入れ、単純な埋め込み類似度だけでなく**単語や概念間の関連性（グラフ構造）**を利用して関連ドキュメントを探索します[72]。具体的には、文書内の重要キーワードをノード、それらの関係をエッジとするグラフを構築し、検索時には頻出度や関係性に重み付けして回答生成に反映するというものです[73]。これにより、単にベクトル距離が近いだけでなく**文脈上本当に関連深い情報**を抽出でき、**より精度の高い回答**が期待されています[73]。Amazonも自社のグラフデータベース（Amazon Neptune）でGraph RAGをサポートする動きを見せており[74]、**グラフ理論と組み合わせた検索**は今後RAG精度向上の切り札の一つになるでしょう。さらに**ハイブリッド検索**（ベクトル検索＋キーワード検索の統合）も引き続き有望です。ベクトル検索は概念マッチが得意ですが完全一致が苦手なため、従来型のインデックス検索と組み合わ

せて漏れを補完する実装も増えています。また**メタデータによる検索性能向上**も地味ながら重要な改善策です [75] [76]。文書にカテゴリーや日付、重要度といったタグ情報を付与して検索に活用すれば、ノイズ削減や結果の絞り込みに役立ちます [77]。これら検索側の工夫によって、**「必要な情報に当たらない」問題を減らす**ことが、啓発期への鍵となります。

- **② LLMとRAGの統合的進化（フィードバックループとエージェント化）**: 現状のRAGはLLMへの単回プロンプトとして検索結果を投げる静的な構造ですが、今後は**LLMと検索が対話的に連携する形**への進化が考えられます。例えば、**マルチステップのエージェント型AI**との融合です [78]。LLMが回答中に追加情報の必要を判断し、逐次的に検索クエリを生成して必要データを集めるような**能動的RAG**が登場すれば、より複雑な質問にも一貫性を保って答えられるでしょう [79]。実際、**Agentic RAG**を提唱する声もあり、計算やAPI利用を組み合わせて**より「考えながら調べるAI」**への発展が模索されています [78]。また**生成結果の検証（バリデーション）**も次の課題です [80]。現状のRAGは「情報を引っ張ってくる」ことはしても**それが正しいか手続きを踏んだかは評価しません** [79]。将来的には、生成内容を別のモデルで事実チェックしたり、複数検索結果との照合で矛盾を検出したりする**AIガードレール**の実装が必要になるでしょう [31] [81]。例えばOpenAIや各社も「ツール使用+検証」のアプローチを研究しており、**LLM自身に計算させたり複数回答させて比較する**など、回答の質保証メカニズムが発展すると考えられます。**継続的学習**も重要なテーマです。現状RAGシステムは構築後に未知の質問が来ても静的対応ですが、**ユーザーからのフィードバックを反映して検索インデックスや生成挙動を調整する仕組み**があれば、使うほど精度が上がる理想に近づきます [82] [77]。実際、運用中も**継続的なキャリブレーション（較正）**が必要との指摘があり [83]、ユーザーの評価やクリックした文書を学習して応答を改善するリインフォースメント学習的な試みも今後出てくるでしょう。要は、**RAGを一度作って終わりではなく、生きたシステムとして適応させていく**取り組みが啓発期には求められます。

- **③ ベストプラクティスの確立と評価手法の標準化**: 技術革新と並行して、**RAG導入・運用のノウハウ蓄積**が不可欠です。幻滅期に失敗した事例から学び、**成功パターン**を確立していくことが啓発期への階段となります。例えば前述のように**「データ前処理・チューニングが肝要」**という教訓から、具体的なガイドラインが整備されつつあります。NTT東日本は**「RAG精度向上のための体制と指標」**として、データクレンジング手法やKPI設定、改善のPDCAサイクル構築などを詳しく解説しています [84] [39]。またDan Giannone氏らは**RAG評価フレームワーク**の必要性を唱え、**Gold標準の質問集と正答セットを用意**し、その上で**検索ステップと生成ステップそれぞれを評価**すべきだとしています [61] [62]。今後、業界横断で使える**RAGシステム評価ベンチマーク**や、精度・再現率以外に**ビジネスKPIとの関連**を測る手法が出てくるでしょう [39]。たとえば「回答の正確性スコア」と「利用者満足度」や「業務時間短縮率」を総合的に評価し、ROIを示すような取り組みです。**標準評価指標が確立されれば、各社が自社のRAG性能を客観比較でき、投資判断もしやすく**なります。これにより経営層の理解も深まり、安定したリソース投下が期待できます。また**ユースケース毎のベストプラクティス共有**も重要です。社内FAQ向けRAG、カスタマーサポート向けRAG、社内知識Wiki向けRAGなど、目的別に見ると最適解は異なる場合があります。各分野で成功した設定（例えばチャンクサイズはFAQでは小さく、技術文書では大きめが良い等）をコミュニティで共有し、再現性の高い手法を体系化することが技術成熟につながります [41] [43]。日本では大手SIerやクラウドベンダーが中心となり、セミナーやブログで情報発信を始めています [37]。今後さらに**学会や業界団体での知見交換**が進み、RAGの教科書的な知識体系が醸成されれば、幻滅期の混乱から一歩進んだ「啓発」の段階に入ったと言えるでしょう。

- **④ コスト構造の改善とソリューション化**: コスト面のハードルについては、技術的な最適化とともに**サービスモデルの進化**が助けになる可能性があります。たとえば埋め込みコストに対しては、**オープンソースの軽量Embeddingモデル**の活用や、**差分更新**（変更部分だけ再ベクトル化）など運用工夫で削減できる余地があります [77] [75]。一部研究では、小さなテキストなら**OSSの埋め込みモデルでも商用並み性能**との報告もあり、必ずしも毎回高価なAPIを使わずともよい可能性があります [75] [76]。また**オンデマンド型のRAGプラットフォーム**も増えてくるでしょう。既に各クラウド事業者は

「自社データをアップロードすればRAGチャットボット構築」というサービスを提供し始めています（例: MicrosoftのAzure OpenAIによるEnterprise Chat、Amazon BedrockによるRAG支援など）。これらを使えば、企業個別にゼロからシステムを組むより**安価かつスピーディに導入**できます。ただしブラックボックス化による評価の難しさもあるため、その点は透明性向上が必要です。さらに**大規模知識ベースとの統合**という方向性もあります。例えば社内に**ナレッジグラフやFAQデータ**が既に整備されているなら、RAGと直接連携させてより効率的に検索させる（無駄な全文検索を減らす）といった工夫でコスト削減・高速化が見込めます。**スケーラビリティ**に関しては、質問内容によっては全量検索でなく局所データベースのみ見るような**階層型検索**なども提案されています。啓発期に向けては、**RAGのコストをいかにコントロールしつつ効果を出すか**が一つのテーマとなり、それを実現するソリューションが市場に出回るでしょう（既に一部コンサルやベンダーが「RAG精度改善サービス」として8週間でPoC実施などのメニューを打ち出しています [85]）。

以上のような技術的・手法的な展望を総合すると、**幻滅期を脱する鍵は「現実に即した改良」と「地に足のついた期待値の再設定」**にあります [30]。過度なバズワードとしてではなく、**問題解決の手段としてRAGを正しく位置付け直す**ことが重要です。Gartnerのコメントにも「幻滅期の技術に対しては、**技術的限界を正しく理解し現実的な期待値を設定する能力**が重要」だとあります [86]。RAGの場合、その限界と可能性の両方を踏まえて、**確実に価値を出せる領域から着実に成果を積み上げる**ことが、次の段階への道筋となるでしょう。

6. RAG導入事例の分析：成功例と課題例に見る実用化のハードル

幻滅期とは言え、既にRAGを現場で活用し成果を上げている企業も存在します。ここでは国内外の**導入事例**をいくつか取り上げ、**成功したケースと困難に直面したケース**を比較分析します。実例から、RAG実用化におけるハードルと克服のポイントが見えてきます。

成功事例①：LINEヤフー「SeekAI」 – 全社員向け業務支援チャットボット

統合後のLINEヤフー社では、**社内ナレッジ活用ツール「SeekAI」**を全社員に展開しています [87]。これはRAG技術を活用し、社内のFAQや業務マニュアルなどから情報を検索・要約して回答するチャットボットです [87]。注目すべきは、本格導入前に**一部部門でトライアル運用**し、実用性と有効性を確認してから全社展開した点です [88]。この段階的導入により、現場のフィードバックを反映し調整を重ねることで、スムーズな受け入れにつなげました。「情報収集時間の削減」「業務効率化」が実際に確認されたことが成功の鍵です [88]。また、LINEヤフーのように**ITリテラシーの高い企業**では、社員側もAI活用に前向きで受容度が高く、運用上のトラブルが少ないことも成功要因と思われます。全社規模で**RAGチャットボットが定着した好例**と言えるでしょう。

成功事例②：東京ガス「社内生成AIアプリ（AIGNIS-chat）」 – 現場主導の内製開発と浸透戦略

東京ガスは**2024年10月に社内向け生成AIアプリケーションを独自開発し運用開始**したと発表しました [89]。Microsoft Azure OpenAI Serviceを活用し、テキスト入力による業務情報検索・文章作成支援を行うものです [90]。東京ガスの取り組みの特徴は、**現場ユーザーの声を反映しながら継続的に機能改善**を進める計画を掲げていることです [91]。実際、導入を通じて**社内活用を拡大し業務の質とスピードを高める**方針を示しており、単なるPoCに留めずDX戦略の一環として育てていく姿勢が伺えます [91]。東京ガスの場合、社内で**「第3の創業」と銘打ったDXと生成AI融合の大号令**があり、トップダウンとボトムアップを組み合わせた推進体制が取られています [92]。**アジャイルに開発して社員を巻き込み、期待値も適切にコントロールしつつ段階的に精度向上**を図ったことが、現場での受容につながったと思われます [71]。また**ユースケースの選定**もポイントで、まずは問い合わせ対応や文書作成支援など**効果が出やすい領域**に絞って適用しているようです [93] [94]。こうした焦点化により、短期間で目に見える成果（例：問い合わせ対応時間の大幅削減）が出やすく、経営にも現場にも成功体験を提示できています [95]。東京ガスの事例は、日本企業における**RAG内製化成功の代**

表例として注目され、同社はデロイトと協働した多様なPoCから知見を得ていることも明かされています [96]。

成功事例③：Morgan Stanley「AskResearchGPT」(グローバル事例) – 専門知識検索に特化したRAG活用

米国投資銀行Morgan StanleyはOpenAIと提携し、**社内の富裕層向け投資リサーチ資料をAIが検索・要約する「AskResearchGPT」**を開発しました [97] [98]。これはまさにRAGの典型で、数万ページに及ぶ金融リサーチ文書をベクトルDB化し、GPT-4がそれらを参照してアドバイザーからの質問に答える仕組みです [99] [98]。Morgan Stanleyの成功要因は、**ドメインを金融リサーチに限定**し、明確な対象ユーザー（社内の財務アドバイザー）にフォーカスした点です。LLMの専門外知識をRAGで補い、専門家が求める粒度の情報を迅速に提供することに注力しました。また約半年かけてGPT-4モデルをチューニングし、出力品質を高めたとされています [100]。さらに**ユーザーインターフェース**にも工夫があり、アドバイザーが信頼して使えるよう**生成回答に常に出典（元のリサーチ文書）を提示**する設計としました [101]。これによりAIの回答に対する懐疑心を和らげ、利用者の安心感を確保しています。結果としてAskResearchGPTは**業務効率を高めつつ誤情報のリスクを抑えるツール**として受け入れられています [102]。グローバル企業の例として、**LLM+社内知識で業務支援**を実現した好例です。ポイントは、**対象領域を限定して高品質データを準備し、ユーザーの信頼を得るためのUX上の工夫**（エビデンス表示など）をしたことにあります。

課題事例①：某社の社内QAボットPoC中止 – 評価指標欠如で成果示せず

具体名は伏せますが、国内のある大手では**社内問い合わせ対応AI**としてRAGチャットボットのPoCを行ったものの、**評価指標が定まらず効果を示せないまま中止**になった事例があります。担当者の話では、当初は「社員の質問にAIがすぐ答える」という触れ込みで期待されたものの、**何をもって成功とするか（回答精度何%ならOKか、どれだけ業務時間短縮できればOKか）が曖昧**なまま進めてしまいました。結果、ユーザーからは「便利だが嘘が混じる」「結局詳しい人に聞き直す」という反応が多く、肝心の**KPI（例：問い合わせ対応工数の削減率）も測っていなかった**ため、プロジェクトオーナーに説得力ある報告ができませんでした。結局、**経営層からROI不明確との指摘**を受けPoCは打ち切りとなり、担当チームは「しっかり要件定義と評価設計をすべきだった」と反省しています。この例は、**目的と評価を定めないまま導入すると幻滅を招くだけ**という教訓を与えています [60] [66]。

課題事例②：小売業の製品問い合わせボット – データ更新の遅れと誤回答で顧客クレーム

ある小売業では、顧客からの商品問い合わせにAIが回答するチャットボットを導入しました（商品の在庫や仕様などを回答）。当初は「24時間自動回答」でコールセンター負荷軽減を期待されましたが、**問題は商品データの更新が追いつかなかったこと**です。RAGの元になる商品情報データベースを手動更新に頼っていたため、新商品の追加や仕様変更がインデックスに即時反映されず、**古い情報をもとにAIが誤答する**ケースが発生しました。例えば「〇〇商品のサイズは？」との問いに旧モデルのサイズを答えてしまいクレームになる、といった具合です。これにより「AIの回答は信用できない」と判断され、結局有人対応に切り替える場面が多発。**運用体制の不備（データ即時更新と検証プロセス不足）**が原因で顧客体験を損ね、信頼を損なった事例です。この会社は、RAG自体より**データ運用フロー**に課題があると理解し、今は更新自動化や回答検証ルールの整備に取り組んでいます。つまり、**リアルタイム性が要求されるユースケースではデータ更新体制まで含めた仕組み作りが必要**という反省点です。幻滅期にある今、このような試行錯誤が各所で起きており、失敗もまた貴重なフィードバックとなっています。

以上、成功例からは**効果的なRAG導入の鍵**として以下が浮かび上がります： - **ユースケースとデータを限定し焦点を絞る**（広げすぎない）。 - **ユーザー巻き込みと段階的展開**（PoC→限定運用→全社展開）。 - **データ品質・更新フローの確立**（常に最新・正確な情報を供給）。 - **ユーザーへのエビデンス提供や期待値管理**（信頼性確保と教育）。 - **明確なKPI設定と効果測定**（経営への説明責任を果たす）。

一方課題例からは、**評価設計の欠如**や**運用体制の甘さ**が失敗を招くことが分かります。要するに、**技術よりもプロジェクト運営能力が試される**面が大きいのです。幻滅期に学ぶべきは、これら実例の知見を共有し再発防止策を講じることです。幸い成功事例も増えてきているので、それらを手本に各社が戦略を練り直せば、RAG導入のハードルは徐々に下がっていくでしょう [103]。実際、先進事例を通じて「自社での活用イメージが具体化した」という声もあり [103]、幻滅期を乗り越えるきっかけが芽生えつつあります。

7. 総合考察：RAGの現状、課題、将来展望と導入検討時の留意点

以上の調査結果を踏まえ、**RAG（検索拡張生成）の現状と課題、そして将来展望**を総括します。また、企業がRAG導入を検討する際に考慮すべきポイントを提言します。

7.1 RAGの現状と主要課題

RAGは現在、ガートナーのハイプサイクルで「幻滅期」に分類されるように、**過度な期待が収束し現実的な課題と向き合う局面**にあります [4] [3]。多くの企業が実践に取り入れ始めた結果、**精度面・運用面の問題**が次々と浮き彫りになりました。その主な課題は以下に集約されます。

- **回答精度・安定性**: ハルシネーションの減少は見られるものの、**検索精度不足やデータ品質のばらつき**によって依然**誤答や不安定な回答**が発生します [32] [37]。これは特に社内データの整備度合いに左右され、統一されていない情報源からは正確な回答を引き出しにくい現状です。
- **実装・運用の困難さ**: RAGは**仕組み自体はシンプル**ですが、裏で動く**データ準備（クレンジング・チャンキング）**と**システム統合**に高度なスキルと労力を要します [41] [7]。さらに継続運用には検索ログの分析やインデックス更新など**地道なメンテナンス**が欠かせず、組織としてそれを支える体制が整っていないと形骸化しがちです [32] [45]。
- **コストとROI**: 埋め込み作業やクラウド利用料など、**新たなコスト構造**に対する認識が不足していました [52]。特に投資対効果（ROI）の算定が難しく、**明確なビジネス価値を示せない**と経営の支持を失いやすいです [57]。幻滅期に投資が絞られるのはこのためで、RAGも例外ではありません。
- **評価とユーザー受容**: RAGシステムの**性能評価指標が曖昧**なまま導入されたケースが多く、結果として**効果検証ができず**改善サイクルが回らない事態が見られます [39] [63]。また、エンドユーザーへの教育や期待調整が不十分だと、初期の誇大宣伝との落差により**ユーザーが失望し利用しなくなる**リスクもあります [32] [64]。

総じて、**RAGは技術的ポテンシャルは高いものの、周辺要素（データ・人・プロセス）が追いつかず期待倒れになっている**のが現状です [2] [5]。特に日本においては、社内文書の電子化・構造化の遅れや、検索エンジン技術の社内知見不足も影響していると考えられます（グローバル企業に比べデータ整備に遅れがある傾向）。しかし一方で、幻滅期とは**進歩の停滞を意味するものではなく、必要な教訓を得て次の段階へ備える時期**でもあります [29] [6]。

7.2 将来展望：RAGは復活し啓発期へ向かうか

今後の展望として、**RAGは十分に幻滅期を脱し「啓発期」へ進み得る**と考えられます。その楽観の根拠は、**既に改善の兆しや新技術の投入が始まっている**ことです。

まず技術面では、前章で触れた**検索アルゴリズムの高度化（Graph　RAG等）**や**LLMとの統合強化（エージェント的動作、結果検証）**などが着実に進むでしょう [104] [78]。これによりRAGの精度・信頼性はワンランク向上し、**「RAGでもう一歩届かなかったケース」が解消**されていく可能性があります。例えばGraphRAGは現在研究段階ですが、成功すれば**検索の抜け漏れを大幅に減らし**、RAGによる回答の完全性を高めることが期待できます [73]。また、大手クラウドベンダーが次々RAG機能を自社サービスに組み込んでいる流れも見逃せません。Microsoft、Google、AWSなどがエンタープライズ向けに**「AI + 検索」のソリューション化**を進めており、これらが標準化すれば導入ハードルは下がります [74]。つまり現在は各社が手探りで構築していたもの

が、**製品・サービスとして洗練**されていき、安定度が増すでしょう。Gartnerも「**現在のAIの進化はインターネット初期の進化に似ている**」と言及し、今後ベンダー投資が活発なことで技術成熟は早まると見ています[105][106]。RAGに関しても、多くのSIerやスタートアップがこの領域に参入しノウハウを蓄積しているため、**1〜2年後には成功率が飛躍的に高まっている**可能性があります。

次に組織・運用面では、幻滅期の今だからこそ各企業は**人材開発と体制作りの重要性**を認識し始めました[107][108]。単に技術導入するだけでなく、それを活かす**人材のスキル・マインドセット**が鍵と理解されてきています[109][108]。例えば、RAG運用には**データエンジニア**や**ナレッジマネジメント担当**が欠かせない、という認識が広まりつつあります。また「失敗を許容し学習を重視する文化」が必要とも指摘されており[30][110]、IT部門のみならず現場を巻き込んだ**学習する組織**への変革が進めば、技術の習熟度も上がります。**幻滅期の技術への継続投資が将来の優位を生む**という洞察[111][6]も示されているため、先進企業ほど中長期視点でRAG改善に取り組むでしょう。実際、「この時期に改善を続ける企業は、競合他社が撤退する中で技術優位性を確立できる」との分析もあります[111]。そうした動きが増えれば、RAG全体の評価も徐々に持ち直していくはずです。

総合すると、**RAGは決して行き止まりの技術ではなく、むしろ生成AIを実用化する上で避けて通れない王道**です。その必要性（LLMの知識不足を補う）は揺るがず、課題はあるもののイノベーションによって解決可能な範囲と考えられます。GartnerもRAGを完全に見限ったわけではなく、**「適切な取り組みを継続すれば実用的な価値を創造できる可能性がある」**としています[112]。幻滅期はむしろ**次の「啓発」に向けた助走期間**と捉え、各組織が地力を養うタイミングと言えます。

7.3 企業がRAG導入を検討する際の留意点

最後に、これからRAGを導入・活用しようとする企業へのアドバイスをまとめます。幻滅期の教訓を踏まえ、以下のポイントに注意すると良いでしょう。

- **(A) 過度な期待の排除と現実的ゴール設定**: 経営者やユーザーに対して、RAGの**得意な部分と限界**を最初に正しく説明しましょう。**「完璧ではないが使い方次第で有用」**というニュアンスを共有し、具体的な目標指標（例：回答精度○%向上、問い合わせ対応時間△%削減）を設定して合意形成することが大切です[64]。成功している企業はこの期待値管理が上手く、トップダウンの号令とボトムアップの現実調整を両立しています。

- **(B) ユースケースの明確化と段階的導入**: RAGは万能ツールではないため、まず**効果が出やすいピンポイントの用途**から始めるのがおすすめです[94][113]。社内FAQ検索、マニュアル自動要約、特定商品の問い合わせ対応など、勝ちパターンを作れそうな領域に絞りましょう。一度成功例を作ってから徐々に範囲を広げていく**スモールスタート＆スケール**の戦略がリスクを下げます[103]。PoC段階では限定ユーザーで試し、フィードバックを取り入れて改善し、それから本格展開する流れが望ましいです[88]。

- **(C) データ整備と検索チューニングの重視**: **データクオリティは命**です。導入前に社内データの棚卸しを行い、必要情報がどこにあるか把握しましょう。その上で、**クレンジング（ノイズ除去）やメタデータ付与**を徹底し、検索インデックスには**厳選した信頼できる文書のみ**を投入するくらいの慎重さが必要です[68][42]。「迷ったら入れておけ」ではなく「必要なものだけ入れる」スタンスが成功率を高めます[69]。また導入後も検索ログを分析し、ヒットしないクエリに対するデータ追加や、逆に不要データの除去など**継続的なインデックス調整**を行う体制を作りましょう[114][39]。検索エンジンの専門家やデータサイエンティストがチームにいると理想的です。

- **(D) KPI設定と効果測定の仕組み**: **計画段階から評価指標を定め**、導入後はそれをトラッキングして経営にレポートする仕組みを用意しましょう[39]。例えば**正答率や回答所要時間、ユーザー満足度**などを定量調査し、ベースラインと比較してどれだけ改善したか示せるようにします[39]。また、「月に

何件の問い合わせ対応を自動化でき、その分コスト削減」というように**ビジネス指標（コスト・時間・売上機会）へのインパクト**も測定しましょう。数値化が難しい場合は、ユーザーアンケートや事例集をまとめ**定性的価値**も伝えます。評価軸が定まれば改善目標も立てやすく、PDCAを回す土台になります [39] [63]。

- **(E) ユーザー教育とUI上の工夫**: エンドユーザーには**RAGの使い方や限界**を周知し、誤回答があり得ること、その際の報告フィードバック方法などを教えます。**「AIの回答をうのみにせず、参考情報として扱う」**というリテラシーを醸成しましょう [30]。UI上も、可能なら**回答と一緒に根拠となる文書やリンクを表示する**ようにすると信頼度が増します [101]。例えば「この回答は社内マニュアルXの内容に基づきました」と示せば、ユーザーは自分で検証できますし、AI回答への不信感も和らぎます。また、ユーザーからのフィードバックを簡単に送れるボタンなどを設け、**現場の声を拾う仕組み**も取り入れてください [115] [39]。

- **(F) 継続的改善と専門チームの設置**: RAG導入は始まりに過ぎず、その後の**継続改善が成功のカギ**です [116]。導入したら放置ではなく、**専門の改善チーム**を置き、ログ分析・モデル更新・ナレッジ追加などを回すようにします [114]。半永久的に続ける必要はないかもしれませんが、少なくとも最初の6ヶ月〜1年は手厚くモニタリングし、システムが安定して期待通り動くようになるまで**チューニングと組織学習**を繰り返しましょう [83] [77]。このフェーズを乗り越えれば、あとの運用は徐々に省力化できるはずです。**幻滅期技術への取り組みでは長期視点と失敗から学ぶ姿勢が不可欠**との指摘もあります [117] [116]。腰を据えた改善努力こそ、最終的な成功と競争優位につながります [111] [6]。

結論: 現在RAGは幻滅期に差し掛かり、一時的に期待が萎んでいる状況ですが、その基本的価値（**LLMと知識を結びつける戦略的重要技術**）は揺らいでいません [5] [118]。むしろこの幻滅期を経て、各企業・業界がより賢明な活用法を身につけることで、次の啓発期には**RAGは実務に欠かせないインフラとして定着**していくでしょう [103]。RAGの現状課題は多角的ですが、それらは十分対処可能であり、既に解決に向けた道筋も見え始めています [37] [73]。企業にとって重要なのは、短期的な過剰期待や失望に振り回されず、**技術の成熟度を正しく見極めて自社の導入戦略・タイミングを判断すること**です [119] [120]。Gartnerのアナリストも「過度な期待や過小評価に陥らず、自社に合った導入戦略と展開のタイミングを見極める必要がある」と助言しています [121]。RAG導入を検討する企業は、ぜひ幻滅期に浮上した課題から学び、現実的かつ戦略的なアプローチで取り組んでください。**いま蒔いた種は、やがて啓発期に大きな果実をもたらす**はずです。

最後に、本レポートの内容を端的にまとめると：**「RAGは一度幻滅を経験した。しかしそれは真の実用化への通過点に過ぎない。課題を直視し克服することで、RAGは再び脚光を浴び、企業の競争力を高める武器となるだろう」**という展望が導き出せます。 [6] [111]

引用・参考資料： - Gartnerジャパン「日本におけるクラウドとAIのハイプ・サイクル：2025年」関連発表 [1] [122] - 吉川剛史「Gartnerハイプサイクル2025が示す未来図」※note記事 [3] [7] [6] - NTT東日本 技術コラム「RAGの精度向上術：企業導入の課題と解決策」 [32] [39] - IT-daytrading「RAG導入の成功と失敗の分岐点」※note記事 [64] [41] [44] - BCG高柳氏「生成AIの精度を高めるRAG」 [38] [9] - CloudNativeブログ「企業のRAGにおける課題点、グラフRAGへの期待」 [52] [37] [73] - Dan Giannone「The Non-Technical Challenges with RAG」※Medium記事 [59] [68] [63] - TechFirmブログ「企業におけるRAG活用事例10選」 [87] [90] [103] - その他、OpenAI/Morgan Stanley事例紹介、各社プレスリリース、業界メディア記事 等 [99] [101]

---

[1] [27] [119] [121] [122] クラウドとAIのハイプサイクル、黎明期にエージェンティックAIやMCPなど17項目をプロット―ガートナー | IT Leaders
https://it.impress.co.jp/articles/-/28217

2 3 4 5 6 7 28 29 30 56 57 86 105 106 107 108 109 110 111 112 116 117 118 120 Gartnerハイプサイクル2025が示す未来図 ― 過熱と幻滅を乗り越える組織変革の羅針盤〜「AI時代の人材開発・組織開発」（151）｜吉川剛史
https://note.com/take_yoshikawa/n/n44f0d374a5b9

8 11 12 13 14 15 16 33 34 35 75 76 77 82 83 〖論文瞬読〗RAGシステムの実装に隠れた7つの落とし穴と教訓〜実践する上で知っておくべきこと〜｜AI Nest
https://note.com/ainest/n/nf911000afd6c

9 18 23 26 38 生成AIの精度を高める「RAG」 | BCG Japan
https://bcg-jp.com/article/4412/

10 17 19 20 22 87 88 89 90 91 103 企業におけるRAG活用事例10選
https://www.techfirm.co.jp/blog/rag-case-study

21 47 LLM・RAGのビジネス導入の落とし穴「回答精度が期待 ... - Laboro.AI
https://laboro.ai/activity/column/laboro/ragllm/

24 25 31 49 50 79 80 81 The State of RAG in 2025: Bridging Knowledge and Generative AI
https://squirro.com/squirro-blog/state-of-rag-genai

32 39 42 84 114 115 RAGの精度向上術：企業導入の課題と解決策 | コラム | クラウドソリューション｜サービス｜法人のお客さま｜NTT東日本
https://business.ntt-east.co.jp/content/cloudsolution/column-661.html

36 RAG 2.0 深入解读 - 知乎专栏
https://zhuanlan.zhihu.com/p/1903437079603545114

37 51 52 53 54 72 73 74 104 企業のRAGにおける課題点、埋め込みコストや精度向上とグラフRAGへの期待 – CloudNative Inc. BLOGs
https://blog.cloudnative.co.jp/26709/

40 41 43 44 46 48 64 71 RAG導入の成功と失敗の分岐点：過剰な期待とデータ前処理の重要性｜IT-daytrading
https://note.com/it_daytrading/n/nc541fee0217d

45 「RAG」は本当に簡単？見えない落とし穴と成功への道筋
https://rag-and.com/news/_Gl4dcnW

55 人見悠大 | IT PM/PMO × 開発 on X: "ガートナーのハイプ・サイクルで
https://x.com/i/status/1956259801274704292

58 59 60 61 62 63 65 66 67 68 69 70 The Non-Technical Challenges with RAG | by Dan Giannone | Medium
https://medium.com/@DanGiannone/the-non-technical-challenges-with-rag-e91fb165565e

78 Agentic RAG systems for enterprise-scale information retrieval - Toloka
https://toloka.ai/blog/agentic-rag-systems-for-enterprise-scale-information-retrieval/

85 ベーシックAzureOpenAI RAG精度改善支援サービス：8週間 実装
https://azuremarketplace.microsoft.com/en-us/marketplace/consulting-services/1614659899734.basic_aoairag

92 93 94 東京ガスの変革「第3の創業」と生成AIの融合 | デロイト トーマツ ...
https://toyokeizai.net/articles/-/853105

95 生成AIを搭載した社内アプリを独自開発・利用開始 - 東京ガス
https://www.tokyo-gas.co.jp/news/topics/20241010-02.html

96 WinActor®導入事例【東京ガス株式会社】「東京ガス」様 ...
https://winactor.com/case/infra/tokyo-gas-201805/

97 How Morgan Stanley Is Training GPT To Help Financial Advisors
https://www.forbes.com/sites/tomdavenport/2023/03/20/how-morgan-stanley-is-training-gpt-to-help-financial-advisors/

98 Morgan Stanley Research Announces AskResearchGPT
https://www.morganstanley.com/press-releases/morgan-stanley-research-announces-askresearchgpt

99 I made a secure GPT-4 for my company knowledge base. - Reddit
https://www.reddit.com/r/Entrepreneur/comments/16gnmaw/i_made_a_secure_gpt4_for_my_company_knowledge_base/

100 Key Milestone in Innovation Journey with OpenAI - Morgan Stanley
https://www.morganstanley.com/press-releases/key-milestone-in-innovation-journey-with-openai

101 Contextually Enriched, Knowledge-Enhanced, and Externally ...
https://medium.com/@adnanmasood/contextually-enriched-knowledge-enhanced-and-externally-grounded-retrieval-models-for-fun-7620dd9f643f

102 Morgan Stanley uses AI evals to shape the future of financial services
https://openai.com/index/morgan-stanley/

113 “利用者の数”と“課題解決の深さ”の2軸で 取り組む東京ガスの生成AI ...
https://www.sedesign.co.jp/ai-blog/interview-tokyogas