

RAG は本当に「幻滅期」か——Gartner¹ の 2025 年版ハイプサイクル発表から読み解く深掘りレポート

Genspark

要約

- 2025 年 8 月、ガートナー・ジャパンは「日本におけるクラウドと AI のハイプサイクル：2025 年」を発表し、RAG (Retrieval-Augmented Generation／検索拡張生成) を「幻滅期 (Trough of Disillusionment)」に位置づけた。一次情報と国内主要メディアの報道を精読すると、RAG が実運用で直面している検索精度・評価・運用複雑性・コストといった構造的課題が、過度な期待とのギャップを顕在化させたことが背景にあると読み取れる。ただし、当該発表は「RAG が幻滅期に入った」という位置づけを示すにとどまり、個別の直接理由や定量指標は明示していない点には留意が必要である。ITmedia @IT²／Gartner プレスリリース ¹／ITmedia エンタープライズ ³
- RAG は、LLM の出力を外部知識で「根拠付け」る枠組みとして 2023～2024 年に急速に普及したが、実際にはリトリバの適合度不足、チャンク設計・再ランク・権限管理・増分更新などの運用複雑性、ベクトル DB・再ランク・長文 LLM の推論に伴うレイテンシとコスト、そして品質評価の難しさが導入障壁となった。一方で、適切に設計・評価・MLOps を整えた事例では問い合わせ削減やドキュメント QA の精度改善などの価値が報告されており、RAG は「効く領域では確かに効く」が、「万能ではない」技術として現実的な位置に収まりつつある。Zilliz blog⁴／Qiita (RAG 評価) ⁵／Nikkei BP (64%削減事例) ⁶
- 幻滅期脱却に向けては、(1) リトリバの高度化 (ハイブリッド検索、学習済み再ランク、知識グラフ統合＝GraphRAG)、(2) 評価フレームワークの標準化・自動化 (RAG Triad／RAGAS／ARES／TruLens)、(3) 長文コンテキスト LLM や選択的生成 (Sufficient Context) の実装、(4) 権限・PII・更新差分などの RAG MLOps／データガバナンス整備、(5) キャッシュと近似探索のチューニングによる SLA とコスト最適化、(6) ドメイン用途への要件定義と A/B 評価の定着が鍵となる。これらは既に産業界・アカデミアで具体的な枠組みとベストプラクティスが蓄積しつつあり、RAG が啓発期に進むための実装論は見えつつある。Google Cloud blog⁷／arXiv 2407.13193 (RAG サーベイ) ⁸／arXiv 2501.13958 (GraphRAG サーベイ) ⁹

- 企業導入の勘所は、(a)「ハルシネーションの完全解決」などの過度期待を排し、RAGが貢献する KPI（一次回答率、平均応答時間、一次解決率、根拠提示率など）を明確化、(b) データ整備・アクセス制御・更新運用を前提設計に組み込み、(c) 検索・生成・再ランク・評価を小さく作って高速に計測・改善、(d) コスト試算と SLA 設計を初期から可視化し「部分最適」ではなく「全体最適」を志向する姿勢である。
Beatrust techBlog[10](#)／Google Cloud blog[7](#)

1. 提示記事の精読：ガートナーが RAG を「幻滅期」に置いた背景を把握

今回の国内一次報道である@IT 記事は、Gartner が「日本におけるクラウドと AI のハイプサイクル：2025 年」を公表し、34 の要素を三観点（AI と産業革命関連／クラウド関連／マイグレーション関連）に整理したうえで、RAG とプラットフォームエンジニアリングが「幻滅期」に移行したと明示している。ただし、記事自身は RAG を幻滅期とした直接の理由や定量的根拠を詳述していない。Gartner プレスリリースおよび他メディアの補足でも、「RAG が幻滅期に入った」という位置づけは一致しているが、個別の失敗要因の列挙や実証データは提示されていない。従って、理由の解像度は一次発表資料では限定的であり、実運用・学術の知見から補完的に読み解く必要がある。ITmedia @IT[2](#)／Gartner プレスリリース [1](#)／ITmedia エンタープライズ [3](#)／Impress Watch[11](#)

2. 基本概念の定義と解説

(a) RAG (Retrieval-Augmented Generation)

RAG は、大規模言語モデル (LLM) に対し、外部のナレッジベース（ドキュメント、FAQ、データベース、ナレッジグラフ等）からクエリに関連する情報を検索 (retrieval) して与え、その根拠に基づいて回答 (generation) させるアーキテクチャである。典型的なパイプラインは、(1) 質問の埋め込み化、(2) ベクトル DB や BM25 等で類似チャンクの取得、(3) 再ランクで関連度の高い根拠を選別、(4) LLM へプロンプトとして根拠を与え生成、(5) 引用や根拠提示を返す、という流れである。目的は、最新・内部・専門の知識を動的参照し、LLM の「曖昧な一般知識」に頼らない信頼性・鮮度・コンプライアンスを高める点にある。利点として、モデル再学習を要せずに知識更新できること、根拠提示 (cite) により説明可能性を改善できること、データ境界・権限に基づくガバナンスを設計しやすい点が挙げられる。
arXiv 2407.13193 (RAG Survey) [8](#)／Zilliz blog (評価概観) [4](#)／Google Cloud blog (RAG 最適化) [7](#)

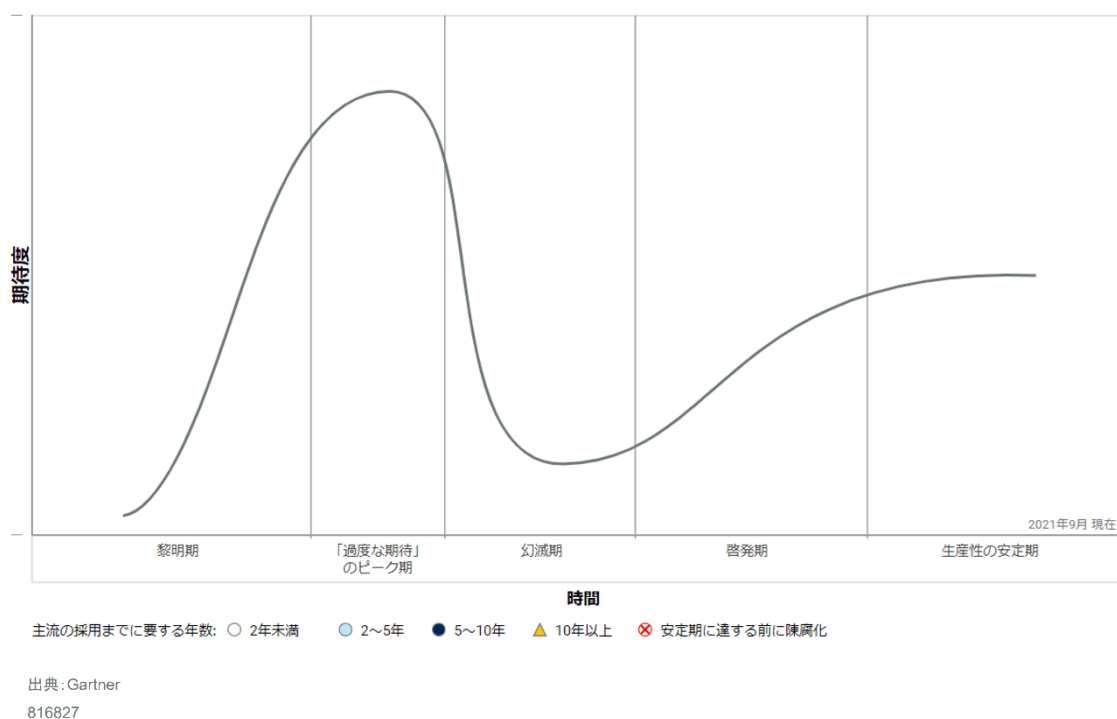
(b) ガートナーのハイプサイクルと「幻滅期」

ガートナーのハイプサイクルは、新興技術が市場に普及・定着するまでの一般的な認識曲線を「黎明期」「過度な期待のピーク期」「幻滅期」「啓発期」「生産性の安定期」の 5 段階に区

分したフレームである。幻滅期は、初期の過剰な期待がしぼみ、PoC や導入の失敗、投資対効果の遅れ、採用ペースの減速などにより幻滅感が広がる段階と定義される。他方で、この時期を通じて実装の現実解が見え、早期採用組の学びを経て実用価値が再評価されると、啓発期へと向かう。Gartner（方法論解説）[12](#)

参考図版（ハイクサイクル）：

ハイク・サイクル



Gartner

出典: Gartner（方法論）[12](#)

3. RAG が幻滅期にあるとされる具体的な技術的・運用的課題

第一に「検索精度」の問題がある。多くの現場で、(i) チャンク分割の粒度、(ii) 埋め込みモデル選定、(iii) 近似最近傍探索のパラメータ、(iv) BM25 とのハイブリッド、(v) 再ランクの有無、といった設計選択が回答品質に大きく影響し、再現性ある高精度を安定的に出すことが難しい。関連性の低いチャンクが混入すれば、生成での要約エラーやハルシネーションを誘発する。包括的サーベイは、RAG の性能がリトリバの設計とデータ品質に支配されること、そして評価フレームの整備が不可欠であることを強調している。arXiv 2407.13193 (RAG Survey) [8](#) / Zilliz blog (評価概観) [4](#) / Google Cloud blog (RAG 最適化) [7](#)

第二に「導入・運用の複雑さ」。RAG は、検索・再ランク・生成の各コンポーネントに加え、

社内データの権限管理（行・列・フィールドレベル）、マルチテナント、PII/機微情報の保護、継続的なデータ更新・増分インデクシング、メタデータ設計、キャッシュ戦略、監査ログ、ガードレールといった実装論が不可欠になる。結果として、PoC での成功を本番 SLA に乗せる際に、性能劣化やセキュリティ・運用リスクが顕在化しやすい。国内の実装ガイドは権限・セキュリティの設計や更新運用の難しさを繰り返し指摘している。Liskul (RAG とは・権限管理の複雑化) [13](#) / amie-ai (RAG の課題と対策) [14](#) / NTT 東日本コラム (導入ステップ) [15](#)

第三に「コストとレイテンシ」。ベクトル DB、再ランク（クロスエンコーダ等）、長文コンテキスト LLM の組合せは、推論コストと待ち時間を増大させやすい。スループットやピーク時の同時実行を考慮したインフラ設計では、クラウド課金・ストレージ・ネットワークも無視できない。対策として、近似探索のチューニング、キャッシュ（クエリ・コンテンツ・根拠）、サマリ前処理、営業時間外の停止などが紹介されるが、ビジネス要件とのトレードオフを伴う。NHN テコラス (Aurora をベクトル DB に・運用コスト工夫) [16](#) / G-gen (主要クラウドの RAG 構成比較とコスト示唆) [17](#) / EXIIS Lab (導入・運用コストの整理) [18](#)

第四に「評価の難しさ」。RAG は検索と生成を跨ぐため、単一指標では品質を捉えづらい。RAG Triad（コンテキストの適合性、回答の忠実性、回答の関連性）や、RAGAS/ARES/TruLens といったフレームが提案され、根拠の正当性や引用一貫性などを定量化する流儀が広がっているが、データセット・タスク依存性が強く、オフライン指標とオンライン満足度のギャップも課題だ。評価を前提にした開発体制（評価駆動の反復）が普及途上である点も、品質安定化の阻害要因となる。Google Cloud blog (RAG Triad) [7](#) / Qiita (RAG 評価フレーム概観) [5](#) / Zilliz blog (評価の全体像) [4](#)

4. 過度な期待（例：ハルシネーションの完全解決）と実運用のギャップ

RAG は「ハルシネーション低減の特効薬」として期待を集めたが、実際には、(i) 取得根拠が不十分／不適合、(ii) LLM が根拠を正しく解釈できない、(iii) プロンプト設計が不十分、といった状況では依然として誤答が生じる。最新研究は「十分な文脈 (Sufficient Context)」が存在する場合に選択的生成でハルシネーションを低減できることを示すが、これは「条件付きの改善」であり、完全解決ではない。現場では「重要質問ほど根拠を厳密に要求」「回答不能時は明示する」「引用必須」などの運用ガードが必要になる。すなわち、RAG は「よく効くが万能ではない」技術であり、設計・評価・運用の成熟が前提となる。Sufficient Context (解説) [19](#) / Allganize 技術解説 (RAG とハルシネーション) [20](#) / arXiv 2407.13193 (RAG Survey) [8](#)

一方で、カスタマーサポートや社内 QA など明確に文書主導のドメインでは顕著な成果も報告されている。Nikkei BP のまとめでは、従来の FAQ を超える応答速度や柔軟性ととともに、導入翌月に問い合わせ件数を 64%削減した事例が紹介されている。これは FAQ 整備や自己解決率向上のコンテキストで RAG がうまく機能した例だが、こうした大幅削減は要

件・データ成熟・運用体制の良好な条件が揃った結果であり、一般化には慎重さが必要である。Nikkei BP（問い合わせ 64%削減の紹介）[6](#)／helpfeel（CS 活用解説）[21](#)

5. 幻滅期を脱し「啓発期」へ移行するための技術革新・アプローチ

第一に、リトリバの高度化である。BM25 とベクトル類似のハイブリッド、dense+cross-encoder の二段再ランク、クエリ拡張、セマンティック・ネガティブの活用、そして知識グラフ統合（GraphRAG）が、リトリバ精度と根拠一貫性の改善に寄与する。GraphRAG の包括的サーベイは、論理整合性・因果・関係性の強化に有効な設計パターンを整理している。これらは RAG の「検索→生成」の上流品質を底上げする基盤となる。Google Cloud blog（最適化論）[7](#)／arXiv 2501.13958（GraphRAG Survey）[9](#)

第二に、評価フレームの標準化・自動化だ。RAG Triad に沿って、コンテキスト適合・回答忠実性・回答関連性を自動測定し、RAGAS や ARES、TruLens 等でリグレッション可能なメトリクスを整備、オンライン A/B と結びつける。RAGCHECKER のような微粒度診断フレームは、どの段が劣化要因か（retrieval か、selection か、generation か）を局所化する助けになる。評価駆動の反復（E2E の CI/CD）を習慣化し、モデル更新・データ更新・プロンプト変更の影響を継続監視することが啓発期の原動力になる。Qiita（RAG 評価フレーム）[5](#)／Zilliz blog（評価の全体像）[4](#)／RAGCHECKER（解説ブログ）[22](#)

第三に、モデル側の進展を取り込む。長文コンテキスト LLM は根拠の落ちこぼしを減らし、選択的生成（Sufficient Context）や「回答不能」ポリシーの厳格化は誤答回避に有効だ。さらに、Agentic RAG（分解→検索→検証→合成の自己改善ループ）などの手法は、探索と根拠検証の反復で安定性を高めうるが、計算コスト増の管理が必要になる。Sufficient Context（解説）[19](#)／Zenn（Agentic RAG 解説）[23](#)

第四に、RAG MLOps／データガバナンスの実装。差分取り込み、アクセス制御（ABAC/RBAC）、監査・PII 処理、メタデータ標準、可観測性（トレース、根拠ログ、評価ダッシュボード）、コスト SLO（p95/p99 レイテンシ）を初期から設計に組み込む。クラウド各社の RAG 参照アーキテクチャやベストプラクティスを横断的に比較し、要件起点で選択することが重要である。G-gen（クラウド比較）[17](#)／Liskul（権限設計）[13](#)

6. 導入事例：成功と課題の比較分析

成功側の代表例として、問い合わせの自己解決率向上・一次解決率の改善が挙げられ、Nikkei BP のまとめでは、導入翌月に 64%の問い合わせ削減に至ったケースが紹介される。こうした成果は、(i) ナレッジの網羅性・最新性、(ii) 検索・再ランクの高適合、(iii) 厳格な引用・根拠提示、(iv) ワークフロー統合（エスカレーション・チケット化）といった設計が揃った場合に顕在化しやすい。逆に、データが未整備でアクセス制御が曖昧、更新が追いつかない、評価が属人的、といった状態では、誤答や不信が蓄積し、運用定着に失敗しやすい。現場のブログや導入記事は、成功・失敗の分岐点として「データ前処理と評価体制」を繰り返し指

摘している。Nikkei BP6/Macnica (失敗からの学び) 24/helpfeel (CS 活用解説) 21
また、特定企業の事例紹介 (CS、社内ナレッジ、FAQ 高度化など) は国内で多数流通しているが、その多くはベンダー発表に依拠するため、測定指標の定義やベースライン、観測期間、母集団の偏りに注意が必要である。例えば「問い合わせ削減率」一つ取っても、(a) 同期間の FAQ 整備・UI 改善など他施策の影響、(b) 季節要因、(c) カテゴリ・チャネルのミックス変更、(d) 内部エスカレーションの再配分、などの交絡がありうる。従って、RAG の純粋効果を評価するには、A/B テストや時系列因果推定などの厳密な枠組みが望ましい。
Nikkei BP6/Zilliz blog (評価の全体像) 4

事例の比較観点 (簡易表)

観点	成功事例で見られた特徴	課題事例で頻出の落とし穴
データ	網羅・重複解消・最新化・メタ付与	断片化・更新遅延・権限曖昧
検索	ハイブリッド+再ランクで高適合	単一手法・パラメータ未調整
生成	引用必須・回答不能ポリシー	根拠欠落・要約過剰・誤断定
評価	RAG Triad+RAGAS/ARES 自動化	主観レビューのみ・回帰不能
運用	変更管理・観測・SLO 運用	PoC 止まり・運用 SLA 未設計
成果	問合せ削減・一次解決率上昇	信頼性低下・棚卸しで停止

出典の一例 (評価・運用の枠組み): Google Cloud blog7/Qiita (RAG 評価) 5/Zilliz blog4

7. 統合的評価: RAG の現状・課題・展望と、企業が考慮すべき点

総じて、Gartner が示す「RAG は幻滅期へ」という位置づけは、RAG が「期待のピーク」を過ぎ、実装と運用の現実解に直面している地合いを反映している。一次資料は直接理由を列挙していないが、現場・研究の蓄積からは、検索精度・評価の難しさ・運用複雑性・コストという四点が構造的課題として共通して観測される。これらは RAG 固有の弱点ではなく、「情報検索×生成」という二段構えのシステムが本質的に抱える設計・評価・運用の難度を示す。裏返せば、ここを技術とオペレーションで解きに行く潮流 (GraphRAG、再ランク、Triad 評価、自動評価、長文 LLM、選択的生成、MLOps) は既に整いつつあり、「啓発期」に向けた実装論は輪郭を得ている。ITmedia @IT2/Gartner (方法論) 12/arXiv 2407.131938

企業が導入を検討する際の実務ポイントは次の通りだ。第一に、ユースケースを定義し、KPI を RAG Triad およびビジネス指標 (一次回答率、CSAT、解決時間、Escalation 率、根拠提示率) で紐づける。第二に、データ整備とアクセス制御、増分更新の設計を先行し、PoC でも本番同等の制約を模擬する。第三に、検索・再ランク・生成・評価を小さく組み合わせ、RAGAS/ARES/TruLens 等の自動評価を導入して回帰可能な改善サイクルを回す。第四に、性能とコストのトレードオフを可視化し、キャッシュや近似探索、回答ポリシーで SLA を満たす。第五に、引用必須・回答不能の明示・根拠表示 UI など、信頼性をユーザー体験に

組み込む。これらを満たした領域からスケールさせることが、幻滅期を機会に変える実装戦略である。Google Cloud blog⁷/Zilliz blog⁴/Qiita (RAG 評価) ⁵

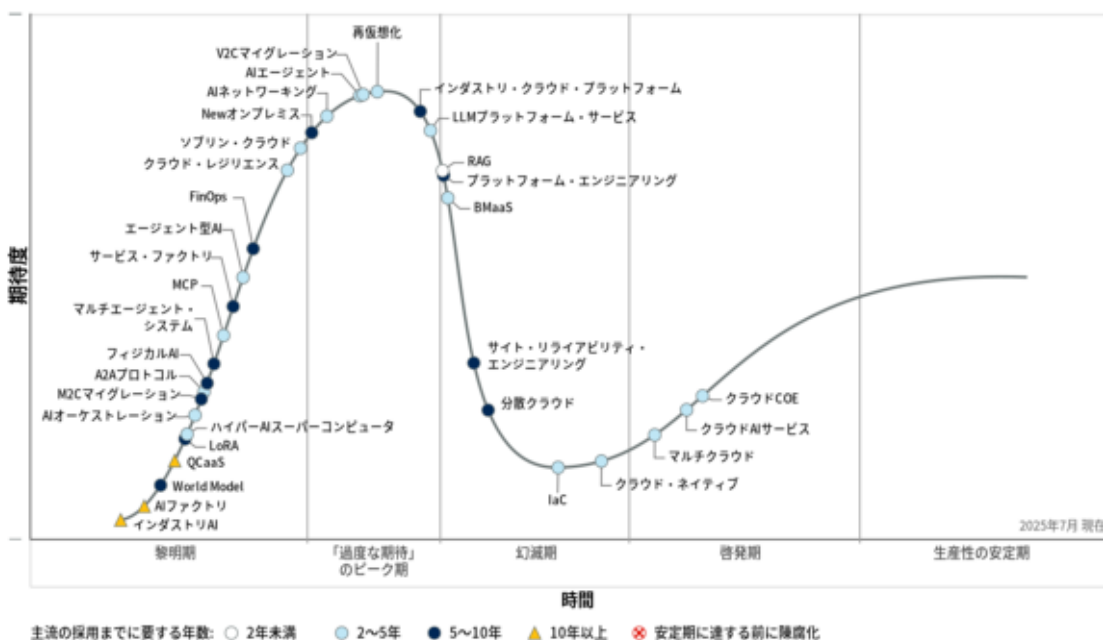
参考：今回発表の位置づけと関連図版・報道

- RAG とプラットフォームエンジニアリングが「幻滅期」へ移行と明記 (国内報道)。ITmedia @IT²
- 発表の一次情報 (34 要素の俯瞰と三観点の整理)。Gartner プレスリリース ¹
- 関連報道まとめ (AI エージェントへの過度な期待への注意喚起)。ITmedia エンタープライズ ³/Impress Watch¹¹

参考図版 (国内報道の図・画像)

- @IT 記事の図版 (日本におけるクラウドと AI のハイプサイクル：2025 年)：

日本におけるクラウドとAIのハイプ・サイクル：2025年

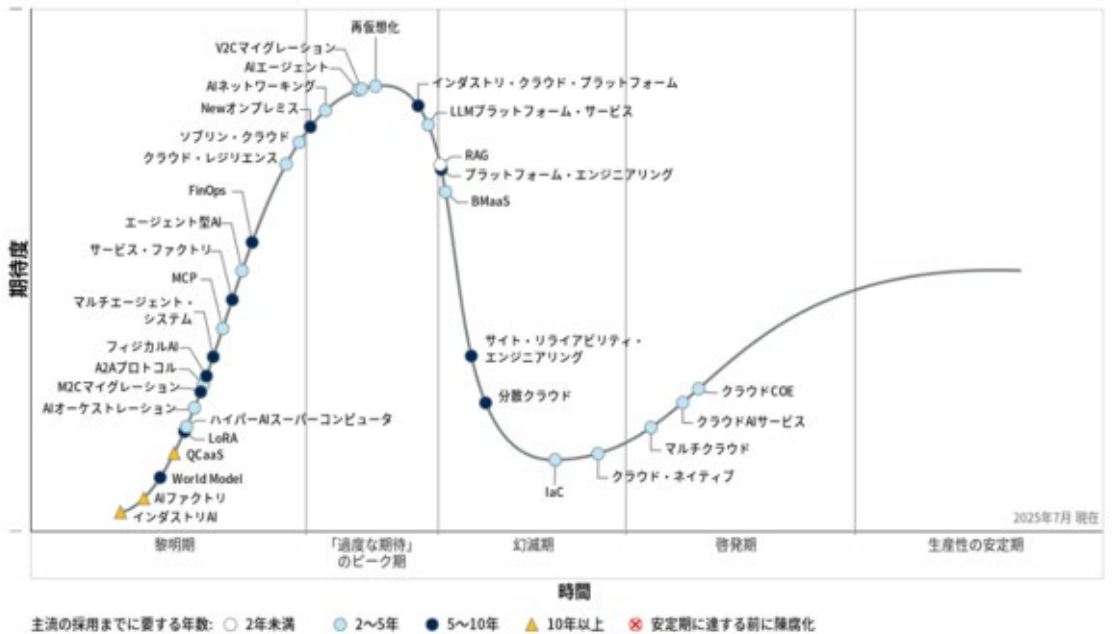


Gartner

出典: ITmedia @IT²

- ITmedia エンタープライズの図版:

日本におけるクラウドとAIのハイブ・サイクル：2025年

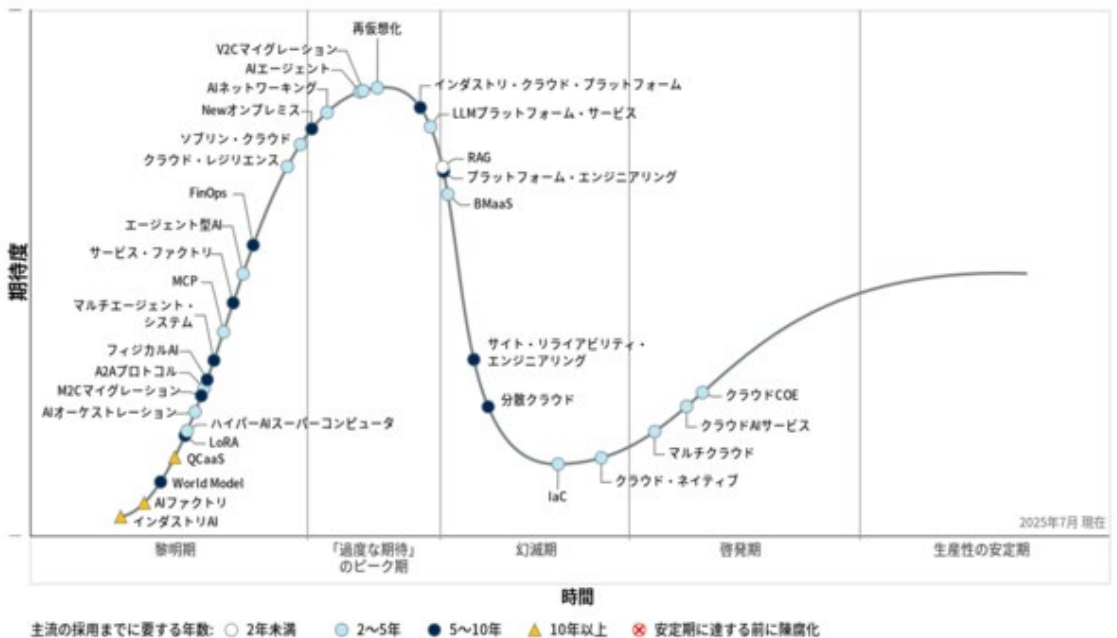


Gartner

出典: ITmedia エンタープライズ [3](#)

- Impress Watch 記事内の図版（ハイブサイクル図）：

日本におけるクラウドとAIのハイブ・サイクル：2025年



Gartner

出典: Impress Watch [11](#)

補足（定義・評価・サーベイへの導線）

- RAG の包括的サーベイ (2024 年)。arXiv 2407.13193[8](#)
- RAG 評価の総覧 (RAG Triad, RAGAS, ARES, TruLens)。Google Cloud blog[7](#)／Qiita (比較) [5](#)／Zilliz blog[4](#)
- GraphRAG (知識グラフ統合) サーベイ。arXiv 2501.13958[9](#)
- Sufficient Context (選択的生成でハルシネーション低減)。解説記事 [19](#)

結論

Gartner の「幻滅期」入りは、RAG の価値が無いという意味ではない。むしろ、RAG を“使いこなす”ための設計・評価・運用の成熟が問われる段階に入ったという宣言だ。ここから啓発期に向かう鍵は、ハイプサイクルの基本に立ち返り、早期採用組の学びを標準化し、測れる形で改善を積み上げることにある。ユースケースを限定し、データと評価を中心に据え、検索と生成の両輪を最適化する。そうして初めて、RAG は「過度な期待」を抜け出し、「持続的な価値」を生む企業基盤として定着していくはずだ。Gartner (方法論) [12](#)／ITmedia @IT[2](#)

(注) 本レポートは、一次発表資料・国内主要報道・学術サーベイ・技術ブログ等の公開情報に基づき、RAG の現状と実装論を多角的に整理した。Gartner の当該発表は、RAG を「幻滅期」と位置づけるものの、個別要因の列挙や定量データは提示しないため、理由部分は実運用・学術の共通知見で補完している点を明記する。Gartner プレスリリース [1](#)／ITmedia @IT[2](#)／Impress Watch[11](#)