AIのハルシネーションを防ぐ実践ガイド

誤回答を減らす設計・運用・検証



Start Challenge
Consulting co., ltd.
挑む力が、世界を照らす

生成AIの誤回答を減らすための設計・運用・検証ガイド

RAG・プロンプト・ガードレール・人による確認を組み合わせた多層防御アプローチ

対象：生成AIの導入担当者／情報システム部門／業務部門／管理職
構成：全19ページ（現状認識・対策の全体像・実践方法8項目・運用設計）

Start Challenge Consulting co., ltd.
発行：2026年8月委
委Confidential

はじめに｜生成AIはなぜ「もっともらしい誤回答」をするのか

生成AIは検索ではなく「次に続く言葉」を確率的に予測する仕組みであるため、文章として自然でも内容が誤っている回答が構造的に発生する。対策の目的は完全な排除ではなく、発生確率と業務影響の抑制である。

① 入力：質問と文脈を受け取る

利用者の質問と、与えられた前提情報を単語列として受け取る。渡されていない情報は参照できない。

② 予測：次の語を確率で選ぶ

学習した統計的パターンから、次に続く確率の高い語を選び続ける。正解を照合する処理は含まれない。

③ 出力：自然な文章が完成する

文法的・文脈的に自然な文章になる。**「自然さ」と「正しさ」は別の指標**であり、自然さは正しさを保証しない。

特に誤りやすい条件

- 学習データに存在しない社内固有の情報（規程・商品・顧客）
- 改定・改正が続く最新の制度、価格、仕様
- 計算、集計、条件分岐を含む質問
- 前提や目的が示されない曖昧な質問

対策の考え方

- ハルシネーションを完全にゼロにすることを目標にしない
- **発生確率を下げる**：入力・参照情報・生成ルールを設計で制御する
- **影響を抑える**：出力の機械検証と、人による確認を業務に組み込む
- 業務のリスクに応じて、対策の厳格さを変える

本資料で分かること

- 01 ハルシネーションの7類型と、企業で発生する原因（P3・P4）
- 02 5層で構成する対策の全体像と、実践方法8項目（P6〜P15）
- 03 業務リスク別の対策レベルと運用チェック項目（P16・P17）

事実：生成AIは次に続く語を確率的に予測して文章を生成する／解釈：自然さと正しさは独立した指標である（当方の整理）／提案：発生確率の低減と影響の抑制を分けて設計する。出典：本資料は提供された構成案（全18ページ）に基づき作成。

ハルシネーションの定義と代表的な7つのパターン

ハルシネーションは「完全な作り話」だけではなく、出典の捏造・情報の混同・情報の陳腐化・解釈の誤りまでを含む。表の下段ほど文面が正しく見えるため、人による目視だけでは検知できない。

#	パターン	代表的な現れ方	検知の手がかり
1	**事実の捏造**	存在しない制度名、数字、人物、組織を実在するかのように記述する	固有名詞と数値を一次情報・マスタと照合する
2	**出典の捏造**	実在しないURL、文献、判例、製品名を根拠として提示する	URL・文献IDの存在を機械的に検証する
3	**情報の混同**	別製品・別年度・別部署の情報を一つの回答に混ぜて説明する	参照資料の対象範囲・版数を回答と突き合わせる
4	**参照範囲外の補完**	社内文書に記載のない内容を、一般知識で補って回答する	回答文と引用文の対応関係を1文ずつ確認する
5	**情報の陳腐化**	改定前の規程、旧価格、旧仕様を最新のものとして回答する	参照資料の更新日・有効期限を回答に表示させる
6	**解釈・結論の誤り**	引用自体は正しいが、条件の適用や結論の導き方を誤る	判断根拠を分けて出力させ、人が論理を確認する
7	**計算・条件分岐の誤り**	集計、税率、期間、例外条件の処理を誤って回答する	計算はAIに行わせず、外部処理で再計算して突合する

留意点　1〜2は文面から気づける場合があるが、3〜7は文章が正しく見えるため、人の目視のみでは検知できない。機械的な検証（P13）と人による確認（P14）を組み合わせる。

事実：提供された構成案に記載の7類型／解釈：類型の並びと「検知の手がかり」は当方による整理／提案：類型ごとに検証手段を割り当てる。出典：本資料は提供された構成案（全18ページ）に基づき作成。

企業でハルシネーションが起きる主な原因

原因はモデルの性能ではなく、入力条件の曖昧さと参照データの不備、および回答ルール・検証・評価という統制の欠如に集中している。いずれも設計と運用で対処可能な項目である。

A. 入力・参照データに起因する原因

① 入力（質問の与え方）

質問の条件や目的が曖昧
対象範囲・立場・判断基準が示されず、AIが前提を推測して補う。

前提情報を渡していない
社内固有の情報を渡さないまま質問し、一般知識で回答させている。

② 参照データ

データが古い・重複・矛盾
改定前の資料が残り、複数版が同時に検索対象となっている。

必要な情報を検索できていない
検索設計が不十分で、根拠となる文書がAIに渡っていない。

B. 回答ルール・検証運用の欠如に起因する原因

③ 生成ルール

「必ず回答すること」を求めている
回答保留の経路がなく、情報がなくても文章を生成させている。

出典や根拠の提示を要求していない
根拠の記述義務がないため、検証できない回答が流通する。

④ 検証・運用

生成結果をそのまま業務システムへ登録
検証処理を挟まず、誤りが基幹データや顧客連絡に流入する。

回答品質を評価する仕組みがない
誤回答が可視化されず、改善のサイクルが回らない。

8つの原因はいずれも「人と仕組みの側の設計事項」であり、モデルの入れ替えでは解決しない。原因の所在に応じて、P6の5層のどこを強化するかを決める。

事実：提供された構成案に記載の8原因／解釈：A（入力・データ起因）とB（統制欠如起因）の2群への分類は当方による整理／提案：原因群ごとに担当部門と対策層を割り当てる。出典：本資料は提供された構成案に基づき作成。

ハルシネーションが企業にもたらすリスク

同じ誤回答でも、下書き作成と法務・医療では損失の性質がまったく異なる。管理すべきは発生確率ではなく「誤った場合の影響度」であり、影響度に応じて対策レベルを設計する。

利用場面	想定されるリスク	誤回答が及ぶ範囲	影響度（解釈）
営業提案	存在しない機能や導入実績を顧客に説明する	顧客との信頼関係、受注後の実現可能性	高
カスタマーサポート	誤った手続きや料金を案内する	顧客対応コスト、返金・補償、苦情の拡大	高
人事・労務	社内規程や制度を誤って説明する	従業員の不利益、労使間のトラブル	高
法務	契約条項や法令を誤って解釈する	契約上の責任、紛争リスク、法令違反	重大
経理・財務	金額、税率、集計結果を間違える	決算数値、税務申告、社外への開示情報	重大
医療・金融	誤回答が重大な判断につながる	人の健康・資産、規制対応、事業継続	重大
社内検索	古い資料を最新ルールとして回答する	業務手順の誤り、下流業務への誤り波及	中

管理の原則　発生確率だけでなく、誤回答が発生した場合の影響度で管理する。影響度「高」以上の業務は、P16の対策レベルに従い人による確認を必須とする。

事実：提供された構成案に記載の利用場面とリスク／解釈：「誤回答が及ぶ範囲」と影響度（中・高・重大）の区分は当方による整理であり、実際の格付けは各社の業務実態に基づき見直すこと／提案：影響度を基準に対策レベルを設計する（P16）。

ハルシネーション対策の全体像｜5層による多層防御

単一の対策では誤回答を止められない。入力・参照情報・生成・出力・運用の5層それぞれに検知点を置き、前段で防げなかった誤りを後段で捕捉する多層防御として設計する。

LAYER 1 入力対策

質問と前提条件を明確にする

役割・対象範囲・回答形式・不明時の動作をプロンプトで定義し、AIに前提を推測させない。

→ 実践方法① ②(P7・P8)

LAYER 2 参照情報対策

正しいデータをAIへ渡す

RAGで社内文書を参照させ、外部システムから確定値を取得する。データ整備と検索設計が前提。

→ 実践方法④ ⑤(P10～P12)

LAYER 3 生成対策

回答ルールをプロンプトで制御する

推測の禁止、「分からない」の許可、出典表示の義務化を回答ルールとして固定する。

→ 実践方法② ③(P8・P9)

LAYER 4 出力対策

根拠確認・形式検証・禁止表現を設定する

回答を業務で使う前に、機械的に検証する処理を必ず通す。違反時は再生成または停止。

→ 実践方法⑥(P13)

LAYER 5 運用対策

人による確認・ログ分析・継続改善

承認フローとエスカレーション先を定め、評価指標で精度を継続的に測定する。

→ 実践方法⑦ ⑧(P14・P15)

質問入力 ▸ 参照情報の取得 ▸ 回答生成 ▸ 出力検証 ▸ 人による確認・業務利用 ▸ ログ分析による改善(前段で漏れた誤りを後段で捕捉する)

設計原則1｜単一の対策に依存しない

プロンプトの工夫だけ、RAGの導入だけでは不十分。各層は独立した検知点として機能させる。

設計原則2｜業務リスクに応じて対策レベルを変える

低リスク業務に全層を適用すると運用が破綻する。影響度に応じた段階設計とする(P16)。

事実:提供された構成案に記載の5層構成／解釈:各層と実践方法①～⑧の対応付け、および処理フローの図示は当方による整理／提案:層ごとに責任部門と検知点を明確化する。出典:本資料は提供された構成案に基づき作成。

実践方法①曖昧な質問とプロンプトを具体化する

曖昧な質問はAIに前提の推測を強いる。役割・対象範囲・参照先・出力形式・不明時の動作を6項目で明示することで、推測による補完の余地を構造的に減らせる。

プロンプトで指定すべき6項目

01　AIの役割を明確にする
「法務レビュー担当として」など立場と観点を固定する。

02　回答対象と対象外を指定する
扱う論点と、扱わない論点の両方を書く。

03　参照してよい情報を限定する
「添付資料の記載内容だけを根拠として」と範囲を閉じる。

04　回答形式を指定する
項目、順序、文字数、表形式などを事前に定める。

05　不明な場合の動作を定義する
「確認が必要」と記載させ、推測での補完を禁止する。

06　推測と事実を分けて表示させる
資料に基づく事実と、AIの解釈を別項目で出力させる。

プロンプトの改善例（契約書レビュー）

改善前｜条件が指定されていない

この契約書の問題点を教えてください。

▸ 立場・観点・根拠範囲が不明なため、AIが一般的な契約論を推測で補い、当該契約に存在しない条項に言及する余地が生じる。

改善後｜役割・根拠範囲・不明時の動作を指定

添付した契約書の記載内容**だけを根拠として**、**発注者側**のリスクを抽出してください。
根拠となる条文番号を示し、判断できない項目は**「確認が必要」**と記載してください。

対応する指定項目

・「発注者側のリスク」＝ 01 役割／02 回答対象
・「記載内容だけを根拠として」＝ 03 参照範囲の限定
・「条文番号を示し」＝ 04 回答形式（検証可能性の確保）
・「判断できない項目は確認が必要」＝ 05 不明時の動作

運用上の要点　プロンプトを利用者個人の工夫に委ねず、業務ごとの定型プロンプトとして管理部門が配布・版管理する。改訂時はP15のテストセットで再評価する。

事実：提供された構成案に記載の6項目および改善前後のプロンプト例／解釈：改善後プロンプトと6項目の対応付け、および改善前の失敗要因の説明は当方による整理／提案：定型プロンプトとして版管理する。

実践方法②「分からない」と回答できるルールを設定する

「必ず回答する」設計は、根拠がない場合に推測での補完を強制する。回答保留・追加質問・人への引き継ぎという逃げ道を用意することが、誤回答を減らす最も直接的な手段である。

設定すべき5つの回答ルール

①推測させない

参照情報に根拠がない場合、一般知識で補完せず、確認できない旨を明示させる。

②追加質問させる

判断に必要な情報が不足している場合は、回答の前に不足情報を利用者へ確認させる。

③回答範囲を明示

回答できた範囲と、回答できなかった範囲を分けて提示させ、利用者に判断させる。

④人へ引き継ぐ

確信度が低い回答は自動で担当者へエスカレーションし、AI単独で完結させない。

⑤矛盾時は停止

参照情報間の矛盾を検知した場合は回答を生成せず、資料の不整合として報告させる。

プロンプト例｜推測による補完の禁止

> 参照情報に回答の根拠がない場合、一般知識で補完せず**「提供された情報では確認できません」**と回答してください。

この一文により、回答不能を「失敗」ではなく「正常な出力」として扱えるようになる。定型文を固定することで、後段の出力検証（P13）で自動検知・自動振り分けが可能になる。

導入時に併せて整備すべき事項

- **回答拒否の許容範囲を合意する** 回答率の低下を品質低下と誤解されないよう、事前に関係部門と基準を共有する。
- **引き継ぎ先を明記する** 「確認できません」で終わらせず、次の窓口・手順を回答に含める。
- **拒否の妥当性を評価する** 回答すべき質問への過剰な拒否も品質問題として測定する（P15）。

要点 「答えられない」と言える設計は機能の欠落ではなく、誤回答の業務流入を止める安全弁である。回答率と正答率はトレードオフとして管理する。

事実：提供された構成案に記載の5ルールおよびプロンプト例／解釈：併せて整備すべき事項、回答率と正答率のトレードオフに関する記述は当方による整理／提案：回答拒否の許容基準を事前合意する。

実践方法③後出典・引用箇所・根拠を表示させる

出典表示の目的は説明責任ではなく検証可能性の確保である。回答文と引用文を分離し、出典の存在と内容の一致を機械的に検証できる形式で出力させる。

出典表示に求める6つの要件

01 **参照した資料名を回答ごとに表示する**　どの文書に基づく回答かを利用者が即座に判別できる。

02 **ページ番号・章番号・URL・更新日を付ける**　照合先を特定し、情報の新旧を判断できる。

03 **回答文と引用文を分ける**　AIの生成部分と資料の原文を混在させない。

04 **出典が存在するか機械的に検証する**　文書ID・URLの実在をシステムで確認する（出典の捏造対策）。

05 **出典の内容と回答が一致しているか確認する**　引用元に書かれていない主張が含まれていないか照合する。

06 **出典がない回答は利用者へ警告する**　根拠なしの回答をそのまま業務利用させない。

留意点｜出典が表示されていること自体は正しさの証明にならない

- **出典自体が捏造される**ことがある（P3類型2）。実在確認は必ず機械的に行う。
- **実在する資料を挙げつつ、記載のない内容を述べる**ことがある（P3類型4）。内容の一致確認まで行う。
- **旧版の資料を出典として示す**ことがある（P3類型5）。更新日と有効期限を必ず併記させる。
- したがって出典表示は**P13の出力検証と組み合わせて**初めて統制として機能する。

標準回答フォーマット（6項目）

#	項目	記載する内容
1	**回答**	結論のみを簡潔に記述する
2	**判断根拠**	結論に至った理由・適用した条件
3	**参照資料**	文書名・文書ID・版数
4	**該当ページ・条文**	照合先を特定できる位置情報
5	**情報更新日**	参照資料の最終更新日・有効期限
6	**確認が必要な事項**	根拠が不足し判断できなかった論点

要点　フォーマットを固定すると、項目3〜5を機械的に検証でき、項目6が高リスク業務における人への引き継ぎトリガーとして機能する（P14）。

事実：提供された構成案に記載の6要件および回答フォーマット6項目／解釈：「記載する内容」列および留意点の記述は当方による整理／提案：フォーマット固定によりP13の自動検証と接続する。

実践方法④憶AGで社内情報を参照させる

RAGは質問に関連する社内文書を検索してAIへ渡す仕組みであり、学習データへの依存を減らし、根拠を利用者へ提示できる。ただし検索結果が誤っていれば誤回答は残るため、単独の解決策ではない。

基本フロー

STEP憶｜質問入力	STEP憶｜関連文書検索	STEP憶｜AIへ提供	STEP憶｜回答生成	STEP憶｜出典表示
利用者が業務上の質問を入力する。	検索基盤が関連する社内文書を抽出する。	検索結果を根拠情報としてAIへ渡す。	渡された情報のみを根拠に回答を生成する。	参照した文書と該当箇所を回答に併記する。

RAGによって得られる効果

- **学習データだけに依存しない** 社内固有の情報を回答の根拠にできる。
- **最新情報を反映しやすい** 文書を差し替えれば回答内容を更新できる（再学習が不要）。
- **回答根拠を利用者へ示せる** 参照文書と該当箇所を提示し、検証可能な状態にできる。
- **主な活用領域** 社内規程・業務マニュアル・FAQ・商品情報の照会業務。

RAG導入後も残るリスク

- **検索が外れる** 根拠文書が渡されず、AIが一般知識で補完する。
- **旧版が検索される** 改定前の資料が残っていると陳腐化した回答になる。
- **渡した情報を誤読する** 引用は正しいが、解釈や条件適用を誤る（P3類型6）。
- **権限を越えた情報が漏れる** 検索結果にアクセス権限が反映されていない。

要点 RAGを導入してもハルシネーションはゼロにならない。効果はデータ整備と検索設計の質に規定されるため、次ページの整備要件と併せて実装する。

事実：提供された構成案に記載のRAGの仕組み・活用領域・基本フロー・限界／解釈：「導入後も残るリスク」の4項目はP3・P11の内容を踏まえた当方の整理／提案：データ整備・検索設計と一体で実装する。

RAGで失敗しないデータ整備と検索設計

RAGの精度はモデルではなく、参照対象の整備状況と検索設計で決まる。「古い資料を消す」「版数と管理部門を付す」「検索できないときは答えさせない」の3点が失敗回避の中核となる。

A. 文書の棚卸しと管理

古い資料を削除または利用対象外にする
改定前の版が検索対象に残っている限り、陳腐化した回答は再発する。物理削除またはインデックス除外を運用に組み込む。

版数・有効期限・管理部門を付与する
文書ごとに責任者を定め、更新の判断者を明確にする。出典表示（P9）で利用者に提示する情報でもある。

PDFの文字化けや表の崩れを確認する
スキャン文書、複雑な表、図中文字は抽出に失敗しやすい。取り込み後のテキストを目視で検品する。

B. 分割とメタデータ付与

文書を適切な大きさに分割する
大きすぎると要点が埋もれてしまい、小さすぎると条件や前提が切り離される。章・条単位など意味の区切りを基準とする。

部署名・商品名・年度などを付与する
絞り込み条件として使えるメタデータがなければ、別部署・別年度の情報が混同される（P3類型3）。

アクセス権限を検索結果に反映する
閲覧権限のない文書が根拠として使われないよう、検索段階で権限を適用する。

C. 検索設計と回答制御

キーワード検索とベクトル検索を併用する
型番・条文番号などの完全一致はキーワード検索、言い換えを含む照会はベクトル検索が有効。双方を組み合わせる。

検索結果が不十分な場合は生成しない
関連度が閾値を下回る場合は回答を生成せず、P8のルールに従い「確認できません」を返す。

検索精度そのものを継続測定する
回答品質の劣化は多くの場合検索側で起きる。検索情報の適合率を指標化する（P15）。

要点 RAGの導入は情報システム部門だけでは完結しない。文書の版管理と廃棄判断は業務部門の責任範囲であり、導入前に管理責任者を確定させる必要がある。

事実：提供された構成案に記載の8項目／解釈：A（文書管理）・B（分割とメタデータ）・C（検索設計）への3群分類、および各項目の補足説明は当方による整理／提案：導入前に文書の管理責任者を確定する。

実践方法⑤ 外部ツール・データベースと連携する

数字と事実をAIの記憶から回答させないことが原則である。文章の生成はAIに任せ、値の取得と計算は既存システムに委ねる役割分担が、計算・混同・陳腐化の誤りを構造的に排除する。

処理ごとの実行主体の切り分け

処理	推奨する実行方法	切り分けの理由
文章の要約・分類	生成AI	言語処理はAIの得意領域であり、誤りも検知しやすい
顧客情報の取得	CRM・データベース	正のマスタが存在する情報は必ず取得して渡す
売上集計	SQL・BIツール	集計条件を再現・監査できる形で実行する
数値計算	計算プログラム	確率的生成では計算の正確性を保証できない
最新情報の取得	API・Web検索	学習時点以降の情報はAIの内部に存在しない
在庫・料金確認	基幹システム	時点で変動する値は都度取得しなければ誤る
契約内容の確定	人による確認	責任を伴う判断は自動化の対象としない（P14）

設計原則

原則1 ｜ 数字と事実を記憶から回答させない

金額、在庫、期日、顧客属性など「正のデータが社内に存在する情報」は、必ずシステムから取得した値を渡す。AIに想起させた時点で誤りの可能性が発生する。

この原則により、P3の類型1（事実の捏造）・類型5（陳腐化）・類型7（計算誤り）を発生源で遮断できる。

原則2 ｜ 取得結果と生成文章を分離して管理する

APIやDBから取得した値と、AIが生成した説明文をデータとして分けて保持する。

これにより、後段の出力検証（P13）で「生成文中の数値が取得値と一致しているか」を機械的に突合でき、業務システムへ登録する値は取得値のみに限定できる。

要点 AIの役割を「言語処理」に限定し、事実・数値の供給は既存システムに担わせる。この切り分けが済んでいない業務では、プロンプトの改善だけでは誤りは減らない。

事実：提供された構成案に記載の処理／実行方法の対応表および2原則／解釈：「切り分けの理由」列とP3類型との対応付けは当方による整理／提案：業務ごとに実行主体の切り分け表を作成する。

実践方法⑥ ガードレールと出力チェックを実装する

生成された回答は、利用者に届く前に必ず検証処理を通す。検知内容を「安全性」「形式」「事実整合」の3系統に整理し、違反時は再生成または停止させる自動制御として実装する。

① 安全性チェック

禁止語の検知
断定的な保証表現、差別的表現、社外提示に不適な表現を検出する。

機密情報の検知
未公開情報、社内限定情報が出力に含まれていないか確認する。

個人情報の検知
氏名、連絡先、口座情報などの含有を検出し、必要に応じ遮断する。

② 形式チェック

回答文字数・必須項目の充足確認
P9のフォーマット6項目が欠落なく出力されているかを確認する。

指定形式への適合確認
JSONなど後続処理が前提とする構造に適合しているかを検証する。

形式違反は後続処理を停止させる
構造が崩れた出力を業務システムへ流さない。

③ 事実整合チェック

数値・日付・URL・商品コードの存在確認
マスタやAPI取得値と突合し、実在しない値を検出する。

回答と参照資料の整合性チェック
引用元に記載のない主張が含まれていないかを照合する。

別のAIによる回答レビュー
生成とは別のモデルに根拠との整合性を判定させ、二重に確認する。

違反検知時の制御フロー

回答生成
AIが回答を出力する

→

検証処理（①〜③）
利用者に届く前に必ず通す

→

適合 ▶ 利用者へ提示／後続処理へ

違反 ▶ 回答を再生成、または停止

→

業務システム登録前の検証
基幹データや顧客連絡に流入する直前に、再度の検証処理を必ず挟む

留意点1｜検証は「AIの外側」に置く
プロンプトによる自己申告は検証ではない。回答を評価する処理は、生成とは独立した仕組みとして実装する。

留意点2｜検知結果をログに残す
違反の種類と頻度を記録し、P15の評価指標およびプロンプト・データ整備の改善入力として活用する。

事実：提供された構成案に記載の8つのチェック項目／解釈：3系統（安全性・形式・事実整合）への分類、制御フローの図示、留意点は当方による整理／提案：検証処理を生成とは独立した仕組みとして実装する。

実践方法⑦後uman後n後he後oopを設計する

人による確認は「念のため目を通す」運用では機能しない。AIだけで完結させる業務を明示的に限定し、承認者・確認項目・差し戻し条件を業務フローとして定義する必要がある。

設計すべき5項目

01 AIだけで完結させる業務を限定する
自動完結を許容する範囲を業務単位で列挙し、それ以外は確認を必須とする。

02 承認者・確認項目・差し戻し条件を決める
「誰が」「何を」「どの場合に戻すか」を定義する。確認項目が曖昧なままでは形式的な承認に陥る。

03 高リスクな回答は必ず担当者が確認する
P16の対策レベルに従い、影響度が高い業務では確認をシステム的に強制する。

04 エスカレーション先を設定する
AIが判断できない場合の引き継ぎ先を明示する。「確認できません」で止めない。

05 修正内容を記録し、改善データとして活用する
人が直した箇所は誤りの実データである。テストセット（P15）とプロンプト改訂に還元する。

人による確認を必須にする業務

- **契約・法務判断**
  条項解釈の誤りが法的責任に直結する
- **採用・人事評価**
  個人の処遇に影響し、説明責任が生じる
- **医療・金融に関する助言**
  健康・資産への影響が大きく、規制対応も伴う
- **顧客への正式回答**
  企業の公式見解として扱われる
- **金額を伴う処理**
  誤りが金銭的損失として即時に発生する
- **外部公開する文章**
  訂正が困難で、レピュテーションに影響する

確認を実効化するための3つの条件

①確認の負荷を設計に織り込む
全件確認が必要な設計は現場で形式化する。検証で振り分けた要確認分のみを人に回す。

②確認すべき箇所を提示する
回答全文ではなく、根拠が弱い箇所・数値・「確認が必要な事項」を明示して確認対象を絞る。

③責任の所在を明文化する
AIの回答を採用した場合の最終責任者を規程上で明確にし、承認記録を保存する。

事実：提供された構成案に記載の5設計項目および確認必須業務6件／解釈：各業務の必須理由と「実効化の3条件」は当方による整理／提案：確認対象を絞り込み、責任の所在を規程上で明文化する。

実践方法⑧後ストデータと評価指標を整備する

モデルやプロンプトの変更で品質は変動する。実業務の質問から正解付きテストセットを作り、変更時に必ず再テストする体制がなければ、精度の劣化は発見できない。

テストセットの整備

01　実際の業務質問からテストセットを作る
想定質問ではなく、利用ログに残る実際の質問文を採用する。表記ゆれや省略を含む生の形が重要。

02　正解・根拠・許容範囲を登録する
正解文だけでなく、根拠となる文書と該当箇所、および許容する回答の幅を定義する。表現の違いを誤りと判定しないため。

03　曖昧・誤解・意地悪な質問も含める
前提が欠けた質問、誤った前提を含む質問、範囲外の質問を意図的に入れ、「回答を拒否できるか」を検証する。

04　変更時に必ず再テストする
モデル更新、プロンプト改訂、参照文書の入替のいずれでも品質は変動する。変更を再テストの起動条件として運用に組み込む。

05　人による修正内容を継続的に取り込む
P14で記録した修正箇所を新規テストケースとして追加し、同種の誤りの再発を検知できる状態にする。

主な評価指標（8項目）

指標	測定する内容	主な確認層
正答率	正解と一致した回答の割合	全体
根拠一致率	回答内容が参照資料の記載と一致する割合	生成・出力
出典表示率	出典が付与された回答の割合	生成
回答拒否の適切性	拒否すべき質問を拒否し、回答すべき質問に答えた割合	生成
検索情報の適合率	検索結果に正しい根拠文書が含まれた割合	参照情報
人による修正率	承認時に修正が入った回答の割合	運用
誤回答の重大度	誤回答が業務に与えた影響の大きさ（件数より優先）	運用
回答時間・運用コスト	対策強化に伴う応答時間と費用の増加	全体

要点　正答率のみを追うと、回答拒否の抑制や応答時間の悪化を見落とす。正答率・根拠一致率・拒否の適切性・重大度の4指標を併せて監視する。

事実：提供された構成案に記載のテストセット整備4項目および評価指標8項目／解釈：「測定する内容」「主な確認層」列、項目05、および指標の併用に関する記述は当方による整理／提案：4指標を併せて監視する。

業務リスク別の対策レベル

すべての業務に同じ対策を適用すると、低リスク業務では過剰な負荷となり、高リスク業務では統制が不足する。「誤った場合の影響度」を基準に4段階で対策レベルを設計する。

レベル	利用例	必要な対策	適用する対策層（P6）
低リスク	アイデア出し、文章の下書き	プロンプト、利用者確認	Layer 1・3 利用者本人が成果物を確認する前提
中リスク	社内FAQ、議事録、社内検索	RAG、出典表示、ログ保存	Layer 1・2・3 ＋ ログに基づく定期評価
高リスク	顧客回答、見積書、契約確認	システム照合、承認フロー	Layer 1〜5（全層） 数値はAPI・DB取得値のみ使用
重大リスク	法務、医療、金融、人事判断	専門家確認、回答制限、監査	Layer 1〜5（全層） ＋ 用途の限定と監査証跡の保存

レベル判定の考え方

- 判定基準は発生確率ではなく**「誤った場合の影響度」**（P5）
- 社外に出るか、金額を伴うか、訂正が可能かで切り分ける
- 同一システム内でも、用途ごとにレベルを分けて適用する

よく見られる運用の失敗

- **過剰統制**：低リスク業務に全層を適用し、利用が定着しない
- **統制不足**：中リスク前提の仕組みが、いつのまにか顧客回答に使われる
- **レベルの固定化**：用途拡大時にレベルを再判定していない

要点 対策レベルは利用申請時に判定し、用途を拡大する際は再判定を必須とする。判定結果と適用対策を台帳として管理する。

事実：提供された構成案に記載の4レベル・利用例・必要な対策／解釈：「適用する対策層」列、レベル判定の考え方、運用の失敗例は当方による整理／提案：判定結果と適用対策を台帳で管理し、用途拡大時に再判定する。

導入・運用チェックリスト

データ・生成出力・運用の3領域について、導入前の確認と稼働後の定期点検に用いる。未充足項目が残る領域は、対策レベル（P16）を引き上げるまで高リスク業務へ適用しない。

A. データ
主管：業務部門（文書管理責任者）

- ☐ **参照資料は最新か**
  改定後の版が検索対象に反映されているか
- ☐ **重複・矛盾・期限切れの資料がないか**
  旧版が残存し、複数版が同時に検索されない状態か
- ☐ **文書の管理責任者が決まっているか**
  更新・廃棄の判断者と、その承認手順が明確か

関連：P10・P11（RAGの参照情報対策）

B. 生成・出力
主管：情報システム部門

- ☐ **推測禁止のルールがあるか**
  根拠がない場合の動作がプロンプトで定義済みか
- ☐ **出典を表示しているか**
  資料名・該当箇所・更新日が回答に含まれるか
- ☐ **不明時に回答を停止できるか**
  検索結果が不十分な場合に生成を止められるか
- ☐ **数値や固有名詞を検証しているか**
  マスタ・API取得値との突合処理が実装済みか

関連：P7～P9・P12・P13（生成・出力対策）

C. 運用
主管：業務部門・管理職

- ☐ **高リスク回答の承認者が決まっているか**
  承認者・確認項目・差し戻し条件が定義済みか
- ☐ **質問と回答のログを保存しているか**
  検証結果と承認記録も併せて保存されるか
- ☐ **誤回答を報告できるか**
  利用者からの報告経路と一次受け窓口があるか
- ☐ **定期的に精度を評価しているか**
  評価指標と実施頻度が定められているか
- ☐ **モデル更新後に再テストしているか**
  更新を再テストの起動条件として運用しているか

運用上の要点 本チェックリストは導入審査時と稼働後の定期点検（例：四半期ごと）で使用し、未充足項目は担当部門と期限を付して管理する。

事実：提供された構成案に記載の3領域12項目／解釈：各項目の確認観点、主管部門の割り当て、関連ページの対応付けは当方による整理（実際の主管は各社の体制に合わせて設定）／提案：導入審査と定期点検の2場面で運用する。

まとめ｜ハルシネーションを前提に安全な仕組みをつくる

ハルシネーションをゼロにすることは難しい。前提として受け入れ、データ・検索・検証・運用を組み合わせた仕組みで発生確率と業務影響を抑えることが、実務上唯一の現実的な解である。

本資料の要点

01 **完全な排除を目標にしない** 生成の仕組みに由来する現象であり、発生確率と影響度の抑制を目的とする。

02 **プロンプトだけに依存しない** データ・検索・検証・運用を組み合わせた多層防御として設計する。

03 **正確な情報源をRAGやAPIで供給する** 数字と事実をAIの記憶から回答させない。

04 **出典を表示し、根拠を確認できる状態にする** 検証可能性の確保が統制の前提となる。

05 **高リスク業務では人による確認を必須にする** 承認者・確認項目・差し戻し条件を業務フローに組み込む。

06 **テストとログ分析を継続する** 変更のたびに再評価し、誤回答を継続的に減らしていく。

推奨する導入ステップ

1 **対象業務とリスクを整理する**
影響度を基準に対策レベルを判定する（P16）

2 **小規模な業務で検証する**
低～中リスク業務から開始し、実データで課題を把握する

3 **正解データと評価指標を作る**
実業務の質問からテストセットを整備する（P15）

4 **RAG・ガードレール・承認フローを実装する**
対策レベルに応じて適用する層を決める（P6）

5 **ログを基に継続改善する**
修正記録と検知ログを改善の入力として循環させる

結論 「AIを正確にする」のではなく、「誤りが業務に影響しない仕組みをつくる」ことが導入判断の要件である。ステップ1のリスク整理から着手することを推奨する。

事実：提供された構成案に記載の要点6項目および導入ステップ5段階／解釈：各ステップの補足および結論の記述は当方による整理／提案：ステップ1（対象業務とリスクの整理）から着手する。

お問い合わせ

本資料の内容に関するご質問、自社業務への適用検討、対策レベルの判定や評価指標の設計についてのご相談を承ります。

Start Challenge
Consulting co., ltd.
挑む力が、世界を照らす

生成AIの業務適用における設計・運用・検証の支援

CONTACT

会社名	Start Challenge Consulting
メールアドレス	info@startchg.com
担当	平出

ご相談の際にお知らせいただきたい事項
対象業務／利用中のAIサービス／参照させたい社内文書の種類と管理状況

本資料の取り扱い 本資料はご検討目的での社内共有を前提としています。社外への転載・二次配布をご希望の場合は事前にご連絡ください。