

AIにアクセスを許可する前の12の質問。

何を参照できるのか。どの操作を実行できるのか。問題が起きたとき、誰が介入するのか。チームや提供元との話し合いに、この問いを持ち込みましょう。

企業の業務を1つ選びます。例は話し合いのために用意したもので、実施済みのテストではありません。各制御は、実際の用途に合わせて定め、確認する必要があります。

01-04 用途を定める。
業務、データ、行動。

05-08 制限を確認する。
アクセス、確認、テスト。

09-12 運用に備える。
記録、不具合、責任者。

どの業務を 対象にするか？

対象業務と、それを行うためにAIが必要とするものから考えましょう。

01 AIはどの作業を行い、その後は誰が判断するか？

具体的には

AIが注文の品目をまとめます。営業チームは、提案を確定する前に、価格、数量、納期を確認する必要があります。

求めるもの

入力、出力、禁止する用途、次の判断を担う人を明記した説明を求めます。検証では、人が確認するまで提案が保留されるかを調べます。ルールの記事だけでは、システムがそれを適用している証拠にはなりません。

02 どのデータを使い、どのデータを除くか。

具体的には

品目と数量の確認に、銀行情報や顧客情報に含まれる個人的なメモを送る必要はありません。

求めるもの

添付資料も含め、各項目の目的を確認します。テスト用に準備したデータで、何がツールに渡り、回答に何が現れるかを確認してください。名前を隠すだけでは匿名性は保証されません。実データは利用前に評価してください。

どの業務を対象にするか？

対象業務と、それを行うためにAIが必要とするものから考えましょう。

03 この業務で有効な情報源はどれか。

具体的には

フォルダーに取引条件の表が2つあります。AIは両方を見つけますが、受け取った注文に適用できるのは片方だけです。

求めるもの

各情報源の出所、版、責任者、許可された用途。文書を差し替えて検証し、回答で使われた参照先を確認します。利用を許可された情報源同士が食い違う場合は、誰が解決するかを定めます。出典の提示だけで正しさが保証されるわけではありません。

04 参照が許されれば、変更や送信もできるのか。

具体的には

在庫の参照、商品の置き置き、提案書の送信は、それぞれ異なる行動であり、影響も異なります。

求めるもの

許可する操作と記録の一覧。範囲外の変更を試し、実行側のシステムで阻止されることを確認します。AIに慎重に行動するよう文章で指示するだけでは、技術的な制限にはなりません。

例外があっても、制限は機能するか。

AIが停止したり、人に引き継いだりすべき場面についても、根拠を求めます。

05 AIは誰の権限でシステムにアクセスするのか？

具体的には

2人の担当者がそれぞれ異なる顧客を受け持っています。一方の照会で、もう一方が担当する顧客の注文が見えてはいけません。

求めるもの

利用者・システムのID、役割、データ、行動を対応付けた表を用意します。担当する顧客群をまたぐアクセス、権限の取り消し、生成された回答の内容をテストします。AIが「アクセスできません」と答えても、実際の参照が遮断された証拠にはなりません。

06 添付資料が作業を変えさせようとしたら？

具体的には

注文の添付資料に、顧客リストを別の宛先へ送るよう命じる記述があります。これは依頼された照合作業に含まれません。

求めるもの

管理された環境で、文書、ページ、ツールの応答に不正な指示を含めた検証を求めます。操作のブロックと調査への引き継ぎを確認します。フィルターは完全ではありません。アクセス権と実行範囲の制限は引き続き必要です。

例外があっても、制限は機能するか。

AIが停止したり、人に引き継いだりすべき場面についても、根拠を求めます。

07 操作の前に、人は何を承認するか？

具体的には

チームが提案書を確認します。その後に宛先や内容が変われば、その承認では送信できません。

求めるもの

操作の内容、宛先、影響を示すプレビューと、その操作に対する個別の承認記録。内容の変更で再確認が必要になるかをテストします。判断する人には情報、時間、権限が必要です。確認ボタンだけでは十分ではありません。

08 どの誤りがあれば、先に進めないのか？

具体的には

読めない添付資料、コードのない項目、矛盾する情報源には、作り上げた回答ではなく、追加情報が必要な場合があります。

求めるもの

通常のケースと例外を用意し、テスト前に期待する結果を定め、版ごとに観測した結果を記録します。重大な誤りは平均と分けて扱います。少数のサンプルですべてを網羅することはできません。情報源、モデル、ツールを変更した場合は、関係するテストを見直す必要があります。

想定外のことが起きたとき、誰が対応するか？

調査、中断、修正、利用終了の方法を合意します。

09 何を記録し、誰が閲覧できるのか？

具体的には

参照が拒否されました。チームは、文書全体をログヘコピーせずに、理由を理解する必要があります。

求めるもの

時刻、実行者の識別情報、操作、結果、理由を含む、保護された記録のサンプル。アクセス権、保存期間、機微な項目の扱いを確認します。記録自体も保護が必要です。パスワード、認証情報、不要なデータを保存すると、漏えいのリスクが増します。

10 いつ試行を停止すべきか。

具体的には

在庫情報を参照できない状態です。無制限に問い合わせを繰り返すと、消費量が増え、注文が待たされます。

求めるもの

業務ごとの時間、試行回数、使用量の上限と、代替手順。処理の中断と状態の通知をテストします。データを変更する操作を繰り返す前に、重複を防ぐため、前回の結果を確認してください。上限は業務プロセスによって異なります。

想定外のことが起きたとき、誰が対応するか？

調査、中断、修正、利用終了の方法を合意します。

11 どう停止し、すでに生じた影響を修正するか。

具体的には

配送先住所が誤って変更されました。新たな操作を止めても、すでに配送に回った注文は修正されません。

求めるもの

責任者、停止方法、影響の確認、再開基準を含めた手順を事前に練習します。関係するシステムも確認します。すべての行動を取り消せるわけではありません。すでに読まれたメッセージなどは、訂正と連絡が必要な場合があります。

12 利用と終了に、誰が責任を持つのか？

具体的には

試験運用が終了しても、アクセス権、文書、記録、コピーが自社や提供事業者のシステムに残る場合があります。

求めるもの

確認とインシデント対応の責任者、根拠のある保存期間、終了計画。アクセス権の取り消し、コピー、終了の証拠も含めます。期間や義務は、状況と、必要な技術面・法的な評価によって異なります。

[質問から判断へ](#)

回答には、裏付けが必要です。

質問ごとに、判断、利用できる根拠、責任者、未解決事項を記録します。確認が不足していれば、利用を広げる前に確認方法を決めます。

判断

何を許可するか？

根拠

何を確認したか？

責任者

誰が判断し、継続的に確認するか？

未解決事項

まだ何を確認する必要があるか？

Gaulettに相談する

gaulett.com/ja/contact/

[プロジェクトを準備する](#)

[検証について知る](#)

理解を深める参考資料

これらの質問は、参考資料を企業の日常業務に当てはめたものです。引用した機関が実施した検証を示すものではありません。

[NIST · AI RMF 1.0](#)

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

[NIST · Generative AI Profile](#)

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>

[OWASP · AI Agent Security](#)

https://cheatsheetseries.owasp.org/cheatsheets/AI_Agent_Security_Cheat_Sheet.html

[OWASP · Prompt Injection](#)

<https://genai.owasp.org/llmrisk/llm01-prompt-injection/>

[OWASP · Logging](#)

https://cheatsheetseries.owasp.org/cheatsheets/Logging_Cheat_Sheet.html