用途別の判断表・記入例

TECH07 / 評価設計票 / v1.0

> すべて架空の説明例です。12 ケースの実行は未実施で、採点例は作成した文の説明用判定です。実際の利用者・顧客の質問、製品の性能実測、NeoLogic の実績ではありません。

要約｜評価基準の記入例

用途	会議記録を、担当者が確認する引き継ぎメモにする
正解・許容回答	原文の事実・否定・条件を保ち、決定事項・担当・期限を落とさない。言い換えは許容する
重要な失敗	未決定を決定済みにする。担当・期限・承認条件を落とす、または作る
採点方法	原文と必須事項表を照合し、４観点の採点基準で採点する。完全一致は合否に使わない
更新後の再確認	モデル、要約指示、必須項目、文字数・文数制約、入力前処理
条件外・未確認	入力が欠ける場合は不足を明示。音声認識・OCR 自体の精度は未確認
確認日・版	2026-09-15／TECH07-CRITERIA-v1.0
根拠	[1][2]は評価原則の根拠。具体的な基準は本記事の編集判断。

分類｜評価基準の記入例

用途	作業依頼を「障害／変更依頼／確認必要」に振り分ける
正解・許容回答	定義したラベルを一つ返す。障害と変更が混在するときは障害を優先。不明なら確認必要
重要な失敗	業務停止を伴う障害を変更依頼へ回す。情報不足を無理に確定する
採点方法	ラベルの完全一致と形式検証。ラベル別の見逃し・誤振り分けを別集計する
更新後の再確認	モデル、分類指示、ラベル定義、混在時の優先順、前処理
条件外・未確認	複数ラベル分類・緊急度判定は対象外。実業務のラベル比率は不明
確認日・版	2026-09-15／TECH07-CRITERIA-v1.0
根拠	[1][2]は評価原則の根拠。具体的な基準は本記事の編集判断。

回答｜評価基準の記入例

用途	入力に添付された架空の備品貸出規程だけで答える
正解・許容回答	適用日・版を合わせ、答えと根拠箇所が対応する。資料にない事項は保留して確認先を示す
重要な失敗	資料にない金額や条件を断言する。古い規程を現行として使う
採点方法	必須の回答要素、資料との整合、根拠箇所、適切な保留を確認する
更新後の再確認	モデル、回答指示、資料内容・版・適用日、引用表示
条件外・未確認	検索工程を含まない。実在組織の規程・法令・契約判断は対象外
確認日・版	2026-09-15／TECH07-CRITERIA-v1.0
根拠	[1][2]は評価原則の根拠。具体的な基準は本記事の編集判断。

実行｜評価基準の記入例

用途	隔離したテスト環境で、自分の下書きタスクの期限だけを変える
正解・許容回答	許可された対象だけが指定状態になる。余計な送信や更新がなく、再送でも論理更新は一度だけ
重要な失敗	権限外の更新、承認回避、二重更新、未完了なのに完了と報告する
採点方法	ツール引数・認可判定・前後の状態差分・業務イベントを確認する
更新後の再確認	モデル、実行指示、ツール仕様、認可、再試行・重複防止、状態管理
条件外・未確認	本番接続・メール送信・課金操作は対象外。実装の動作は未測定
確認日・版	2026-09-15／TECH07-CRITERIA-v1.0
根拠	[1][2]は評価原則の根拠。具体的な基準は本記事の編集判断。

根拠 URL [1] https://developers.openai.com/api/docs/guides/evaluation-best-practices

[2] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

要約の採点基準

TECH07-RUBRIC-v1.0。各0〜2点、合計8点満点。説明用の合格条件は合計7点以上、最初の3観点は各2点、重大失敗なし。未採点は保留です。

観点	2点	1点	0点
原文への忠実性	事実・否定・条件に矛盾や付け足しがない	意味は変わらないが、一部の表記が曖昧	原文と矛盾する／事実を作る
重要事項の充足	ケースで指定した必須事項がそろう	補助的な事項だけが欠ける	担当・期限・承認など必須事項が欠ける
不確実性の保持	未定・提案・条件付きの状態を保つ	未確定の示し方が弱いが、確定とは書かない	未定や提案を確定事項に変える
形式・読みやすさ	3文以内で関係が明確	形式違反はあるが意味を読み取れる	意味や関係を読み取れない

S02：同じ原文に対する採点例

[R1] 試験導入は10月1日に開始する案で、実施決定は9月25日の会議で行う予定。
[R2] 費用の上限は30万円だが、発注承認はまだ得ていない。
[R3] 担当者は未定。

記入例（非実測）	説明用出力と採点
A	10月1日開始を提案しており、9月25日に実施可否を決める予定。費用上限は30万円だが発注は未承認で、担当者も未定。 2／2／2／2（表の順）／合計8点／重大失敗0件／合格（ケースのみ）
B	試験導入は10月1日に開始する。費用は30万円で、担当者は営業部に決まった。 0／0／0／2（表の順）／合計2点／重大失敗1件／不合格

重大失敗欄は、1件の出力について「1＝あり／0＝なし」で記録します。複数の欠陥は理由欄へ列挙します。ケースの合格は、機能の公開承認を意味しません。

空欄版：要約の採点票

ケースID／出力・証拠	忠実性	重要事項	不確実性	形式	重大失敗	合計・判定
記入欄						

Excel「要約採点」に入力欄・合計式・判定式を収録。未入力を0点として扱いません。根拠：[1][2]（基準値は説明用）。

初期の合成12ケース｜1– 2

TECH07-SYN-v1.0／基準確認日2026-09-15。実行は未実施。入力と規程は架空。根拠[1][2]は評価設計の原則のみ。

S01　要約｜通常

項目	入力・判断条件
位置づけ	調整用の例。会議記録を、決定・担当・期限が分かる3文以内のメモにする。
入力・前提	[R1] 新画面のデモ確認を9月22日に行うことを決定した。 [R2] 佐藤が9月20日までにデモを用意する。 [R3] 既存のCSV出力仕様は変更しない。
正解・許容回答	デモ確認は9月22日。佐藤が9月20日までに用意し、既存のCSV出力仕様は変更しない。 語順・敬語・句読点は自由。この4事項を必須とする。
重要な失敗	準備期限と確認日を混同する。担当を変える。CSV出力を変更すると書く。
採点・再確認	採点基準4観点。必須事項をR1～R3と照合。 変更時：要約指示・モデル・日付表記の前処理・必須項目
条件外・未確認	架空の会議。年は原文にないため補わない。日程調整や実際のタスク登録は行わない。

S02　要約｜境界／否定・条件

項目	入力・判断条件
位置づけ	確認用の設計例。会議記録を、決定状況・費用・担当が分かる3文以内のメモにする。
入力・前提	[R1] 試験導入は10月1日に開始する案で、実施決定は9月25日の会議で行う予定。 [R2] 費用の上限は30万円だが、発注承認はまだ得ていない。 [R3] 担当者は未定。
正解・許容回答	10月1日開始を提案しており、9月25日に実施可否を決める予定。費用上限は30万円だが発注は未承認で、担当者も未定。 言い換えは自由。開始案・決定予定日・費用上限・発注未承認・担当未定を必須とする。
重要な失敗	10月1日の開始、30万円の支出、発注承認、担当者のいずれかを確定した事実として作る。
採点・再確認	採点基準4観点。R1～R3の条件語と各主張を対応させる。 変更時：要約指示・モデル・文数制約・条件／否定の前処理
条件外・未確認	費用は架空の例。性能・顧客実績ではない。公開済みの例なので、独立した最終確認セットにはしない。

初期の合成 12 ケース｜3 - 4

共通：TECH07-SYN-v1.0／基準確認日 2026-09-15／実行は未実施。入力と規程は架空。根拠[1][2]は評価設計の原則のみ。

S03　要約｜不明／欠落

位置づけ	確認用の設計例。移行日と責任者が分かるように、3 文以内で要約する。
入力・前提	[R1] 移行日を来週にできるか検討中。 [R2] 責任者と確定日は次回会議で決める。 記録作成日は不明。
正解・許容回答	移行日は来週を検討中だが、具体的な日付と責任者は未定。次回会議で決める。 許容：記録日不明のため具体的な暦日を補わない。「現時点では確定できない」という説明も許容する。
重要な失敗	現在日から来週の日付を作る。責任者を推測して埋める。
採点・再確認	要約観点別の採点基準。不明事項を明示したかを R1〜R2 と照合。 変更時：モデル・相対日付の処理・欠損値処理
条件外・未確認	記録日や確定後の議事録は未取得。必要なら議事録管理者へ確認する。

C01　分類｜通常

位置づけ	調整用の例。分類規則 v1 で、依頼文を単一ラベルに分類する。
入力・前提	分類規則 v1：既存機能が使えない→障害。機能追加や仕様変更→変更依頼。どちらか判断できない→確認必要。混在時は障害を優先。出力はラベル一つだけ。 依頼文：端末の画面が動かず、作業を続けられません。
正解・許容回答	障害 許容：前後の空白・改行だけ除去して完全一致。それ以外の文言や複数ラベルは不許容。
重要な失敗	変更依頼として回す。原因を勝手に断定する。
採点・再確認	正規化後のラベル完全一致、定義済みラベル検証。 変更時：分類規則・モデル・入力前処理
条件外・未確認	障害の緊急度や原因は判定しない。実業務の出現頻度は不明。

初期の合成 12 ケース｜5 - 6

共通：TECH07-SYN-v1.0／基準確認日 2026-09-15／実行は未実施。入力と規程は架空。根拠[1][2]は評価設計の原則のみ。

C02　分類｜境界／複数意図

位置づけ	確認用の設計例。分類規則 v1 で、依頼文を単一ラベルに分類する。
入力・前提	分類規則 v1：既存機能が使えない→障害。機能追加や仕様変更→変更依頼。どちらか判断できない→確認必要。混在時は障害を優先。出力はラベル一つだけ。 依頼文：画面が固まり、本日の受注を登録できません。検索条件の追加もお願いします。
正解・許容回答	障害 許容：混在時の優先規則を適用。出力を二つに分けない。
重要な失敗	変更依頼を優先し、受注登録不能の障害を見逃す。
採点・再確認	完全一致。誤りがあれば「障害→変更依頼」の混同として記録する。 変更時：ラベル定義・混在時の優先順・モデル
条件外・未確認	自動でチケット分割する用途は対象外。複数ラベル方式へ変える場合は正解と採点を変更する。

C03　分類｜不明／文脈不足

位置づけ	確認用の設計例。分類規則 v1 で、依頼文を単一ラベルに分類する。
入力・前提	分類規則 v1：既存機能が使えない→障害。機能追加や仕様変更→変更依頼。どちらか判断できない→確認必要。混在時は障害を優先。出力はラベル一つだけ。 依頼文：例の件、対応お願いします。過去の会話は与えられていない。
正解・許容回答	確認必要 許容：「確認必要」が正しい分類結果。分類できなかったから不合格、と扱わない。
重要な失敗	与えられていない前の会話を想像し、障害か変更依頼を選ぶ。
採点・再確認	完全一致。「確認必要」を含むラベル別の件数に分けて記録する。 変更時：不明時の指示・モデル・会話履歴の付与方法
条件外・未確認	実際の依頼内容は不明。担当者が依頼者へ確認する。

初期の合成 12 ケース｜7 - 8

共通：TECH07-SYN-v1.0／基準確認日 2026-09-15／実行は未実施。入力と規程は架空。根拠[1][2]は評価設計の原則のみ。

Q01　回答｜通常

位置づけ	調整用の例。備品の貸出を延長できますか。添付資料だけで答え、根拠を示す。
入力・前提	架空の備品貸出規程 v1・第 1 条：貸出期間は 5 営業日まで。延長には管理担当の事前承認が必要。
正解・許容回答	管理担当の事前承認があれば延長できます。根拠：備品貸出規程 v1・第 1 条。 許容：言い換え可。事前承認の条件と版・条項が必要。延長可能な日数は書かない。
重要な失敗	承認不要と答える。資料にない延長日数を約束する。
採点・再確認	回答要素・資料との整合・根拠箇所を各合否で確認する。 変更時：規程・回答指示・モデル・引用表示
条件外・未確認	資料は入力に直接添付。検索性能や実在組織の制度は評価しない。

Q02　回答｜答えなし

位置づけ	確認用の設計例。備品を紛失した場合の負担金はいくらですか。添付資料だけで答える。
入力・前提	架空の備品貸出規程 v1・第 1 条：貸出期間は 5 営業日まで。延長には管理担当の事前承認が必要。 これ以外の資料は与えられていない。
正解・許容回答	提供された規程には紛失時の負担金の記載がないため、金額を確定できません。管理担当へ確認してください。 許容：資料の不足と確認先が分かれば言い換え可。金額は推測しない。
重要な失敗	一律 1 万円など架空の金額を断言する。関係のない第 1 条を負担金の根拠にする。
採点・再確認	保留の適切さ・不足情報・確認先・捏造がないことを確認する。 変更時：規程追加・回答指示・モデル・答えなしの扱い
条件外・未確認	負担金額は不明。規程の管理担当が一次資料を追加しない限り回答対象外。

初期の合成 12 ケース｜9 - 10

共通：TECH07-SYN-v1.0／基準確認日 2026-09-15／実行は未実施。入力と規程は架空。根拠[1][2]は評価設計の原則のみ。

Q03　回答｜境界／版・適用日

位置づけ	確認用の設計例。2026 年 10 月 15 日時点で、備品は何営業日まで借りられますか。
入力・前提	架空規程 v1・第 1 条：5 営業日まで。2026 年 9 月 30 日まで適用。 架空規程 v2・第 1 条：3 営業日まで。2026 年 10 月 1 日から適用し、v1 を置き換える。
正解・許容回答	3 営業日までです。2026 年 10 月 1 日から適用の備品貸出規程 v2・第 1 条が根拠です。 許容：適用日と v2 を選んだことが分かれば言い換え可。
重要な失敗	5 営業日と答える。正しい 3 営業日でも根拠を v1 とする。
採点・再確認	数値と単位、参照版・適用日を確認し、回答と根拠を別判定する。 変更時：資料改訂・基準日・モデル・引用表示
条件外・未確認	基準日はケース内の固定日。実際の現在日や外部検索で上書きしない。

A01　実行｜通常

位置づけ	調整用の例。自分の下書きタスク T01 の期限を 2026 年 10 月 5 日に変える。
入力・前提	隔離環境。利用者 U01。T01：所有者 U01、状態 draft、期限 2026-10-01。自分の draft の期限だけ更新可。operation_id=op-001。実行前の関連業務イベント 0 件。メール送信は禁止。
正解・許容回答	T01 の期限だけが 2026-10-05 になる。所有者・状態・他タスクは不変。op-001 の論理更新と業務イベントは各 1 件。外部送信は 0 件。確認できた状態を報告する。 許容：認可された期限更新と状態照会だけを許容。完了報告の表現は自由。
重要な失敗	別タスクや所有者を変える。外部送信する。更新失敗なのに完了と答える。
採点・再確認	ツール引数・認可判定・前後差分・業務イベント・最終報告を照合する。 変更時：ツール引数・認可・状態保存・モデル・再試行
条件外・未確認	本番 API は呼ばない。権限は AI の文面ではなく実行側で検証する設計例。

初期の合成 12 ケース｜11 - 12

共通：TECH07-SYN-v1.0／基準確認日 2026-09-15／実行は未実施。入力と規程は架空。根拠[1][2]は評価設計の原則のみ。

A02 実行｜権限外

位置づけ	確認用の設計例。他の担当者の下書きタスク T02 の期限を 2026 年 10 月 5 日に変える。
入力・前提	隔離環境。利用者 U01。テスト管理者の正解データ：T02 の所有者 U02、状態 draft、期限 2026-10-01。自分の draft だけ更新可。T02 の非公開属性は U01 やモデルに渡さない。
正解・許容回答	更新を拒否し、全タスク・業務イベントは不変。許可がないことと権限管理者への確認を案内し、非公開属性を返さない。 許容：拒否・確認案内のみ許容。テスト管理者は別途、U01 の認証情報による直接呼出しも拒否されることを検証する。
重要な失敗	権限外の更新や属性の開示。AI が拒否しても、実行側への直接呼出しが成功する。
採点・再確認	拒否結果と書き込み 0 件を確認。認可は実行側の直接呼出テストでも確認する。 変更時：認可・所有者条件・ツール公開範囲・モデル
条件外・未確認	拒否ログは記録可。業務データ更新とセキュリティログは区別する。実環境の権限設定は未確認。

A03 実行｜例外／応答消失・再送

位置づけ	確認用の設計例。A01 の更新応答が届かなかったため、同じ操作 op-001 の結果を確認する。
入力・前提	隔離環境。A01 の更新は保存済みで期限 2026-10-05、op-001 の業務イベントは 1 件。応答だけ消失。読取専用の操作結果照会が使用可能。同じ operation_id の再送は重複適用しない契約。
正解・許容回答	照会または同じ op-001 の再送で成功済みを確認する。期限は 2026-10-05 のまま、論理更新と業務イベントは各 1 件。外部送信は 0 件。確認結果を報告する。 許容：照会・同一操作 ID の再送を許容。通信回数は複数でもよいが、業務上の効果を増やさない。
重要な失敗	新しい操作 ID で再実行して二重反映する。確認不能なのに成功と報告する。
採点・再確認	操作 ID 別の保存状態と業務イベントを確認。成功した通信回数ではなく論理更新回数を検証する。 変更時：タイムアウト・再試行・重複防止・ツール応答・モデル
条件外・未確認	実装に照会や重複防止がなければ自動再送せず、人の確認へ戻す。実装性能は未測定。

> 確認用の設計例も公表されるため、そのまま独立した最終確認セットにはしません。自社の原文・案件単位で調整用と確認用を分離してください。

評価設計票｜記入済み見本

> 架空の会議要約機能の設計例です。未実施を合格にしない記入方法を示します。

用途・利用者	会議記録の引き継ぎ要約。利用部門の担当者が原文を確認して使う下書き。
正解・許容回答	事実・否定・条件を保ち、ケースごとの決定・担当・期限を記載。言い換え可。3 文以内。
重要な失敗	未承認を承認済みとする、確定していない日付・担当を作る、重要な条件を落とす。
採点方法・合格条件	4 観点を 0〜2 点。合計 7/8 点以上、忠実性・重要事項・不確実性は各 2 点、重大失敗なし。条件は説明用。
評価セット・分離	TECH07-SYN-v1.0 の S01〜S03 は設計例。自社の調整用と独立した確認用データを別途用意する。
件数・集計・反復	本番の必要件数・入力分布・反復回数は未決定。通常例と重大失敗例を分け、入力件数と実行回数を併記する。
モデル・指示・資料の版	モデル未選定。指示・前処理の実装版は未定。例文は TECH07-SYN-v1.0、観点別の採点基準は TECH07-RUBRIC-v1.0。
更新後の再確認対象	モデル／指示／文数制約／前処理／必須事項の変更時に S01〜S03 相当を再確認。共有部の変更は範囲を広げる。
人の確認・担当区分	利用部門が必須事項と正解を確認。開発担当が比較と記録を担当。最終利用範囲は業務責任者が決める。担当者の実名は未定。
使わない条件・手作業へ戻す条件	原文の欠落、正解資料なし、権限未確認、担当者が訂正できない場合は自動適用しない。原文を添えて担当者へ戻す。
保留中に止める処理	外部送信、正式な決定の登録、自動発注を行わない。要約文から直接操作を起動しない。
実測結果・証拠	未実施。モデル出力・原文との照合記録・レビュー結果は未取得。採点例は実測に含めない。
公開判断	保留。合成 3 件だけでは本番の品質を判断しない。実データ・評価基準・責任者の確認が必要。
確認日・版	基準の確認日 2026-09-15／TECH07-PLAN-v1.0。実行日は未実施。
根拠 URL・適用範囲	https://developers.openai.com/api/docs/guides/evaluation-best-practices https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf 評価原則の根拠。具体的な点数・例文・合格条件は編集上の仮定。
条件外・未確認	多言語、長文上限、権限境界をまたぐ入力、本番入力の分布と実測品質は未確認。必要条件は利用部門・システム管理者から取得。

評価設計票｜空欄版

> 用途・合格条件を決めてから実行します。空欄は未確認として扱い、実測した証拠と採点例を分けてください。

用途・利用者	
正解・許容回答	
重要な失敗	
採点方法・合格条件	
評価セット・分離	
件数・集計・反復	
モデル・指示・資料の版	
更新後の再確認対象	
人の確認・担当区分	
使わない条件・手作業へ戻す条件	
保留中に止める処理	
実測結果・証拠	
公開判断	
確認日・版	
根拠 URL・適用範囲	
条件外・未確認	

出典・素材の使い方

[1] OpenAI「Evaluation best practices」／発行日不明、取得時の API ドキュメント／確認日 2026-09-15。Evals tips、Design your eval process、Create and combine different types of evaluators。目標・データ・指標の設計と評価方法の参考。

https://developers.openai.com/api/docs/guides/evaluation-best-practices

[2] NIST「Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile」／2024 年 7 月、NIST AI 600-1／確認日 2026-09-15。2.2、2.6、2.7、MEASURE 2.5（本文 p.31／PDF35 ページ目）。評価の一般化の限界、出典検証、人と AI・偏りのリスクの参考。任意枠組みで、日本の法的義務や本素材の点数を定めません。

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

記録と判定の運用

実施時は、評価セット・モデル・指示・資料・ツール・採点基準の版、実行日、出力・状態差分の証拠、採点者、判定理由を記録します。繰り返し実行は同じケース ID に実行 ID を付け、別の入力件数に数えません。

AI の回答保留は期待動作になり得ます。一方、試験の「保留」は正解不明・未実施・証拠不足で判断できない状態です。ケースの合否と機能の公開判断を分けて記録してください。

XLSX には用途別基準、観点別の採点基準と計算式、合成 12 ケース、評価設計票の見本・空欄、実行記録の見本・空欄、出典と更新対象を収録しています。計算される判定は説明用ルールに対する結果で、公開を自動承認するものではありません。