

DENSO

Crafting the Core

SPI Japan 2025

生成AIによるプロセスセルフアセスメント支援の 実現性評価

2025年10月23日

池永直樹

株式会社デンソークリエイト



Certificate of
SOMRIE
Professional grade

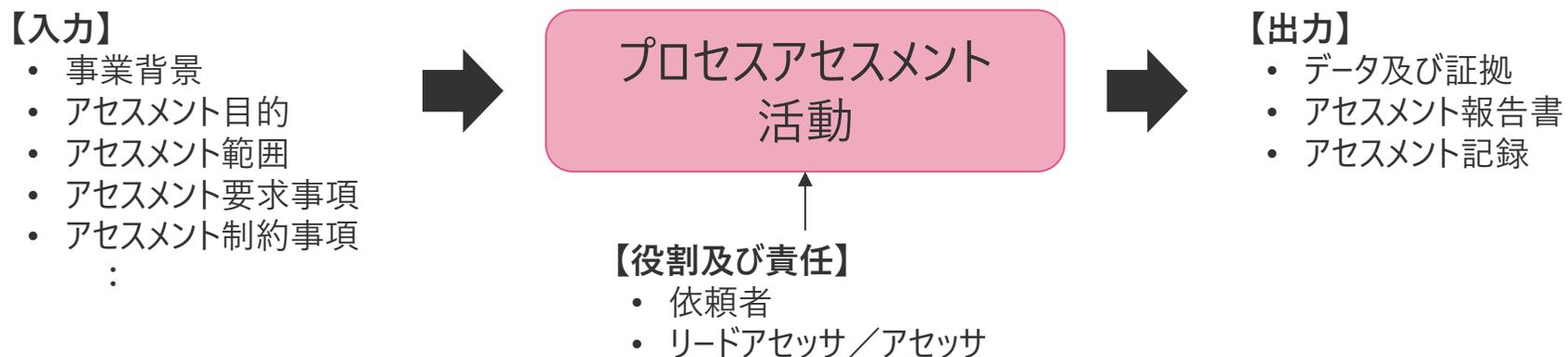
アジェンダ

1. 背景
2. 問題と課題
3. 解決策の提案
4. 評価
5. 評価結果と考察
6. 生成AIモデルの比較
7. まとめと今後の展望

背景：プロセスアセスメント（1）

■ プロセスアセスメントとは

- 開発プロセスの状態や改善点など、現状把握の有効な手段の一つ^[1]
- 自組織／プロジェクトの仕事のやり方に、改善すべき点があるか、伸ばす点があるか、組織／プロジェクト目標の達成に対してどういう状態にあるかを診断する、人間でいう健康診断のようなもの^[2]
- 繰り返し実施することによって、仕事のやり方の問題点の発見やプロセス改善の成果の確認に役立つ^[2]
- プロセスアセスメントモデル（PAM）を用いる
- アセッサと呼ばれるプロセスの専門家によって実施される



背景：プロセスアセスメント（2）

■ 自動車業界におけるプロセスアセスメント

- Automotive SPICE：業界標準のプロセスアセスメントモデル
 - Automotive SPICE 以外にも多くの拡張モデルが存在する^[3]
- 2022年には、グローバルで年間約1500回のプロセスアセスメントが実施されている^[4]



Hardware SPICE 

Mechanical SPICE 

Cybersecurity SPICE 

SPICE for IT-Services 

Data Management SPICE 

Agile SPICE 

Organization SPICE 

Modeling & Simulation SPICE 

背景：セルフアセスメントの必要性

■ プロセスアセスメント実施時の課題

- プロセスアセスメントの実施には、十分なインタビュー時間が必要である
- アセッサーの人数は限られている



組織内の多数のプロジェクトを網羅的に、
かつ頻度高くアセスメントすることは現実的ではない

■ セルフアセスメントとは

- ソフトウェア技術者自らが、ソフトウェア開発プロセスのチェックリストなどを用いて、自己診断の形式でプロセスアセスメントを行う方法

**セルフアセスメントが、アセッサーによるプロセスアセスメントを
補完する役割として利用できる**

問題と課題：セルフアセスメント実施における問題

開発現場の技術者によるセルフアセスメントの結果が 開発プロセスの実態を表したものになりにくい

■ 要因

- 開発現場の技術者はそのドメイン知識には精通しても、プロセス知識が不足していることが多い
- プロセスアセスメントモデルはそもそも抽象度が高く、開発現場の技術者には扱いづらい
- プロセスアセスメントモデルを基に作ったプロセスのチェックリストを用いても、技術者のプロセスに関する知識不足を十分補完できない



チェック項目の意味を誤解釈して、又は、達成度合いを正しく判断せず自己診断してしまう

問題と課題：セルフアセスメント展開時の課題

目的

プロセス改善活動をより活性化させる



手段

自己診断形式でのプロセスアセスメント = セルフアセスメント



問題

開発現場の技術者によるセルフアセスメントの結果が開発プロセスの実態を表したものになりにくい



課題

セルフアセスメント実施者のプロセスの知識レベルに寄らず開発プロセスの実態を表した診断結果を得る

解決策の提案：着想

■ ソフトウェア開発への生成AIの適用

- 要件定義、議事録管理、プログラム開発(コード生成、ペアプログラミングなど)、レビュー、テストなどソフトウェアエンジニアリングへの適用が進んでいる^[5]

■ プロセスアセスメントへの生成AIの適用

- 生成AIがアセッサーを支援するアセスメントツール^[6]などが登場し始めている



生成AI

特徴

- 自然言語処理
- 知識の広さ
- 文章読解力
- コンテンツ生成



特徴

- 自然言語中心
- プロセス知識
- 文章読解
- 評価、判断



セルフアセスメント

**セルフアセスメント実施者のプロセス知識が不十分でも
生成AIのサポートがあれば自己診断の精度が向上するのではないか？**

解決策の提案：セルフアセスメントへの生成AIの活用方法

■ 活用方法

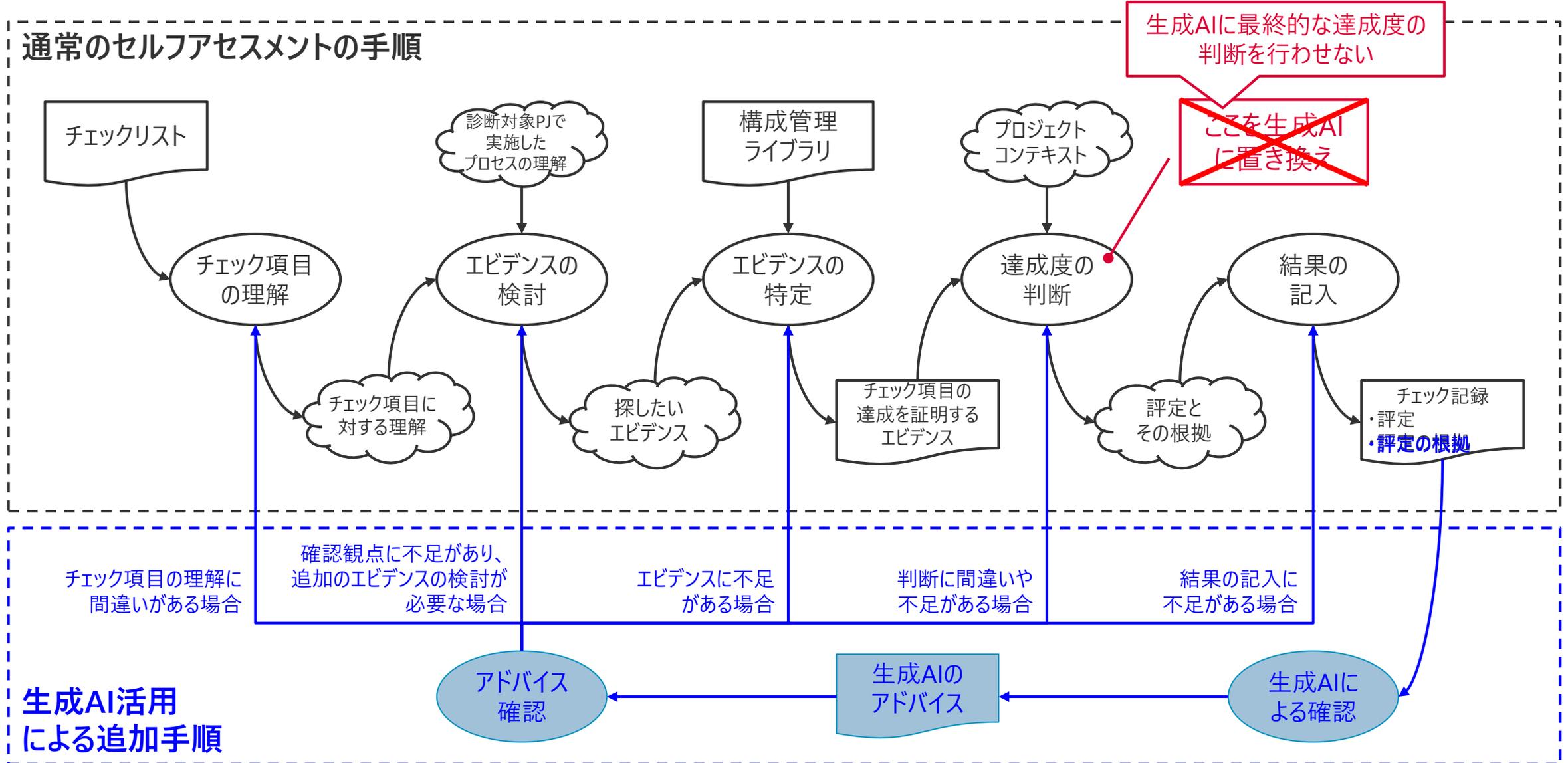
セルフアセスメントの実施を **生成AIに支援させる** 方式

- ✓ チェックリストを用いて実施することを前提とする
- ✓ チェック項目に対する回答（評定の根拠）を基にして生成AIにアドバイスさせる
- ✓ 最終的な達成度の判断はセルフアセスメント実施者に任る

■ 採用理由

- セキュリティ監査に対する生成AIの性能を評価した**先行研究^[7]**で、**監査性能は不十分**と報告
 - 成果物そのものを生成AIに入力し適合／不適合を回答させるものであったが、監査性能は十分ではなかった
- プロセスアセスメントはより**複合的な判断**が必要
 - 成果物だけでなく、アセスメント目的やプロジェクトコンテキストを考慮した上で出来栄えを診断する必要がある
- 開発現場の**プロセスに関するスキル向上**も重要
 - セルフアセスメントを通じて、開発現場のプロセスに対する知識や意識を向上させることも重要である

解決策の提案：生成AIを活用したセルフアセスメント手順



解決策の提案：チェックリスト

The screenshot shows a software development process checklist application. The main content area contains a table with the following columns: 名前 (Name), No., 設問の補足 (Supplement to the question), 評価 (Evaluation), and 評価の根拠 (Basis for evaluation). The table lists several checklist items under the 'プロジェクト管理プロセス' (Project Management Process) category.

Annotations and instructions are as follows:

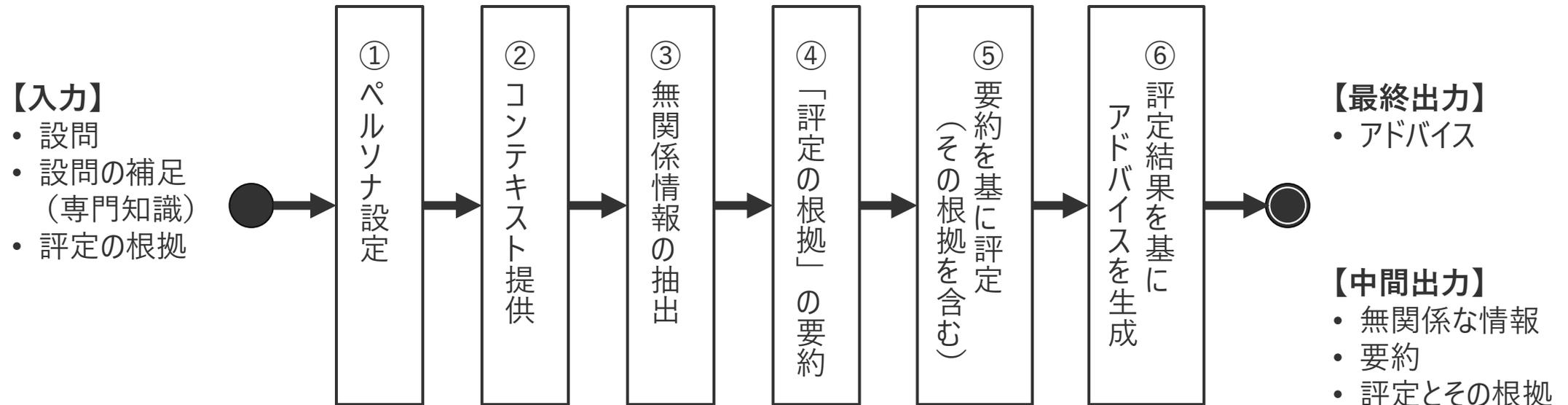
- Blue box:** Highlights the '名前' (Name) column.
- Red box:** Highlights the '設問の補足' (Supplement to the question) column.
- Red box:** Highlights the '評価' (Evaluation) column.
- Red box:** Highlights the '評価の根拠' (Basis for evaluation) column.
- Yellow box:** Contains the text: '自己診断時にセルフアセスメント実施者のプロセスの知識を補う目的で用意 解説書のように読み物として記述' (Prepared for the purpose of supplementing the knowledge of the self-assessment implementer's process during self-diagnosis. Described as a reading material like a manual).
- Yellow box:** Contains the text: 'セルフアセスメント実施者が入力する' (Entered by the self-assessment implementer).
- Pink box:** Contains the text: 'F/L/P/Nのいずれかを選択する' (Select one of F/L/P/N).
- Pink box:** Contains the text: 'F/L/P/Nを選択した根拠を記入する' (Enter the basis for selecting F/L/P/N).

解決策の提案：プロンプト

■ プロンプトテクニック

- 生成AIの活用においてプロンプトパターンは数多く考案されている
 - Prompt engineering^[8]の戦術、Prompt Engineering Guide^[9]のテクニック、プロンプトパターンカタログ^[10] など
- ペルソナ設定、コンテキスト提供、出力テンプレート、Chain of Thoughtなどを組み合わせてプロンプトを作成した

■ プロンプトの全体像



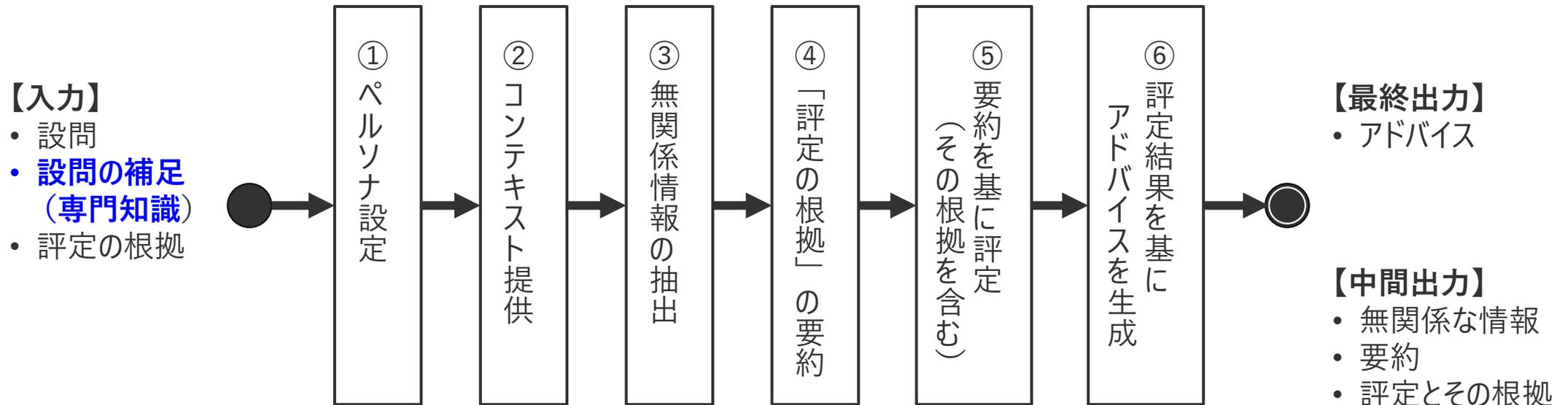
解決策の提案：プロンプト - プロンプトテンプレート -

プロンプト	「プロンプトの全体像」との対応
<p>あなたはソフトウェア開発プロセス（Automotive SPICE、ISO/IEC12207、CMMIなど）の専門家です。</p>	① ペルソナ設定
<p>プロジェクト活動について、プロジェクトメンバー自身がチェックリストを用いてソフトウェア開発プロセスを自己診断しています。 この自己診断は、プロセスの定義状況を問うものではなく、プロジェクトでのプロセスの実施を問うものです。</p>	② コンテキスト提供
<p>[設問] に続く問いについて、[評定の根拠] に続く回答を得ました。得た回答に対して、プロンプトで指示するStep1,2,3,4の手順を実施して下さい。</p>	
<p>【Step1】 ノイズとなる情報を除去するために、[評定の根拠] から [設問の補足] も考慮して、[設問] に無関係な情報を抽出し、箇条書きで回答して下さい。 # 除外した[設問]に無関係な情報 ・xxx : yyy (xxxは除外した情報、yyyはその理由。ない場合は「なし」のみ記載)</p>	③ 無関係情報の抽出
<p>【Step2】 「#要約の考慮点」を考慮して、[評定の根拠] を要約してください。 #要約の考慮点 ・Step1の「# 除外した[設問]に無関係な情報」を要約に含めない。 ： (省略) # 要約 xxx (400字以内)</p>	④ 「評定の根拠」の要約
<p>【Step3】 Step2の要約結果を元に、[設問の補足] および「#達成度評価の考慮点」を考慮して、達成度を評価してください。 達成度は、F（十分達成）、L（おおむね達成）、P（部分的に達成）、N（未達成）、NR（評定不能）で回答して下さい。 FLPNは、ISO/IEC33020のプロセス属性評定の尺度のことです。NRは、要約結果の情報では達成度を評価できない場合に選択してください。 また、達成度の判断理由も回答してください。 #達成度評価の考慮点 ・Step2の要約結果がプロセス定義に関する内容のみの場合、NRと評定する。 ： (省略) # 達成度 F/L/P/N/NR のいずれか。 # 達成度の判断理由 xxx (400字以内)</p>	⑤ 要約を基に評定（その根拠を含む）
<p>【Step4】 Step3の達成度および判断理由をもとに、「#アドバイスの考慮点」を満たすように、正しい評定に向けたアドバイスをしてください。 #アドバイスの考慮点 ・Step3の達成度がNRの場合、NRと判断した理由のみを出力して下さい（この場合、箇条は一つ）。 ： (省略) # アドバイス 1.xxx (優先度が高い順に最大5個まで。各箇条は100字以内)</p>	⑥ 評定結果を基にアドバイスを生成
<p>[設問] <設問の内容> [設問の補足] <設問の補足の内容> [評定の根拠] <評定の根拠の内容></p>	【入力】

解決策の提案：プロンプト - ポイント1：専門知識 -

■ 専門知識

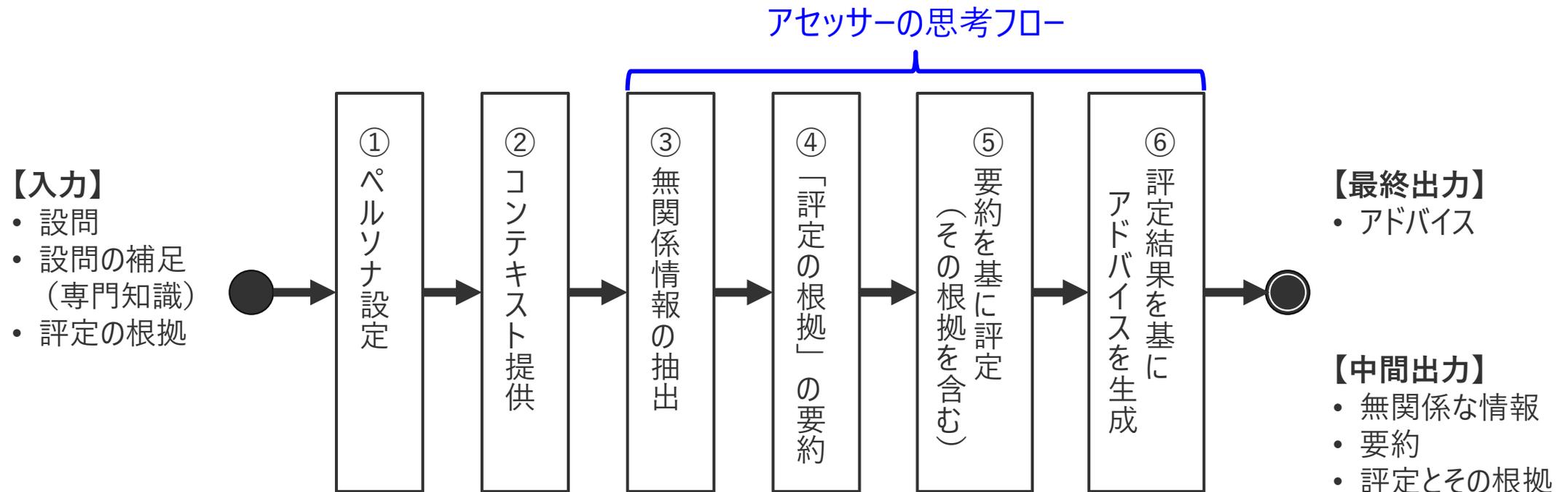
- 汎用AIから有用なアドバイスを得るには、プロセス知識及びその組織や製品分野におけるプロセスの特徴などの専門知識が必要である
- チェック項目の構成要素の一つである「設問の補足」がそれに該当するので、これを生成AIに入力する



解決策の提案：プロンプト - ポイント2：アセッサーの思考ステップ -

■ アセッサーの思考ステップ

- 初期のプロンプトは要約（④）させた後にアドバイス生成（⑥）の流れとしていたが、指示をいくら調整してもアドバイスの質が向上しなかった
- そこで、アセッサーの思考フローに基づき推論ステップを分割し、要約を基に評価（⑤）するステップを追加した

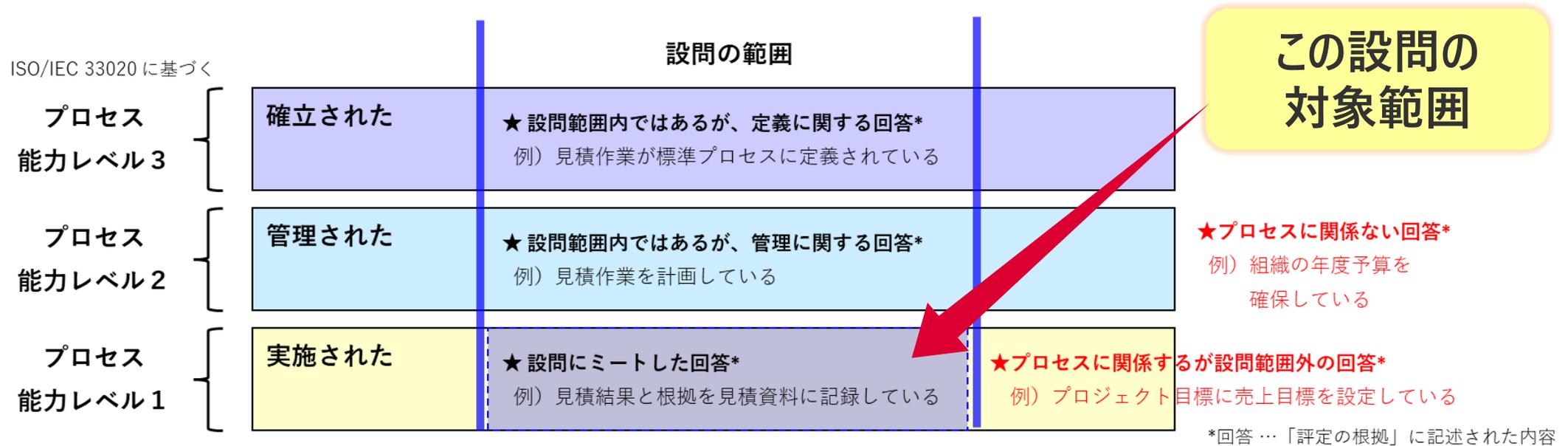


解決策の提案：プロンプト - ポイント3：多層的アプローチ（1） -

■ 「設問」と「評定の根拠」に記述された内容 との関係

- 設問に関係のない情報が「評定の根拠」に記述される可能性がある
- 設問に関係のない情報が有用な情報として扱われず評定され、アドバイスが出力されていた

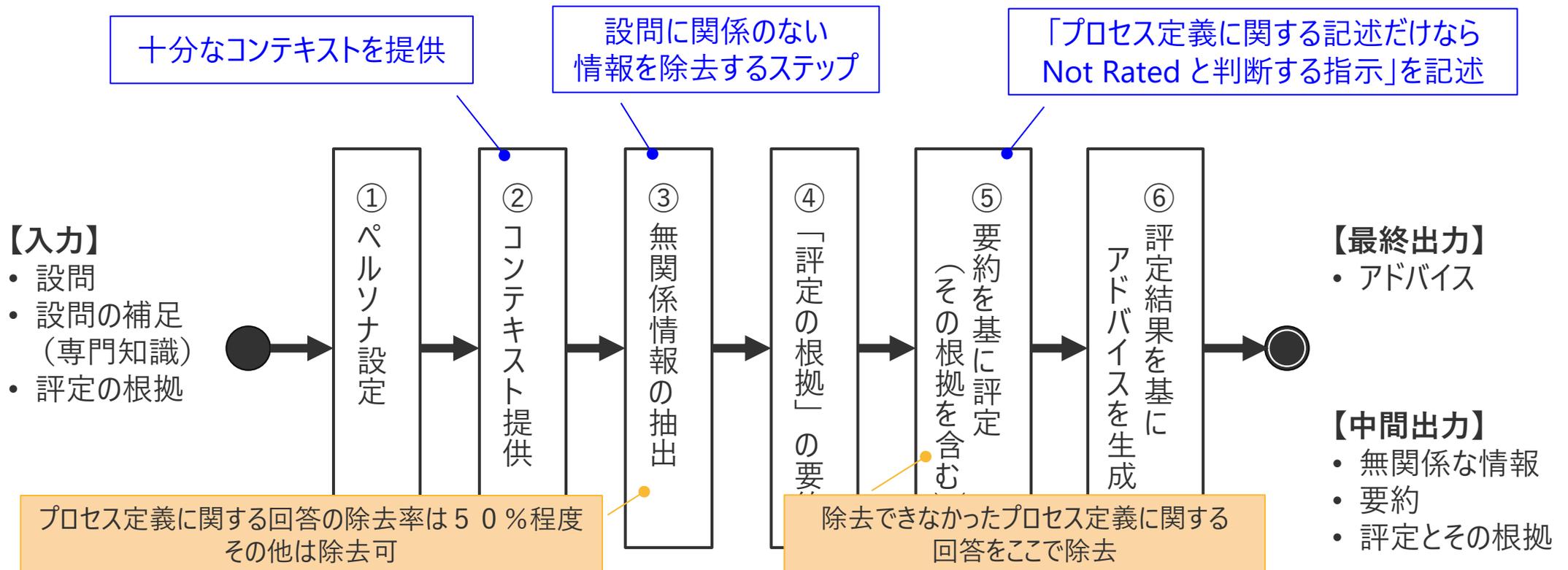
設問の例) プロジェクトに必要な人的リソース、インフラストラクチャ、コストを見積もっていますか？



解決策の提案：プロンプト - ポイント3：多層的アプローチ（2） -

■ 多層的アプローチ

- 検出するために複数の仕掛けを追加し、いずれかのステップで不要な情報として扱わせるようにした
- 特に“設問範囲内であるがプロセス能力レベル3に該当する情報”の除去が困難であり、
⑤に専用の指示を追加した



評価：評価観点

生成AIによるセルフアセスメント支援が実現可能であるかを3つの観点で評価する

観点1

生成AIが出力するアドバイスが**アセッサーと同レベル**であるか

生成AIが出力したアドバイスに見逃しと不要な指摘がないかをアセッサーが評価する。

- 正しいアドバイスをしない ⇒ 見逃し
- 間違った／無関係なアドバイスをする ⇒ 不要な指摘

観点2

生成AIの評定が**アセッサーの感覚と合っている**か

「評定の根拠」に基づく生成AIの評定とアセッサーの評定の一致度を評価する。

観点3

セルフアセスメント実施者にとって**納得感のあるアドバイスが出力される**か

ISO25010 の利用時品質を用いて
確認項目を定義し、
セルフアセスメント実施者が評価する。

利用時品質	確認項目
有効性	理解できたか
効率性	分量は適量であったか
満足性	自己診断を見直すきっかけになったか

評価：評価方法と環境

■ 評価方法

- セルフアセスメントを実施する立場であるプロジェクトマネージャー（又はリーダー）を対象に試行する
- 実際のプロジェクトを診断してもらう
- 生成AIが出力したアドバイスを、アセッサー及びセルフアセスメント実施者が評価する
- アセッサーは intacs 認定 Automotive SPICE Principal Assessor 資格^[11] を保有する筆者が担当する

■ 評価環境

- GPT-4o（Azure OpenAI） ※評価開始時点で自組織で利用可能な最新モデル

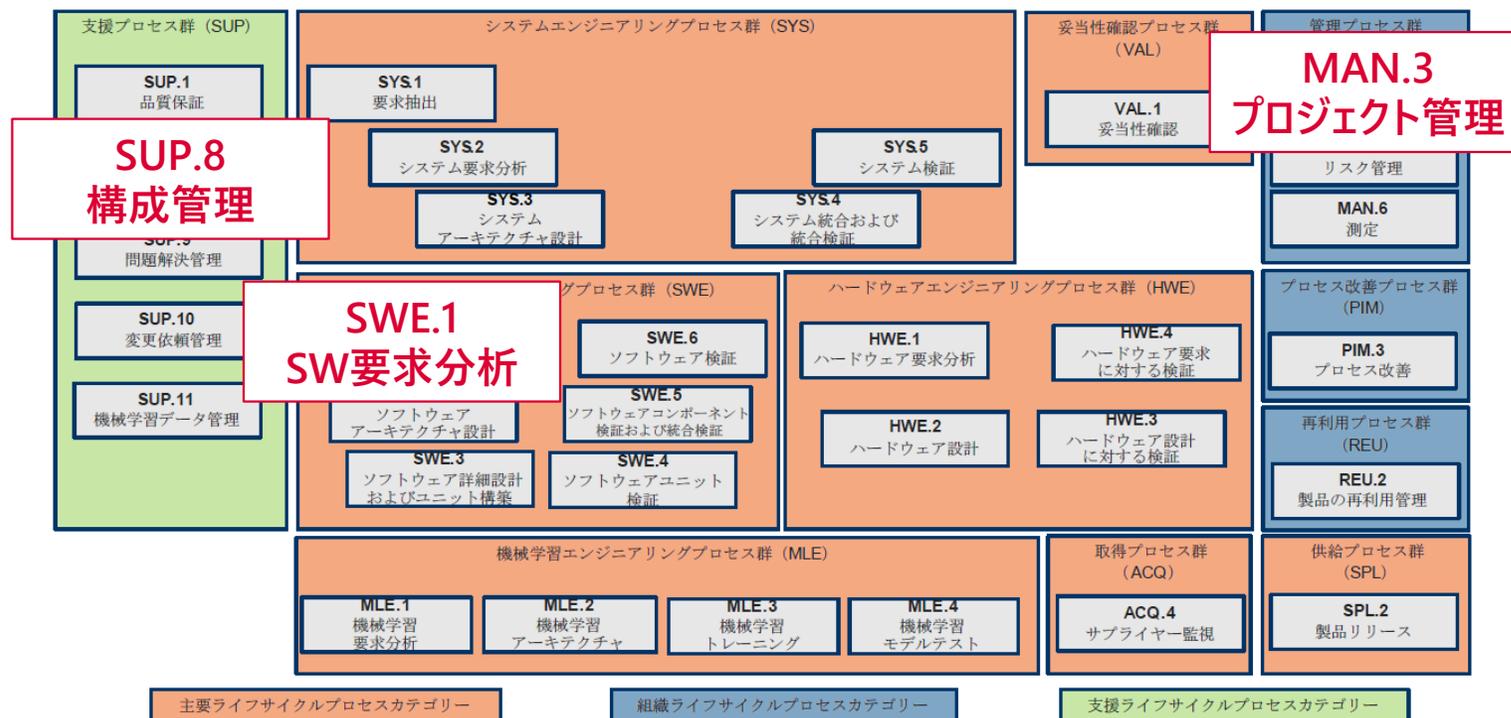
評価：対象プロセスとチェック項目

Automotive SPICE V4.0 から下図の3プロセスを選択した

Automotive SPICE 4.0 実践ガイドブック^[12] に掲載のチェック項目を使用した

- 合計：33項目（MAN.3：12項目、SUP.8：12項目、SWE.1：9項目）

能力レベル1を対象とした



評価：評価データ

合計 161 件の評価データを用いた

- セルフアセスメント実施者 7 名が実際のプロジェクトを対象に自己診断した回答（# 1 – 7）
- 筆者が恣意的に作成した回答（test）

	#1	#2	#3	#4	#5	#6	#7	test	合計
MAN.3	0	12	12	12	0	12	11	6	65
SUP.8	0	12	12	12	12	12	0	0	60
SWE.1	9	9	0	0	9	0	9	0	36
合計	9	33	24	24	21	24	20	6	161

評価結果と考察：生成AIの出力例

ユーザー

あなたはソフトウェア開発プロセス（Automotive SPICE、ISO/IEC12207、CMMIなど）の専門家です。
：（省略 ※「解決策の提案：プロンプト - プロンプトテンプレート -」スライド参照）

[設問]
プロジェクトに必要な人的リソース、インフラストラクチャ、コストを見積もっていますか？

[設問の補足]
人的リソースとは、人員および工数である。インフラストラクチャとは、ツール、ツールのライセンス、テスト機材などである。
：（省略）

[評定の根拠]
・プロジェクトに必要な人的リソース、コストは、「見積もり計画書」で見積もっている。
・プロジェクトに必要なインフラストラクチャは、「4.開発環境」「5.リソース計画」で見積もっている。



生成AI

#回答

Step1: 除外する[設問]に無関係な情報の抽出
・なし：設問に無関係な情報は含まれていません。

Step2: 要約
プロジェクトに必要な人的リソースとコストは「見積もり計画書」で見積もられ、
インフラストラクチャは「4.開発環境」および「5.リソース計画」で見積もられている。

Step3: 達成度評価
達成度
P（部分的に達成）

達成度の判断理由
「プロジェクトに必要な人的リソース、インフラストラクチャ、コストを見積もっていますか？」という設問に対し、人的リソースとコスト、およびインフラストラクチャの見積もりが「見積もり計画書」、「4.開発環境」、「5.リソース計画」で行われていることが確認できる。しかし、見積もりの手法や根拠についての具体的な説明がないため、客観的に説明可能な見積もりが実施されているかどうかの確認が不十分である。

Step4: アドバイス

1. 見積もり手法やそれに基づく根拠が客観的に説明可能であることを確認してください。
2. 見積り結果及び根拠と要件、ワークパッケージなどとの間のトレーサビリティが確立されていることを確認してください。

評価結果と考察：評価観点1（1）

観点1 生成AIが出力するアドバイスがアセッサーと同レベルであるか

	チェック項目	アドバイス項目数*
見逃し (適切なアドバイスをしない)	45 / 161件 (28%)	46個
不要な指摘 (間違った／無関係なアドバイスをする)	50 / 161件 (31%)	62個

* 一つのチェック項目に対して複数のアドバイス項目が出力されることから、
以下のようにアドバイス項目ごとに分類している

設問	生成AIが出力したアドバイス項目	アドバイスの分類
プロジェクトに必要な人的リソース、インフラストラクチャ、コストを見積もっていますか？	1. 見積もり手法やそれに基づく根拠が客観的に説明可能であることを確認してください。	適切な指摘
	2. 見積り結果及び根拠と要件、ワークパッケージなどとの間のトレーサビリティが確立されていることを確認してください。	不要な指摘
	【分析時メモ】 管理・支援プロセスの活動の見積りについてのアドバイスが出力されなかった	見逃し

評価結果と考察：評価観点1（2）

■ アセッサー（筆者）との比較

➤ 筆者と比べると、

- 見逃し（正しいアドバイスをしない） : 45 / 161件（28%）
- 不要な指摘（間違った／無関係なアドバイスをする） : 50 / 161件（31%）

とそれぞれ3割程度多かった

➤ 一方で、初級アセッサー* と同等レベルであった

* 初級アセッサー：筆者が所属する組織の intacs 認定 Automotive SPICE Provisional Assessor 資格^[1]を保有し、プロセス改善業務経験は十分ありアセスメント経験が数回程度のメンバー

	生成AI	初級アセッサー
見逃し	28%	28%
不要な指摘	31%	4%

**アセッサー（筆者）と比較すると劣るが、
初級アセッサーと同等レベルであり実務で活用可能なレベルである**

評価結果と考察：評価観点 2

観点 2 生成AIの評定がアセッサーの感覚と合っているか

生成AIとアセッサーの評定結果の混同行列

アセッサー \ 生成AI	F	L	P	N	Not Rated
F	35	4	4	1	0
L	20	14	10	3	1
P	0	5	17	3	0
N	1	0	3	12	1
Not Rated	1	2	5	3	16

- ✓ 完全一致  : 58% (94 / 161件)
- ✓ 完全一致  + 1水準違い  : 86% (139 / 161件)
- ✓ Not Rated と評定すべきなのにしなかった  : 7% (11 / 161件)
- ✓ Not Rated と評定すべきでないのにした  : 1% (2 / 161件)

■ 一致度の高さ、及び、偶然かどうか（加重カッパ係数^[13]）

➤ カッパ係数 = 0.7494、Z値 = 16.526、p値 = 2.372e-61

⇒ 一致度は高く、統計的にも有意である（偶然ではない）

カッパ係数^[14]

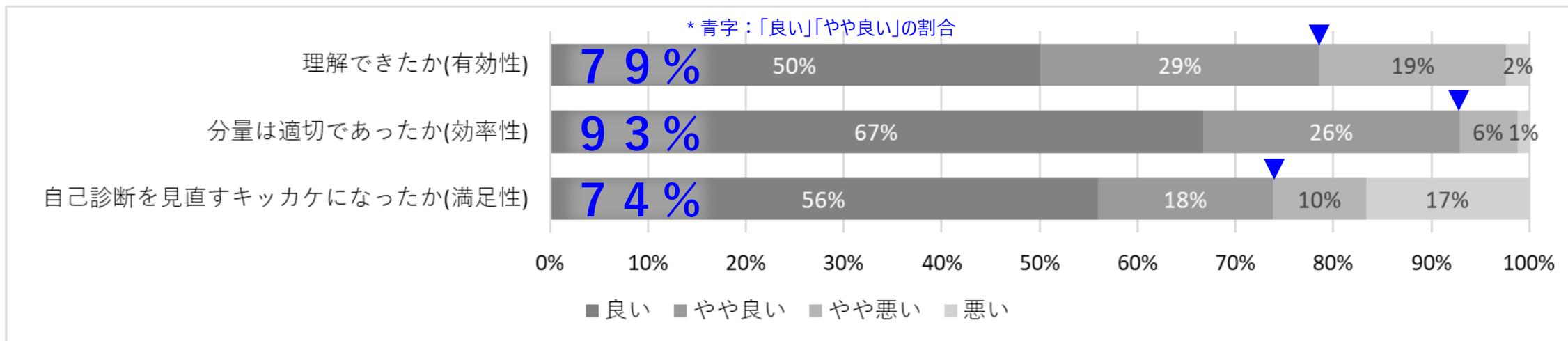
< 0.00	: Poor
0.00-0.20	: Slight
0.21-0.40	: Fair
0.41-0.60	: Moderate
<u>0.61-0.80</u>	: <u>Substantial</u>
0.81-1.00	: Almost Perfect

「完全一致 + 1水準違い」は86%であり、アセッサーに近い感覚で評定できる

評価結果と考察：評価観点3（1）

観点3 セルフアセスメント実施者にとって納得感のあるアドバイスを出力できるか

生成AIのアドバイスに対するセルフアセスメント実施者の評価



■ 各確認項目について合格と言えるか

➤ “良い”が“悪い”を明らかに上回っていると判断する基準値：2 / 3（特別多数決*）

⇒ 有効性：79%、効率性：93%、満足性：76%であり、2 / 3（67%）を上回っている

*特別多数決： 地方自治体などで重要事項を決議する場合に採用される、通常の過半数より高い割合の賛成が必要な意思決定方法

評価結果と考察：評価観点3（2）

■ セルフアセスメント実施者の主なコメント

➤ ポジティブなコメント

- 間違いに気づいた／質問に回答できないことが分かった
- 知識が得られた／気づきを得た

➤ ネガティブなコメント

- できていないと記述していることに対して再度確認せよとアドバイスされる
- アドバイスが曖昧

➤ その他

- メモとして最小限しか書いておらず、書いていないことについてアドバイスが出力されていた

➡ 自己診断結果の精度向上
スキル向上
に対する効果あり

➡ 有効性、満足性の評価を
下げる要因

➡ 想定範囲内

いずれの確認項目も基準値（2／3）を上回っており、肯定的なコメントも得られていることから、生成AIが出力するアドバイスは納得感があると判断できる

生成AIモデルの比較：観点1の比較結果

評価観点1,2について、より新しい生成AIモデル（GPT-4.1）を使った結果を示す
GPT-4oと同じプロンプトを使用した

観点1 生成AIが出力するアドバイスがアセッサーと同レベルであるか

	GPT-4o	GPT-4.1
出力されたアドバイス項目の総数	152個	220個
見逃し		
チェック項目数	45 / 161件 (28%)	19 / 161件 (12%)
アドバイス項目数	46個	20個
不要な指摘		
チェック項目数	50 / 161件 (31%)	43 / 161件 (27%)
アドバイス項目数	62個	62個
ステップ③で抽出された不要な情報数	9個	9個

生成AIモデルの比較：観点2の比較結果

観点2 生成AIの評定がアセッサーの感覚と合っているか

GPT-4o

アセッサー \ 生成AI	F	L	P	N	Not Rated
F	35	4	4	1	0
L	20	14	10	3	1
P	0	5	17	3	0
N	1	0	3	12	1
Not Rated	1	2	5	3	16

GPT-4.1

アセッサー \ 生成AI	F	L	P	N	Not Rated
F	29	6	6	0	3
L	13	9	16	2	8
P	0	4	17	2	2
N	0	1	2	9	4
Not Rated	1	2	1	1	23

		GPT-4o	GPT-4.1
評定	完全一致	58% (94 / 161件)	54% (87 / 161件)
	完全一致 + 1水準違い	86% (139 / 161件)	81% (130 / 161件)
	NRと評定すべきなのにしなかった	7% (11 / 161件)	3% (5 / 161件)
	NRと評定すべきでないのにした	1% (2 / 161件)	10% (17 / 161件)
統計	一致度の高さ (カッパ係数)	かなり高い (0.7494)	かなり高い (0.6473)
	Z値、p値	偶然ではない (Z値 = 16.526、p値 = 2.372e-61)	偶然ではない (Z値 = 11.185、p値 = 4.833e-29)

生成AIモデルの比較：分析

■ 観点 1

- 出力されるアドバイス項目の総数が1.5倍に増加：152個→220個
- 見逃したアドバイス項目数が60%減少：46個→20個
 - 特に、「設問の補足」の内容を基にしたアドバイスの漏れが大きく減った（40個→16個）
- 不要な指摘のアドバイス項目数に変化なし：62個→62個
 - 継続：20%、解消：40%、新規：40%
 - アドバイス項目単位では傾向を特定できなかった

■ 観点 2

- 評価の一致度はどちらもかなり高く、偶然でもない
- Not Rated でないのに Not Rated と評価するチェック項目が増加：2件→17件
 - 「設問の根拠」が「設問の補足」の内容に全く触れられていない場合に Not Rated と評価されることが多い
 - プロジェクトレベルでの決めごとを問う設問に対して Not Rated と評価されることが多い

生成AIモデルの比較：考察

■ 考察

- より新しい生成AIモデル（GPT-4o → GPT-4.1）を使用すると**アドバイスの質が向上**する
- 一方で、生成AIの能力向上により、入力（指示、設問、設問の補足、評定の根拠）をより厳密に解釈できるようになったためか、評定については**柔軟性が低下**した
- 不要な指摘の変化は傾向が掴めていないため対策が難しい

■ 対策

- プロンプト
 - 多層的アプローチでの「設問にミートしない回答」に対する仕掛けを調整する
- チェックリスト
 - 「設問」の問いかけ方を見直す

**新しい生成AIモデルを使用するだけでアシスタントとしての能力が向上する
より能力を引き出すにはプロンプトや設問などの調整が必要**

まとめと今後の展望

■ まとめ

- ソフトウェア技術者自らが自己診断の形式でプロセスアセスメントを行う方法（セルフアセスメント）があるが、「開発プロセスの実態を表したものになりにくい」という問題がある
- そこで、セルフアセスメントの実施を生成AIに支援させる方式を考え、その実現可能性を評価した
- 試行の結果、生成AIは初級アセッサーと同等レベルで納得感のあるアドバイスの出力が可能であり、**生成AIによるセルフアセスメント支援が実現可能であると判断できる**

■ 今後の展望

- アドバイスの質向上
 - さらに新しい生成AIモデル（GPT-5など）の活用
 - プロンプトチューニング、設問の見直し
- 実運用に向けた取り組み
 - セルフアセスメントを手軽に実施できる環境（ツール）の実装
 - 標準プロセス定義書との統合（生成AIに与える専門知識として標準プロセスの定義内容を用いる）

DENSO

Crafting the Core

参考文献

1. IPA/SEC, プロセス改善ナビゲーションガイド ～虎の巻編～, 2009/2/25
2. IPA/SEC, プロセス改善ナビゲーションガイド ～プロセス診断活用編～, 2007/3/30
3. SPICE Center Presents SPICE Assessment Models and Reference Models, <https://intacs.info/index.php/spice-center>
4. Jan Morenzin, Automotive SPICE® News and data from VDA QMC, 1st Asia SPICE Conference
5. AIを用いたソフトウェア開発, <https://www.ipa.go.jp/digital/ai/software-engineering.html>
6. Assessor Academy, AXIOM Co-Assessor -次世代型アセスメントツール-, <https://assessor.co.jp/axiom/>
7. 多田麻沙子,徳本晋,栗田太郎,石川冬樹, ISO27017 に基づくクラウドセキュリティ監査業務に対するLLMの性能, ソフトウェア・シンポジウム2024
8. OpenAI prompt engineering, <https://platform.openai.com/docs/guides/prompt-engineering>
9. Prompt Engineering Guide, <https://www.promptingguide.ai/>
10. Jules White et al., "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT", 2023.
11. international Assessor Certification Scheme, <https://intacs.info/>
12. ビジネスキューブ・アンド・パートナーズ, Automotive SPICE 4.0 実践ガイドブック, 日経BP, 2024年1月22日
13. カッパ係数 (Cohen's kappa statistic) の情報まとめ, https://qiita.com/kazuo_reve/items/384c29d46cb3a18f827a
14. J. Richard Landis and Gary G. Koch, The measurement of observer agreement for categorical data., Biometrics, Vol. 33, No. 1 (Mar., 1977), pp. 159-174