

生成AIを活用したソフトウェア開発プロセスの

セルフアセスメントアシスタント(AI-ProSaA)の提案

第40年度（2024年度）ソフトウェア品質管理研究会
研究コース1 ソフトウェアプロセス評価・改善
2025年3月7日（金）

研究員 : 池永 直樹（株式会社デンソークリエイト）
intacs®認定 Automotive SPICE Competent Assessor
主査 : 田中 桂三（オムロン株式会社）
副主査 : 中森 博晃（パナソニック コネクト株式会社）

アジェンダ

- 背景
- 解決策の提案
- 実験
- 実験結果と考察
- まとめ
- 今後の展望

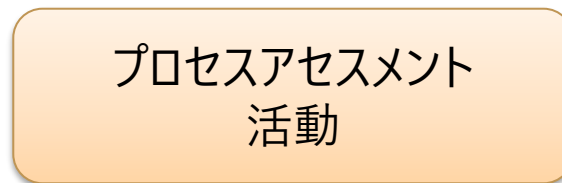
背景：プロセスアセスメント（1）

■ 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^[2]
- 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^[3]
- **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^[3]
- **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がプロセスの診断を実施する

【入力】

- 事業背景
- アセスメント目的
- アセスメント範囲
- アセスメント要求事項
- アセスメント制約事項
-



【出力】

- データ及び証拠
- アセスメント報告書
- アセスメント記録



【役割及び責任】

- 依頼者
- リードアセッサ／アセッサ

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
- 2022年には、グローバルで年間約1500回実施されている^[5]

■ 監査とプロセスアセスメントの違い

- プロセス監査

主に、ルールや規則に対する不適合事項を摘出する

- 決められたことが確実に実施されているか／守られているか

- プロセスアセスメント

プロセスアセスメントモデルをモノサシに使い、**組織の目標やプロジェクトの特性を勘案**して仕事のやり方を診断し、仕事のやり方の強みや弱みを把握する

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、インタビューする十分な時間が必要
- 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重要である

背景：セルフアセスメント実施における問題

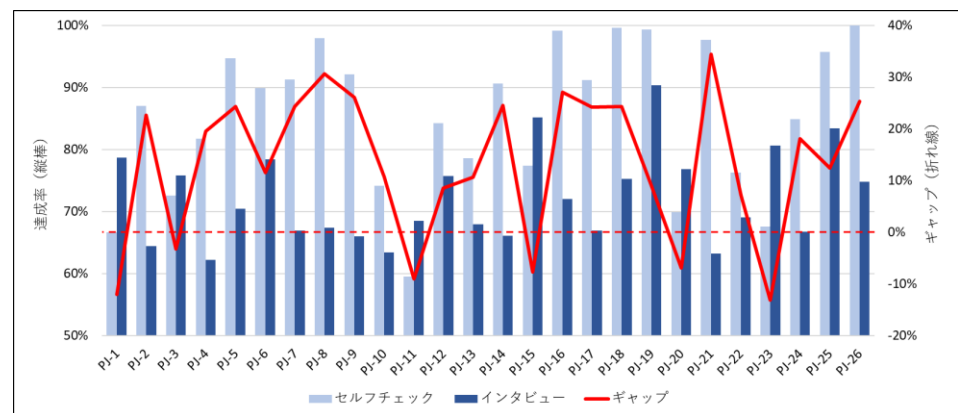
開発現場の技術者によるセルフアセスメントの結果が開発プロセスの実態を表したものにならない

■ 要因

- 開発現場の技術者は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

■ データ：セルフチェックと専門家によるチェック結果の差

- 専門家と同じ結果
($\pm 8.75\%$ とした)となったのは
19%(5/26 プロジェクト)
- 全体的にはセルフチェック
結果を甘く付ける傾向



図：セルフチェック結果及び専門家によるインタビュー形式でのチェック結果の点数とそのギャップ

背景：セルフアセスメント展開時の課題

目的

プロセスアセスメントを活用したプロセス改善活動の活性化



手段

自己診断の形式で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らない



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：仮説

■ 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^[6]

■ ソフトウェア開発プロセスへの生成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^[7]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

仮説

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

解決策の提案：AI-ProSaA (アイプロッサー)

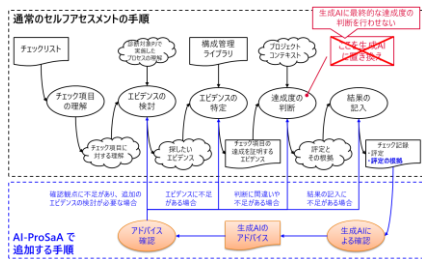
生成AIを活用したセルフアセスメントアシスタント

Generative AI-utilized Process Self-assessment Assistant

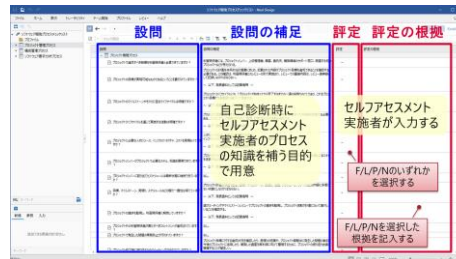
■ 特徴

- 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
- チェック項目に対する回答（評定の根拠）に対して生成AIにアドバイスさせるセルフアセスメントを支援する方式

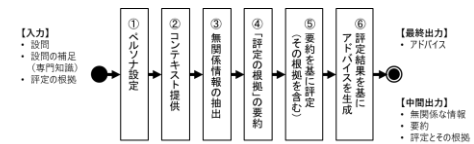
■ 構成要素



生成AIを活用するプロセス



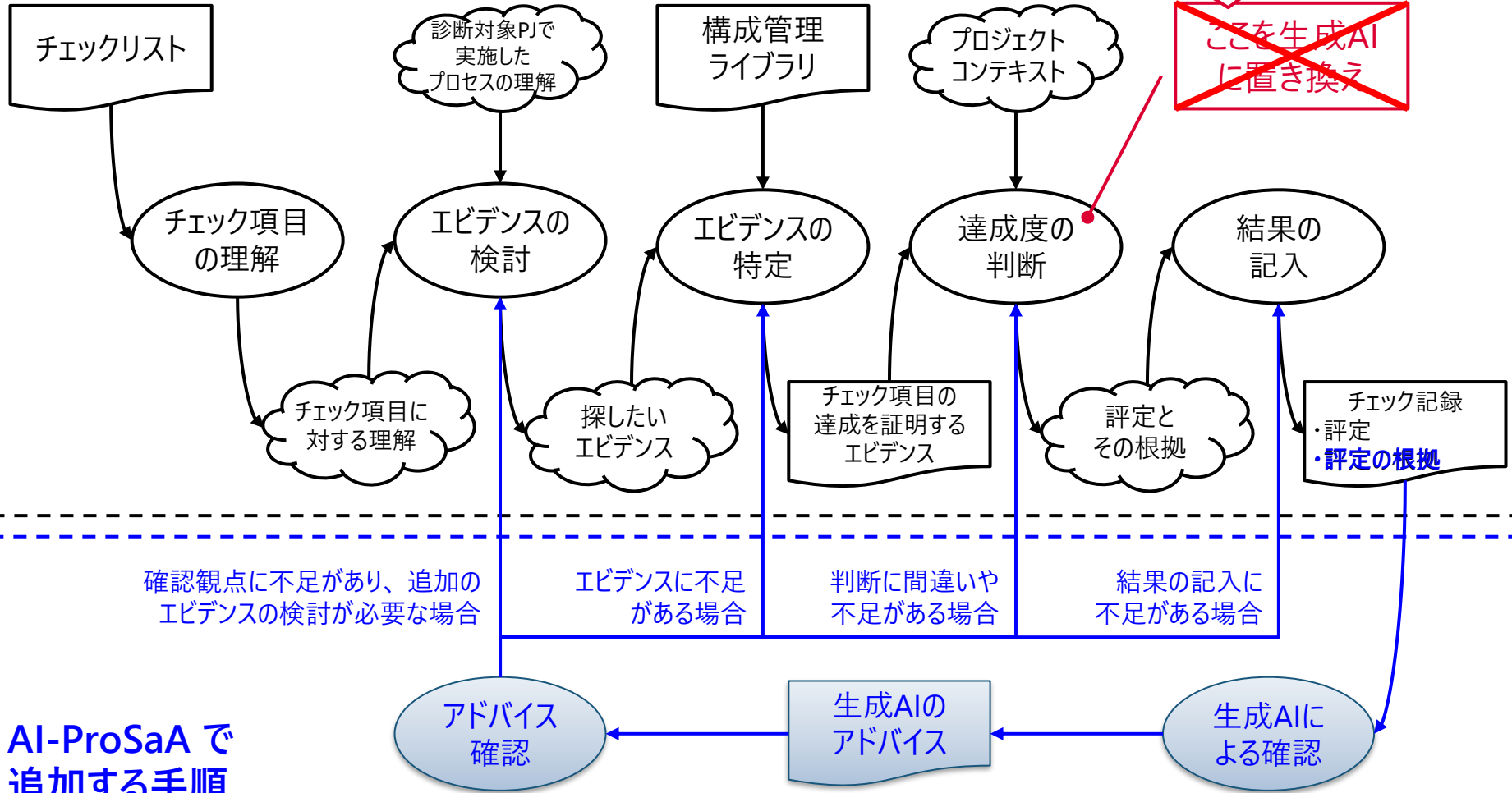
チェック項目の構成



プロンプトテンプレート

解決策の提案：生成AIを活用するプロセス

通常のセルフアセスメントの手順



AI-ProSaA で追加する手順

解決策の提案：チェック項目の構成

ソフトウェア開発プロセスチェックリスト - Next Design

ファイル ホーム 表示 トレーサビリティ チーム開発 プロファイル レビュー ヘルプ

ソフトウェア開発プロセスチェックリスト

- プロジェクト管理プロセス
- 構成管理プロセス
- ソフトウェア要求分析プロセス

Excel出力用

設問	設問の補足	評価	評価の根拠
<input type="checkbox"/> プロジェクト管理プロセス	設問の補足	評価	評価の根拠
<input type="checkbox"/> プロジェクトで達成すべき目標を利害関係者と同意できていますか？	利害関係者には、プロジェクトメンバー、上級管理者、顧客、委託先、開発環境のサポート窓口、関連する他プロジェクトなどが考えられる。 プロジェクトの計画を体系的な計画書にまとめ、記載された内容がプロジェクト目標を達成できることを確認する必要がある。この確認は、利害関係者とのレビューの形で実施され、レビューでの審議内容は、レビュー議事録として記録しなければならない。 ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> プロジェクトの目標が実現可能なものであるということを裏付けていますか？	プロジェクトの計画を体系的な計画書にまとめ、記載された内容がプロジェクト目標を達成できることを確認する必要がある。この確認は、利害関係者とのレビューの形で実施され、レビューでの審議内容は、レビュー議事録として記録しなければならない。 ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> プロジェクトのマイルストーンやそれらに基くライフサイクルは明確ですか？	プロジェクトの計画を体系的な計画書にまとめ、記載された内容がプロジェクト目標を達成できることを確認する必要がある。この確認は、利害関係者とのレビューの形で実施され、レビューでの審議内容は、レビュー議事録として記録しなければならない。 ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> プロジェクトライフサイクルを通じて実施する活動は明確ですか？	プロジェクトの計画を体系的な計画書にまとめ、記載された内容がプロジェクト目標を達成できることを確認する必要がある。この確認は、利害関係者とのレビューの形で実施され、レビューでの審議内容は、レビュー議事録として記録しなければならない。 ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> 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？	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？ ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> プロジェクトメンバーがプロジェクトに必要なスキル、知識を獲得できていますか？	プロジェクトメンバーがプロジェクトに必要なスキル、知識を獲得できていますか？ ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> プロジェクトメンバーに割り当てたスケジュールは最新状態で維持できていますか？	プロジェクトメンバーに割り当てたスケジュールは最新状態で維持できていますか？ ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> 目標、マイルストーン、見積り、スケジュールなどの間で一貫性は保っていますか？	目標、マイルストーン、見積り、スケジュールなどの間で一貫性は保っていますか？ ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> プロジェクトの進捗を監視し、利害関係者に報告していますか？	プロジェクトの進捗を監視し、利害関係者に報告していますか？ ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> プロジェクト外の利害関係者が果たすべきコミットメントが達成されていますか？	プロジェクト外の利害関係者が果たすべきコミットメントが達成されていますか？ ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> プロジェクトで発生した問題の再発防止が行われていますか？	プロジェクトで発生した問題の再発防止が行われていますか？ ～ 以下、発表資料としては記載省略 ～	-	
<input type="checkbox"/> プロジェクト終了時に振り返りを実施し、メンバーが引き継がれていますか？	プロジェクト終了時に振り返りを実施し、メンバーが引き継がれていますか？ ～ 以下、発表資料としては記載省略 ～	-	

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

セルフアセスメント実施者が入力する

F/L/P/Nのいずれかを選択する

F/L/P/Nを選択した根拠を記入する

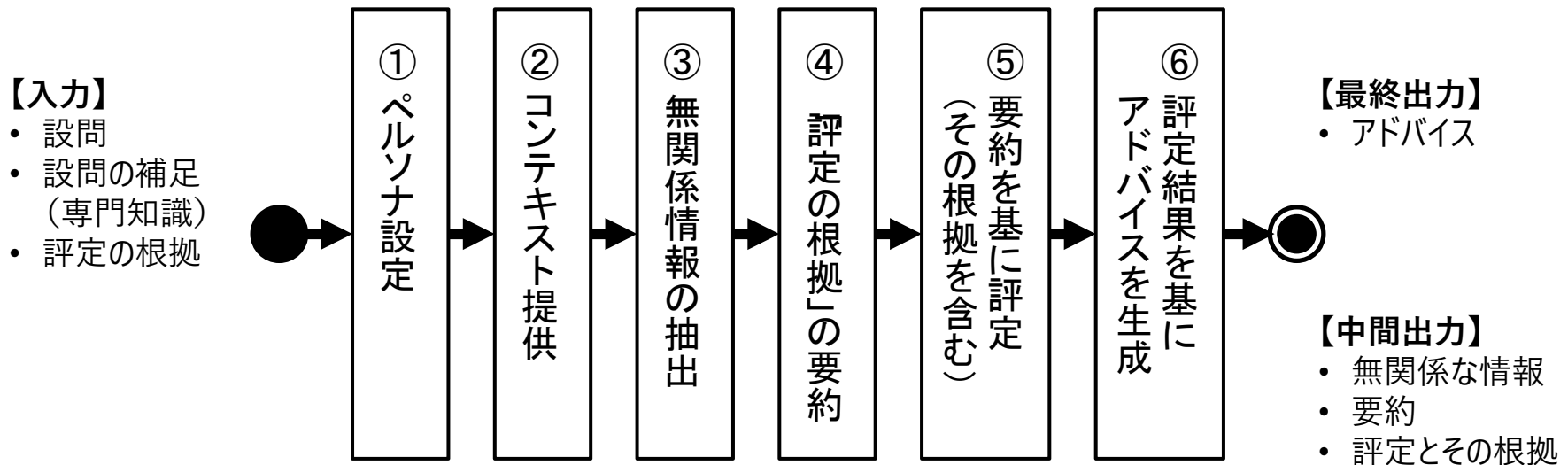
解説策の提案：プロンプトテンプレート（1）

■ 利用する生成AIモデル

- gpt-4o

■ プロンプトの全体像

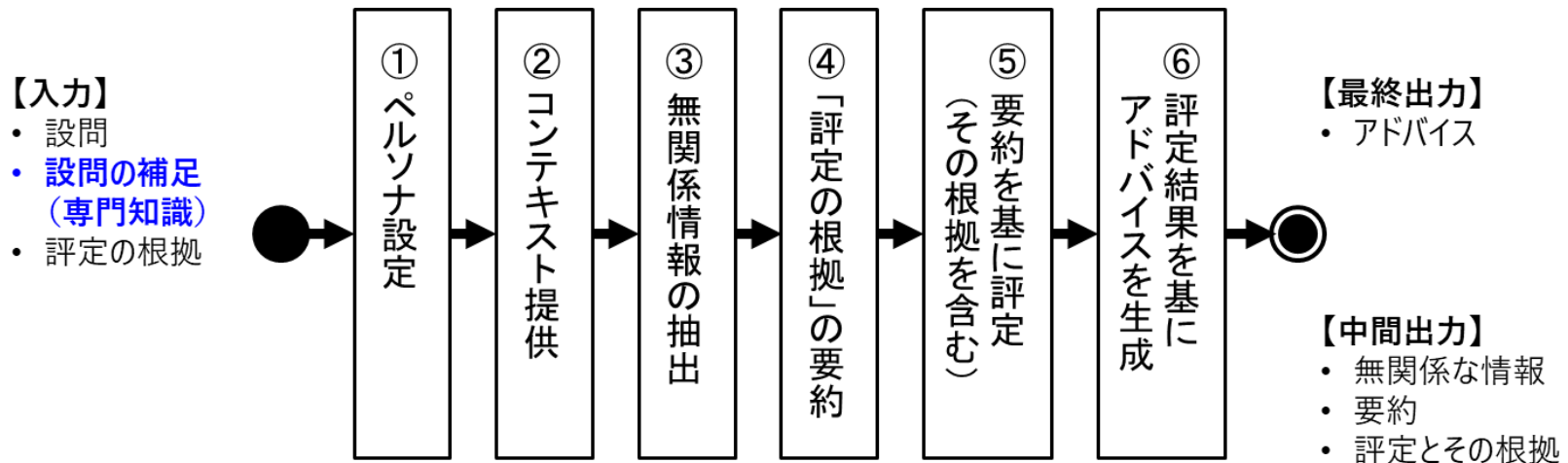
- ペルソナ設定、コンテキスト提供、出力テンプレート、Chain of Thoughtなどを組み合わせてプロンプトテンプレートを作成



解説策の提案：プロンプトテンプレート（2）

■ ポイント1：専門知識

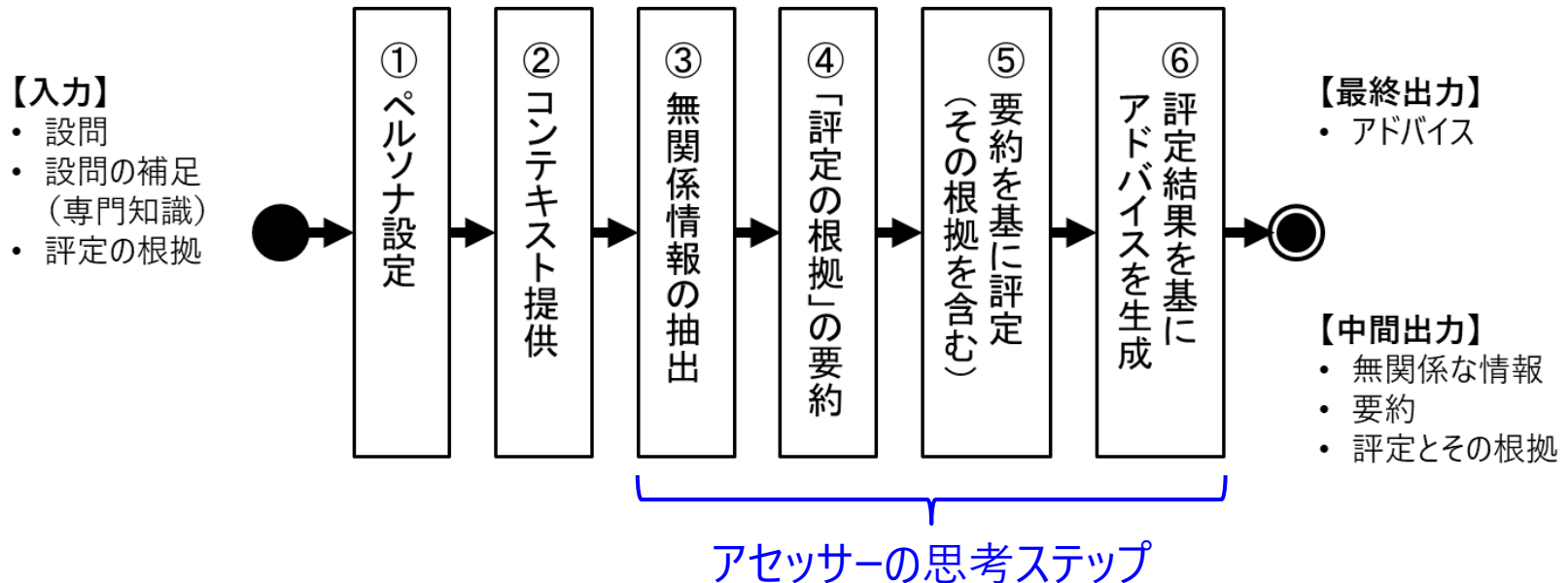
- 汎用AIから有用なアドバイスを得るには、プロセス知識やその組織や製品分野におけるプロセスの特徴などの**専門知識が必要**
- チェック項目の構成要素の一つである「設問の補足」がそれに該当するので、これを生成AIに入力



解説策の提案：プロンプトテンプレート（3）

■ ポイント2：アセッサーの思考ステップ

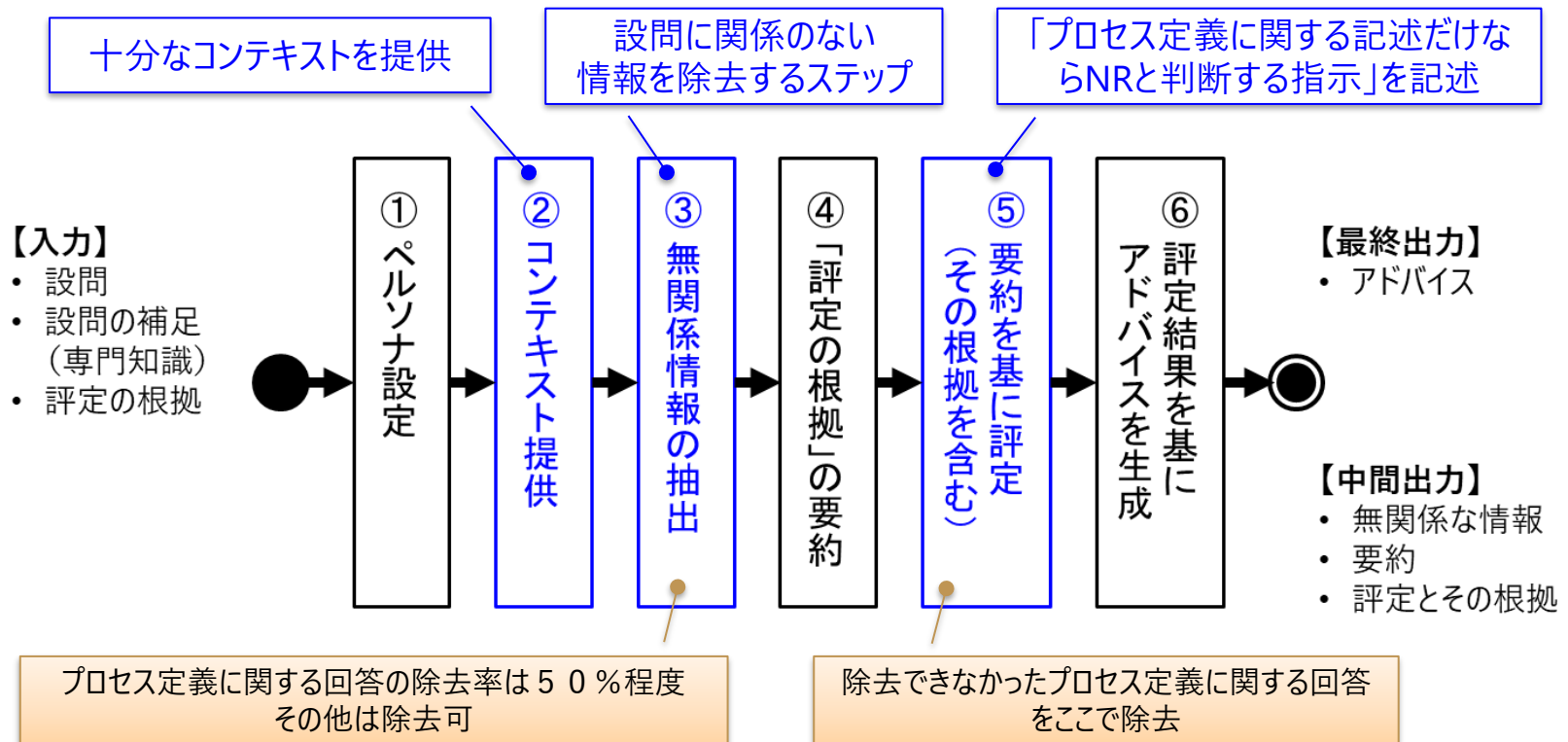
- 初期のプロンプトは要約（④）させた後にアドバイス生成（⑥）の流れとしていたが、指示をいくら調整してもアドバイスの質が向上しなかった
- そこで、アセッサーの思考フローに基づき推論ステップを分割し、要約を基に評価（⑤）するステップを追加



解説策の提案：プロンプトテンプレート（4）

■ ポイント3：多層的アプローチ

- 設問に関係のない情報が「評価の根拠」に記述される可能性がある
- それを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



解決策の提案：研究課題

AI-ProSaA がセルフアセスメントの支援に有効であるかを評価する。

RQ1

アセッサー（筆者）と同じレベル（見逃し・不要な指摘ゼロ）でアドバイスできるか

RQ2

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2 / 3 以上提供できるか

納得感の有無を比べたときに明らかに上回っていることを確認するために、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「特別多数決」の2 / 3を基準値とする。

RQ3

生成AIのアドバイスを受けて診断結果が1 / 3 以上見直されるか

以下のような場合があり、診断結果の見直しはRQ2より少ないと考えた。

- 何らかの気づきはあるが評価が変わる程ではない
 - 評価の根拠に記入しなかったことに対してアドバイスを受けたので評価は変わらない
- そこで、納得感を得たアドバイスのうち半数以上が見直しされれば有効なアドバイスを出力できたと考え、基準値をRQ2の半分の1 / 3とする。

実験：評価観点

RQ1

アセッサー（筆者）と同じレベル（見逃し・不要な指摘ゼロ）でアドバイスできるか

- アドバイスの内容に見逃しがあったチェック項目
- アドバイスに不要な指摘が含まれているチェック項目

RQ2

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2 / 3以上提供できるか

ISO25010 の利用時品質を評価観点として用いる。

利用時品質	アセッサー	セルフアセスメント実施者
有効性	正確かつ完全であるか	理解できたか
効率性	分量は適切であるか	分量は適量であったか
満足性	診断を見直すキッカケになりそうか	診断を見直すキッカケになったか

RQ3

生成AIのアドバイスを受けて診断結果が1 / 3以上見直されるか

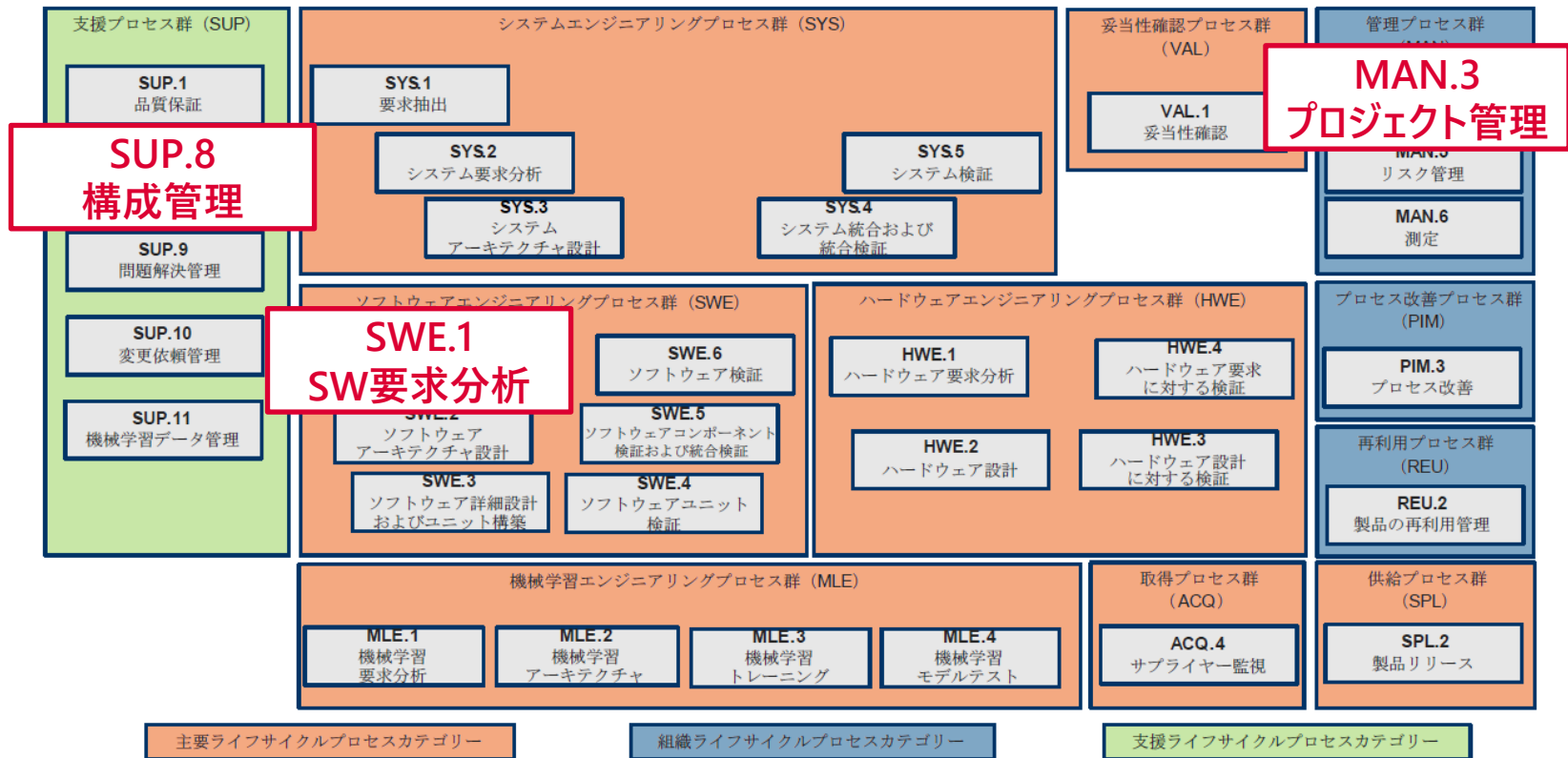
アドバイスを受けてセルフアセスメント実施者が診断結果を見直したか。

実験：実験対象プロセスとチェック項目

Automotive SPICE V4.0 から以下の3プロセスを選択

Automotive SPICE 4.0 実践ガイドブック^[9]に掲載のチェック項目を使用

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



出典：Automotive SPICE PAM V4.0 日本語版

実験：実験データ

合計 161 件の実験データを用いた

- 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

実験結果：RQ1

RQ1

アセッサー（筆者）と同じレベル（見逃し・不要な指摘ゼロ）でアドバイスできるか

	チェック項目	アドバイス項目数*
見逃し	28% (45 / 161件)	46個
不要な指摘	31% (50 / 161件)	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. 見積もりに使用された方法やデータの詳細を見積もり計画書等に明確に記載し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

考察：RQ1（1）

■ アセッサーとの比較

- 筆者(Competent Assessor)と比較すると見逃し・不要な指摘は多い
 - 「筆者がアドバイスを見たときに、見逃し・不要な指摘がゼロであること」が期待値
- 一方で、初級アセッサー*と同等レベルであった
 - 初級アセッサーと同等レベル = 実務で活用可能なレベル

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

* 初級アセッサー：筆者が所属する組織の 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

RQ1に対し完全に有効だったと言えない（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ることが確認できた

考察：RQ1（2）

■ 「見逃し」についての考察

分析

- いずれのチェック項目でも同じ内容を見逃す傾向が確認できた
- 見逃した内容はいずれも「設問の補足」に記述されている内容

対策

解説書のように読み物として記述した「設問の補足」でなく、
診断観点リストとして要点に絞って入力する

予想効果

	実験結果	予想効果
チェック項目	28% (45 / 161件)	6% (9 / 161件)
アドバイス項目数	46個	10個

考察：RQ1（3）

■ 「不要な指摘」についての考察

分析

#	分類	説明	個数(比率)
1	オウム返し	「評定の根拠」に“〇〇ができていない”と記述している内容をそのまま“〇〇を確認して下さい”と出力	28(45%)
2	過剰な要求	その設問の範囲内であるが過剰と考えられる内容	25(40%)
3	無関係	その設問の範囲外で明らかに無関係と判断できる内容	2(3%)
4	見当違い	その設問の範囲内であるが「評定の根拠」の内容に対して明らかに見当違いと判断できる内容	7(12%)

対策

分類#1 : プロンプトチューニングで対策できる

分類#2-4 : 対策が困難

⇒ 分類#2はやりすぎにはなるが、設問の範囲内であるため出力されても問題ではない

⇒ 分類#3,4はアドバイスの質を落とすものであるが、セルフアセスメント実施者のプロセス知識に関わらず明らかに不要な内容と判断できると考えられることから実害はない

予想効果

	実験結果	予想効果
チェック項目	31% (50 / 161件)	18% (29 / 161件)
アドバイス項目数	62個	34個

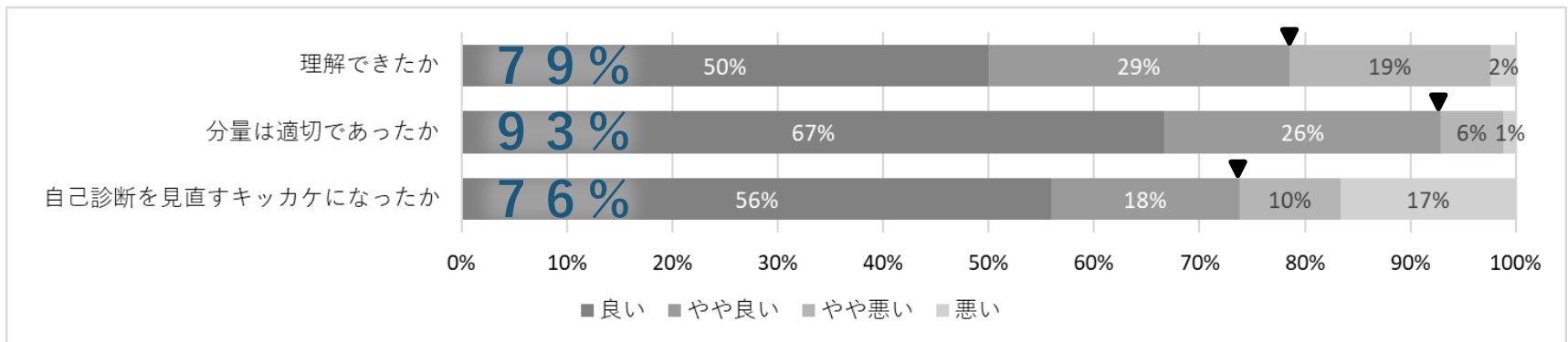
実験結果：RQ2

RQ2

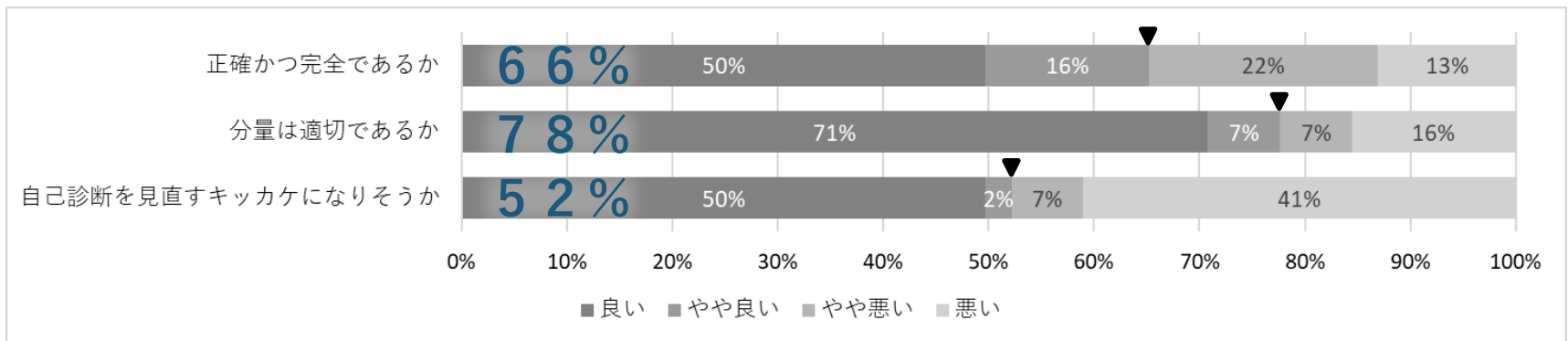
セルフアセスメント実施者にとって納得感のあるアドバイスを
2 / 3 以上提供できるか

* 青字：「良い」「やや良い」の割合

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



生成AIのアドバイスに対するアセッサー(筆者)の評価



考察：RQ2

■ セルフアセスメント実施者の評価

- 「やや良い」以上
 - 有効性：79%、効率性：93%、満足性：76%
- 得られたコメント
 - 肯定的⇒「知識が得られた」、「間違いに気づいた」
 - 否定的⇒「できていないと回答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」

■ アセッサー(筆者)の評価

- セルフアセスメント実施者より悪いのは、見逃しと不要な指摘に気づいているからである

**セルフアセスメント実施者の「やや良い」以上の評価が
基準値の2 / 3をいずれも上回っていることから、
RQ2に対し有効だったと判断する**

実験結果と考察：RQ3

RQ3

生成AIのアドバイスを受けて診断結果が1 / 3 以上見直されるか

セルフアセスメント実施者が自己診断結果を見直した割合：32%

- 見直し結果の妥当性
 - 実施された開発プロセスの実態を表した結果に修正されていることを確認
- 得られたコメント
 - 肯定的⇒「気づきを得た」「質問に回答できないことが分かった」
 - 想定通り⇒「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」

基準値の1 / 3 に対して僅か1%の差異(近似値)と捉え、RQ3に対し概ね有効だったと判断する

考察：妥当性への脅威

- 実験データは主に筆者所属組織の要員のものであり、十分な標本とは言い切れない
 - 他のセルフアセスメント実施者の回答を使用した場合に傾向が変わる恐れがある
- 今後、セルフアセスメント実施者を増やしより多くの実験データでの傾向を確認したい

まとめ

背景

プロセスアセスメントを活用し、プロセス改善活動の活性化したい
セルフアセスメントを展開したいが、
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らない
問題がある

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られる

解決策

セルフアセスメント実施を支援する目的での**生成AIの活用**として
アシスタント方式である AI-ProSaA を考案した

結論

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、自己診断の見直しにも役立つことが確認できた
⇒ 妥当性への脅威はあるが、**課題を解決できる**と判断した

今後の展望

以下に取り組み、生成AIによるアドバイスの質をアセッサー（筆者）のレベルまで向上させる

1. 見逃し・不要な指摘に対する対策の実施
2. OpenAI o3 など新しい生成AIモデルの活用

謝辞

論文作成にあたり、ご指導・ご協力いただいた
皆様に深くお礼申し上げます。

研究コース 1

指導員の皆様

研究員の皆様

実験に協力してくださった方々

参考文献

1. international Assessor Certification Scheme, <https://intacs.info/>
2. 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
3. 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
4. 堀田勝美, 日本におけるプロセスアセスメント活動, 情報処理学会 短期集中セミナー, 2020/1/10
5. Jan Morenzin, Automotive SPICE® News and data from VDA QMC, 1st Asia SPICE Conference
6. AIを用いたソフトウェア開発, <https://www.ipa.go.jp/digital/ai/software-engineering.html>
7. Assessor Academy, AXIOM -次世代型アセスメントツール-, <https://assessor.co.jp/axiom/>
8. 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
9. Business Cube & Partners, Automotive SPICE 4.0 実践ガイドブック 入門編, 日経BP, 2024/1/22
10. OpenAI prompt engineering, <https://platform.openai.com/docs/guides/prompt-engineering>
11. Prompt Engineering Guide, <https://www.promptingguide.ai/>
12. Jules White et al., "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT", 2023.