

AI エージェントによる要求仕様改善における観点表粒度の影響

Impact of Viewpoint-Table Granularity on AI-Agent-Assisted Requirements Specification Improvement

○中川 桂 田口 真義¹⁾ 早乙女 暖²⁾

東京海上日動システムズ株式会社

○Kei Nakagawa Masayoshi Taguchi¹⁾ Yutaka Saotome²⁾

Tokio Marine & Nichido Systems Co., Ltd.

Abstract This study evaluates two viewpoint tables in AI-agent-assisted review of requirements specifications. An earlier requirements-review study showed different outcomes among no-table, common-table, and product-family-specific-table conditions, but did not make every viewpoint-to-proposal relationship traceable. We therefore compared a 22-item common table with a 39-item product-family-specific table using a public electric-pot specification and a workflow that processed every viewpoint, linked proposals to existing requirement identifiers, and applied quality gates. Three researchers independently evaluated 25 safety-related items, and final scores were determined by reviewing disagreements and applying a median-based aggregation rule. Both conditions covered all 25 items, while mean finding-effectiveness scores were 1.56 and 1.92, respectively. The results suggest that explicit workflow control supports coverage and that product-specific knowledge can improve proposal specificity. Human judgment remains essential for adoption.

1. はじめに

ソフトウェア開発現場では、熟練開発者の退職や新規人材の確保難により、人的リソース不足が深刻化している。一方、製品・サービスの高度化に伴い、要求仕様は複雑化・大規模化しており、確認すべき項目数や観点は今後さらに増加すると考えられる。人的リソース不足は人口構造の変化等に起因する構造的課題であり、個々の企業努力のみで抑制することは困難である。このような状況下で品質を継続的に維持・向上させるには、膨大な要求仕様の確認に生成 AI を適用し、業務効率化を図ることが有効な施策の一つと考えられる。特に、網羅的な観点に基づく要求仕様の確認や補完など、反復的かつ機械的な業務では、作業負荷の軽減と成果物品質の平準化が期待できる。

従来、要求仕様書の作成は、熟練開発者をはじめとする専門知識を有する担当者が中心となり、対象製品群に関する暗黙知も参照しながら要求仕様を整理し、チームレビューを経て確定していた。しかし、人的リソース不足が進行する中では、従来の運用のみで要求仕様の品質を維持・向上させることは困難になると考えられる。

本研究では、従来のチームレビューに先立ち、生成 AI によるレビュー工程を導入した要求仕様の改善プロセスを対象とする。具体的には、レビューアが確認すべき事項を整理した観点表を生成 AI に与え、その内容に基づいて要求仕様をレビューさせる手法を検討する。

先行研究「生成 AI を利用するシステムの安全性評価を支援するテスト観点表の提案」では、生成 AI を利用するシステムの安全性評価に関する知見を 18 のテスト観点として体系化し、人と生成 AI によるテストケース作成に適用している。その結果、観点表を参照した場合、作成されるテストケースの件数および観点数が増加し、抜け漏れや属人性を低減できる可能性が示された。一方で、題材との関連性が低いテストケースも増加したことから、観点を与えるだけでなく、対象との関連性を確認する処理の必要性が報告されている[1]。

また、先行研究「生成 AI を用いた要求仕様改善における観点表粒度の違いとその影響」では、要求仕様レビューへの観点表適用を検討し、観点表なし、対象製品分野に依存しない汎用的な共通観点表、および対象製品の機能・状態・操作・異常事象まで具体化した製品群別観点表の 3 条

連絡先：東京海上日動システムズ株式会社 中川 桂

〒206-8510 東京都多摩市鶴牧 2-1-1 多摩東京海上日動ビル TEL：042-356-2415

Tokio Marine&Nichido Systems Co.,Ltd., 2-1-1 Tsurumaki, Tama-shi, Tokyo 206-8510, Japan

1) リコーITソリューションズ株式会社 〒224-0035 神奈川県横浜市都筑区新栄町 16-1

2) キヤノン株式会社 〒146-8501 東京都大田区下丸子 3-30-2

Keywords：生成 AI, 要求仕様, AI エージェント, 観点表, 粒度, 指摘有効性

件において、Copilotのエージェントによる評価を実施し、改善結果について比較した。そのうえで、各条件における改善結果を比較・分析した。その結果、25評価項目に対する観点網羅率は48%、68%、84%、記述妥当性は0.68、0.96、1.28となり、製品群別観点表を用いた条件で最も高い結果を得た[2]。ただし、二つの観点表は粒度に加え、観点数および記載内容も異なるため、この差を粒度のみの効果として解釈することはできない。また、すべての観点を評価する指示は与えていたものの、各観点と改善提案の対応関係を追跡できなかったため、観点表の処理状況と改善結果を分離して評価する課題が残った。

そこで本研究では、観点表に記載されているすべての観点を明示的に処理し、既存の要求仕様との対応付けおよび品質確認を行うAIエージェントワークフローを構築する。本研究で扱うResearch Question（以下、RQ）は次の二点である。なお、本研究では観点表ありの2条件に対象を絞り、観点表の有無は再評価しない。

RQ1：すべての観点を処理する同一ワークフローの下で、共通観点表と製品群別観点表が同一の25評価項目を漏れなく扱えるか。

RQ2：同一ワークフローで処理した場合、二つの観点表から生成される改善提案の指摘有効性はどのように変化するか。

本研究の成果物は、既存の要求仕様に対する改善提案である。生成AIは要求仕様の不足や曖昧さを指摘し、人が検討すべき事項を提示する役割を担う。最終的な仕様の採否は、妥当性、コスト、実現可能性等を踏まえて人が判断する。また、本研究における安全性は、複数観点から不足を確認するための題材であり、安全性という品質特性自体の優位性を検証するものではない。

2. 提案アプローチ

本研究では、要求仕様の改善を支援する手法として、生成AIに対し、レビュー対象となる要求仕様書および観点表を入力情報として与える。そして、観点表に記載された観点に従って生成AIによるレビューを実施し、その結果を参考情報として人が要求仕様書を改善するアプローチを用いる。

2.1 観点表の粒度と役割

本アプローチにおいて用いる観点表は、要求仕様書レビュー時に確認すべき事項を整理したものであり、生成AIが要求仕様を評価する際の参照条件として機能する。本研究では、共通観点表と製品群別観点表の2種類を用いる。両者は、観点数、記述の具体性、記載内容が異なるため、本研究ではこれらを含む観点表構成の違いとして扱う。本研究で適用した2種類の観点表の概要を表1に示す。

表1 2種類の観点表の位置付け

観点表	観点数	記述の特徴
共通観点表	22	製品分野に依存せず、安全性の確認事項を一般的・抽象的に記述
製品群別観点表	39	対象製品群の機能・状態・操作・異常事象まで具体化して記述

共通観点表の作成においては、対象製品分野に関する事前知識を前提とせず、生成AIとの対話を通じて観点の抽出および整理を実施した。また、製品群別観点表の作成においては、対象製品分野に関する調査を実施した上で、研究チームによるレビューおよび修正を反復的に行った。これにより、製品群別観点表は共通観点表より具体性を向上させるとともに、熟練開発者の暗黙知を反映した内容の再現を図った。

2.2 AIエージェントワークフロー

本研究におけるAIエージェントは、自然言語による指示を入力として受け取り、既存の要求仕様書や観点表を参照するとともに、検証用スクリプトの実行および実行結果の評価を行いなが

ら、複数の処理工程を段階的に遂行する生成 AI 環境である。実験には OpenAI Codex を用いた。前年の研究では、観点表に記載されたすべての観点の評価を指示していたが、一部の観点到評価漏れが発生しており、網羅的な確認には至らなかった。この課題を解決するため、本研究では各観点到一意な観点 ID を付与し、各観点 ID を処理単位として順次実行する構成を採用した。これにより、観点ごとの評価実施状況を明確化し、評価漏れの抑制を図った。

また、要求仕様書を機械処理可能な形式へ正規化した後、各要求に対して REQ-XXXX 形式の一意な要求 ID を付与した。要求 ID は、各要求を識別するとともに、生成された改善提案と既存要求との対応関係を管理し、改善提案の根拠を追跡可能とするために導入したものである。観点到、関連する要求 ID、既存要求で確認できる内容、不足または曖昧な内容、改善提案、人が決定すべき事項を分けて保存する。観点表と要求仕様書だけでは数値や設計条件を確定できない場合は、「設計判断要」または「証拠不足」という不確実性ラベルを付け、AI が確定要求として書かないように指示した。本研究で適用した AI エージェントワークフローを図 1 に示す。

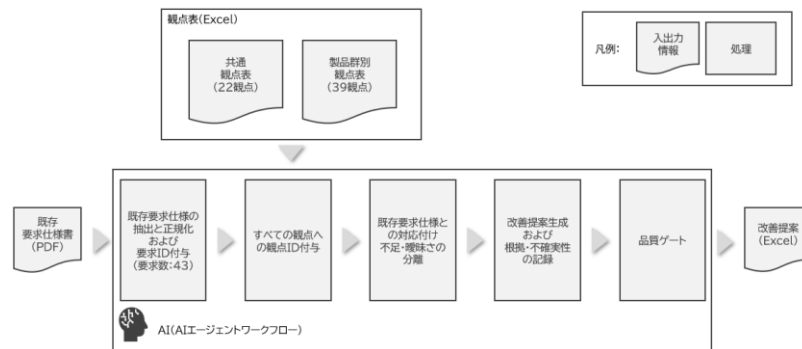


図 1 AI エージェントワークフロー

2.3 品質ゲート

品質ゲートは、生成された改善提案に対して表 2 に示す検査項目に基づく機械的な検査を実施し、改善提案の品質を担保するための工程である。具体的には、観点的処理漏れ、既存の要求仕様との対応関係の欠落、観点本文の単純な転記、および要求仕様に根拠を持たない値や条件の追加を検出することを目的とする。なお、品質ゲートは、処理の成立性を確認するための工程であり、生成された改善提案の指摘有効性を保証するものではない。品質ゲートによる検査結果の詳細は第 5 章に示す。

表 2 品質ゲートの主な検査項目

検査項目	判定内容
観点網羅性	観点表に含まれるすべての観点 ID が処理されている
根拠追跡性	生成された改善提案から関連する既存要求 ID を追跡できる
製品文脈・具体性	対象製品の機能・状態等に即し、一般的な指示だけで終わっていない
転記抑制	生成された改善提案と観点本文の語句一致率が 75%未満である
根拠のない追加防止	確定できない値・条件を追加せずに不確実性が明示されている

3. 実験

3.1 実験目的と対象

実験対象は、組込みシステム教育教材「話題沸騰ポット GOMA-1015 型 要求仕様書 第 3 版」である[3]。家庭用電気ポットは身近な製品である一方、加熱や温度制御を伴い、状態・操作・異常時の安全を多面的に検討できる。また、公開教材であり、整理途上の要求として不足や曖昧さが残るため、高度な事前知識を持たない研究メンバーでも改善提案を比較しやすい。これらの理由か

ら、観点表の粒度差と処理手順を確認する題材として選定した。

本実験の目的は、観点表の有無を比較することではなく、観点表を用いる条件において、共通観点表と製品群別観点表を同一の AI エージェントワークフローで処理し、生成結果を同一基準で比較することである。

入力 PDF を Markdown 形式へ正規化し、前提、制約、例外、注記、数式、未確定事項を保持した。正規化後の文書から 43 件の要求を識別し、REQ-XXXX 形式の要求 ID を付与した。二つの条件には、同一の正規化文書と要求 ID を入力した。

3.2 実験条件

表 3 の 2 条件を設定した。対象要求仕様書、正規化文書、要求 ID、AI エージェントへの処理指示、品質ゲート、評価項目は共通とし、入力する観点表だけを変更した。処理対象となる観点数と実行回数は表 3 に示す。

表 3 実験条件

入力条件	処理対象・実行回数
共通観点表 (22 件)	22/22 件, 1 回
製品群別観点表 (39 件)	39/39 件, 1 回

両条件の差分は入力する観点表のみである。ただし、二つの観点表は観点数、記述の具体性、記載内容が異なるため、本実験は粒度のみの効果を分離する設計ではない。

3.3 処理手順

各条件には第 2 章で定義した同一のワークフローを適用した。入力観点表の観点 ID 集合を事前に固定し、観点 ID ごとの出力 ID 集合と照合したうえで、表 2 の品質ゲートを適用した。これにより、最終的な改善提案だけでなく、入力観点が処理され、根拠要求と対応付けられたことも確認対象とした。

具体的には、各条件について、入力観点 ID 集合の固定、正規化済み要求仕様書と要求 ID 一覧の入力、観点 ID ごとの関連要求探索、改善提案と人が判断すべき事項の記録、品質ゲートの適用、人間評価用資料の生成を行った。

3.4 実験出力

各条件の実行結果として、観点 ID ごとの追跡記録、品質ゲート結果、人間評価用資料を生成した。観点 ID ごとの追跡記録には、対象観点、関連する既存要求 ID、既存要求で確認できる内容、不足または曖昧な内容、改善提案、人が判断すべき事項、設計判断要または証拠不足のラベルを含めた。

4. 評価方法

4.1 評価項目と指標

安全性に関する同一の 25 評価項目を用意した。この 25 項目は前年研究と同じ基準を用いるための評価項目であり、表 1 の入力観点とは役割が異なる。入力観点は生成 AI が改善提案を作るための手掛かり、25 評価項目は生成結果を両条件で比較するための基準である。したがって、観点 ID と評価項目 ID は一対一に対応しない。

主要指標は観点網羅率と指摘有効性であり、判定・算出方法を表 4 に示す。観点網羅率は入力観点の処理率ではなく、生成結果が共通の 25 評価項目をどこまで扱ったかを示す。カテゴリ平均は項目数の少ないカテゴリの影響が大きくなるため、主結果には用いず補助分析とした。

本研究では、前年研究の記述妥当性に相当する評価観点を、改善提案が要求レビューで利用できる程度を示す指標として「指摘有効性」と呼ぶ。

全体の指摘有効性平均は、カテゴリごとの平均を再平均せず、25 項目を同じ重みで集計した。一方、カテゴリ別分析は、どの種類の評価項目で条件差が生じたかを探索するために用いた。カテゴリごとの項目数は 1~4 件で異なるため、カテゴリ平均を条件全体の優劣や効果量として扱わない。

表 4 評価指標

指標	判定・算出方法
観点網羅率	対応する改善提案が明確に存在する評価項目数 ÷ 25 × 100
指摘有効性 0 点	対応する提案がない，または要求改善の指摘として利用できない
指摘有効性 1 点	関連する提案はあるが，具体性・前提が不足し，追加確認を要する
指摘有効性 2 点	人による要求レビューで利用できる具体性を持つ
指摘有効性平均	25 項目の最終合議スコア合計 ÷ 25

指摘有効性は，提案を製品仕様へそのまま採用できるかを判定する指標ではない．2 点であっても，対象製品での必要性，コスト，実装可能性，規格との関係などは別途検討が必要である．本研究では，改善提案が人による要求レビューを開始するための具体的な検討材料となるかを評価する．

4.2 人間評価と合議

執筆者 3 名が，他者の回答を参照せず，各条件 25 項目，合計 50 件を独立に評価した．評価資料には，条件，評価項目，関連する改善提案，既存要求，根拠要求 ID，不足内容，人が判断すべき事項，評価基準を同一行へ配置した．独立評価後に不一致項目を抽出し，根拠要求，改善提案，評価基準を確認した．

合議では，評価基準の適用に大きな異論がない場合は，3 名の評価値の中央値を最終値とした．具体的には，同じ点数を付けた評価者が 2 名以上いる場合はその点数を採用し，0 点，1 点，2 点に分かれた場合は中央値である 1 点を採用した．これは，評価者 3 名が対象製品である電気ポットの開発経験を有しておらず，製品仕様としての妥当性を精緻に判定することが難しいため，本研究では順序尺度上の代表値として最終スコアを定めることを優先したためである．

4.3 統計的分析

指摘有効性の条件差を補助的に確認するため，同一の 25 評価項目に対する最終合議スコアを対応のあるデータとして扱い，Wilcoxon 符号付順位検定を行った．指摘有効性は 0～2 点の順序尺度であり，正規分布を仮定せず，同一項目における条件差を順位に基づいて確認できるため，本検定を用いた．有意水準は 5%とし，差が 0 の項目を除いて検定した．

なお，本検定は今回の 25 評価項目における条件差を確認する補助情報であり，他製品への一般化や差の大きさを直接示すものではない．また，観点網羅率は両条件とも 25 項目すべてを網羅したため，統計的検定の対象とはしない．

5. 結果と考察

5.1 実験結果

5.1.1 観点網羅率および指摘有効性の結果

表 5 に，観点網羅率および指摘有効性の結果を示す．共通観点表と製品群別観点表は，いずれも 25 項目中 25 項目を網羅し，観点網羅率は 100%であった．

表 5 観点表適用形態別の観点網羅率および指摘有効性

観点表	網羅項目	観点網羅率	0点/1点/2点	指摘有効性平均
共通観点表	25/25	100%	0/11/14	1.56
製品群別観点表	25/25	100%	0/2/23	1.92

指摘有効性の平均は，共通観点表が 1.56 点，製品群別観点表が 1.92 点であった．評価項目ごとの比較では，製品群別観点表が高かった項目は 10 件，共通観点表が高かった項目は 1 件，同点は 14 件であった．

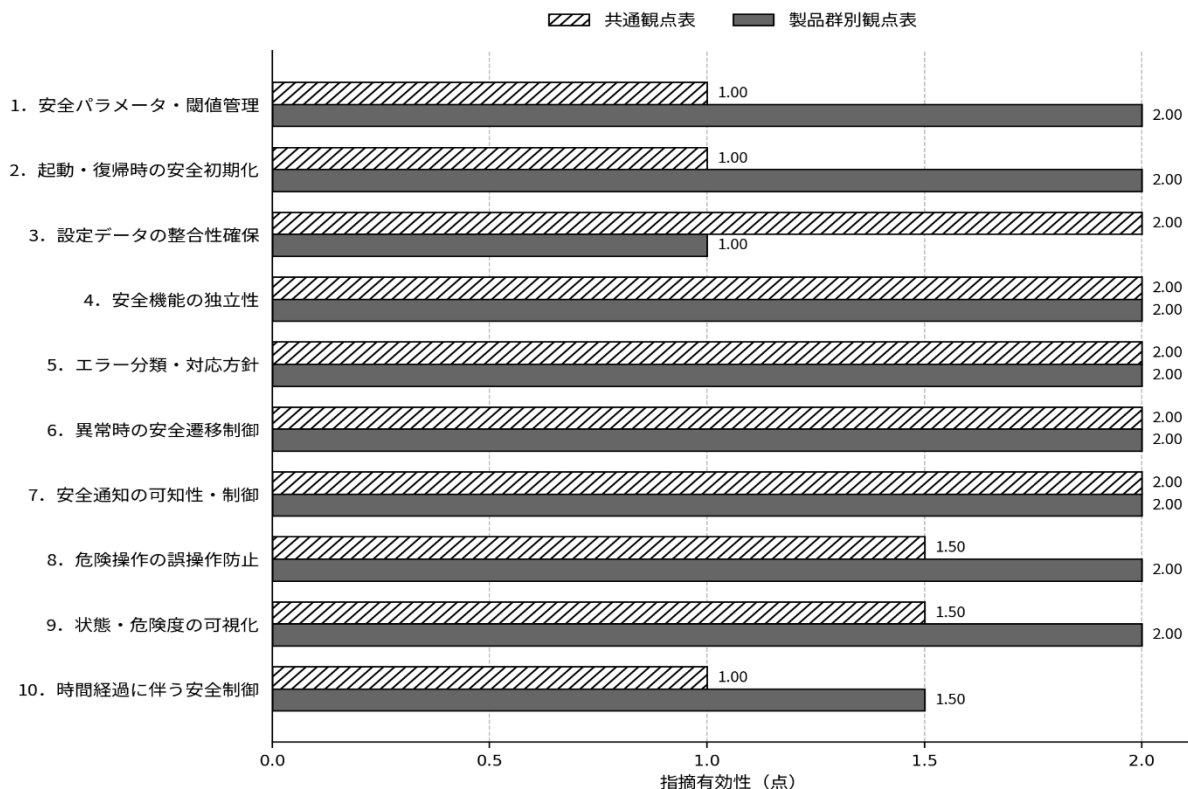


図2 カテゴリ別の指摘有効性スコア

図2に、カテゴリ別の指摘有効性を示す。10カテゴリのうち、製品群別観点表が高かったカテゴリは5カテゴリ、共通観点表が高かったカテゴリは1カテゴリ、同点は4カテゴリであった。

5.1.2 統計的分析

観点網羅率については、共通観点表と製品群別観点表のいずれも25項目すべてを網羅し、すべての項目で判定が一致した。このため、検定の対象となる条件差がなく、統計的検定は実施しなかった。

25評価項目別の指摘有効性は、Wilcoxon 符号付順位検定を用いて比較した。検定統計量 W は6.0、 p 値は0.0117であり、製品群別観点表は共通観点表より有意に高かった。

5.2 考察

RQ1：すべての観点を処理する同一ワークフローの下で、共通観点表と製品群別観点表は同一の25評価項目を漏れなく扱えるか。

品質ゲートによる検査の結果、共通観点表の22件および製品群別観点表の39件の観点がすべて処理された。さらに、表5に示したとおり、両条件とも25評価項目すべてに対応する改善提案が存在し、観点網羅率は100%であった。これらの結果より、観点IDを処理単位として管理し、品質ゲートで観点の処理漏れの有無を検査する今回のワークフローは、いずれの観点表を用いた場合も、入力したすべての観点を処理したうえで、25評価項目を漏れなく扱えたといえる。品質ゲートは、改善提案の内容の良し悪しを評価するものではないが、少なくとも入力観点の処理漏れの有無を機械的に検査する役割を果たしたといえる。一方で、観点網羅率は両条件で同じ100%となったため、観点表による違いは現れなかった。

RQ2：同一ワークフローで処理した場合、二つの観点表から生成される改善提案の指摘有効性にどのような違いがあるか。

指摘有効性平均は、共通観点表の 1.56 点に対し、製品群別観点表は 1.92 点であり、製品群別観点表の方が統計的に有意に高かった。両条件とも 25 評価項目すべてに対応する提案が得られたが、追加確認を要する提案（最終合議スコア 1 点）は、共通観点表の 11 件に対し、製品群別観点表では 2 件にとどまった。製品群別観点表の指摘有効性が共通観点表より高かった要因の一つは、提案の具体性にあると考えられる。製品の機能、状態、操作および異常事象を具体化したことで、改善対象、条件および期待動作が明確となり、要求レビューで利用しやすい提案につながった可能性がある。ただし、共通観点表と製品群別観点表には、記述の具体性に加え、観点数および扱う内容にも違いがある。このため、指摘有効性の差を、観点の粒度を細かくしたことのみの効果として説明することはできない。なお、カテゴリ別では、一部に共通観点表が上回る結果もみられた。

また、実験者からは、両観点表において、元の要求仕様で必要とされる範囲を超える提案が生成される場合があったとの所感が得られた。ただし、その妥当性は人による判断を要するため、生成結果のみから過剰提案と断定することはできない。本実験では過剰提案を判定および減点する共通基準を設けておらず、その件数も定量的に測定していない。そのため、過剰提案については指摘有効性とは区別し、補足的な所見として扱う。

6. まとめ

本研究では、共通観点表と製品群別観点表を用いた生成 AI による要求仕様レビュー支援を対象に、すべての観点を明示的に処理し、品質ゲートで観点の処理漏れの有無を機械的に検査する AI エージェントワークフローを構築した。そのうえで、観点網羅率および指摘有効性を評価した。

その結果、両条件とも入力したすべての観点が処理され、25 評価項目すべてに対応する改善提案が得られ、観点網羅率は 100% となった。今回のワークフローを用いることで、観点表の種類によらず、入力観点の未処理を残さずに確認事項を扱える可能性を示した。また、指摘有効性平均は、共通観点表の 1.56 点に対し、製品群別観点表は 1.92 点であり、製品群別観点表を用いた条件の方が統計的に有意に高かった。本実験の条件では、製品群固有の機能、状態、操作および異常事象を具体化した製品群別観点表を用いることで、人による要求レビューで利用しやすい改善提案が得られる可能性が示された。ただし、この差は、製品群別観点表における観点の具体性のみの効果として説明することはできない。

実務で利用する際は、共通観点表は、組織横断で共有すべき基本的な確認事項を広く扱い、観点表の作成・維持負担を抑える用途に適している。一方、製品群別観点表は、熟練者が有する製品群固有の具体的な暗黙知を形式知化し、生成 AI への入力条件として与えることで、より具体的な改善提案が得られると考えられる。ただし、製品群別観点表には作成、更新および維持のコストも伴うため、すべての対象で製品群別観点表を利用するのではなく、対象、目的、必要な具体性および維持コストに応じて使い分ける必要がある。また、生成 AI の改善提案は要求仕様へそのまま採用するものではなく、人が必要性、費用、実現可能性および他の要求との整合性を確認したうえで、採否を判断する候補として利用することが望ましい。

以上を踏まえ、今後の展望を以下に示す。

- ・他の製品および品質特性に適用対象を広げる
- ・観点数と記述の具体性を分けた比較を行う
- ・観点表の作成・維持コスト、過剰提案および提案の確認時間を含めた評価を実施する

謝辞

本研究は、第 41 年度ソフトウェア品質管理研究会（SQiP）研究コース 5 において実施した研究を拡張したものである。同研究活動を進めるにあたり、石川冬樹主査、徳本晋副主査、栗田太郎アドバイザーには、丁寧なご指導を賜りました。ここに深く御礼申し上げます。また、SQiP 研究会において本研究を共に進めていただいた多田麻沙子氏に、心より感謝申し上げます。

参考文献

- [1] 田口 真義, 伊藤 弘毅, “生成 AI を利用するシステムの安全性評価を支援するテスト観点表の提案”, ソフトウェア品質シンポジウム 2025, 2025. https://www.juse.jp/sqip/symposium/2025/timetable/files/B2-2_ronbun.pdf
- [2] 早乙女 暖, 田口 真義, 多田 麻沙子, 中川 桂, “生成 AI を用いた要求仕様改善における観点表粒度の違いとその影響”, 第 41 年度ソフトウェア品質管理研究会 研究コース 5 成果報告, 2026.
- [3] 組込みソフトウェア技術者・管理者育成研究会 (SESSAME) 事務局, 話題沸騰ポット (GOMA-1015 型) 要求仕様書第 3 版, https://www.sesame.jp/workinggroup/WorkingGroup2/POT_Specification.htm, 2002