多角的視点を持つマルチエージェントシステムによる要件定義レビュー

Multiple Agents with Different Roles for Reviewing Requirements Definition

井上 祐寛 [1]　　松永 嵩 [1]　　綾塚 祐二 [1]

Takuhiro Inoue[1],　　Takashi Matsunaga[1],　　Yuji Ayatsuka[1]

[1] 株式会社クレスコ 技術研究所
[1] Technology Laboratory, CRESCO, LTD

We propose using multiple agents to review the requirements definition to improve quality. Each agent plays a different role, such as project manager, business analyst, or system architect, and performs cross-checking from different perspectives to reduce oversights, false detections, and contradictions. As a first step, we constructed a review agent that works as a project manager using an large language model (LLM) and evaluated its output.

1. はじめに

ソフトウェア開発プロジェクトにおいて、要件定義は全体の成否を左右する最も重要な初期工程の一つである。この段階での不備、特に要求の抜け漏れや曖昧な記述は、後続の設計・開発工程で大規模な手戻りを引き起こし、開発コストの増大と納期の遅延に直結する。このリスクを低減し要件定義の品質を向上するため、従来より多様な視点を持つステークホルダー（たとえば、ビジネス部門、開発部門、品質保証部門など）によるレビューが不可欠とされてきた。

しかし、多様な視点を確保するためには、専門知識を持つ多くの担当者のリソースを確保する必要があるが、人数が増えるほどスケジュール調整は困難となる。また、担当者の専門領域に起因する視点の偏りといった属人的な課題も常に存在し、網羅的な品質担保の障壁となっている。

本研究では、これらの課題を解決するため、Large Language Model (LLM) を活用し、個々の LLM エージェントに対して「プロジェクトマネージャ (PM)」「ビジネスアナリスト」「システムアーキテクト」といった異なる専門的役割（ロール）を割り当てることで、多角的な視点を持つマルチエージェントによるレビューシステムを提案する。

LLM の役割演技（ロールプレイング）能力を活用し、各エージェントがそれぞれの専門的視点から要件定義書を並列でレビューする。これにより、単一の視点では見逃されがちな課題（例：機能要件の矛盾、非機能要件の欠落、ユーザビリティやテスト容易性の課題）を網羅的に洗い出し、また、エージェント間の相互照合により誤検出や矛盾の低減を目指す。

本稿は、その最終目標に向けた第一段階の研究報告である。本稿では、まずシステムの基盤として「プロジェクトマネージャ(PM)」の役割を付与した単一エージェントが動作する実験環境を構築する。そして、エージェントによるレビューの実行手順と、その生成物の品質を定量的に評価するための「レビュー手順と評価の枠組み」を整備する。この枠組みの有効性を検証するために実施した、仕様書を用いた実験結果についても併せて報告する。

2. 関連研究

本研究は、「LLM の役割演技」、「LLM マルチエージェントシステム（MAS）」、そして「LLM による多視点評価」という、近年急速に進展している三つの研究領域を融合し、ソフトウェア工学の「要件定義レビュー」というドメインに適用する試みである。

2.1. LLM の役割演技（LLM Role-Playing）

LLM に特定の役割（ロール）を割り当てる「ロールプレイング」は、LLM の能力を引き出す有効な手法として確立されている。Tseng らのサーベイ[1]では、LLM が割り当てられた役割に基づき、その環境や文脈に適応した応答を生成する能力（LLM Role-

Playing）が示されている。本研究ではこのロールプレイング技術を応用し、ソフトウェア開発の各ステークホルダーとしての専門的視点を持つエージェントを構築する基盤とする。

2.2. LLM マルチエージェントシステム (MAS)

単一の LLM では解決困難な複雑なタスクに対し、複数の LLM エージェントが協調して問題解決にあたるマルチエージェントシステム（MAS）の研究が進展している。Guo らのサーベイ[2]によれば、MAS は各エージェントに特化した役割と知識を与えることで、集合的な知性を活用するアプローチである。

特にソフトウェア開発領域では、MetaGPT[1] や AutoGen[2] といったフレームワークが提案されている。これらは、エージェントに「プロダクトマネージャ」、「プログラマー」、「テスター」といった役割を割り当て、コーディングやテストといった開発プロセス自体を自動化する試みである。これらの先行研究が主に「開発プロセスの自動化」に焦点を当てているのに対し、本研究は MAS のアーキテクチャを「要件定義レビュー」という特定の品質保証タスクに応用する点に特徴がある。

2.3. LLM による多視点評価

LLM に多様なペルソナを与え、評価タスクに適用する研究も進んでいる。南雲らの研究[3]では、ビジネスアイディアの評価において、多様な属性を持つペルソナが、それぞれ独自の評価基準を用いて評価を行うデルファイ法が提案されている。さらに、ファシリテーター役の LLM が各ペルソナの評価を集約・要約するプロセスも示されている。

本研究の最終目標は、この多視点評価の手法を「要件定義レビュー」というドメインに特化させることである。南雲らの研究が「属性（例：年代、性別）」に基づくペルソナを用いたのに対し、本研究では「職務的役割（PM、アーキテクト等）」に基づく専門的視点を用いる点で独自性がある。

3. 提案手法：レビュー手順と評価の枠組み

本稿では、PM 役割の単一エージェントによるレビュー環境と、その生成物の品質を定量的に評価する枠組みを設計・構築した。この枠組みは、図 1 に示す通り、二段階の LLM エージェント・パイプラインで構成される。この二段階構造の採用は、LLM の「生成」と「品質検証」の役割を分離し、生成されたレビューを定義した点数付けに従って評価可能とするためである。これにより、レビュー指摘の品質を定量的に判断できることを確認し、評価プロセスの安定化を図る。

3.1. 第一段階：レビュー生成エージェント (PM Agent)

第一段階は、仕様書をレビューし、指摘事項を生成するエージェントである。

・役割

エージェントには、LLM のロールプレイング能力を活用し、明確な役割を付与した。具体的には、「IT 企業入社 30 年目（開発 20 年、プロジェクトマネージャ 10 年）の経験豊富なプロジェクトマネージャ」と定義した 。

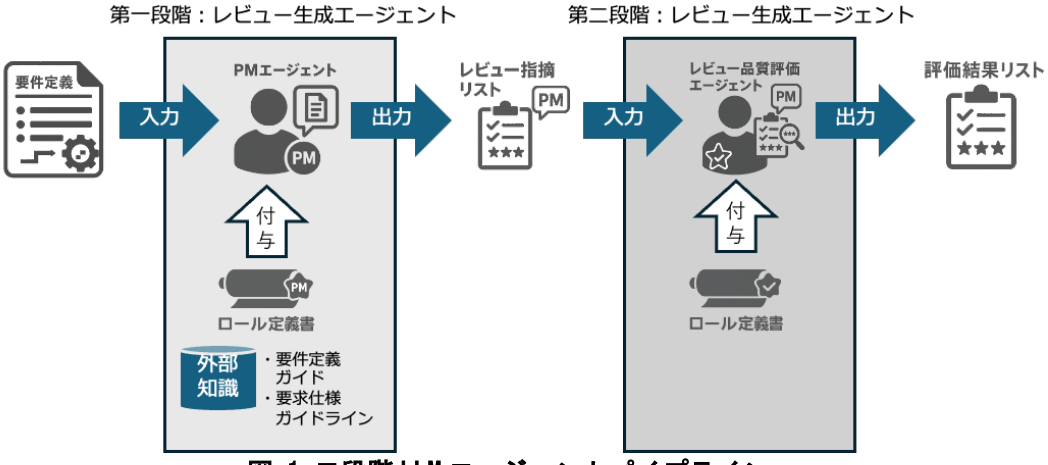

図 1. 二段階 LLM エージェントパイプライン

1 MetaGPT https://github.com/FoundationAgents/MetaGPT

2 AutoGen https://microsoft.github.io/autogen/0.2/

・タスク
役割定義に基づく専門的視点から、入力された要件定義書（本研究では「話題沸騰ポット要求仕様書（GOMA-1015 型）第 3 版[3]」）をレビューする。その際、出力形式は「修正番号」、「指摘箇所」、「指摘した記述」、「指摘の理由」、「指摘箇所の添削例」に統一することとした 。

・基盤モデルと知識
エージェントの基盤モデルには、OpenAI の ChatGPT モデル群の一つであり、複雑な推論に強いとされる`o3`モデルを採用した。また、エージェントの専門性を担保し回答の質を向上させるため、外部知識として IPA（情報処理推進機構）の「ユーザのための要件定義ガイド」および JUAS（日本情報システム・ユーザー協会）の「要求仕様ガイドライン」を付与した。

・レビュー抽出手法
レビューの生成方法として、一度に多数（例：100 件）を要求する手法と、対話的に少数ずつ深掘りする手法を比較した。その結果、一度に多数を要求すると、修正理由や修正案が簡素化される傾向が観測された。そのため本研究では、「他にありますか？」「もっとありませんか？」と対話的に深掘りする手法を採用し、エージェントが持つ指摘事項を網羅的に（本実験では 40 件）抽出した

3.2. 第二段階：レビュー品質評価エージェント（Scoring Agent）

第一段階で生成されたレビューの「品質（＝指摘の妥当性や重要度）」を定量的に採点するエージェントである。この採点は、指摘の重要度を明確化し、レビュー後の仕様書修正の優先順位付けに必要な情報を提供するために行う。

・役割
採点エージェントにも、レビュー生成エージェントと同様のペルソナ（経験豊富な PM）を付与した

・タスク
第一段階で生成された個々のレビューコメントに対し、要件定義書と照合した上で「辛口」で 0〜5 点の整数採点を行うよう指示した。

・採点基準

評価の客観性を担保するため、厳格な採点基準を定義した（表 1）

表 1.採点基準

点数	判断基準（概要）	詳細な意味合い
5	致命的	基本仕様・安全要件の欠落や重大法令違反。製品化停止レベル、最優先で是正。
4	重大	法規・規格違反リスク、事故・故障・市場クレームに直結。設計手戻り大。
3	中度（要修正）	機能・保守・安全・規格適合に実質影響。放置不可。
2	軽度	品質・操作性へ間接影響。回避策・部分記載あり。
1	ごく軽微	誤字・体裁、機能・安全に影響せず即日修正可。
0	指摘無効	すでに仕様書に記載／指摘不要／誤認。

・評価の安定化
LLM による採点は、同一の入力に対しても「ばらつき」を生じる可能性がある。この問題を軽減し評価の信頼性を高めるため、本研究では同一のレビューコメントに対して採点エージェントを 5 回独立して実行し、その平均点を最終的な品質スコアとして採用した。

4. 実験：評価枠組みの適用と検証

4.1. 実験目的

本稿で構築した「レビュー手順と評価の枠組み」（3 章）が、要件定義レビューにおいて有効に機能するかを実証する。具体的には、第一段階の PM エージェントが実用的な指摘を生成できること、および第二段階の採点エージェントがその指摘品質を妥当に定量化できることを検証する。

4.2. 実験設定

実験対象として、ハードウェアの仕様書（「話題沸騰ポットの仕様書 V3」）を用いた 。この仕様書は、実際の製品開発で用いられるレベルの詳細度を持ちつつ、レビューの観点からは (意図的な)曖昧さや欠落箇所が含まれている 。

3 組込みソフトウェア管理者・技術者育成研究会(SESSAME)　https://www.sessame.jp/
電子ポットを題材にした組込みシステム分析・設計のための要求仕様書

実験手順として 3 章で定義した二段階評価パイプラインを適用した。

（第一段階）PM エージェントが対象仕様書をレビューし、対話的な深掘りによって 40 件の指摘事項を生成した。

（第二段階）採点エージェントが、生成された 40 件の指摘事項をそれぞれ 5 回ずつ独立して採点し、平均点を算出した 。

4.3. 実験結果

本枠組みによる評価結果は、ハードウェア仕様書のレビューとして極めて妥当なものであった。採点エージェントは、プロジェクトリスクに直結する「致命的な欠陥」の指摘には一貫して高得点（平均 4.8 点以上）を付与し、一方で「改善提案」レベルの指摘には低得点を付与する、明確な傾向を示した。

(1) 高スコアの指摘

表 2 はスコア（平均 4.8 点以上）を獲得した重要指摘の例である。安全性、基本機能などに関する致命的な欠落が含まれ、PM エージェントが、その役割通り、プロジェクトの根本的なリスク(安全、基本機能不全)を最優先で特定できていることを示す。

(2) 低スコアの指摘

表 3 はスコア（平均 3.0 点未満）となった指摘の例である。主に詳細なユーザビリティ（使い勝手）に関する仕様の改善提案が含まれる。これらは「あれば望ましい」ものであり、PM の視点からは（表 2 の致命的欠陥と比較して）優先度が低いと判断されている。採点エージェントがこの重要度の差を正しくスコアに反映できていることを示す。

5. 考察

本研究で構築した二段階の評価枠組みを適用した結果、提案手法の有効性と、今後のマルチエージェント化に向けた明確な課題が示された。

5.1. PM エージェントの実用性

実験結果（表 2）が示す通り、構築した PM エージェントは、安全性（感電、破裂リスク）、基本仕様（電源、貯水容量、防水等級）といった、プロジェクトの致命的な欠陥を優先的に特定する能力を有していることが実証された。これは、PM という役割定義と、外部知識の付与が有効に機能したことを示唆し

表 2：高スコア（平均 4.8 点以上）を獲得した重要指摘の例

修正理由	修正案	平均点
アース端子/漏電保護など感電対策記載無し	三極プラグ＋漏電ブレーカ内蔵を必須、安全規格(IEC 60335)準拠を追記	5
蒸気排気路/過圧逃がし構造記載無しで破裂リスク	0.05 MPa で開く弁構造を追加し試験プロトコルを図示	5
最大/最小貯水容量(L)が仕様に見当たらない	公称 1.5L、最小加熱水量 0.3L など容量レンジを追加	4.8
接水樹脂部の食品衛生法・BPA フリー等の材質要件が欠落	口金・蓋パッキンは厚生労働省告示第 370 号適合材を指定	4.8
防水・防滴(IP)等級が未規定で台所利用時に安全性不足	最低 IPX1、推奨 IPX4 の筐体設計と試験条件(IEC60529)を明記	4.8
AC100V/50-60Hz など電源電圧・周波数・許容変動の記載無し	定格 100V±10% 50/60Hz、突入電流・待機電力上限を明記	4.8

表 3：低スコア（平均 3.0 点未満）となった指摘の例

修正理由	修正案	平均点
タイマは 1 分刻みのみで 30 秒以下の短時間設定が不可能	秒単位の加算/長押し高速算モードを追加	2.6
タンは 1 分加算のみで長押連続入力・チャタ対策時間未定義	押し 0.8s 以上で 10 分刻み算、離した瞬間に停止と明	2.8
イマ上限「最大 1 時間」の拠不明	ースケースに基づき上限値妥当性を説明 or 可変設定	2.4
ザー音量固定で夜間利用に慮不足	ニューで 60dB⇔70dB 切、または消音＋LED 通知追加	2.8

ており、単一のエージェントであっても実用的なレビューエージェントとして機能しうることを示している。

5.2. LLMによる品質評価枠組みの有効性

本研究の評価枠組みの中核である「レビュー品質評価エージェント（辛口採点エージェント)」は、その妥当性を実証した。採点エージェントは、PMエージェントが生成した指摘事項に対し、致命的な欠陥（表2）には一貫して高い平均点（4.8点以上）を付与し、ユーザビリティ等の改善提案（表3）には低い平均点（3.0点未満）を付与した。これは、LLMが定義された採点基準に基づき、指摘事項の品質（重要度や妥当性）を定量的に評価可能であることを示している。また、採点を5回繰り返して平均する手法は、LLM固有の「ばらつき」を抑え、評価の信頼性を担保する有効な枠組みであると結論付けられる。

5.3. 単一エージェントの限界とマルチエージェントシステムの必要性

本実験の考察で最も重要な点は、単一エージェントの限界が明確になったことである。PMエージェントは、その役割通り「プロジェクトの重大リスク」を優先する一方で、表2に示すような「詳細なユーザビリティ（タイマー刻み、ブザー音量)」に関する指摘の優先度は低く、見落とされる可能性が考えられる。

これは、単一の視点ではレビューが偏るという従来（人間によるレビュー）の課題を裏付けるものであり、本研究の最終目標であるマルチエージェント・レビューシステムの必要性を強く示唆している。PMエージェントが見落とす可能性のある「ユーザビリティ」の観点は、まさしく「UXデザイナー」や「ユーザー」といった異なる役割を持つエージェントが補完すべき領域である。

6. 結論と今後の展望

6.1. 結論

本研究では、LLMによる多角的な要件定義レビューシステムの構築に向けた第一段階として、「プロジェクトマネージャ（PM)」役割の単一エージェントによるレビュー手順と、その品質をLLMで定量的に評価する評価の枠組みを構築・整備した。

提案する枠組みは、以下の二段階パイプラインで構成される。

・第一段階（レビュー生成）
明確な役割（PM）と外部知識を付与された「PMエージェント」が、対話的な深掘りを通じて仕様書の指摘事項を生成する。

・第二段階（品質評価）
同様の役割を持つ「辛口採点エージェント」が、厳格な基準に基づき、複数回（5回）の採点平均によってレビュー品質を安定的に定量化する。

実験により、本枠組みが有効に機能することを実証した。PMエージェントは安全性や基本仕様に関する重大な欠陥を優先的に特定し(平均スコア4.8点以上)、採点エージェントがその品質を妥当に評価できることを確認した。

同時に、PMエージェントの指摘がその役割に偏ることで、ユーザビリティ等の観点が見落とされる可能性という単一エージェントの限界も明確になった。この結果は、本研究の最終目標であるマルチエージェントシステムの必要性を強く裏付けるものである。

6.2. 今後の展望

今後の研究では、本稿で明らかになった単一エージェントの限界を克服すべく、提案するマルチエージェント・レビューシステムの構築を進める。

まず、PMに加え、ビジネスアナリスト、システムアーキテクトなどの多様な役割を持つエージェントの実装に取り組む。そして、各エージェントが独立したレビュー結果を共有し、矛盾を解消するための協調メカニズムの構築を主要なテーマとする。

また、システムの信頼性向上と機能拡張に向けた研究にも並行して取り組む。これには、自己肯定バイアスやハルシネーションの抑制効果を定量的に測定する信頼性の客観的検証が含まれる。将来的には、より高度な論理推論や判断を可能にするための新たな技術的アプローチも継続して検討していく。

参考文献

[1] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. "Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization." [cite_start]*arXiv preprint arXiv:2406.01171v3 [cs.CL]*, 2024. [cite: 3109, 3110-3116]

[2] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. "Large Language Model based Multi-Agents: A Survey of Progress and Challenges." [cite_start]*arXiv preprint arXiv:2402.01680v2 [cs.CL]*, 2024. [cite: 2006, 2007-2010]

[3] 南雲陸, 佐々木. "LLM を活用したペルソナベースのデルファイ法による多視点アイディア評価." [cite_start]*第 39 回 人工知能学会全国大会 (The 39th Annual Conference of the Japanese Society for Artificial Intelligence)*, 2025. [cite: 1860, 1861-1864, 1867]

[4] Zefang Zong, Jingwei Wang, Yunke Zhang, et al. "Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models." [cite_start]*arXiv preprint arXiv:2501.09686v3 [cs.AI]*, 2025. [cite: 2131, 2132-2137]