生成 AI を用いたレビュー承認の自動化に関する研究 —Human-AI Agreement Zone（HAZ）の定量的活用—

松井 高宏 [1]
[1] 合同会社 DMM.com

概要：AI の業務適用が進展する現代において、「どこまで安全に AI に判断を任せられるか」は品質保証の観点から極めて重要な課題である。本研究では、人と AI の判断が一致する領域を Human-AI Agreement Zone（HAZ）と定義し、それを信頼性スコアとして抽出することで、AI が主体的に判断できる範囲を定量的に示す自動化構造を提案する。対象は年間 50 万件の商品レビュー審査業務であり、本手法により約 60%を AI が自動承認し、従来最大 7 日を要していた公開処理時間を 10 分以内に短縮、品質事故はゼロであった。承認判断のプロセスには Agentic Workflow を導入し、処理を 5 段階に分解。さらに、判断確定ステップ（Step）とレビュー本文の品質（Grade）から信頼性スコアを算出し、そのスコアに基づいて自動承認可能な範囲（HAZ）を明示した。実運用データ 20 万件に基づく評価では、信頼性スコア 0.15 以下の領域において人と AI の一致率が 100%（95%信頼区間下限でも 99.997%）を達成した。HAZ は、AI に安全に判断を委ねるための新たな品質指標として機能し、従来の Human-in-the-Loop（HITL）に代わる設計指針となり得る。本研究は、信頼性スコアモデルと HAZ の構造化によって、生成 AI 時代における品質保証と現場実装を両立させる先駆的事例である。

キーワード：生成 AI、レビュー自動承認、信頼性スコアモデル、Human-AI Agreement Zone（HAZ）、Agentic Workflow

Automation of Review Approval with Generative AI —Quantifying Human-AI Agreement Zone (HAZ)—

TAKAHIRO MATSUI[1]
[1] DMM.com LLC, Japan

1. 研究背景と目的

近年、スマートフォンの普及とユーザー生成コンテンツ（UGC）の増加により、SNS や EC サイトではレビューやコメントのモデレーション（内容審査）負荷が著しく増大している。モデレーションとは、UGC の内容を検閲・審査し、ガイドラインや法規制に基づいて適切性を判断する業務であり、サービス品質の維持に不可欠である［1］。

特に、誹謗中傷や偽情報の拡散といった社会的課題が深刻化しており、投稿監視の自動化は国際的にも急務とされている。日本でも総務省「プラットフォームサービスに関する研究会」が開催され、有害情報への対応強化や品質保証の仕組みづくりが求められている［2］。

一方、商品レビューの審査では、単純な NG ワード検出では対応できず、皮肉や曖昧な表現など、文脈を踏まえた判断が求められる。たとえば、「この掃除機は驚くほど静かだ。隣の家まで音が聞こえるほどに。」といった皮肉や、「効果があるような、ないような…」といった曖昧な言い回しには、高度な文脈理解と一貫した品質基準が必要となる。

近年の生成 AI、特に大規模言語モデル（LLM）の進展により、こうした複雑な判断タスクへの AI 適用が現実味を帯びてきた。しかし、LLM の出力にはハルシネーションや一貫性の欠如といった課題があり、PoC（概念実証）止まりの AI 導入事例が多く、判断や承認といった責任を伴う業務には適用されにくい。その最大の要因は、「どこまで AI に任せてよいか」という境界が未定義であることにある。

実際、2025 年にはコードエディタ AI「Cursor」の開発企業において、AI チャットボットが誤った情報を提供し、企業の評判を損なう事例が発生した［12］。このように、AI 判断の透明性と責任所在の不明確さは、品質事故や信頼失墜につながる現実的なリスクとなっている。

本研究では、この課題に対し、人と AI の判断が一致し、合意可能とみなせる領域（Human-AI Agreement Zone: HAZ）に着目し、これを信頼性スコアとして定量的にモデル化することで、安全性を担保した自動化と責任分担の明確化を両立する。これにより、品質保証と業務効率を同時に実現するとともに、AI 活用における安全な判断基準の設計指針を提示する。

2. 関連研究

従来のモデレーション手法は、ルールベースや機械学習による分類モデルが中心であり、特定のキーワードや定型パターンに依存していた。しかし、商品レビューのように文脈や表現が多様な領域では、こうした手法では十分な対応ができず、誤判定による品質事故や利用者信頼の低下が課題となっていた。近年では、大規模言語モデル（LLM）の登場により、より柔軟かつ文脈的な判断を行うモデレーション手法が模索されている。中でも、Prompt Engineering [13] による出力制御、思考過程を段階的に言語化する Chain-of-Thought（CoT）[4]、およびプロンプトを分割して段階的処理を行う Agentic Workflow [8]は、AI 判断の説明可能性や一貫性向上の観点から注目されている。本研究は、これらの要素技術を組み合わせ、人と AI の判断一致と責任分担を前提に品質保証と安全性を両立する設計として実務適用した。これにより、現場適用可能で理論的裏付けのある品質管理モデルを提示している。

3. 対象データの設計

本研究が対象とするのは、年間約 50 万件のレビューが投稿される大規模 EC サイトにおける、商品レビューのモデレーション業務である。レビュー投稿数は過去 10 年間で 10 倍以上に増加し、人的対応の限界が急速に顕在化している。加えて、これらの審査作業に最大 7 日間の審査待ちが発生し、月間 150 時間以上の人的工数を要していた。この業務には専任スタッフが複数名で対応しており、投稿量の増加に伴ってサービス品質の維持と品質保証を両立できる抜本的な対策が求められていた。さらに、不適切投稿の割合は全レビューの約 10%に達し、誹謗中傷や購入非推奨、文言不明など、20 種類以上の違反パターンを想定した複雑なルール設計が必要とされていた（図 1）。

このような増加傾向とルールの複雑性が、自動化を極めて困難にしていた。そこで本研究では、増大する審査負荷に対応すべく、レビュー判定の自動化に取り組む。自動化の試みとその課題、さらにそれを克服するための構造化手法については、次章以降で記載する。

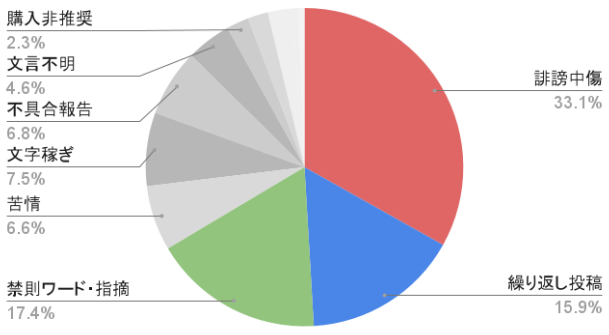

図 1 不適切投稿の割合

Figure 1. Proportion of inappropriate posts.

4. Agentic Workflow の設計

商品レビューのモデレーション自動承認を目指し、まず自社規約を大規模言語モデル（LLM）に読み込ませ、人の判断を代替できるかを PoC（概念実証）で検証した。しかし、皮肉や曖昧表現に対する誤判定が頻発し、人と AI の判断一致率は 70〜80%にとどまった。理由説明にも一貫性がなく、ハルシネーション（事実に基づかない出力）も多発したため、品質保証や安定運用の観点からも実運用は困難と判断した。

特に、審査や承認判断のような複雑かつトークン消費量の多いタスクでは、LLM がハルシネーションを引き起こしやすいことが知られている。Guerreiro ら（2023）は、ニューラル機械翻訳の大規模調査において、タスクの複雑さや不確実性がハルシネーションの主要な要因であると指摘している [11]。

これらの課題を踏まえ、AI の判断プロセスを段階的に分割・構造化し、各ステップの出力を定量化する Agentic Workflow 方式を採用した（図 2）。これは従来の LLM による「一括判断方式」から、「段階的判断と順次精査」へと変更するものであり、判断精度と説明可能性を高めることで、安全かつ品質を担保した自動化基盤の設計を可能にしている。各ステップは、Claude 3.7 への個別プロンプトによって実行され、出力は順に連結されて次の判断工程へと受け渡される。

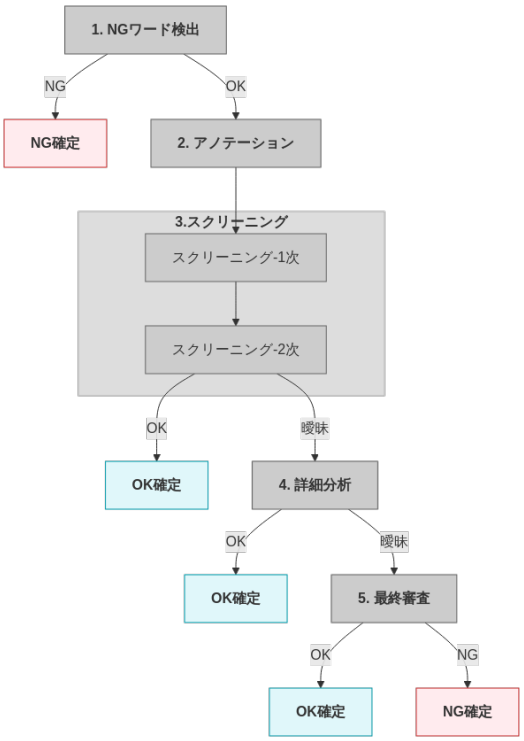

図 2 Agentic Workflow

Figure 2. Agentic Workflow.

レビュー承認プロセスは Agentic Workflow に基づき、(1) NG ワード検出（禁止語句の抽出）、(2) アノテーション（リスク要素の付与）、(3) スクリーニング（問題有無の一次判定）、(4) 詳細分析（文脈の深掘り判定）、(5) 最終審査（統合評価）の 5 段階に分割されている。

この設計は、医療診断におけるリスク要因の段階的抽出・精査の構造を参考にしている。また、(2)〜(5) の判断工程には Claude 3.7 Sonnet [3] を用い、各ステップで Chain-of-Thought（CoT）形式による理由付けを行っており、特にスクリーニング工程では誤判定や見落としを防ぐため、同一レビューに対して二重チェックを実施している。

5. 信頼性スコアモデルと HAZ の定義

前章で示した Agentic Workflow により、レビュー承認の判断プロセスは構造化・安定化された。しかし、各レビューの判断がどれだけ確信をもって下されたかを定量的に把握することは容易ではない。たとえば、初期段階で判断が確定したレビューと最終審査まで進んだレビューでは判断難易度が異なるにもかかわらず、従来は結果が一律に扱われてきた。これは品質保証やリスク管理の観点から不十分である。そこで本研究では、Agentic Workflow の出力に基づき、各レビューに信頼性スコア S を付与する評価モデルを設計した。これにより、AI 判断の確からしさを数値化し、「どこまで AI に任せられるか」を品質基準として判断できるようにした。

5.1 信頼性スコアモデリング

レビューごとに 0.0〜1.0 の範囲で信頼性スコア S を付与し、AI による判断の確からしさを定量評価した。S が低いほど信頼性が高く、自動承認が妥当と判断される。一方、S が高い場合は判断が曖昧または困難とみなされ、手動確認が推奨される。スコア S は、Step と Grade という 2 つのパラメータに基づく理論モデルから算出する。これらは品質保証の観点で重要な要素であり、判断の難易度とリスク度合いを同時に反映する。

入力パラメータ

- Step（判断確定ステップ）: Agentic Workflow の中で判断が確定した段階を表す指標。Step が小さい場合は AI が容易に判断できたことを示し、大きい場合は詳細分析や最終審査まで進んだ困難な判断を示す。
- Grade（レビューの品質）: レビュー本文の悪質性の程度を数値化した指標。Grade が低いほど高品質であり、高いほどリスクが大きいことを示す。

スコア S は以下の関数により算出する

算出式

$$S = f(Step, Grade)$$

さらに、Step と Grade の組み合わせをもとにスコア範囲別のカテゴリを設定し、判断難易度とレビュー品質の両面からレビューを分類した。

スコア・カテゴリ

$$
\begin{aligned}
S = \{ \\
&0.00 - 0.05 = f(\text{スクリーニング}, S\ (\text{高品質})) \\
&0.05 - 0.10 = f(\text{スクリーニング}, A\ (\text{良質})) \\
&0.10 - 0.15 = f(\text{スクリーニング}, B\ (\text{普通})) \\
&0.16 - 0.30 = f(\text{詳細分析以降}, C\ (\text{該当なし})) \\
&0.31 - 0.70 = f(\text{詳細分析以降}, D\ (\text{曖昧})) \\
&0.71 - 1.00 = f(\text{詳細分析以降}, E\ (\text{違反})) \\
&1.0 \quad\quad = f(\text{NG ワード検出}, F\ (\text{重大違反})) \\
\}
\end{aligned}
$$

このモデルにより、両パラメータを総合的に評価し、信頼性スコアとして定量化できるようになった。

5.2 HAZ の定義

信頼性スコアによる評価だけでは、「AI に責任を委ね得る責任範囲」を明確に定義するには不十分である。そこで人と AI の判断が実質的に一致しているとみなせる領域を Human-AI Agreement Zone（HAZ）と名付け、これに基づく自動化戦略を構築した。

HAZ は次のように定義される。

$$HAZ = \{ r \mid Human(r) \approx AI(r) \}$$

すなわち、HAZ とは「人と AI の判断が実質的に一致しているとみなせるレビューの集合」である。

ここで r は 1 件のレビュー、*Human(r)* および *AI(r)* はそれぞれ人間と AI の判定結果を示す。記号「≈」は、運用上「合意できる」と判断される一致を意味し、完全一致だけでなく合理的な許容一致も含まれる。これにより、両者の判断に基づいて AI による判断自動化の適用範囲を定義できるようになった。図 3 は、人と AI の判断集合の重なりをベン図で示したものであり、その交差部分を HAZ と定義している。

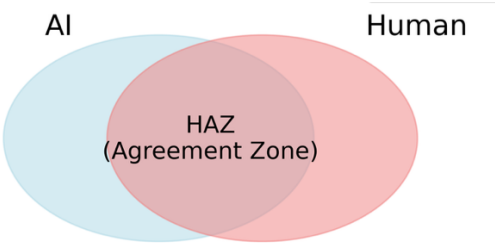

図 3 HAZ の概念図

Figure 3. Human-AI Agreement Zone

6. HAZ 導入前評価

本章では、HAZ の導入に必要な一貫性を確保するために実施した信頼性向上の施策と、それによって得られた評価結果をまとめる。まず 6 カ月にわたる信頼性向上施策（6.1）を示し、その後の評価結果（6.2）を記載する。

6.1 信頼性向上に向けた改善施策

本研究では、HAZ 導入に向けて判断精度と一貫性の確保を目的として、6 カ月にわたり様々な改善施策を講じてきた。その中でも、特に効果の高かった 4 つの施策について、以下に整理して示す。

(1) プロンプト最適化
人と AI の判断差異を週次で分析し、誤判定の傾向をプロンプトに反映する Human-in-the-Loop（HITL）方式による継続的な改善を行った。

(2) UK（Unknown）判定の導入
AI による判断が困難なケースでは、無理に OK/NG 分類を行うことなく、「UK（Unknown）」として保留とし人に判断を委ねた。UK は全体の約 10%に抑制した。

(3) モデル更新
LLM のバージョン更新により、文脈把握力と推論安定性を向上させた。特に Claude 3.5 以降の改善が顕著であり、現在は Claude 3.7 を採用している。

(4) Many-shot In-Context Learning
皮肉や多義的表現といった微妙なニュアンスを捉えるため、Few-shot [9]に代えて 100 件以上の代表例を含む Many-shot 形式を採用した［10］。

これらの施策により Agentic Workflow によるレビュー判定の信頼性は段階的に改善された。特に、Recall・Precision・Accuracy といった主要指標において安定的な向上が見られ、HAZ 導入に必要な前提条件となる判断精度と一貫性の確保が実現された（図 4）。

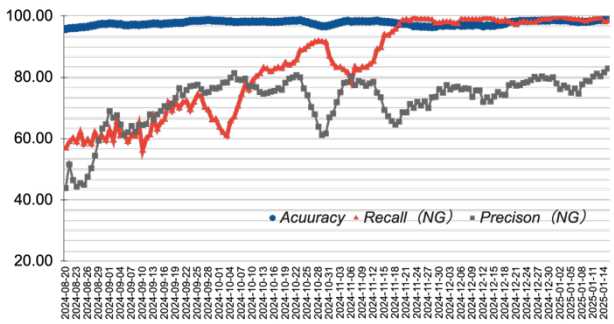

図 4 判定精度の変化（Recall / Precision / Accuracy）
Figure 4. Change in Classification Metrics Over Time

6.2 導入前評価結果

6.1 の改善施策を経て、20 万件以上のレビューを対象に評価を実施したところ、Accuracy および Recall は約 98%、Precision は 80%となった。Precision がやや低いのは、誤承認（False Negative）を避けるため、意図的に過検出（False Positive）を許容する安全側設計に基づいているためである。本評価では、社会的リスクの低減と品質保証の観点から、不適切レビューの見逃し防止（Recall）を最重要指標とし、精度のバランスよりも Recall 最優先の目的設計とした。そのため F1 スコアなどの中庸指標は採用していない。

表 1 判定精度の評価結果（HAZ 導入前）
Table 1. Performance Metrics（before HAZ）

Accuracy	Recall	Precision
98%	98%	80%

この結果により、AI 判断は人による審査作業の品質水準に到達し、自動承認（HAZ）導入の基盤が整った。しかし、全体の約 2%では依然として人と AI の判断にわずかな差異が残っていた。ここに重大なリスクは確認されなかったものの、AI が単独で承認判断を行う体制には前例がなく、判断の揺れに対する責任所在も社会的に明確ではなかった。

そこで本研究では、「どこまで AI に任せられるか」という問いに品質指標として答えるため、各レビューに信頼性スコアを付与し、自動化が許容される範囲（Human-AI Agreement Zone: HAZ）を定量的に導出する戦略を採用した。次章では、この戦略を基盤として構築した HAZ の定量評価結果を示す。

7. HAZ 導入後評価

前章の信頼性評価結果から、HAZ の導入は品質保証上も十分可能であると判断された。本章では、導入前に得られた判定精度を基に、信頼性スコア別の評価を行い、自動承認における安全性と HAZ 閾値設定の妥当性を検証する。

7.1 スコア別の評価結果

信頼性スコア S に基づき、人と AI の一致率をスコア区間ごとに分析した。評価指標には、6 章と同様に Accuracy（人と AI の一致率）を用いた。その結果、スコア $S \leq 0.15$ の領域では人と AI の判定が完全に一致し、一致率は 100%（95%信頼区間下限でも 99.997%）となった。この領域は、統計的にも偶然による一致ではないことが確認されており、品質保証上「誤判定リスクが実質ゼロ」とみなせる水準である。これにより、HAZ の閾値設定が自動化の安全性確保に有効であることが裏付けられた。

表 2 判定精度の評価結果（HAZ 導入後・スコア別）
Table 2. Performance Metrics (After HAZ, by Score)

Range	Accuracy	Recall	Precision
S <=0.05	100%	-	-
S <=0.10	100%	-	-
S <=0.15	100%	-	-
S <=0.30	99.8%	-	-
S <=0.70	99.8%	-	-
S <=1.0	98%	98%	80%

注：Recall / Precision は、スコア帯内で AI による NG 判定自体が存在しないため定義できず、「–」とした。

なお、Recall および Precision は「不適切（NG）レビューを正しく判定できたか」を示す指標であるが、スコア $S \leq 0.70$ の範囲では AI による NG 判定自体が発生しておらず、指標値は算出できなかった。このことから、実際に NG 判定が発生するのは $S > 0.70$ の範囲に限定されている。

7.2 HAZ の閾値設定と合理性

スコア $S \leq 0.15$ の領域で一致率（Accuracy）が 100% に達したことから、本研究ではこの範囲を Human-AI Agreement Zone（HAZ）と定義し、自動承認の対象とした。一方、$S \leq 0.70$ の範囲でも一致率は 99.8% と非常に高い水準を示したが、この段階ではわずか 0.2% の誤判定の中に重大なリスクが含まれる可能性を完全には排除できない。

本研究では、AI 判断に対する社会的信頼を確保するため、導入初期においては「完全一致」が統計的に確認されたスコア $S \leq 0.15$ の領域に限って自動承認（HAZ）を適用した。この判断はモデル性能の最大化を目的としたものではなく、これは、説明責任と社会的受容性を重視した、保守的かつ段階的な設計方針に基づいた判断である。

このスコアと一致率の関係を視覚的に示したのが図 5 である。図中には、スコアごとの一致率（赤線）と、そのスコア以下に累積的に含まれるレビュー数の割合（灰色の棒グラフ）を示している。特に $S \leq 0.15$ の範囲（HAZ 領域）では一致率が 100%に達し、かつレビュー全体の約 70%を占めていることがわかる。これにより、HAZ の導入が高精度（100%一致）と広い適用範囲（70%カバー）の両立を実現していることが視覚的にも確認できる。

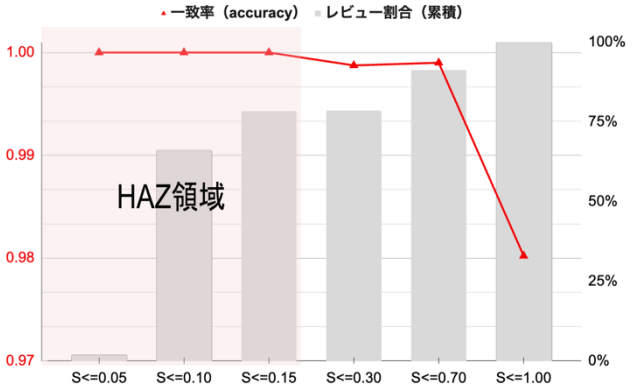

図 5 スコアごとの一致率とレビュー数の割合（累積）
Figure 5. Agreement Rate and Cumulative Review Coverage by Score

注：赤線は一致率、灰色の棒は累積レビュー数の割合を示す。HAZ 領域（$S \leq 0.15$）は一致率 100%、レビュー全体の約 70%を占める。

7.3 HAZ 導入の意義

HAZ は、従来の Human-in-the-Loop（HITL）型プロセスで曖昧だった「AI にどこまで任せられるか」という境界を、初めて客観的かつ定量的に可視化した品質管理指標である。これまで AI 判断の信頼性を裏付ける具体的基準は存在しなかったが、HAZ により「人と AI が合理的に合意できる領域」を提示することで、運用現場の説明責任を強化し、社内外ステークホルダーからの信頼を獲得できる。

すなわち HAZ は、単なる技術的概念ではなく品質保証と説明責任を両立するための合意境界線として、意思決定自動化の客観的かつ説明可能な基盤を提供するものである。

8. 実運用とモニタリング

HAZ の導入により高精度な自動承認が実現したが、AI と人の判断が将来にわたって常に 100%一致することは統計的にも原理的にも不可能である。特に自動承認されたレビューは、人による評価結果が残らないため、一致率の継続的な直接測定は困難となる。そこで本研究では、自動化後の品質維持とリスク低減を目的に、PDCA サイクルを前提とした補完的モニタリング体制を構築した。

■モニタリング体制の構成

(1) クレームのモニタリング：ユーザーや社内からの苦情をトリガーとして対象レビューとスコア分布を確認し、異常兆候を早期検知。

(2) 管理者の定期サンプリング：HAZ 領域のレビューをランダム抽出し、管理者による後追い評価を実施。自動承認の健全性を継続監査。

(3) 通報機能の導入：ユーザーが不適切レビューを直接報告できる仕組みを実装予定。利用者との協働による健全性維持を図る。

これらの仕組みにより、信頼性スコアに依存しすぎない、持続可能かつ適応性のある運用体制を実現した。この体制は、HAZ の本質である「人と AI による信頼の共有領域」を維持・強化するものであり、品質保証プロセスにおけるフィードバックループとして機能する。実際、運用開始から 4 か月以上経過した現時点で、ユーザー・社内双方からの苦情は一件も発生していない。この結果は、HAZ による自動承認が、現場での社会的受容性と実効的安全性を兼ね備えていることの実証的エビデンスである。

9. 実運用評価

本提案手法を年間約 50 万件のレビューを扱う大規模 EC サイトに導入し、92,803 件（約 2 か月強）の実運用データを用いて効果検証を行った。結果、全レビューの 60.8%（60,431 件）が AI による完全自動承認となり（図 6：赤＝AI 承認件数、灰＝人承認件数）、処理能力と品質維持の両立を実証した。理論上はスコア $S \leq 0.15$（HAZ）の範囲で約 70%の自動承認が可能であったが、導入初期段階ではリスクマネジメントと社内合意形成を優先し、適用範囲を最大 60%程度に制限している。

承認速度の分析では、従来最大で 7 日を要していた処理時間が AI ではすべて 10 分以内に収まり、平均処理時間は 99.5%削減された。一方、人による承認では 33.7%のケースで 1〜3 日を要していた（図 7：横軸は承認時間、縦軸はレビュー全体に占める割合。赤＝AI 承認、灰＝人承認。AI は 10 分未満に集中、一方人承認は長時間帯に幅広く分布）。

さらに、導入後の社内審査部門ヒアリングでは「作業負担が大幅に軽減された」「当初は実現困難と考えていたが、安全性を担保したまま達成できたのは画期的」と高評価を得た。これらの結果は、HAZ の導入が品質維持・処理効率化・現場満足度の 3 要素を同時に満たすことを示している。

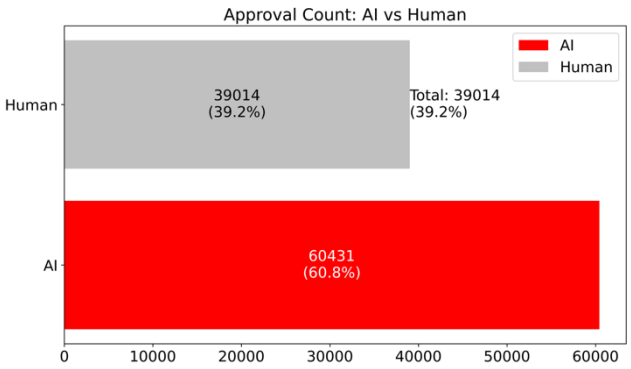

図 6 承認数の比較（AI と人）

Figure 6. Approval Count (AI vs Human).

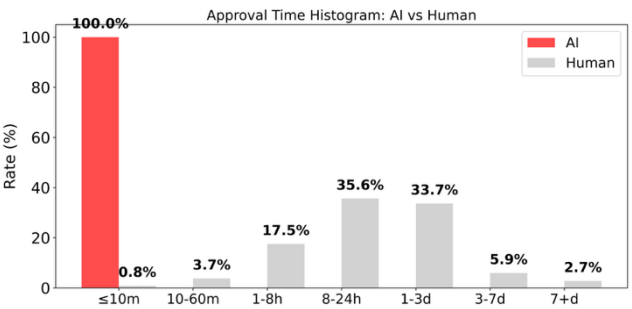

図 7 承認速度のヒストグラム（AI と人）

Figure 7. Approval Time Histogram (AI vs Human).

10. 考察

本研究で提案した Agentic Workflow による段階的プロンプト設計と信頼性スコアリングは、業務自動化と品質保証の両立に寄与した。特に、Human-AI Agreement Zone（HAZ）の導入は、AI 活用における判断の信頼性・説明可能性・責任分担を明確化し、従来の Human-in-the-Loop（HITL）アプローチを品質保証の観点から進化させた枠組みと位置づけられる。

HAZ は、「人が AI と関与すべきか」を判断するための客観的な品質基準であり、信頼可能な判断領域を定義することで、自動化範囲を安全に拡張する構造である。人と AI の判断一致を定量的に評価する枠組みとしては、Kural らが人と AI の判断類似性を情報理論的に測定し、信頼感との関係を分析したが、具体的な一致閾値や運用上の適用条件は提示していない[6][7]。本研究では、信頼性スコア S に基づき完全一致領域（S≤0.15）を定義し、実運用に耐える閾値設定と適用範囲の合理性を示した点で、理論的枠組みを品質マネジメント実務へと展開している。

11. 今後の課題と展望

HAZ（Human-AI Agreement Zone）と段階的判断プロセス（Agentic Workflow）により、AI 判断の信頼性と運用リスクを両立する枠組みを構築した。本研究では今後は、以下の 3 つの方向で進化を目指す。

(1) HAZ の適用範囲拡張
導入初期は、人と AI の判断が完全一致するレビューのみを HAZ に含め、社会的受容性と説明責任を優先した保守的な設計とした。今後は軽微な差異（permissible agreement）を許容し、スコア S ≤ 0.70 程度まで拡張することで、自動化率 80%以上を目指す。

(2) Agentic Workflow の品質設計強化
現在のワークフローは NG 検出に重点を置いているが、これは社内ルールが NG 基準のみで構成されているためである。そのため OK 判定は消去法で行われている。今後は OK 基準を明示し、NG 検出に依存しないレビュー自動承認フローへの再設計が必要となる。

(3) 共創に基づく将来ビジョン
人と AI の関係性を 5 段階の共生モデルとして整理した（表 3）。現在は Lv3（合意型）に該当し、HAZ の導入により一部自動化が実現している。次の段階である Lv4（改善型）では、AI が人との判断差異を学習し、プロンプトやワークフローを自律的に改善する構成が想定される。現時点でも、LLM による誤判定パターンの抽出や改善提案の生成を行っており、部分的な半自動化はすでに実現している。一方で Lv5（自律型）は、人の関与を完全に排除する構成である未来像が想定されているが、本研究では採用しない方針である。常に人が最終的な判断権を持つという設計思想を前提とする。この共生モデルは、AI 活用の成熟段階と人の責任分担を構造的に整理する枠組みであり、今後の AI 設計ガイドラインにも資する可能性がある。

表 3 人と AI の共生モデル
Table 3. Levels of Human-AI Collaboration

レベル	タイプ	概要
Lv1	指示型	人が AI に都度指示
Lv2	補助型	AI が補助し、人が最終判断
Lv3	合意型	人と AI の合意領域を自動処理
Lv4	改善型	AI が人の判断差異を学習・改善
Lv5	自律型	AI が完全自動で実施

注：本研究はあくまで Lv3（合意型）の構造実装とその定量評価を主眼としており、Lv4 以降の構造は今後の研究展開に委ねる。範囲を絞ったのは、HAZ という新規枠組みの社会実装と評価に焦点を明確化するためである。

12. 総括

HAZ（Human-AI Agreement Zone）を用いて、AI と人間の判断が一致する領域を定義し、信頼性スコアに基づいた自動承認構造を設計することで、レビュー業務の安全な自動化を実現した。実運用では、60%以上の完全自動化と 99%以上の処理時間短縮を達成し、品質事故はゼロであった。

HAZ は「どこまで AI に任せられるか」という問いに対し、定量的かつ説明可能な合意領域を提示する新たな品質指標であり、従来の Human-in-the-Loop（HITL）では曖昧だった責任の所在や判断基準を明確化した。

本事例では、完全一致およびスコアに基づいて HAZ を定義したが、実際の運用においては、システムの性質や文脈に応じて人と AI の判断差異を許容する「許容一致（permissible agreement）」の範囲を設計することも可能である。また、スコア以外の品質指標を基に合意領域を定める設計も想定される。「合意範囲の設計」は今後の AI 社会実装において最も重要な品質設計課題の一つとなるだろう。

このような人と AI の役割分担に基づく判断の共有は、効率性だけでなく、安心感・説明責任・一貫した品質基準といった人間中心の価値を両立する社会実装の方向性を示している。HAZ の枠組みは「人と AI の良き関係」の実践モデルとして、多様な分野への応用と理論的な深化が期待される。さらに、この設計思想は医療診断、法務審査、製造業の品質判定など、人と AI が共同判断を行うあらゆる意思決定領域に一般化可能であり、「人と AI の合理的合意」を定量的に可視化し、品質保証と責任分担の明確化する設計指針として機能し得る。

参考文献

[1] UGC モデレーションの法的・社会的観点
Sheng, J. “Automated Content Moderation,” Georgetown Law Technology Review, 2022.
[2] 総務省プラットフォームサービスに関する研究会：「誹謗中傷等の違法・有害情報への対策に関するワーキンググループ（第 5 回）」（2023）
[3] Claude 3.7 Sonnet のモデル仕様に関する公式情報
Anthropic (2024), "Claude 3.7 Sonnet and Claude Code", Official Documentation.
[4] Chain-of-Thought（CoT）手法に関する研究
Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, Denny Zhou: "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", arXiv:2201.11903, 2022
[5] Human-in-the-Loop（HITL）の包括的レビュー
Holmberg, J. et al., “Human-in-the-loop machine learning: A state of the art,” Artificial Intelligence Review, 2022.
[6] 判断一致度を用いた信頼性定量評価
Kural, M. et al., “Quantifying Divergence for Human-AI Collaboration and Cognitive Trust,” arXiv preprint, arXiv:2312.08722, 2023.
[7] AI との協調設計と透明性に関する研究
Seeber, I. et al., “Designing Transparency for Effective Huma-AI Collaboration,” Information Systems Frontiers, 2022.
[8] Agentic Workflow の概念と応用
IBM, “What are Agentic Workflows?” IBM Think Blog, 2025.
[9] Few-Shot Learning
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: “Language Models are Few-Shot Learners,” in *Proc. of Advances in Neural Information Processing Systems* (NeurIPS), 2020. https://doi.org/10.48550/arXiv.2005.14165
[10] Many-Shot In-Context Learning
Wang, S., Xie, H., Zhang, C., Lin, Z., Liu, C., et al.: “Many-Shot In-Context Learning via Augmented Demonstrations,” arXiv preprint, arXiv:2404.11018, 2024. https://doi.org/10.48550/arXiv.2404.11018
[11] Guerreiro, N. M., Voita, E., & Martins, A. F. T. (2023). *Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation*. Proceedings of EACL 2023.
[12] Benj Edwards, “Company apologizes after AI support agent invents policy that causes user uproar,” Ars Technica, Apr. 17, 2025. https://arstechnica.com/ai/2025/04/cursor-ai-support-bot-invents-fake-policy-and-triggers-user-uproar/
[13] Liu, P., et al. (2023). "Pre-train Prompting: What Works, What Doesn't, and What's Next." https://arxiv.org/abs/2107.13586