生成AI を用いたレビュー承認自動化に関する研究

— Quantifying the Human-AI Agreement Zone (HAZ) —

合同会社DMM.com
PF第１開発部 松井高宏

ー　自己紹介　ー

- 所属： DMM.com PF 第 1 開発部
- 業務： 商品レビューのプロダクト開発
- 役職： チームリーダー

長年の現場経験を活かし、生成AIを使って
10ヶ月で自動化、14ヶ月で論文化しました。
その成果をお伝えできることを嬉しく思います。

― Executive Summary ―

レビュー承認業務の自動化に関する研究

研究背景	● AI主導の自動化は難しく、多くは「支援型」に留まる ● 自動化にはAIの誤判断を防止する仕組みが必要
対象業務	● 商品レビューのWeb公開可否の判断をAIで自動化 （誤承認は重大リスク）
提案手法	● 人とAIの判断が一致する領域（HAZ）を定義 ○ 「安全に自動化できる基準」として運用
成果	● 年間50万件の60%を自動化 ● 公開時間99.5%短縮＆品質事故ゼロ

Agenda

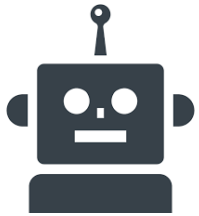
1. 背景と課題
2. 生成AIの導入
3. 生成AIの初期検証
4. 精度改善
5. 自動化基準の策定
6. 自動化の成果
7. 今後の展望とまとめ

1章. 背景と課題

5

研究背景：AI活用と自動化の壁

- 国内外でのAI活用は補助や要約などの「支援型」に留まる
- 「AI主導」の自動化はAIの誤判断が損害に直結する業務では難しく、
  AIの誤判断を防ぐ安全性の確保が必要である



例：ChatBotの誤回答により顧客損害が発生
（例：Cursor、Air Canada社が提訴）

AI自動化に挑戦した業務：レビュー審査

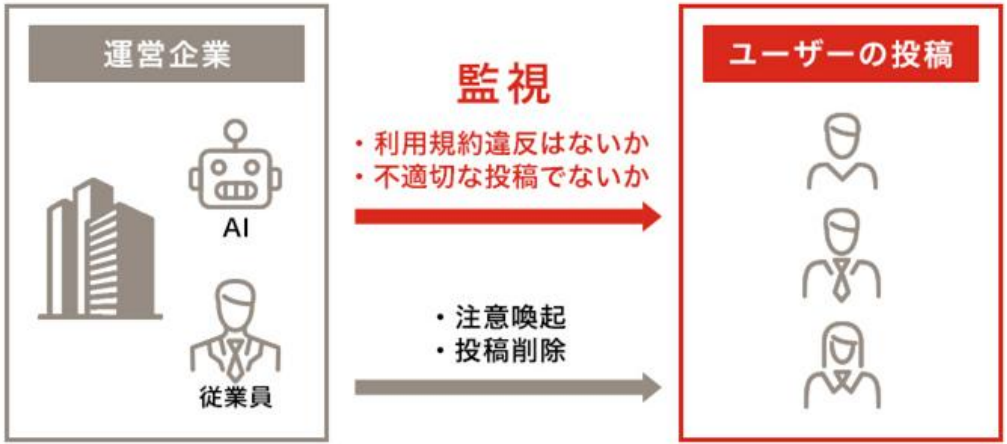
- 投稿レビューの公開可否を全件人手で判断（年間50万件）
- 運営部3名で月150時間を目視チェックを実施
- 多くのWebサービスが抱える共通する「コンテンツモデレーション」の一例

自動化できなかった理由

- 文脈や微妙なニュアンスが含まれる
- 誤承認は品質事故となり信頼を損なう

コンテンツモデレーションとは

- 不適切投稿を監視する業務、安全なWEBサービスの環境の提供が目的
- 企業ブランド保護が重要になる一方、作業負担が増大
- 総務省のPF研究会にてワーキンググループも毎年開催されている



研究目的

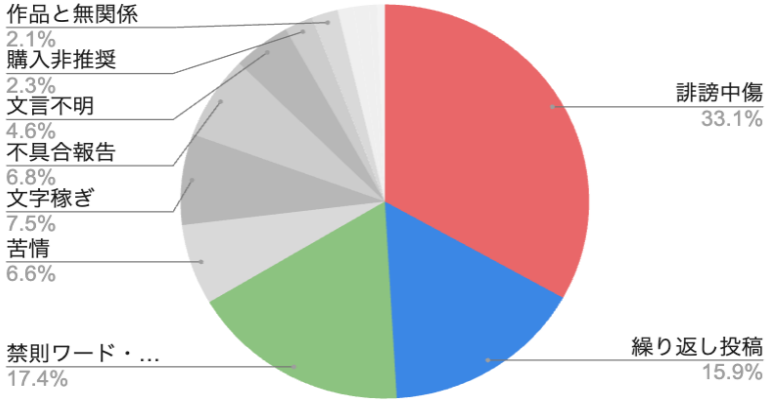
- 本研究では、
  公開ミスが重大なリスクとなるレビュー承認業務を対象に、
  AIの誤判断を防ぎ、安全な自動化を実現する。

2章. 生成AIの導入

審査の難しさ

- 公開できないレビュー
  全体の1割にあたり、誹謗中傷・苦情・購入否定など20種類以上

- 文脈・意図を含んだ総合審査が必要
  - 「下手すぎ、二度と出ないで」→ 公開NG（誹謗中傷）
  - 「下手だが、改善余地あり」　→ 公開OK（批判的だが許容）



生成AI導入の意義

- 直近で生成AIの言語理解能力が大幅に向上。
- 効率化・迅速性・拡張性の観点からも有効と判断し導入を検討

📌 導入メリット

効率化	承認作業の大半をAIに任せ効率化 ✂
迅速化	最大1週間の待機時間がリアルタイム化
基準の統一	人による判断のバラつきをなくした審査 📏
拡張性	投稿増加にも安定対応 📈

生成AI導入の意義

—— 継続性 × 説明可能性（XAI） ——

- 継続性
  - AI専門家でなくても運用・改善できる継続性を実現
  - BEエンジニア中心でも現場全員で品質向上に取り組める

- 説明可能性（XAI）
  - 「なぜNGか」を理由まで明示、現場・関係者が納得できる
  - AIの品質を保証する際に必要な「説明責任」も果たせる

3章. 生成AIの初期検証

初期検証：AIと人の一致度

- AIで自動審査が可能か、様々なモデルの有効性を検証
- 人とAIの判断（OK/NG）がどの程度一致するか（＝一致率）を計測

📌 有効性検証（2024年5月）

モデル	一致率	特徴
Claude 3 Opus	82.0%	高精度だが高コスト。長文理解に強い
Claude 3 Haiku	81.5%	軽量・高速。攻撃表現の検出に強い
GPT-4.0	78.5%	安定しているが、曖昧表現への過剰許容あり
GPT-3.5	70.0%	精度にバラつきがあり

初期検証：プロンプト例

```
# 役割
- あなたはレビューを審査するエージェントです。

#　評価プロセス
- レビュー情報の内容を把握してください。
- 判断項目を順に評価してください。
- 該当する可能性がある場合は、NGと出力します。

#　判断基準
  - 誤解を招く可能性のある表現
  - 過度に攻撃的/下品な表現
```

```
#　出力形式
<output>
  <result>判定結果</result>
  <score>スコア</score>
  <reason>理由の説明</reason>
  <category>該当カテゴリ(N001)</category>
</output>
```

初期検証：自社規約例

・当社は、投稿者が次に該当する内容を含むレビューを投稿することを禁止します。
（1）当社又は第三者の著作権、商標権等の知的財産権を侵害するレビュー
（2）当社又は第三者の財産権、人格権、名誉権等を侵害するレビュー
（3）第三者の個人情報又はプライバシー情報を侵害するレビュー
（4）公序良俗に反するレビュー
（5）法令に反するレビュー
（6）犯罪的行為、犯罪的行為に結びつくレビュー及び犯罪的行為を助長するレビュー
（7）虚偽の情報を含むレビュー

初期検証：見えた課題

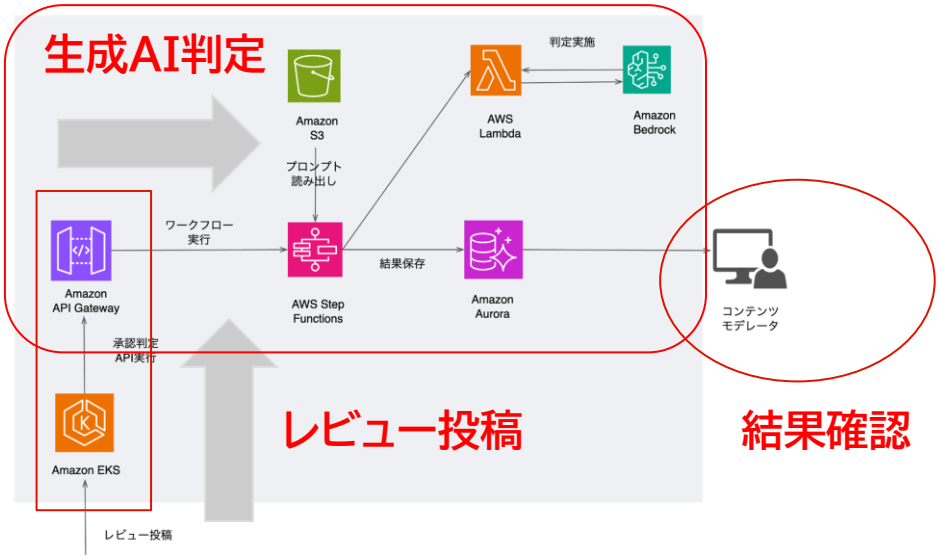
- ある程度の有効性を確認できたが、70〜80%では自動化に不十分
  - 特に文脈依存の高い複雑な判断は、ハルシネーションも発生しやすい

- 次章では、まず精度改善のため、AIの審査構造を見直す必要がある

4章. 精度改善

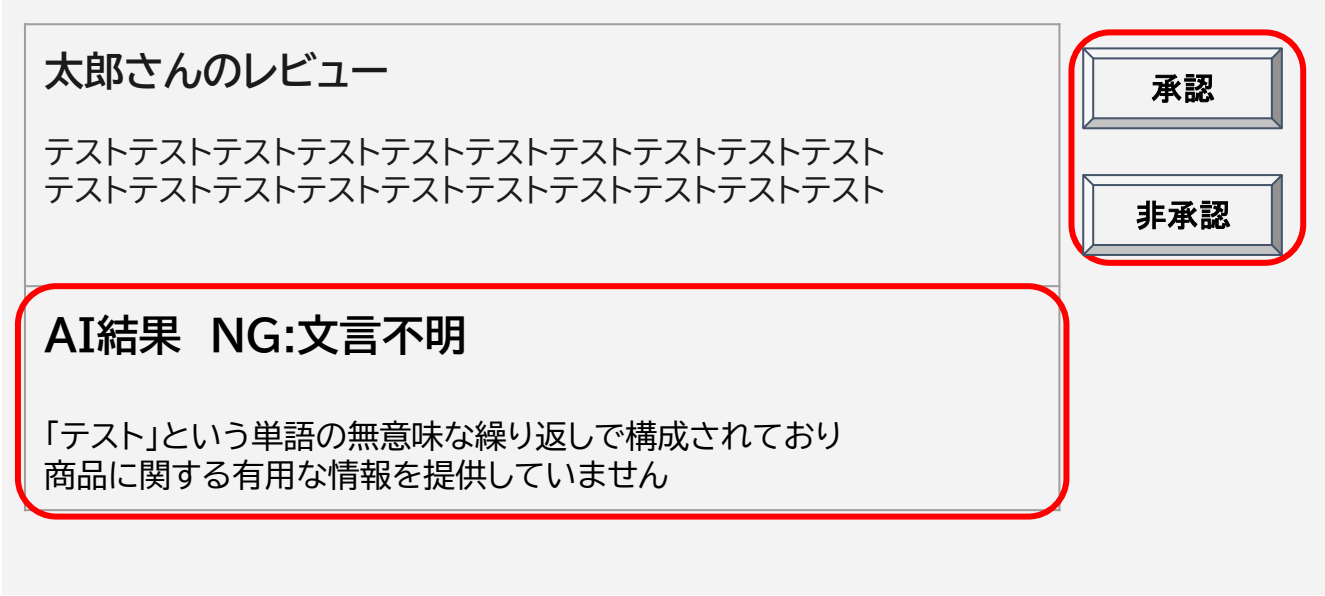
AI審査システムの構築

- AI審査を安定運用するため、AWS Bedrock（Claude）を基盤に構築
  - Step Functions／Lambda／Bedrock を連携
- レビュー投稿 → 自動承認 → 管理画面で結果を確認可能な仕組みを実現



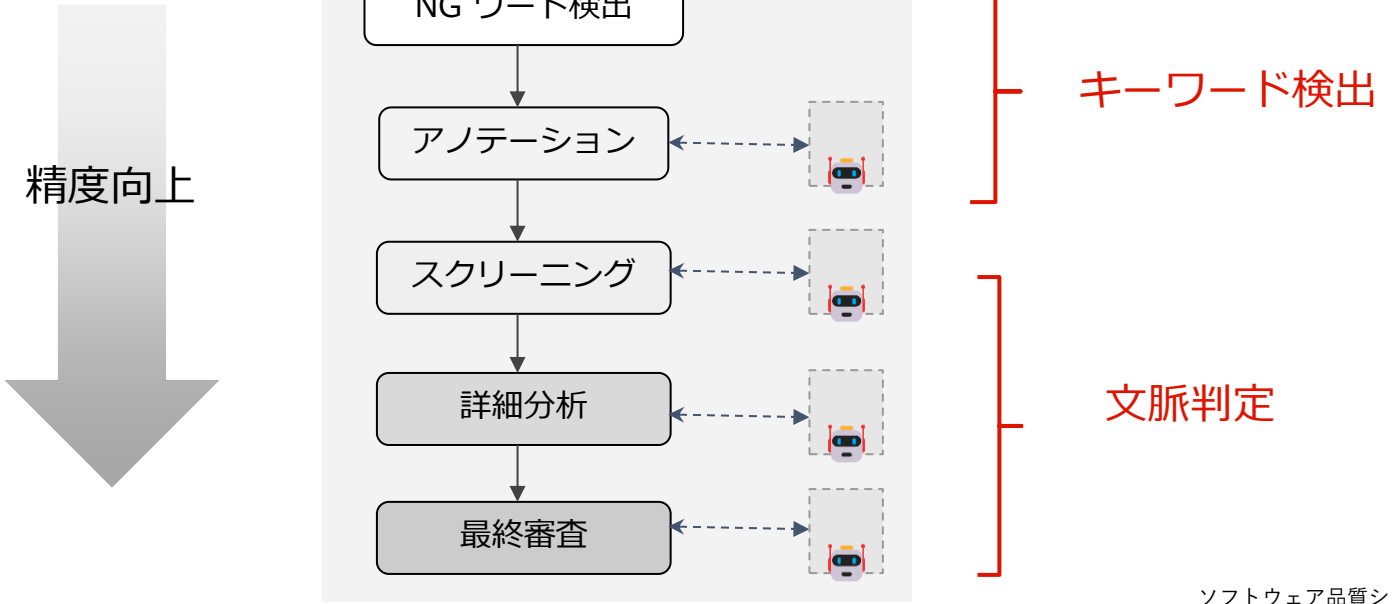
AI審査システムの構築

- レビュー本文、AI判定結果、判定理由を一覧で表示し、
  必要に応じて人が確認・修正できるUIを備える



審査構造分離の設計（＝AWS StepFunctions）

- 審査構造を五段階に分離、各段階でAIの役割を明確化
  - キーワード検出と文脈判定の構造、判断確定した段階で終了する構造



精度改善結果

- 各段階で誤判定を吸収し、一致率は80% → 95%に向上

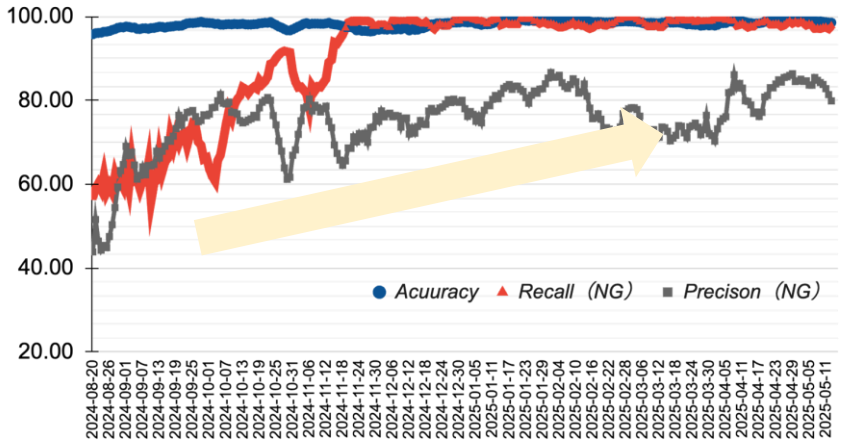
各ステップ例

段階（ステップ）	概要	処理内容
NGワード検出	禁止語チェック	「○ね」→ 即NG
アノテーション	語彙マーキング	「やばい」→ 「*やばい*」
スクリーニング	違反カテゴリ検出	「*やばい*」→ 誹謗中傷カテゴリへ
詳細分析	カテゴリ分析	誹謗中傷カテゴリを深く分析
最終審査	最終結論	NG、OKの最終判断

精度改善の継続と効果

- 6ヶ月にわたり、運営部と様々な改善策により精度向上
  （＊）AIにはOK/NG判定に加え「判定の根拠や理由」も生成、運営部が日々チェック
- 最終的に一致率は 95% → 98% にまで向上、各評価指標も高水準

*精度推移



*最終評価

評価指標	値
Accuracy（一致率）	98%
Recall（NG検出率）	98%
Precision（NG適合率）	80%

＊主要な精度指標

- 最重視したのはRecall（NG検出率）：AIの自動化に最も重要な指標
- 規約違反レビューを誤って公開しないために重視

1. Accuracy（一致率）
   AIと人の判断がどれだけ一致したかを示す割合
2. Recall（NG検出率）
   AIが不適切なレビューをどれだけ見つけられたかを示す割合
3. Precision（NG適合率）
   AIがNGと判定したレビューのうち、実際にNGだった割合

改善事例の紹介

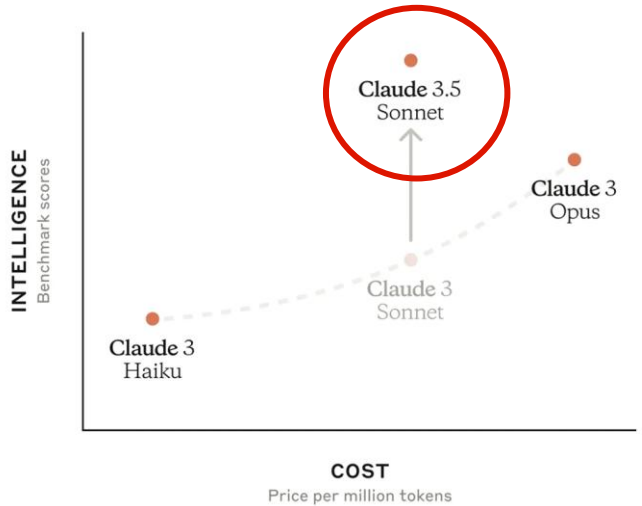
ex1. プロンプト最適化

- 人とAIで判断が分かれたケースの正しい基準を設定

改善事例の紹介

ex2. モデルのバージョンアップ対応

- 新モデル登場で安価で性能の高いモデルが使用可能となった
  - Claude4 Sonnetのモデルを適用（現在）



改善事例の紹介

ex3. 不確実性の対策

- AI判定結果に「OK/NG」とは別に「UK (Unknown)」のカテゴリを設定
- AIの判断が難しいレビューは人に委ね、誤判定リスクを大幅に低減

例：判断が難しい事例

- 動画再生をしないと判断できないケース
- 真偽不明な情報を含むケース

Unknownカテゴリの導入

改善事例の紹介

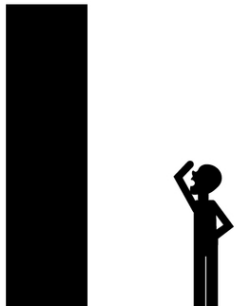
ex4. ニュアンスの学習

- 誹謗中傷の曖昧な概念を大量に学習（Many-Shot In-Context Learning）
  - 「いまいち」といった個人的な感想・批評 = OK
  - 「最悪、バカ」など強い侮辱のある表現 = NG
  - 「センス疑う」など批判的だが断定しにくい表現 = UK

OK	「内容はイマイチです」「セリフが微妙です」
NG	「最悪、バカだと思う」「二度と出すな」
UK	「作品のセンス疑う」「演出は悪いが、面白い」

＊ 「自動化＝安全に自動化できる範囲」はどこか？

- 残る 1〜2% の判断差異が自動化の障壁となった
  - 判断のばらつきに「誰も根拠を示せない」状態が残された
- そこで精度改善に加え「安全に自動化できる基準」の策定が必要になった



5章.自動化基準の策定

安全性と品質重視の設計
HAZとは？

信頼スコア の導入

- AIの判断結果（OK/NG）に信頼スコアを付与
  - 0.0～1.0の範囲で定量化、低いほど安全で高いほど危険

＊ 信頼スコア [0.0～1.0] = f（Step, Grade）

- Step：確定ステップ
  - *Agentic Workflow*の判断が確定した段階「**判断の難しさ**」を示す
- Grade：レビュー品質
  - レビューの曖昧さ・攻撃性に応じ「***S～Fで6段階***」で品質を分類

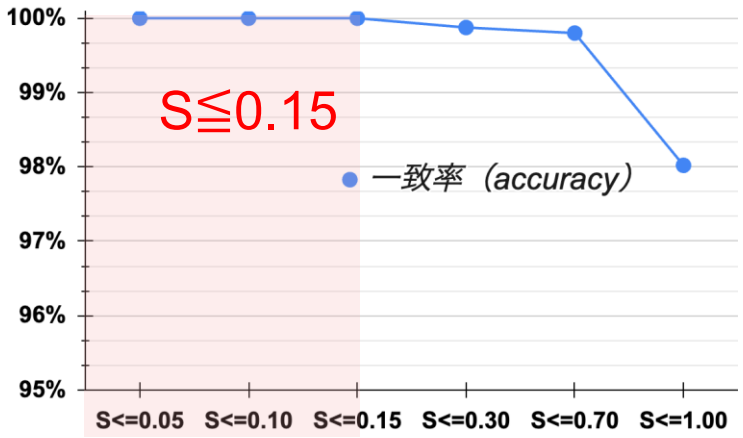
信頼スコア のスコア帯

- 各スコア帯の意味と、対応するレビューの特徴は以下の通り
  （※単なるネガ・ポジ指標でない）

確定ステップ	レビュー品質	信頼スコア
スクリーニング	S:高品質　「説明書通り使えます」	0~0.05
	A:良質　「使い勝手は普通です」	0.06~0.10
	B:普通　「悪くはないです」	0.11~0.15
詳細分析以降	C:該当なし　「想像と違った」	0.16~0.30
	D:曖昧　「なんか微妙」	0.31~0.7
	F:違反　「バカ、○ね」	0.71~1.0

信頼スコア の分析

- 20万件の分析では、S≦0.15のスコア帯は「人とAIの判断は観測上一致」
  - 軽微な人の判断揺れあったが、公開問題なしの判断（95%信頼下限で99.997%）
- また、この帯域は全体の70%を占める（＝大きな自動化の可能性）



スコア帯と一致率

安全に自動化できる範囲を定義

- 信頼スコア S＝0.15を基準に人とAIの審査を分離することが可能
  - AIが審査可能なこの範囲を「人とAIの合意領域（＝HAZ）」と定義
  - HAZ外は必ず人が確認するため、品質リスクをゼロに抑えられる

HAZ：Human-AI Agreement Zone

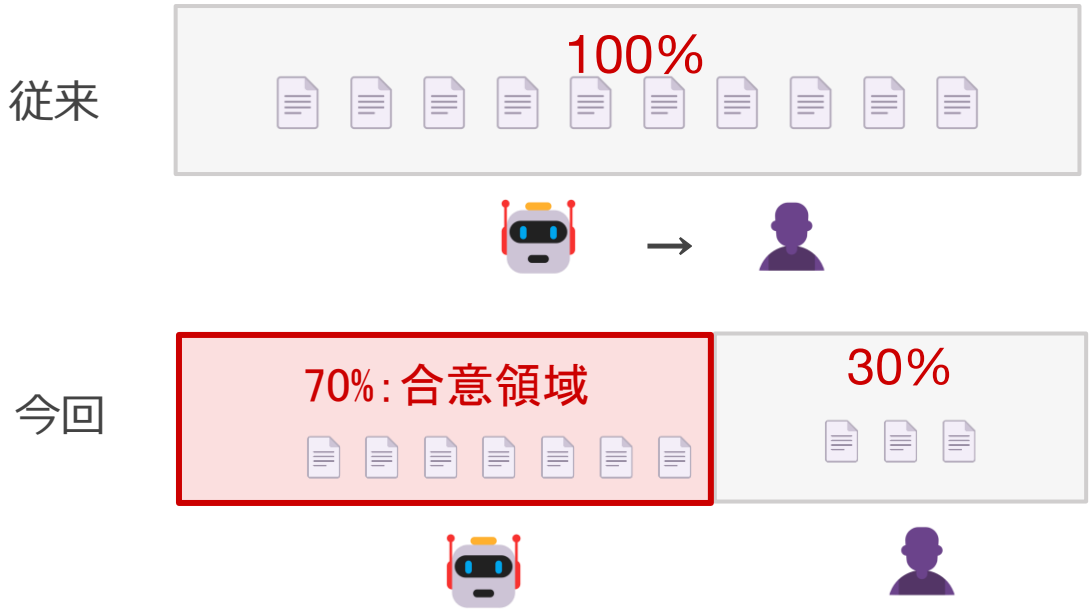
私が提唱した概念

人とAIが合意した範囲において

AIによる安全な自動化が可能なフレームワーク

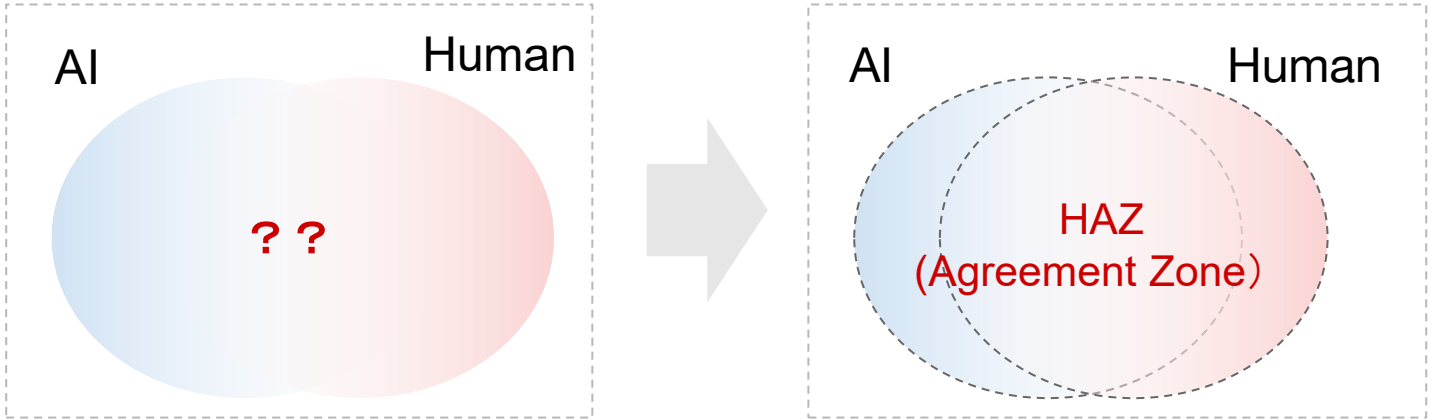
安全と自動化を担保する構造（*HAZ*）

- 従来：AI判定を人が全件チェック（HITL） → 業務量が削減されない
- 今回：合意領域をAIが自動化（HAZ） → 業務量の削減が可能



HAZの社会的意義

- 本研究は曖昧なAI判断の責任範囲を可視化
- 人手に頼ったモデレーション業務の自動化に成功
- このアプローチはAI自動化を求める多くの分野に有効（医療・画像診断・異常検知等）



HAZは、提案にとどまらず

現場を大きく変えました

次は、その変化の実態をお見せします

6章. 自動化の成果

自動化方針

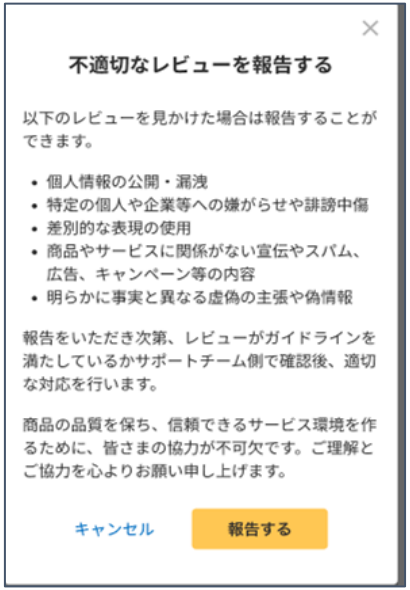
- 万が一に備え、自動公開後に不適切レビュー検知も可能な設計とした

通報機能

- 利用者が不適切レビューを通報できる機能

モニタリング

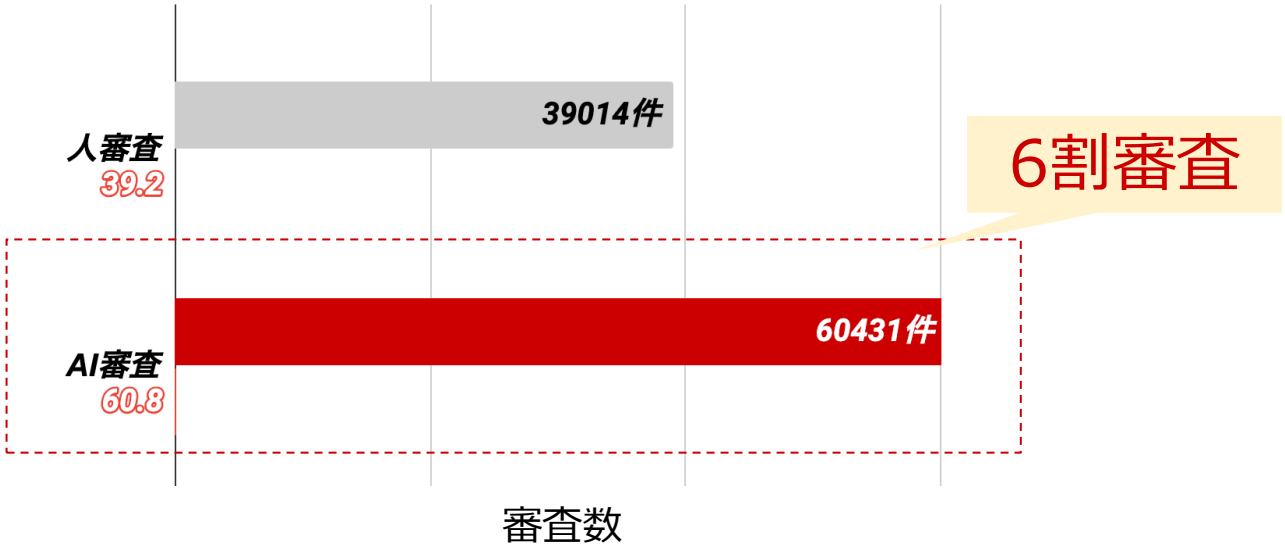
- 不適切レビューがないか後日人が確認



自動化実績（2ヶ月強）

審査数

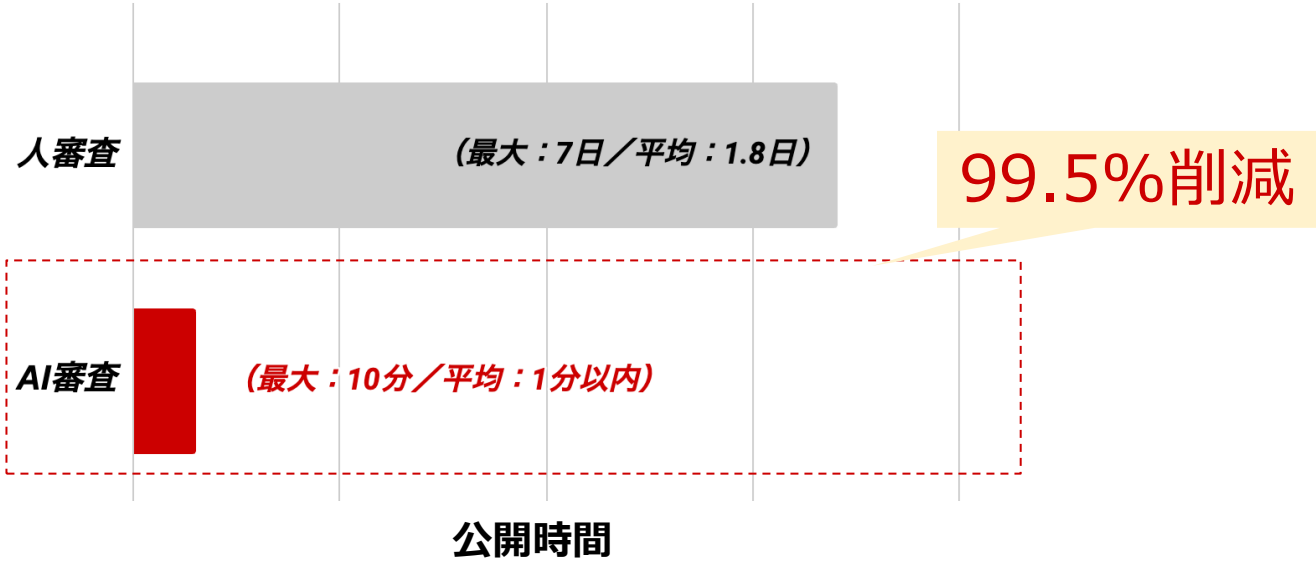
- AI審査：約9万件の審査の6割（初回導入のため商品限定）



自動化実績（2ヶ月強）

Web公開までの時間

- 人審査：最大7日　→　AI審査：　10分以内



社内の声の変遷

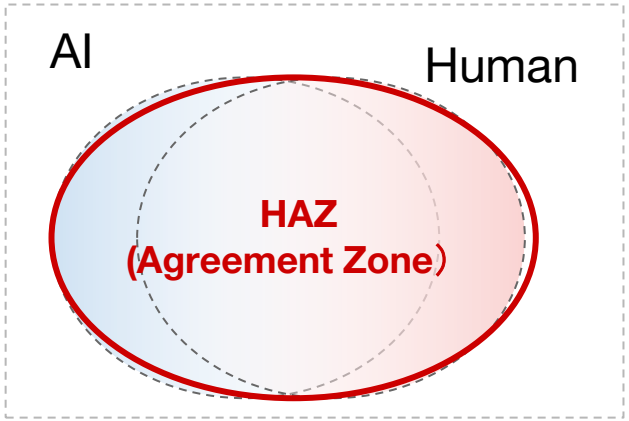
- 導入検討：「AI自動化など絶対に無理、責任取れない」
  ↓
- 検証期間：「精度は高いが実運用が心配」→ HAZ適用
  ↓
- 運用開始：「無理と思ったが、実現がすごい！クレームもゼロ！」

自動化率60%を達成しつつ、品質事故ゼロを維持

7章. 今後の展望とまとめ

（１）HAZの領域拡張

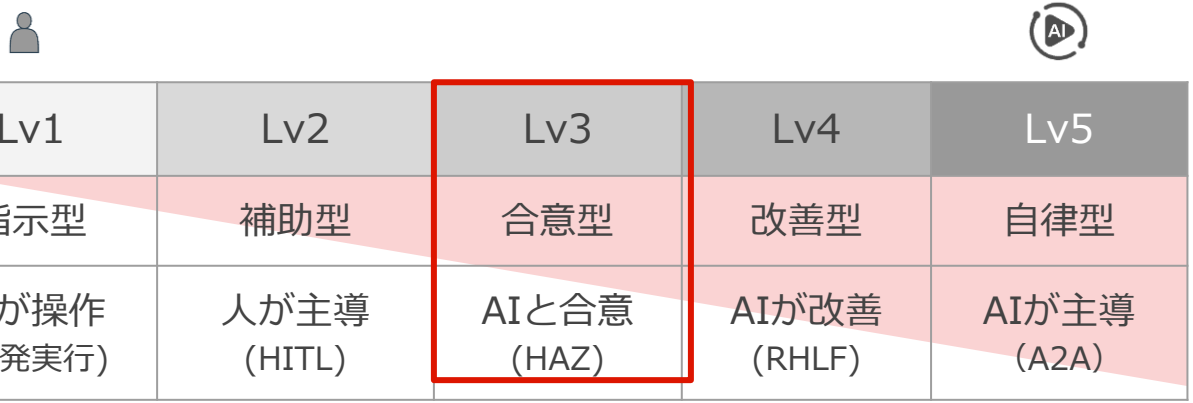
- 人とAIが「完全一致」する範囲（S≦0.15）から
  軽微な差異を含む「許容一致」（S≦0.70）を対象に自動化を拡大
- 許容一致＝人とAIの判断の差異を「揺れ」として扱い、より大きな自動化を可能にする



差異を前提とした合意運用
（permissible agreement）

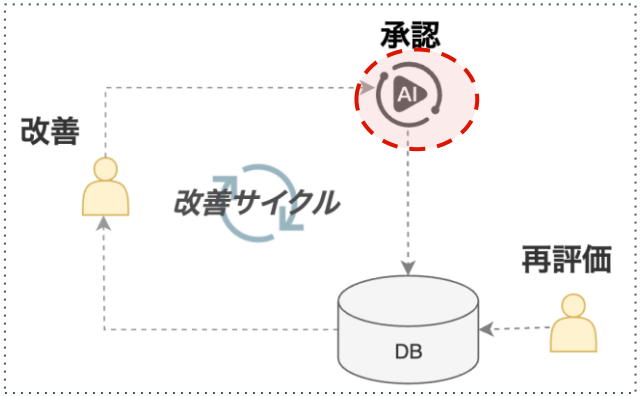
（2）人とAIの共生モデル

- 人とAIの関係性に着目した「共生モデル」を定義
  - レベルが上がるごとに、AIの実行比重が大きくなる構造
- 今回の自動承認システムはLv3（合意型）へ到達
  - 「人とAIが合意できる範囲」を明確化し、安全な自動化を実現

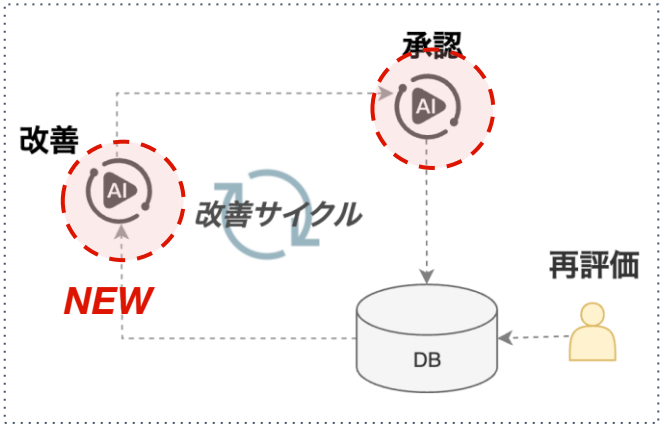


（2）人とAIの共生モデル

- Lv3（合意型）→ 今後はLv4（改善型）を目指す
  - 人が行っている改善サイクルをAIに引き継ぎ、自動化範囲を拡大
  - プロンプト改善やHAZ閾値調整を通じ、安全性を維持しつつ自律度を高める



Lv3（合意型）



Lv4（改善型）

まとめ

自動化の成果

- HAZにより、安全性（品質）を担保しながらモデレーション自動化を実現

自動化成功の鍵

1. ステップ分割による審査の構造化
2. 信頼スコアによる定量化
3. 安全に自動化できる範囲の特定（HAZ）

こうした取り組みを通じ、品質を担保しつつ、
自動化の社会実装を可能にする新モデルを示した

ご清聴ありがとうございました。

APPENDIX

Q&A集（1/2）

HAZの閾値（S=0.15）はどのように決めたのか？	20万件の実データと6ヶ月に渡る運営部との調整と検証で、 安全なラインとして社内合意したラインです。 またAIの根拠があっているかどうかということを重視しました
残る2%の自動化できない領域にはどんな特徴が多いか？	人によって感じ方が分かれるグレーな部分が残り2%です。
ユーザー（投稿者やレビュワー）から反発やクレームはなかったか？	ユーザーからの反発はない、また実は人の方がミスが多いことも分かっています。
誤判定が生じた場合の運用フローは？	通報とモニタリングで、何かあればすぐ運営が対応できる体制です。
他の分野や業務にHAZは適用可能か？	OK/NGのような明らかな基準がある業務なら、AIと人で役割を分けやすい。 ・製造の品質検査：明らかにOKな製品はAIで判定 ・医療診断：軽症や問題ない症例はAI、難しいものは医師が確認 ・画像検査・異常検知：AIが自信ある部分だけ自動化 ・文書の仕分け：間違いようのないものはAI、曖昧なものは人が見る ・投稿監視　：明確にOKな投稿だけAIで自動承認 ・金融の与信審査：安全な案件はAI、それ以外は人が最終判断
このやり方が他の分野や別案件でも通用する根拠は？	汎用的なフレームワークではないが自動運転で実施されている。 高速道のみ自動化するなど。 「自信のある範囲だけ自動化する」仕組みは他分野でも有効です。

Q&A集（2/2）

一致度とスコアの関係、どちらが原因でどちらが結果？	ーまず一致度を見ましたが、それだけでは十分でなく、同じOKにも濃淡がありました。そこでどれだけOK寄りかを定量化するためにスコアを導入した
スコア算出プロセス・処理時間はどれくらい？ 大量データでも実用的？	1件あたり1分以内で処理できて、並列処理が可能であるため 現行の仕組みであっても今の100～200倍まで対応できます。
従来法との定量的な比較をもう少し詳しく教えて	従来のHITLよりも、自動化の領域が大きく広がりました。
導入によるコスト効果や、人的リソースの削減度合いは？	今は6割以上自動化できていて、ほぼ1人で回せるくらい負担が減りました
今後の展望や課題は？	HAZをもっと拡大して、8割くらいまで自動化したいと考えています。
HAZの範囲を広げることはできるか？	人とAIが合意できれば、HAZの範囲はもっと広げられます。
AIモデルの選定や、モデルの具体的な種類は？	ClaudeやGPTなど複数で比較して、コスト、安定性で一番良いものを選びました
AIに責任を持たせないという視点と、今後の自動化拡大についてどう考えるか？	AIの責任範囲を明確すること、人が主体であることを原則として自動化を広げる。