ASDoQ大会2026　ポスター発表

ASDoQ文書品質モデルによるAIレビューの結果を安定させる実行環境の検討

― 構造化と収束ループの効果

森口 功造　株式会社川村インターナショナル

0 研究の概要

1. AIによる文書レビューは、同一条件で実行しても、検出される欠陥の数が試行ごとに変動する。

2. 当社の機械翻訳で用いる技術を応用し、評価基準（ASDoQでの名称は「測定基準（暫定名）」）を構造化して、文書を1文単位で点検・修正を収束するまで反復する手法を構築した。

3. 検出率は82%（従来71%）に向上し、試行間のばらつきは6pt（従来18pt）に縮小した。

4. 文書全体を構造化して問題の候補を与え、全文を入力すると、文書全体を読まなければ判定できない欠陥の検出率が約20pt向上した。

1 背景とリサーチクエスチョン

ASDoQサマーワークショップでは、DADAプロセスに沿って、ASDoQ文書品質モデルに基づき仕様書をレビューするAIエージェントを開発した。DADAプロセスでは、計画・レビュー・検証などの役割を担う複数のエージェントを決まった順に動かす（オーケストレーション）。しかし、同一のSkill・同一の環境であっても、試行ごとに結果が変動した。

検出率：ワークショップのSkill（WS）5回試行で**41〜50%**。別方式のSkill（A）10回試行で**56〜94%**（行単位の欠陥）。

当社は機械翻訳で、文章を構造化したデータに変換し、1文ずつ修正して結果が収束するまで反復する技術を用いている。この技術を応用し、文書品質モデルを構造化したうえで、役割を分けずに修正を収束させる手法でも、同様の処理ができるのではないかと考えた。

RQ：試行間のばらつきが小さく、かつ検出率の高いAIレビューの実行環境は構築できるか。

2 仮説

仮説1　構造化
評価基準を「確認事項・手がかり語・修正方針」に分解して再記述（構造化）し、文書を1文単位で同一の手順により点検すれば、結果は安定する。

仮説2　全文の構造化
評価基準の構造化に加え、文書全体を読んで得た矛盾や問題の候補を構造化データとして与えれば、文単位を超えた欠陥の検出に役立つ。

仮説3　収束ループ
修正した文を再点検し、前回と同一の結果が得られた時点で終了する処理を反復すれば、結果は安定する。

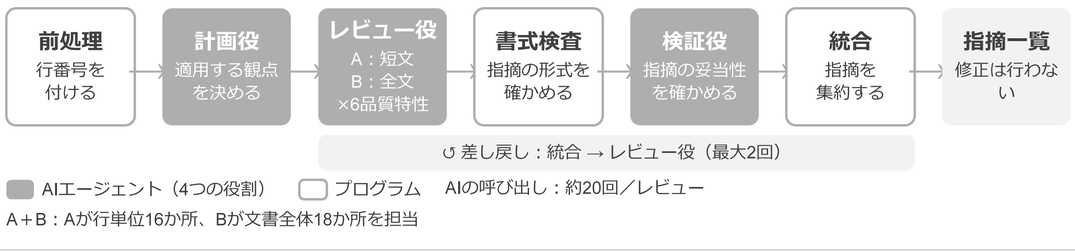
3 測定方法

34か所の欠陥を意図的に埋め込んだ要求仕様書を用い、AIが検出した欠陥の数を計測した。

- 評価基準：ASDoQ文書品質モデル活用WGが検討中の「ASDoQ文書品質モデル 再構成版 v0.1」（173項目）を使用した
- 対象文書：会議室予約システムの要求仕様書（370行）。欠陥は行単位16か所、文書全体18か所
- 検出の定義：AIの指摘が欠陥の行を指し、かつ欠陥の種類（品質副特性）が一致すること
- 各方式を3〜10回試行し、平均値と、最小値〜最大値の差（ばらつき）を求めた

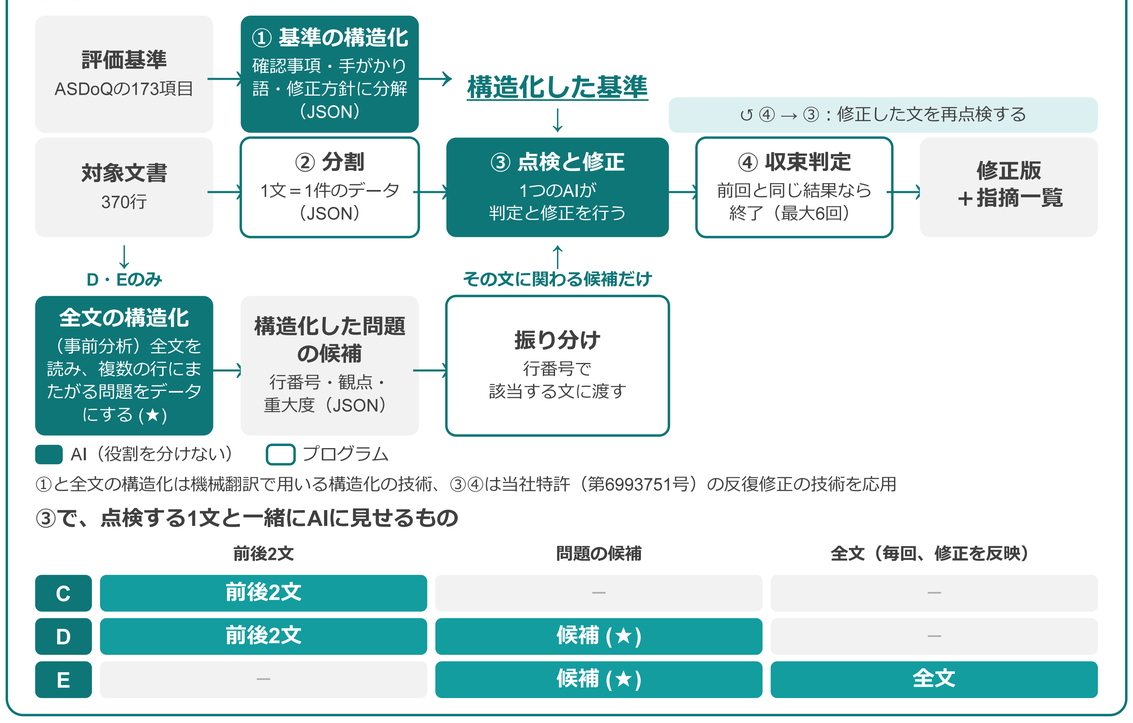
4 比較した方式：パイプラインの比較

DADAプロセスのAIエージェント（A＋B）



A＋B：Aが行単位16か所、Bが文書全体18か所を担当

構造化＋収束ループ（本手法）



①と全文の構造化は機械翻訳で用いる構造化の技術、③④は当社特許（第6993751号）の反復修正の技術を応用

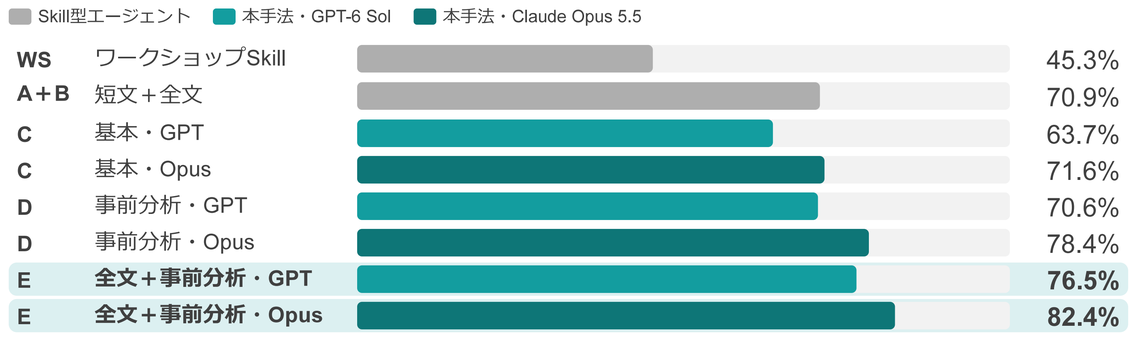
③で、点検する1文と一緒にAIに見せるもの

	前後2文	問題の候補	全文（毎回、修正を反映）
C	前後2文	–	–
D	前後2文	候補 (★)	–
E	–	候補 (★)	全文

① 構造化の例
「〜しない場合は〜しない」などの二重否定を避け、肯定形を用いて直接的に動作を定義している。
↓
確認事項：肯定形で動作を定義しているか
手がかり語：「しない場合は」「しない」
修正方針：肯定形に書き換える
重大度：Critical

★ 全文の構造化の例（E × GPT-6 Sol）
203行：システムは、メールサーバへの送信のタイムアウトを60秒とする。
264行：システムは、メールサーバへの送信のタイムアウトを30秒とする。
↓
行番号：203、264
品質副特性：無矛盾性/重大度：Critical
内容：同じ送信処理のタイムアウトが矛盾する

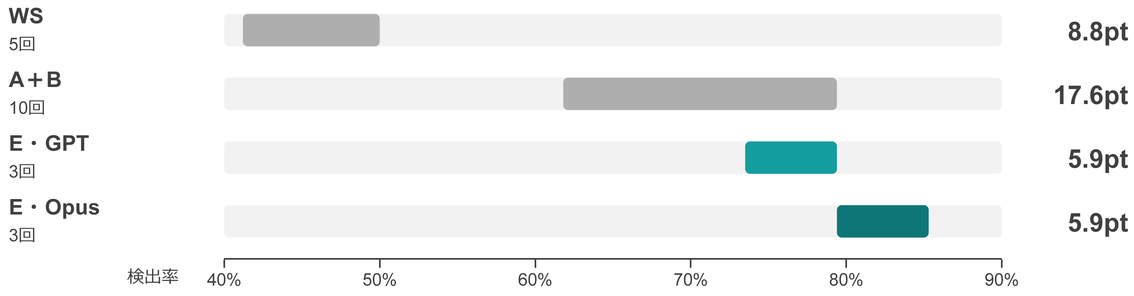
5 結果1　検出率：E が最も高い



34か所に対する検出率（平均）。品質副特性を問わず行のみで照合すると、E は88〜91%、A＋B は79%。

6 結果2　ばらつき：E は試行間の差が小さい

同一文書に対する複数回の試行における検出率の最小値〜最大値



pt：百分率どうしの差（80%と74%の差は6pt）。E では34か所中32か所で3回の結果が一致した。修正文の多くは2回目の点検で収束した。

7 結果3　全文入力の効果

文書全体を読まなければ判定できない欠陥（18か所）の検出率



全文を入力すると検出率は約20pt向上した。行単位の欠陥は、Opusでは前後2文のみでも94%であった。

8 結果4　精度・費用・処理時間

	WS	A＋B	E・GPT	E・Opus
検出率	45%	71%	77%	82%
処理時間/回	約3〜7分	約20分	約40秒	約4分
費用/回	約$1〜2	約$20〜40	$1.42	$2.60
誤った修正/回	—	—	1.3件	0件

E は実測値（1回＝370行を最大6回点検）。A＋B は API 単価による推定値（定額プランで実行）。WS は筆者の試行による概算値。
WS・A＋B は指摘のみで、修正は行わない。

9 結論：最良の方式は E

用途	推奨する組み合わせ	検出率	ばらつき	費用/回	処理時間/回
精度を重視	**E × Claude Opus 5.5**	**82.4%**	**5.9pt**	**$2.60**	**約4分**
費用と処理時間を重視	**E × GPT-6 Sol（修正は人が確認）**	**76.5%**	**5.9pt**	**$1.42**	**約40秒**
（参考）従来	A＋B（短文＋全文）	70.9%	17.6pt	約$20〜40	約20分

仮説1 構造化：支持
検出率が向上し（71→82%）、ばらつきが縮小した（18→6pt）。

仮説2 全文の構造化：支持
問題の候補を与えると、文書全体の欠陥の検出率が向上した（52〜54→65〜67%）。

仮説3 収束ループ：一部支持
大半の文は1〜2回で収束した。反復は修正の確認と打ち切りに寄与したが、安定化の主因は構造化と考えられる。

用語

構造化：文章で記述された基準や文書を、項目単位のデータに変換すること
オーケストレーション：役割の異なる複数のAIエージェントを決まった順に動かすこと
収束ループ：修正した文を再点検し、結果が変わらなくなるまで反復する処理
品質副特性：ASDoQ文書品質モデルにおける17の観点（簡潔性、無矛盾性など）
全文を構造化したJSONファイル：文書全体の要求・用語・値とそれらの関係を、項目単位で記録したデータ

10 今後の課題

- 全文を構造化したJSONファイルを Vector Store（AIが検索・参照できる保管領域）に格納し、修正時に参照する。全文を入力できない長大な文書への適用を想定する。
- 反復なし（1回のみ）との比較により、収束ループの効果を分離して評価する。また、文書の種類と試行回数を増やして検証する
- 判定に特化したモデル（文章を生成しない型付き判定）を用い、判定と修正を分けた場合の再現性を評価する

検証用スキルダウンロード

A＋Bと本手法のSkill、評価用文書を大会参加者に限定配布する。
本手法のSkillはAPIで当社のサービスに接続して動作する（検証用キーは無償で発行）。

参考文献
[1] システム開発文書品質研究会（ASDoQ）ASDoQ文書品質モデル活用WG：ASDoQシステム開発文書品質モデル再構成版 v0.1（検討中）
[2] 山本雅基：DADA Process（AI×人間 協調開発テンプレート）, https://github.com/yamaPiT/DADAProcess（MIT License）