# 生成 AI 時代の医療プロモーション資材における構造・論理分離アプローチ:監査可能性の確立と検証プロセスの「質と量」の両立

A Twin-Track Architecture for Deterministic Verification of Medical Regulatory Documents

小泉 直子[*1] 八角 潤一[*1] 藪 紘明[*1] 後藤 誠[*1] 後藤田 達哉[*2] 田中 和哉[*2*3]
Naoko Koizumi, Junichi Yasumi, Hiroharu Yabu, Takashi Goto, Tatsuya Gotoda, Kazuya Tanaka

[*1] 株式会社 博報堂メディカル Hakuhodo-Medical Inc.
[*2] scheme verge 株式会社 scheme verge, Inc.
[*3] 政策研究大学院大学 GRIPS

The rapid advancement of generative AI has accelerated and scaled the production of documents, including pharmaceutical promotional materials. However, medical information is life-critical and demands high accuracy and auditability; LLM-only approaches cannot sufficiently guarantee reproducibility of regulatory rules or eliminate hallucinations. This study proposes a verification support architecture based on a Twin-Track Strategy that separates document structure extraction from logic-based verification. Deterministic document structuring and rule verification using logic trees ensure reproducibility and explainability. In practical evaluation, the primary review process was reduced from 10–25 hours to 10–30 minutes. Furthermore, analysis of inter-reviewer variability suggests that delegating explicit rule verification to AI enables human reviewers to focus on higher-order judgments requiring contextual understanding

## 1. はじめに

近年，大規模言語モデル（LLM）をはじめとする生成 AI の発展により，医療情報を含む各種文書コンテンツの作成は，速度・量ともに飛躍的に向上している．一方で，医療情報は人の生命・健康に直接影響を及ぼし得るクリティカルな領域であるため，高い正確性と説明可能性が要求され，誤情報やハルシネーション（幻覚）の排除は不可欠である[01]，[02]．このため，生成 AI が生み出すコンテンツについては，最終的な妥当性確認を人間が担う Human-in-the-Loop 型の運用が一般的である．しかし，生成のスケールが拡大する一方で，人間による検証プロセスは物理的制約を受けやすく，検証工程のボトルネック化が実務上の課題となっている[01]，[03]．医療用医薬品のプロモーション資材や添付文書との整合性が求められる関連文書の作成・レビューにおいては，規制要件やルールへの厳密な適合が求められると同時に，製品特性や臨床成績の設計，疾患領域固有の文脈理解に基づく内容評価が不可欠である．このようなレビュー業務は，複数の判断レイヤーを内包しており，専門性を持つ人間にのみ可能な作業な上に,作業負荷が高い．本研究は，生成 AI 時代における医療情報の信頼性担保という課題に対し，「読み取り（構造抽出）」と「検証（ルール適合性確認）」を明確に分離する Twin-Track Strategy に基づく検証支援基盤を提案する．確率的モデルに依存しない決定論的な文書構造化およびルールベース検証を中核とし，人間は文脈理解や表現の妥当性評価といった判断が本質的に必要な領域に集中できる役割分担の実現を目指す．なお，本研究は，医療情報コンテンツ制作における人間中心の AI エコシステムの可能性を検討した先行研究[04]の続報として位置づけられる．先行研究では，ルールベース AI と LLM を組み合わせた「人間中心の AI エコシステム」の全体像を示したのに対し，本稿では，検証プロセスに着目し,再現性・監査可能性を保証するためのアーキテクチャ設計原理（Twin-Track Strategy）を明確化することを主目的とする.さらに検証を AI に委譲することで人間の検証負荷をどのように軽減し,人間の専門性をどこに集中させるべきかを検討した.

連絡先：小泉 直子，株式会社 博報堂メディカル，東京都港区赤坂 6-2-4 S-GATE 赤坂 6 階，https://www.hakuhodo-medical.co.jp/contact/

## 2. 背景と課題

本節では，本研究の問題設定の背景として，生成 AI の普及による検証負荷の増大，LLM による文書構造化の限界，および人間のレビュー業務が内包する構造的課題について整理する．

### 2.1 生成能力の爆発的向上と検証の危機

生成 AI の普及により，文書やコンテンツの生成速度と量は急激に増大している[01]．一方で，医療情報は誤情報の影響が深刻なリスクとなるため，高い正確性とハルシネーションの排除が絶対条件である[02]．しかし，生成 AI が生み出す情報の量と速度に対し，人間による検証プロセスは物理的に追いついておらず，従来の Human-in-the-Loop 型のレビュー体制は検証工程のボトルネック化を招いている[01]，[03]．このように，「生成のスケール」と「検証のスケール」の非対称性は，医療情報の信頼性担保における構造的課題となっている．

### 2.2 LLM による文書構造化の限界（予備実証）

近年，LLM を用いて文書を直接 JSON 形式等の構造化データに変換する試みが報告されている［05］．しかし，LLM は確率的生成に基づくため，長文文書におけるセクション境界の揺らぎや階層構造の欠落が生じやすいことが知られている［06］．本研究においても，初期検討として LLM による文章構造化を試みたが，タイトル・サブタイトル・本文の区分が実行ごとに変動し，その区分によって適応範囲が厳密である後続のルールチェック結果の再現性に悪影響を及ぼすことが確認された．

### 2.3 検証タスクの多層性と人間のばらつき

医療プロモーション資材のレビューは，「ルールチェック」と総称されることが多いが，実際には以下の異なる性質のタスクが絡み合っている．これらの判断レイヤーが混在したままレビューが行われることで，レビュー実施者（レビュワー）間の指摘内容

にばらつきが生じやすくなる. このばらつきは個人能力の問題だけではなく, 検証タスクの構造的特性に起因する課題である.

(1) 明文化された規制ルールへの適合性確認
(2) 表現の分かりやすさや強調・誹謗中傷・曖昧な表現の是正
(3) 医薬品の特性や固有の文脈理解を前提とした内容評価
(4) 誤記・不統一等に関する品質管理(QC)・校正

## 3 提案手法

### 3.1 設計方針

LLM の中でも特にビジョン言語モデルによる直接情報抽出のトレンドとは異なり, 我々は再現性を担保するために特化したルールベース抽出エンジンを開発した. 本研究では, より洗練されたプロンプトエンジニアリングの追求ではなく, アーキテクチャの分離に求める. 具体的には,「読み取り(構造抽出)」と「検証(ルール適合性確認)」を機能的に分離し, 非構造化テキストの処理を第 1 フェーズに閉じ込めることで, 検証フェーズを透明で決定論的なロジックに委ねる設計を採用する. この構造分離により, LLM の利便性(要約・説明生成等)を補助的に活用しつつも, 規制要件に求められる再現性および監査可能性(説明可能性も含む)を担保することが可能となる.本システムのアーキテクチャを示す(図 1).

### 3.2 論理検証エンジン

ドキュメントが JSON にシリアライズされると, 検証エンジンがロジックツリーを適用する. これは製薬工業協会作成:医療用医薬品製品情報概要等に関する作成要領(略称:作成要領)のルール[07]を表現する決定木であり, 数学的に形式化可能な検証ロジックを実現する.

各ルールは以下の一般式で表現される:

$$A = C_1 \; op \; C_n$$

$$C_i = f_i \; \{L_i, D_i(S_i)\}$$

| 記号 | 名称 | 説明 |
|---|---|---|
| $A$ | Alert | ルール全体の評価結果(true/false) |
| $C$ | 条件ブロック | キーワード照合条件を定義する検査単位 |
| $f$ | 評価関数 | all, some, not_all, not_any, exists |
| $L$ | 単語リスト | 検索対象のキーワード群 |
| $D$ | 文書 | 検査文書 $D_T$ または参照文書 $D_R$ |
| $S$ | セクション | 資料内の特定の場所 |
| $op$ | 演算子 | and, or, none |

ここで, $A$ は評価結果, $C$ は条件ブロック, $f$は評価関数を表す. 各条件ブロックは, 評価関数$f$, 検索語リスト$L$, 対象文書$D$, および対象セクション$S$ により定義される.

### 3.3 網羅性(パターン・ロジックツリー体系の根拠)

本研究で採用した 25 パターンのロジックツリー体系は, 医薬品資材審査の実運用ルールを実装可能な形に整理する中で導出されたものである. 具体的には, 複数の医薬品資材審査に関連する手順書(計 15 文書, 約 200 手順)を対象として, 各手順を「評価関数×論理演算＋特殊パターン」の組合せとして整理・分類した結果, 25 種類の基本パターンに集約可能であることが確認された. さらに, これらのパターンは実運用における検証ルール群(320 ルール)として具体的に定義化されており, 医薬品資材審査業務における規制ルールの記述を, 有限個の形式化可能な論理パターンとして表現できることを示している. 本体系は, 作成要領[07]の記述構造を形式化するための実装上の枠組みとして位置づけられるものであり, 今後の運用拡張や新規ルール追加に応じて拡張され得るものである.

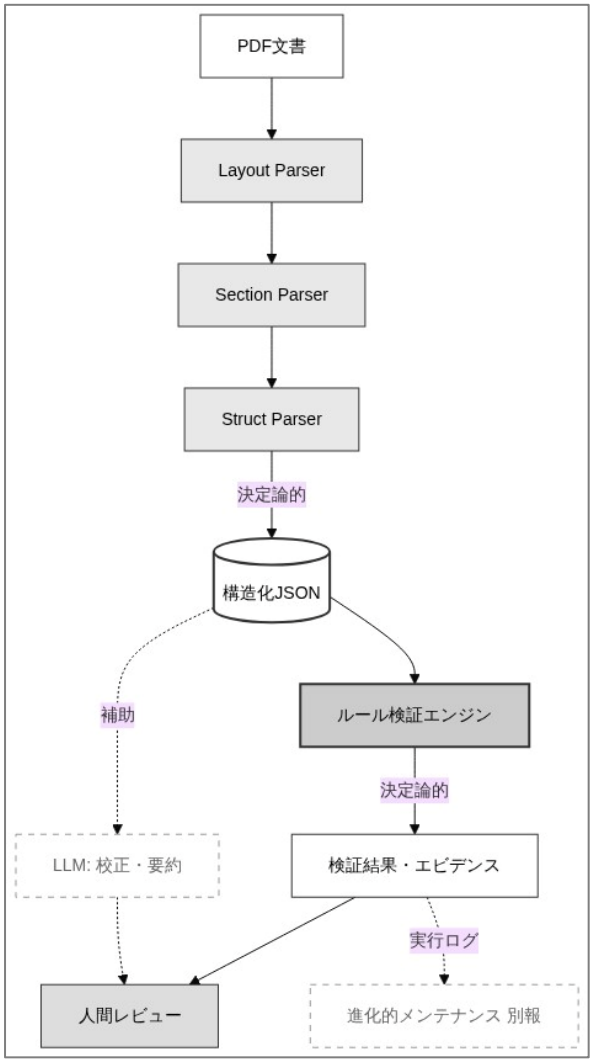


**図 1:本システムのアーキテクチャ**

### 3.4 既存研究との位置づけ

本研究は, 文書解析および規制対応 AI の 2 つの研究領域にまたがるものであり, 関連研究との差異を以下に述べる.

① 文書レイアウト解析との差異:
Layout Parser [08]や DocLayNet [09]等の, 汎用的な文書セグメンテーションを対象とするものに対し, 本研究は医療プロモーション資材(総合製品情報概要等)に特化し, PMDA が規定するセクション構成, 標準テンプレートに準

拠した最大 7 階層の構造をドメイン固有の知識として直接モデル化する点に新規性がある.

② ビジネスルールエンジンとの差異:
Drools [10] や IBM ODM 等の汎用ルールエンジンは宣言的なルール記述を提供するが, 総合製品情報概要に特有のテキスト照合パターン(部分一致, セクション横断参照, 動的キーワード抽出等)への適用は限定的である. 本研究のロジックツリーは, 作成要領 [07] に基づく検証パターンを形式化し, 規制ドメインに最適化している点に新規性がある.

③ 医療用医薬品規制における AI 研究との関係:
医薬品分野の NLP 応用は, 有害事象検出 [11] や臨床試験マッチング [12] が中心であり, 医療情報の規制要件に基づく検証を主対象とした研究は少ない. また, 添付文書と資材間の情報整合性という製薬業界固有の品質保証課題を扱う点に新規性がある.

## 4 有用性評価の観点

### 4.1 構造化の安定性と再現性

LLM を用いた文書構造化と比較して, 決定論的構造化では, 同一入力文書に対するセクション分割結果の揺らぎが解消された. この結果, ルールチェックの前提となる検査対象セクションの特定が安定し, 同一文書に対して常に同一の検証結果が得られる再現性が担保された. 本知見は, 検証プロセスにおける監査可能性および説明責任の観点から重要である.

### 4.2 人間との役割再配分による有用性

医療プロモーション資材のレビュー作業は, 実務上, 複数の判断レイヤーを同時に含んでいる(2.3 節). 同一資材に対する 3 名のレビュワーの指摘結果には, ばらつきが認められた(表 1). これは, 上記の異なる判断レイヤーが「ルールチェック」という 1 つの作業として包括して実施されていることに起因する構造的特性を反映していると考えられる. 本システムにより, (1) 明文化された規制ルールへの適合性確認を決定論的ロジックに委譲することで, 人間のレビュワーは (2) 表現の分かりやすさや強調・誹謗中傷・曖昧な表現の是正, および (3) 医薬品の特性や個の文脈理解を前提とした内容評価といった, 人間の判断が本質的に必要な領域に認知資源を集中できる[13]. この役割再配分は, 人間中心の検証プロセスを支援する実務的有用性の 1 つである.

### 4.3 業務工数の削減効果

従来, 医療プロモーション資材のレビューには, おおよそ 5 営業日を要し, 実作業時間として 10-25 時間程度を要していた. 本システムを用いることで, 決定論的構造化およびルールベース検証が自動化され, 一次チェックは 10-30 分以内に完了する. これは, 実作業時間ベースで 約 95–99% の削減に相当する. 従来 5 営業日かかっていた 一次チェック工程を当日中に完了可能とすることで, 後続の人手による高度判断工程への早期着手や早期・高頻度のチェック実施が可能となる. 本結果は, 検証作業のリードタイム短縮に寄与するだけでなく, 人間のレビュワーが高度な判断を要する作業に時間を配分できることを意味しており, 実務運用上の有用性が高い. なお, 同時投稿の別報 [14] にて詳述するが, 本システムは実運用データにおいて 85.4% の再現率を達成しており, 重要項目の見落とし防止が実用レベルで可能であることを確認している.

### 4.4 レビュワー(人的作業)間の指摘のばらつき

同一の資材(総合製品情報概要)に対して, 経験年数の異なる 3 名のレビュワーが独立にレビューを実施した結果, 抽出された全 81 件の指摘のうち, 3 名全員が一致した項目は全体の 23.5%にとどまり, 顕著なばらつきが認められた(表 1). 項目別の内訳を見ると, 製品の特性や臨床成績の設計等, 資材固有の文脈理解を要する指摘(n=32)においては, 3 名全員が一致した事例は皆無(0.0%)であり, 90.6%(29 件)が 3 人のいずれか 1 名のみによる指摘であった. これは, 高度な内容判断が個人の解釈や経験に強く依存することを示唆している. 一方, 明文化された規制ルールへの適合性に関する指摘(n=27)では, 70.4%(19 件)が 3 名間で一致しており, 解釈の余地が比較的に少ない確認作業においては相対的に高い再現性が確認された.

| レビュワー間の指摘一致マトリクス, n(%) | | | | |
|---|---|---|---|---|
| 検証カテゴリー | 3 人一致 | 2 人一致 | 1 人のみ | 合計 |
| (1)明文化された規制ルールへの適合性 | 19 (70.4%) | 4 (14.8%) | 4 (14.8%) | 27 |
| (2)表現のわかりやすさや強調・誹謗中傷・曖昧な表現の是正 | 0 (0.0%) | 1 (16.7%) | 5 (83.3%) | 6 |
| (3)医薬品の特性や固の文脈理解を前提とした内容評価 | 0 (0.0%) | 3 (9.4%) | 29 (90.6%) | 32 |
| (4)誤記・不統一等に関する品質管理(QC)・校正 | 0 (0.0%) | 0 (0.0%) | 16 (100.0%) | 16 |
| 合計 | 19 (23.5%) | 8 (9.9%) | 54 (66.7%) | 81 |

| レビュワー間の指摘一致度(件数) | | |
|---|---|---|
| | B と一致 | C と一致 |
| A の指摘数 | 25 | 20 |
| Bの指摘数 | | 23 |

**表 1:レビュワーによる指摘の一致状況**

## 5 考察

### 5.1 LLM による文書構造化の限界

本研究は, 文書構造化と検証を決定論的処理のみで実現し, LLM に依存しない設計が再現性と監査可能性の担保に有効であることを示した. 特に, セクション分割精度がルールチェック精度に直結する点は重要な知見である. セクション分割に非決定論的手法を導入した場合, その精度のゆらぎが検証結果全体に波及するため, 本研究では決定論的構造化を選択した.

### 5.2 過検出(False Positive)の取り扱い

本システムの検証ロジックは決定論的ルールのみで構成されるため, LLM ハルシネーションに起因する過検出(FP)は原理的に発生しない. 観測された FP はルール定義の粒度差やコー

ドの不具合など，原因が特定可能かつシステム修正可能なものであった[14].

### 5.3 人間との役割分離

明文化された規制ルールへの適合性確認は，人手によるチェックにおいて再現率が 70% 超と他の検証カテゴリーとの結果と比較して高く，解釈の余地が少ない検証カテゴリーであると考えられる．一方で，それでもなお約 3 割の不一致や見落としが残存しており（表 1），人間による確認のみでは一定の限界が存在することが示唆された．この点において，本システムが決定論的ロジックにより明文化されたルール検証を担うことで，人的確認では取りこぼされやすい部分を補完し，検証プロセス全体の再現性と網羅性を高め得ることが示されたと考えられる．

### 5.4 組み合わせ爆発と運用課題の解決アプローチ

決定論的アプローチにより監査可能性は完全に担保される一方で，ルールの拡張に伴い検査空間が指数関数的に増大する「組み合わせ爆発（Combinatorial Explosion）」が実運用上の課題となる．人手による保守が限界を迎えるこの問題に対し，我々は運用面での解決策として，「進化的メンテナンス（Evolutionary Maintenance）」アーキテクチャの導入を進めている．

具体的には，システムが自身の実行ログからエラー原因（ノイズ，設定ミス，バグ）を統計的に診断し，修正パッチを自律的に生成・検証する仕組みである．一つの製品での修正を他製品へ自律的に波及させ，保守コストの線形増加を抑制する工夫を取り入れている． この自律的な保守運用の詳細なメカニズムおよび実証評価については，同時投稿の別報 [14] にて報告する．

### 5.5 リミテーションと今後の展望

本研究では，構造分離型アーキテクチャの実務的有用性として，再現性の担保，人間の役割再配分，および業務工数削減の観点から評価を行った．一方で，情報抽出・検証タスクにおける一般的な性能指標である True Positive（TP），False Negative（FN）等の定量的指標については，本稿の範囲外とした．これは，決定論的ルールの拡張に伴う「組み合わせ爆発」という課題に対し，現時点の評価のみではシステムの持続可能性を十分に議論できないためである．現在，この課題に対して「進化的メンテナンス」を導入し，運用を通じて動的に精度を維持・向上させる新たな評価枠組みの実証を進めている．そのため，定量的性能評価については，この自律的な運用解決策（Autonomous Maintenance）の進展と併せて，同時投稿の別報 [14]にて，詳細に報告する予定である．

さらに，レビュワー間のばらつき分析は，実務に関与する 3 名を対象とした探索的検討であり，経験年数やバックグラウンドの差異による個人差・バイアスを含む可能性がある．本結果は統計的一般化を目的とするものではなく，実務における判断構造の一端を示す知見として位置づけられる．

## 6 まとめ

本論文では，医療プロモーション資材における構造分離型アーキテクチャの実践的有用性を示した．文書検証を決定論的手法で実装したことで再現性・監査可能性を担保しつつ，人間の確認精度を高める設計が有効であることが示唆された．本研究は，生成 AI 時代の医療プロモーション資材を検証・実装する際の実践的設計指針を提供するものである．

## 参考文献

[01] Hamed, A. A., et al.: Challenges and Opportunities in Verifying AI-Generated Scientific Content, iScience, Vol. 27, No. 2, Article 108782 (2024).

[02] Kim, Y., et al.: Medical Hallucinations in Foundation Models and Their Impact on Healthcare, arXiv:2503.05777 (2025).

[03] Chung, P., et al.: VeriFact: Verifying Facts in LLM-Generated Clinical Text with Electronic Health Records, arXiv:2501.16672 (2025).

[04] 小泉直子ほか. 医療用医薬品プロモーションツール制作における AI 活用による価値創造とエンパワーメント. 人工知能学会全国大会（第 39 回）, 2025.
（doi.:10.11517/pjsai.JSAI2025.0_1M3OS47a02）

[05] OpenAI. Function Calling and Structured Outputs in LLMs, 2023.

[06] Ji, Z., et al. Survey of Hallucination in Natural Language Generation. ACL, 2023.

[07] 日本製薬工業協会：医療用医薬品製品情報概要審査会 医療用医薬品製品情報概要等に関する作成要領（2023 年 10 月改定）.

[08] Shen, Z., et al.: Layout Parser: A Unified Toolkit for Deep Learning Based Document Image Analysis, Proc. of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 3753–3762 (2021).

[09] Pfitzmann, B., et al.:DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis, Proc. of the 28th International Conference on Computational Linguistics (COLING), pp. 1098–1108 (2022).

[10] The Drools Project: Drools: A Business Rule Management System, Red Hat Inc., https://www.drools.org/（**参照日**：2026 **年** 2 **月**）

[11] Sarker, A., Gonzalez-Hernandez, G.et al.: An overview of natural language processing in pharmacovigilance, Drug Safety, Vol. 41, No. 8, pp. 775–789 (2018).

[12] Luo, Y., et al.: Matching Patients to Clinical Trials Using Natural Language Processing, Proc. of AMIA Annual Symposium, pp. 1200–1209 (2017).

[13] Amershi, S., et al. Guidelines for Human-AI Interaction. CHI, 2019.

[14]八角潤一ほか. 進化的メンテナンス：医療用医薬品文書検査ルールのデータ駆動型自動保守フレームワーク. 人工知能学会全国大会（第 40 回）,2026.（同時投稿）