特集《調査》

特許審査における AI の活用

特許庁審査第一部調整課審査企画室 企画調査官　**五十嵐　康弘**

要　約

特許庁では、2017 年度に人工知能技術の活用に向けたアクション・プランを公表して以来、様々な実証事業を行ってきた。本稿では、時代の流れに沿って特許審査における AI の活用を機械学習の活用、エンコーダモデルの活用、生成 AI の活用に分けて述べる。

その中でも、急激に進化している生成 AI については、2024 年度に行った調査事業で特許文献の要約、ドシエ情報（パテントファミリーの出願経過書類等）の要約、表の構造化といった特許審査の実務的なタスクについて調査を行い、実務での活用も期待される結果が得られた。これらについて述べるとともに、調査期間中の生成 AI の進展も考慮して、今後の展望についても言及する。

目次

1．はじめに

特許庁では、2016 年度に行った特許庁業務全体についての AI 技術の適用可能性の検討を受けて、2017 年度に

人工知能技術の活用に向けたアクション・プランを公表し、様々な実証事業を行ってきた。

本稿では、特許審査における AI の活用について述べ、目まぐるしく進歩する生成 AI を念頭に、今後の展望についても触れる。

また、AI に関する専門用語が使われる部分もあるが、生成 AI サービス等で詳細な説明を参照されたい。

本稿は、これまで実施した調査事業の結果を踏まえ、筆者の見解に基づいて記載したものであり、特許庁としての意見・見解を表明するものではない。

２．特許審査実務における AI 技術の適用状況

２．１　特許審査フローの概観

特許審査実務における AI 技術の適用状況について述べるにあたり、前提として特許審査の流れを示す。図 1 に示す通り、特許審査は分類付与、本願発明の理解、先行技術調査、特許性の判断というタスクを含んでいる。

図 1 は模式化されたものであって、実際の審査は一方向に進むものではない。例えば、スクリーニングの結果によって先行技術文献検索に戻ったり、特許性の判断で十分な心証を形成できず先行技術調査に戻ったりするなど、一連のプロセスは、出願時における技術水準の理解、先行技術文献の理解等を深めながら進めていく。

図 1　特許審査のフロー

２．２　AI 技術の進展の概観

AI 技術の適用状況について述べるにあたって、AI 技術の進展について 2010 年ごろからの技術の流れを簡単の整理しておく。

特に、特許審査においては、AI 技術の中でも自然言語を扱うモデルと関わりが深いため、以下、AI 技術については自然言語を中心とした説明になる。

①　機械学習期（～2012 年）

この時期は、サポートベクターマシン（SVM）、決定木など、あらかじめ特徴量を設定した上で、大量のデータによる学習を行い、未知のデータに対する推論を行う技術が実用化された。

②　ディープニューラルネットワーク期（2012～2016 年ごろ）

画像認識分野で圧倒的な性能を見せたことでディープニューラルネットワーク（DNN）を用いた深層学習が注目を集めた。自然言語分野では、リカレントニューラルネットワーク（RNN）を用いた機械翻訳が実用化された。

③　大規模言語モデルと応用拡大期（2017～2021 年ごろ）

DNN を用いたモデルの中でも、Transformer アーキテクチャを基盤とした BERT や GPT などの大規模言語モデルが注目を集めた。特に、BERT は転移学習が可能なモデルであるため、事前学習済みのモデルを基に様々な分野での応用が進められた。

また、2020 年に発表されたスケール則により、モデルの大規模化が急速に進んだ時期でもあり、2025 年現在から振り返ると、BERT 等のモデルを大規模と呼ぶかは様々な立場がある。本稿では、以降、出力結果が符号化された数値列となる BERT 等のモデルをエンコーダモデル、出力結果が自然言語となる GPT 等のモデルを、生成 AI と呼ぶ。

④　生成 AI による普及期（2022～2025 年現在）

ChatGPT を始めとする会話型の UI を持つ生成 AI が一般ユーザにも普及し、自然言語、画像、音声、動画などの入出力が可能になった。

また、学習データに含まれるパターンや特徴に基づく推論で回答を生成するだけでなく、明示的に思考するプロ

セスを組み込み、論理的な推論を経て回答を生成するモデルや、回答を生成する際に検索等の手段により外部の情報源を参照するモデル、外部ツールを操作するエージェント的なモデルなど、単に推論で回答を生成するだけにとどまらない進化を見せている。

前述のアクション・プランを公表した時期は 2017 年であり、機械学習・DNN を用いた AI 技術が様々な分野で応用された時期であった。

その後、特許庁では、上述の新たな技術の登場に対応しながら、様々な実証を行ってきたので、それらについて紹介していく。

２．３　機械学習の活用

図 1 の特許審査のフローでいう「分類付与」や「先行技術調査」といったタスクに対して、分類付与支援や、先行技術調査における検索式作成支援、特許文献のランキング表示などが実務に導入されている。

これらの実証の詳細については、参考文献[(1)-(3)]にまとめられているが、概要としては、分類付与支援では SVM などの技術が、特許文献のランキング表示では LightGBM（決定木の一種）などが用いられており、各実証事業ではモデルの選定、学習データの作成、精度の検証などが行われた。

２．４　エンコーダモデルの活用

2.3. で述べた取り組みは、いずれもモデル・特徴量を設定して特定のタスクに最適化するよう学習を行った上で、学習済みのモデルで推定を行う AI 技術を活用した取組であった。

一方、2018 年に発表された BERT は、転移学習が可能なモデルであり、汎用的知識を学習する事前学習に加えて、比較的少ないデータによるファインチューニングで多様なタスクに適用できることから、高い注目を集めた。

日本語においても、Wikipedia 等から大量のデータを収集して事前学習した BERT が公開されており、自然言語で記述された出願書類や特許公報等を扱う特許の実体審査での活用が期待された。

特許公報等の文章は自然言語であるものの、様々な分野の技術用語や抽象化された用語が多用されるため、特許審査実務において転移学習可能なモデルを活用するには特許公報等で事前学習を行うことが有用であると考えられた。

そこで、特許庁では、2022～2023 年度にかけて、BERT 及び BERT から派生した各種モデル（ALBERT[(4)]、DeBERTa[(5)]、Longformer[(6)]）に対して、特許文献の事前学習を行うことで特許事前学習モデルを構築し、これらに対して、実践的なタスクを設定して精度評価を行う調査を行った（以下、区別のため、特許文献の事前学習を行った BERT を「特許 BERT」、Wikipedia 等から事前学習を行った BERT を「非特許 BERT」という。）。

これらのモデルは、エンコーダモデルのため、自然文の入力に対して、複数の数値を出力する。複数の数値を高次元のベクトルととらえた場合、多数の文献群を高次元空間に分布する点群ととらえれば、空間を適切に区切ることで分類付与が可能であり、2 つのベクトルの内積を算出することで文章間の類似度の算出が可能である。

以下、これら二つの用途を意図したタスクについて述べる。

（1）　機械分類付与タスク

分類付与とは、先行技術調査において目的の文献を効率的に見つけるため、特許文献等に特許分類（FI、F ターム）を付与する業務である。本調査では、12 の技術分野における 240 個の FI に対して付与の要否を推定するタスクを、機械分類付与タスクとして設定した。

機械分類付与タスクにおいては、特許事前学習モデルに対して Transformer 層以降のパラメータを更新してファインチューニングを行い、非特許 BERT をベースラインとして評価を行った。

その結果、DeBERTa が Precision、Recall、F 値のすべてにおいて最高精度となり、ベースラインである非特許 BERT に対して有意な精度向上が見られた。

この結果を受けて、特許庁で外国文献に対して行っている機械分類付与のモデルとして、従前から活用していた

SVM、LightGBM 等に替わる手法として、順次、DeBERTa の導入を進めている。

（2） 類似文章ランキングタスク

先行技術調査では、目的の文献を効率的に見つけるため、文献群から審査対象の発明と類似したものから目視でスクリーニングを行っている。そこで、先行技術調査の高度化・効率化に向けたアプローチとして、審査対象の発明を複数の要素に分割したテキストをクエリとして、類似候補文章群を、類似性の高い順にランキングして並べるタスクを、類似文章ランキングタスクとして設定した。

特許事前学習モデルに対して、TripletLoss（比較対象となる文章と Anchor として、同じラベルが付与されている文章を Positive、異なるラベルが付与されている文章を Negative とした 3 つの文章組に基づいて学習を行う手法）、ArcFace（あるラベルが付与されるクラスに対する特徴ベクトルを取得し、クラス内分散を小さくし、異なるクラスとの分散が大きくなるように学習を行う手法）の 2 手法を用いてファインチューニングを行った。

精度評価においては、Recall@k（クエリに対して出力されたランキング結果上位 k 件に含まれる正解数が、総正解数のうちどの程度の割合含まれているかを示す指標）を用いた。

その結果、Recall@10、Recall@100、Recall@1,000 において、TripletLoss によってチューニングを行った DeBERTa が最高精度となった。

特許事前学習モデルは、従前から検索結果のランキングに用いている BM25（単語の出現頻度や文章の長さに基づいて文章の類似度を算出するアルゴリズム）と比べて、語順を考慮した文章の扱いや、同義語の扱いという点で優れており、活用が見込まれる。

一方、DeBERTa を含むエンコーダモデルは、先行技術文献の長さに比して入力クエリ長の制限が厳しく（おおよそ 1,000 文字程度が上限となる。）、実務利用へは課題も残る。

３．生成 AI の活用（2024 年度調査事業）

３．１　調査の概要

生成 AI は、大量のデータを学習し、そのパターンや特徴をもとに、文章、画像、音声、動画などを新たに生み出す技術である。

生成 AI の活用を検討するにあたって、従来の AI にはなかったいくつかの要素が問題となった。以下に、代表的なものを挙げる。

①　精度評価の困難さ

生成 AI で文章を生み出す場合、一文字一文字の出力プロセスは入力に対応する一定のベクトルが出力されるものの、それを確率分布として扱って連続的に文章を出力した場合には、同じ入力に対して同じ出力が得られるとは限らない。また、エンコーダモデルの活用では、タスクの設定に対して評価尺度を定めやすく、精度評価を行うことが比較的容易であったのに対して、生成 AI では、出力の評価尺度の設定が難しい。その違いは、記号選択式のテストを採点するのと、自由記述式のテストを採点するのとで、採点者の負担が異なるのと似ている。

生成 AI についても、多くのベンチマーク指標が知られているものの、実際のタスクにおける精度は、業務特有の知見が必要となるため、人手評価が必要となり、評価の負担がボトルネックとなり得る。回避するためには、機械的な評価手法で、人手評価と近い評価を得られないかといった工夫も必要である。

②　モデル進化による課題の陳腐化

生成 AI については、日々の進歩が類を見ないほど早く、課題設定時点におけるフラグシップモデルを用いて調査を行ったとしても、調査が終わるころには、モデル自体の進化によって課題が解決されてしまい、調査で見出された手法が陳腐化する可能性がある。

モデルの進化を予想することは困難だが、課題設定において業務の性質を考慮する必要がある。

③　ハルシネーションに関わる問題

生成 AI 特有の問題として、ハルシネーションの発生がある。生成 AI は、プロンプトとして入力する情報に基

づいて、情報を変換する、情報を増やす、情報を減らすといったタスクを行う。

特に、情報を増やすタスクにおいては、プロンプトに入力した情報を基に、事前学習したパターンや特徴による出力を付加するため、ハルシネーションが発生しやすい。

逆に、情報を変換するタスク（例えば翻訳、語調の変換など）、情報を減らすタスク（例えば要約など）では、プロンプトに入力した情報を加工するために、事前学習したパターンや特徴から、文法的な知識や、重要と思われる情報を選別する能力を発揮するため、ハルシネーション自体は防ぎやすいと思われる。ただし、出力結果だけを見ても、入力情報から、どのような情報が失われたかを判別できないという別のリスクが不可避的に発生する。

（1） タスクの設定

これらの①～③に挙げた問題を考慮して、特許審査のタスクのどの部分で活用するかという点が、検討の重要なポイントとなった。

そのような状況の中、大別して以下の 3 つのタスクを扱うこととした。

① 特許文献の要約タスク

② 他庁ドシエ情報の要約タスク

③ 表の構造化タスク

いずれも、課題設定時において技術的な実現可能性が期待でき、生成 AI のモデルが進化したとしても、結果の有用性が損なわれにくいと考えられるタスクであり、ハルシネーションの問題にも対処しやすいものである。

また、精度評価については、人手評価と共に機械評価を行って、人手評価に近い評価を得られる、すなわち人手評価に代替し得る機械評価手法も検討した。

（2） モデルの選定

一般的に API 経由で利用するモデル（以下、クラウドモデル）は、大規模かつ高性能なフラグシップモデルであり、オンプレミスでは構築できないものの、現状の技術的な限界を把握する上で重要となる。

一方、オンプレミスで構築できるモデル（以下、オンプレミスモデル）は、現実的に保有できる GPU の性能からモデルサイズに限界があり、フラグシップモデルよりも性能が限られる反面、情報セキュリティの観点、利用するモデルの安定性といった点で長けている。

また、一定程度の時間的なラグはあるものの、オンプレミスモデルもフラグシップモデルを追うように性能の向上を続けており、現状でオンプレミスモデルが十分な結果を出せないとしても、遠からずフラグシップモデルでの検討結果を活かせるようになる可能性も高い。

そこで、クラウドモデルで技術的な可能性を探りつつ、いずれ同様の性能が得られることを期待してオンプレミスモデルも検討対象に含めて、各タスクを行うこととした。

多数存在する生成 AI のうち、様々な比較検討の結果、クラウドモデルとして、GPT-4o（以下、「GPT」という。）、Claude 3.5 sonnet（以下、「Claude」という。）、gemini-1.5-pro（以下、「Gemini」という。）の 3 モデルを、オンプレミスモデルとして、Llama-3.1-70B-Japanese-Instruct-2407（以下、「Llama」という。）、tsuzumi の 2 モデルを中心に、各タスクの特性に合わせたモデルを用いた。

3．2 特許文献の要約タスク

要約は、生成 AI の用途として、よく知られたタスクである。特許審査では、審査官は大量の先行技術文献を目視でスクリーニングしており、多大な時間を要している。生成 AI によって、審査官目線で有用な要約を作成できるのであれば、目視調査が効率化され有用である。

しかし、上述したように、要約文のみからは、入力情報に対してどの情報が失われたかを判別できないため、要約の有用性の評価は重要である。具体的には、先行技術文献の中の見るべき情報が凝縮された要約となっているか調査が必要となる。

（1） 調査手法

本調査では審査官が実際に審査した出願（様々な技術分野から 38 件を選定）について、実際に審査で引用した先行技術文献 1 つと、それ以外の先行技術文献 13 件を、それぞれ前述の 5 モデルを用いて 400 字程度に要約した。クラウドモデルについては各々プロンプトチューニングを、オンプレミスモデルについてはファインチューニングを行った上で要約を作成した。

これらの要約文を用いて、要約文の読みやすさを評価するとともに、要約文のみから審査官が引用した先行技術文献（以下、「正解文献」という。）を特定できるかというブラインドテストを行った。

評価は人手評価と共に、機械評価（G-Eval, FineSurE のような LLM による包括的な評価）も行い相関関係を調査した。

（2） 結果

要約文の読みやすさの観点では、人手評価では総じてクラウドモデルの評価が高く、オンプレミスモデルは多少の差はあるものの、いずれも十分と思われる評価が得られた。

一方、正解文献を特定できるかという観点（以下、「特定率」という。）では、Claude の 68% についで、オンプレミスモデルの Llama が 63% と続いた。

14 文献の中から 68% の精度で適切な文献を特定できるという結果は、高い確率で必要な情報を 400 字の中に収めているものの、正確性が求められる審査においてこれのみをもってスクリーニングを行っていくことは難しいと考えられる。

また、人手評価と機械評価とを比較してみると、人手評価で指定された文献や、正解文献を、機械評価でも比較的上位とする傾向がみられたが（約 70% で上位 5 位以内と判定）、機械評価で人手評価を代替することは難しいという結果になった。

（3） 考察

前述のとおり、要約タスクでは、情報を減らす加工を行うため、要約の過程でどのような情報が失われたかを判別できないという点が問題となる。

しかし、要約前の文献が 1 万字以上の情報量であった点を考慮すれば、400 字の要約で 68% の特定率があることは、業務効率化の可能性を秘めていることがわかる。

実用化に向けては、要約文とともに要約前の文章を表示するなど一次情報へのアクセス性を損なわないような UI の導入、業務フローの中で人手評価のフィードバックを得て、精度改善につなげられる仕組みの導入など、実装の工夫も必要になると考えられる。

また、今回の検証では、400 字という制約にしたが、要約文の文字数を増やすことで特定率の向上と、業務効率化の双方のバランスをとれる可能性がある。

また、要約文の別の用途として、エンコーダモデルの入力として要約文を用いることも可能と考えられる。BERT や DeBERTa といったモデルは入力長の制約から、先行技術文献を複数に切り分けて処理するなどの工夫が必要だったが、複数に切り分ける方法では、切断された箇所の文脈情報が失われる恐れがあったため、生成 AI を用いた要約文がエンコーダモデルの有用性を向上させるなど、新たな可能性が期待される。

3．3 他庁ドシエ情報の要約タスク

特許審査では、図 1 のフローには現れないところにも、審査官が時間を要しているタスクがある。

パリ条約に基づく優先権主張等によって、同内容の出願が他国にも行われる、いわゆるパテントファミリーが存在する場合には、審査官は、他庁の審査経過を参照して他庁の出願経過書類（以下、「ドシエ情報」という。）を確認する必要がある。

ドシエ情報のファイル形式は pdf ファイルであり、テキストデータが埋め込まれていない外国語のドキュメント

である。また、他庁の書類の様式は様々であり、言語も国によって異なるという特徴もある。

現状では、一つのパテントファミリーに対して、多数のドシエ情報から、参照する価値がある書類（特に、先行技術文献の内容等を含む書類）を一つ一つ確認しており、時間を要している。

そこで、本調査では、ドシエ情報の各書類を個別に翻訳、要約した上で、他庁での審査経過の概要を時系列にまとめる要約タスクを設定し、調査を行った。

（1） 調査手法

本調査では、日本特許庁とヨーロッパ特許庁の両方に出願されており、ヨーロッパ特許庁で審査が完了している100 件を対象として、ヨーロッパ特許庁のドシエ情報を収集し、特に審査官が注目するドシエ情報（サーチレポートや、日本の拒絶理由通知書、意見書、手続補正書に対応する書類等）について以下の処理を行った。

主な手順として、PDF ファイルからのテキストの抽出、日本語翻訳、翻訳されたテキストの書類単位での要約、書類単位での要約を時系列に並べて要約した出願単位の要約を行った。要約は、前述の 5 モデルを用い、主に出力結果の形式を学習させることを目的にプロンプトチューニングを行った。

これらの要約文に対して、読みやすさの評価（1〜5 の五段階、高いほど優れている。）、正確性の評価（1〜5 の五段階）、引用文献の網羅性の評価（要約で抽出できた引用文献数の、全体の引用文献数に対する割合）を評価した。

評価は人手評価と共に、機械評価（LLM as a judge, G-Eval, FineSurE）も行い相関関係を調査した。

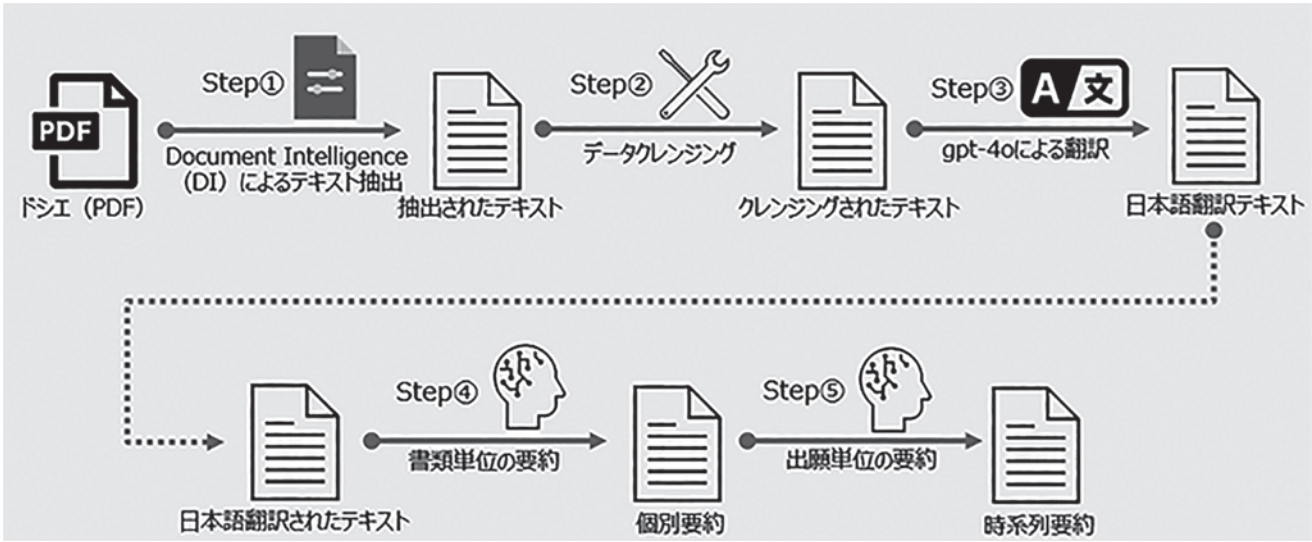

図 2　ドシエ情報の要約の流れ

（2） 結果

人手評価の結果、読みやすさの評価、正確性の評価では、クラウドモデルがいずれも 3〜4 の平均値となったのに対して、オンプレミスモデルでは、2 を下回る結果となり顕著な性能差が見られた。また、引用文献の網羅性についても、クラウドモデルが 8 割程度の引用文献を抽出して出願要約を作成したのに対して、オンプレミスモデルでは 2 割を下回る程度しか引用文献を抽出できない結果となった。

また、人手評価と機械評価とを比較してみると、両者には弱い相関がみられるものの機械評価で人手評価を代替することは難しいという結果になった。

（3） 考察

この調査の人手評価では、上記結果に加えて、集約したい観点を明確にした要約タスクにおいてクラウドモデル

はプロンプトに含まれる情報の文脈を誤解したり、事前学習した内容から誤った内容を付加したりするようなハルシネーションを起こす恐れはほぼなさそうに見える、という心証が得られた。

情報を削るという点においては。クラウドモデルは上述のとおり 8 割程度の引用文献を抽出している。特に、審査過程において重要と思われる引用文献については、ほぼ網羅されており、他庁の審査経過を参照する業務において効率を向上させられるという感触を得た。

もちろん、要約とともに、一次情報へのアクセス性を担保することは重要であり、3.2.（3）と同様、業務フローの中で人手評価のフィードバックを得られる仕組みの導入など、実装の工夫も重要となる。

また、本検証では、パテントファミリーが存在する割合が多く、実務上審査官が頻繁に参照するため、ヨーロッパ特許庁のドシエ情報を対象にして行った。今後、海外庁への出願行動の変化があったり、特定分野におけるパテントファミリーの存在傾向が変化したりした場合、ヨーロッパ特許庁以外のドシエ情報についても対応が必要な可能性がある。生成 AI においては、言語の壁は比較的超えやすいことから、さらなる活用が期待できる。

また、画像情報をテキスト情報と同時に扱えるマルチモーダルモデルや、複数の画像情報を総合して考慮できるモデルが登場しており、書誌情報と複数の書類を直接扱うなど、その時々の技術水準に合わせてより良い手段を採用することも可能になると思われる。

３．４　表の構造化タスク

本タスクは、特許文献に含まれる表に関して設定されたものである。課題は大きく分けて 2 点あり、表自体がイメージデータとして存在し、テキスト情報がないため、検索やスクリーニングにおいて扱いが難しいという点、表の内容を理解するために特許文献の本文も併せて参照する必要があるため、表の情報を単純にテキスト化できたとしても有用な情報にならない可能性があるという点である。

本タスクでは、イメージデータとして存在する表を、テキストとして扱えるようにし、かつ、本文も考慮した上で有用な情報を生成できるかを検証した。

（1）　調査手法

本調査は、画像中の表を json 形式に変換して構造化するタスクと、構造化されたデータを明細書とともにインプットして、単位・凡例・注釈等の情報を加味したキャプショニングを行うタスクの 2 つのタスクに分けて調査を行った。

調査に用いるモデルは、クラウドモデルとして、3.1.（2）にあげた 3 モデルを、また、オンプレミスモデルは、視覚読解に対応している必要があることから、tsuzumi、Qwen2-VL-72B-Instruct の 2 モデルを用いた。

これらのモデルに対して、クラウドモデルについてはプロンプトエンジニアリングを、オンプレミスモデルについてはファインチューニングを行った上で、10,080 件のデータに対して json 形式への変換を行い評価を行った。

構造化の評価は、人手評価は 499 件について五段階評価を行い、機械評価はルールベースでの評価と LLM による包括的な評価を全件について行った。なお、LLM による包括的な評価は、GPT-4o で人手評価と同様の五段階評価とともに判断根拠を出力させた。

また、キャプショニングの評価は、100 件について、人手でキャプショニングの正確性（表の内容を正しく説明できているか）、明細書本文を考慮したキャプショニング（単位、凡例など、明細書本文を考慮した説明文になっているか）の 2 観点について、いずれも五段階評価を行った。

（2）　結果

json 形式への変換タスクでは、人手で作成した正解との比較の結果、ルールベースでの機械評価よりも、LLM による包括的な評価の方が人手評価に近い傾向を示した。そこで、全件（10,080 件）の評価は、LLM による包括的な評価を用いて、各モデルの評価を行った。

その結果、Gemini が最も高く 3.09（1〜5 の五段階評価の平均値、以下同様）、GPT が 2.93、Claude が 2.49 と続

いた。Claude やオンプレミスモデルでは json 形式の出力を途中で終了してしまい、正常に出力できていないケースが多く見られ、評価が下がる傾向が見られた。

また、エラー分析を行ったところ、罫線が省略された表や、二重線、点線が用いられた罫線、多層ヘッダ、ページ分かれなど、様々な要因によってスコアが低くなることも明らかになった。

キャプショニングタスクについては、正確性の観点では Claude が最も高く 3.11、明細書本文を考慮したキャプショニングについては GPT が最も高く 3.21 という結果であった。

（3） 考察

上記結果に見られる通り、json 形式への変換タスクは、Gemini、GPT で正常に変換できるケースが多く、キャプショニングと併せて情報を変換することで、もともとイメージデータであった表から、明細書本文を考慮したキャプショニングに変換できることが確認できた。

しかしながら、表のイメージデータには、複雑な構造を持つものもあり、精度の向上には困難があることも明らかになった。

イメージデータとして存在する表をテキスト情報に変換して検索に用いようとする場合、検索データベースへの蓄積など、大規模な作業が生じる。生成 AI の進展によって更なる改善手段が見いだされた場合には、再蓄積など負担が生じる点、マルチモーダルモデルの進化が激しい点も考慮すると、現状で実務に用いるかは、生成 AI の進化を注視して判断する必要がある。

4．おわりに

最後に、生成 AI の活用の今後について展望を述べる。

本調査事業は、2024 年 9 月～2025 年 3 月の 7 か月間にわたり実施したが、その間にも、新たなモデルが多数登場し、回答精度が向上しただけではなく、2.2. で言及したように、論理的な推論を経て回答するモデル（いわゆる、リーズニングモデル）や、回答を生成する際に検索等の手段により外部の情報源を参照するモデル、外部ツールを操作するエージェント的なモデルなど、従来の生成 AI とは異質な方向への進化も見られた。

ただし、これらの進化について共通していえることは、進化速度は圧倒的ではあるものの、あくまで汎用的な用途を意図して進化を続けており、特許審査などの分野に特化した領域においては、実用化のための試行錯誤が必要であるという点である。

例えば、リーズニングモデルでは、洗練された論理的思考を用いて推論を進められるものの、特許審査で重視される網羅的な調査を目指した思考過程を採用するものではない。また、外部の情報源を参照するモデルの多くはインターネットで公開された情報の検索を用いており、技術文献の検索に最適化されてはいない。当然ながら、目的が異なれば、学習する思考過程や、参照すべき情報も異なるため、特許審査に適用するためには試行錯誤が必要であり、引き続き検討が必要になる。

一方、多くの目的に対して十分な性能となり、進化が緩やかになった面もある。例えば、事前学習のモデルの規模や学習に用いるデータ規模、コンテキスト長といった面においては、特許審査の目的において十分な性能が得られた状況である。

生成 AI の活用を検討していく上では、日々更新される最新の情報の収集のみならず、必要な性能が得られるようになり進化が緩やかになった面を十分に生かすという視点も重要になると思われる。

十分な性能が得られるようになった点を活かして特許審査への適用を検討しつつ、同時に進化の方向性、成長点を見定め、分野に特化した実用化検討の開始時期を適切に見極める必要がある。

（参考文献）

(1) 富永泰規、「外国特許文献への分類付与に関する機械学習活用可能性調査について」、Japio YEAR BOOK 2017、p.212-216（2017）
(2) 近藤裕之、「特許文献への分類付与と付与根拠の推定」、Japio YEAR BOOK 2018、p.16-22（2018）

(3) 後藤昌夫、「特許文献のランキングへの機械学習技術の適用」、Japio YEAR BOOK 2020、p.84-89（2020）
(4) Iz Beltagy, Matthew E Peters, and Arman Cohan, Longformer：The Long-Document Transformer, arXiv, https://arxiv.org/pdf/2004.05150（2020）
(5) Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, DeBERTa：Decoding-enhanced BERT with Disentangled Attention, arXiv, https://arxiv.org/abs/2006.03654（2020）
(6) Zhenzhong Lan et al, ALBERT：A Lite BERT for Self-supervised Learning of Language Representations, arXiv, https://arxiv.org/pdf/1909.11942（2020）

（原稿受領 2025.4.24）