一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
LOIS2025-77(2026-03)

時計描画検査の定量的評価に対するLLMの適用可能性評価

牧　昌宏[†]　吉田　圭佑[†]　平井　駿[†]　佐伯　幸郎[††]　中田　匠哉[†]
陳　思楠[†]　佐藤　厚[†††]　児玉　直樹[††††]　中村　匡秀[†,†††††]

† 神戸大学　〒657-8501 神戸市灘区六甲台町 1-1
†† 高知工科大学　〒782-8502 高知県香美市土佐山田町宮ノ口 185
††† 新潟県立吉田病院　〒959-0242 新潟県燕市吉田大保町 32-14
†††† 新潟医療福祉大学　〒950-3198 新潟県新潟市北区島見町 1398 番地
††††† 理化学研究所・革新知能統合研究センター　〒103-0027 東京都中央区日本橋 1-4-1
E-mail: †{maki,yoshikei,hirashun,tnakata}@es4.eedept.kobe-u.ac.jp, ††saiki.sachio@kochi-tech.ac.jp,
†††chensinan@gold.kobe-u.ac.jp, ††††masa-n@cmds.kobe-u.ac.jp

あらまし　近年，認知症早期発見に向けたスクリーニングの重要性が高まっている．本研究では，従来の画像処理では困難な文脈理解に優れたマルチモーダル LLM を用い，時計描画検査（CDT）の自動採点手法を評価した．実験では，GPT-4o 等のクラウドモデルに加え，プライバシー保護を想定したローカルモデルも選定し，人手採点との整合性を検証した．結果として，円や数字の配置といった大域的認識は人間と同等であったが，針の長さなどの幾何学的判定はプロンプト改善後も精度が低かった．現状の単一モデルでの完全代替は困難であり，実用化にはモデル統合や描画過程データの活用が必要であることが明らかとなった．
キーワード　認知症，神経心理学検査，描画検査，時計描画検査，LLM

Evaluating feasibility of LLM for quantitative assessment of clock drawing test

Masahiro MAKI[†], Keisuke YOSHIDA[†], Shun HIRAI[†], Sachio SAIKI[††], Takuya NAKATA[†], Sinan CHEN[†], Atsushi SATO[†††], Naoki KODAMA[††††], and Masahide NAKAMURA[†,†††††]

† Kobe University　Rokkodai-cho 1-1, Nada-ku, Kobe, Hyogo 657-8501 Japan
†† Kochi University of Technology　185 Miyanokuchi, Tosayamada, Kami City, Kochi 785-8502, JAPAN
††† Yoshida Hospital　32-14 Yoshidadaibo cho, Tsubame, 959-0242 Japan
†††† Niigata University of Health and Welfare　1398 Shimami cho, Kita-ku, Niigata, 950–3198 Japan
††††† Riken AIP　1-4-1 Nihon-bashi, Chuo-ku, Tokyo 103-0027 Japan
E-mail: †{maki,yoshikei,hirashun,tnakata}@es4.eedept.kobe-u.ac.jp, ††saiki.sachio@kochi-tech.ac.jp,
†††chensinan@gold.kobe-u.ac.jp, ††††masa-n@cmds.kobe-u.ac.jp

Abstract　In recent years, the importance of screening for early detection of dementia has grown significantly. This study evaluated automated scoring methods for the Clock Drawing Test (CDT) using multimodal large language models (LLMs), which excel at contextual understanding—a challenge for conventional image processing. The experiment selected cloud models like GPT-4o alongside locally deployed models designed for privacy protection, verifying their consistency with human scoring. Results showed that global recognition aspects, such as the placement of circles and numbers, matched human performance. However, geometric judgments, like the length of the hands, remained inaccurate even after prompt refinement. It became clear that complete replacement with a single model is currently difficult. Practical implementation requires model integration and the utilization of drawing process data.
Key words　Dementia, Neuropsychological Testing, Drawing Test, Clock Drawing Test, LLM

1. はじめに

日本は急速な超高齢社会にあり，医療需要の増加と医療従事者の不足が同時に進んでいる．この人手不足を補うため，電子的な手続きやオンライン運用を取り入れる医療 DX が各所で進んでいる．一方で，高齢者の増加に伴い認知症の患者も増えており，早期に気づき，症状が軽いうちに対応することが重要である．そのためには，短時間で実施でき，信頼性の高いスクリーニング検査が求められる．

描画検査である時計描画検査（CDT）[1] や立方体模写検査（CCT）[2] は，臨床で広く使われている代表的な方法であり [3]，従来は「描き終わった図」を専門家が目視で採点するのが一般的であった．しかし，認知機能の定期評価を広く続けるには，専門家による個別対応だけでなく，集団で効率よく実施できる体制が重要であり，そのための枠組みも提案されている．

我々はこれまで，機械学習によるブラックボックス的な自動採点に頼らず，書き順やストローク速度などの描画過程をわかりやすく可視化して医療従事者に提供する Web アプリ EVIDENT を提案・実装し [4], [5]，さらに集団実施に適した EVIDENT2.0 へと拡張した [6]．これらのシステムを用いた実証実験を通じて，デジタル描画データの収集 [7] と運用を行っている一方で，人手による評価プロセスには以下の 2 つの課題が存在していた．

P1: 採点者間における評価基準のばらつきの解消

P2: 採点およびフィードバック生成の即時化

特に CDT の採点においては，局所的な特徴量を定義する従来の画像処理よりも，数字や針の相対的な位置関係や全体的な整合性を俯瞰して解釈する能力が不可欠であり，高い文脈理解能力を持つ LLM の適用が有効であると考えられる．本研究では上記の課題をふまえ，様々な LLM を活用した即時かつ客観的な自動採点手法の構築と評価を目的とする．具体的には，「複数モデルによる基礎性能の検証と，その結果に基づくプロンプトエンジニアリングによる精度改善」という一連のアプローチを提案する．まず，クラウド型モデルに加え，医療現場での実運用におけるプライバシー保護を想定したオープンソースモデルを選定し，人手による採点結果との一致度を検証する．これにより，精度と実用性の両面から各モデルの特性差を明らかにする．続いて，その検証で得られた技術的課題に対し，LLM に対する指示を最適化することで，幾何学的な認識精度の向上や過剰検知の抑制が可能かを検証し，実用化に向けた解決策を提示する．これらのアプローチに基づき評価実験を行った結果，円や数字の配置といった空間認識においては人間と同等の高い再現性が確認された一方で，針の長さの比較などの幾何学的測定においては精度が著しく低下することが示された．また，モデルごとに推論速度や採点傾向（厳格さや頑強性）に明確な差異があり，現状の単一モデルのみでは人間の完全な代替は困難であることが明らかとなった．今後は，各モデルの特性を活かして複数のモデルを組み合わせるアンサンブルアプローチの導入や，静止画だけでなく描画過程データを統合した解析手法の構築を進めたいと考えている．これにより，医療従事者の手を介さず，迅速かつ高精度に認知機能低下を発見するスクリーニングシステムの実現を目指す．

2. 準　　備

2.1 高齢化社会と認知症

現在，日本は超高齢社会によって医療需要が高まり，そのために医療従事者の不足が問題視されている．厚生労働省の試算によると，2030 年には医師が約 3 万人不足すると予測されている [8]．このような医療現場の人手不足を補うために，デジタル化を用いた業務の効率化，すなわちデジタルトランスフォーメーション（DX）が推進されている．その例として電子カルテやオンライン問診表などが挙げられ，実際に使用する病院も徐々に増えてきている．また，現在は普及していないものの，オンライン診療の導入も期待されている．これにより訪問診療時の移動時間が短縮し，人手不足の解消や医療の地域格差の解決につながると考えられている．

さらに，高齢化社会の進展に伴い，認知症は日本における大きな社会問題となっている．内閣府発表の高齢社会白書によると，2040 年には認知症高齢者数が 584 万人を超えるとされ，これは 65 歳以上の高齢者の約 6 人に 1 人が認知症を患うとの推計である [9]．認知症は早期発見と早期治療が重要であり，そのためにさまざまな認知症検査が行われている．中でも認知症の前段階とされる軽度認知障害（Mild Cognitive Impairment, MCI）の段階での診断が重要であり，そのためには定期的な認知機能検査が必要である．

2.2 神経心理学的描画検査

神経心理学は，言語，記憶，認知，行為，前頭葉の機能など，脳の中枢神経系が果たす役割を解明し，その障害に伴う症状への対処を目的とする学問分野である [10]．この分野における一つの重要な診断手法が，描画を用いた検査法である．これらの検査では，被験者が指示に基づいて特定の絵や図形を紙に描くことを通じて，認知機能の状態を評価する．描画検査は，認知症や脳機能障害の評価に広く使用されており，以下のようなものが存在する．

2.2.1 時計描画検査

時計描画検査（Clock Drawing Test, CDT）は，被験者が認知症かどうか調べるスクリーニング検査として用いられる描画検査である．被験者の前に A4 の紙とペンを置き，用紙のサイズに見合った 10 時 10 分（又は 11 時 10 分）を指す丸時計を書くように口頭で指示する [11]．図 1 に時計描画検査の描画例を示す．図 1 の描画例では円い時計が描かれており，短針が 10 を，長針が 2 を指し，10 時 10 分を表している．

2.2.2 立方体模写検査

立方体模写検査（Cube Copying Test, CCT）は，時計描画検査と同様に認知症のスクリーニング検査として用いられる描画検査である．立方体模写検査では，図 2 に示される，紙に書かれた立方体透視図を被験者に対し視覚的に提示し，A4 の白紙に模写するよう口頭で指示をする．図 2 に立方体模写検査の描画例も示す．図 2 の描画例ではフリーハンドで描かれているため直線に歪みはあるが，立方体は正確に模写されている．

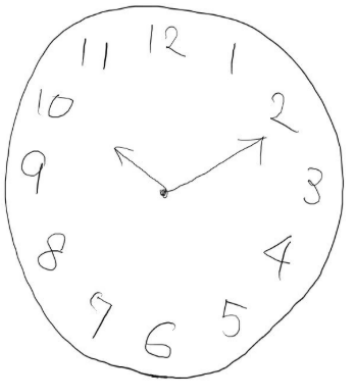

図 1　時計描画検査の描画例

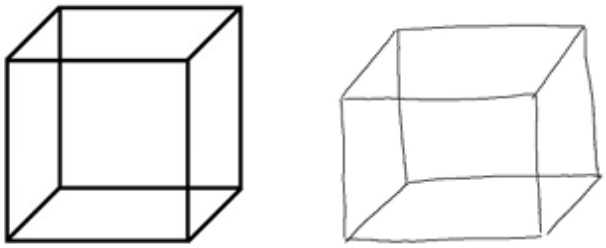

図 2　立方体模写検査の見本 (左) と描画例 (右)

またこれらの描画検査以外にも「犬も歩けば棒に当たる」といった慣用句を書いてもらう書字検査や数字や言葉の順に線を結んでいくトレイルメイキングテストなどが存在する．

2.3　時計描画検査における採点項目

本研究で扱う時計描画における採点項目は図 1 のようになっておりこれらの採点項目では主に時計の要素となる円，数字，針の部分に関する採点が行われておりこの採点より空間把握能力や認知能力を測定している．

表 1　時計描画における採点項目

No	採点要素
1	円が小さすぎて内容に支障はない？
2	すべての数字が円内にある？
3	分割や tic marks がない？
4	数字の間隔は正常？
5	数字は 1-12 だけ描かれている？
6	数字の順番は正しい？
7	針が 2 本だけ？
8	針が矢印または棒状？
9	10 時と 11 時の間の 10 時寄りの所に時針がある？
10	分針が時針より長い？

実際に採点項目 4 では図 3 のように 12，1，2 で数字の間隔が小さくなっているなど数字の間隔が偏って書かれているものが採点としては不適となり，採点項目 7 では時間として指定されている 10 時 10 分より 10 時方向の針を時針，10 分方向の針を分針と捉えそのうえで図 4 のように分針より時針の針が長い時や明確に長いといえない場合は採点として不適となっている．特にこの「針の長さの比較」は，幾何学的な厳密さが求められるため，人間による目視判断でも揺れが生じやすく，本研究における自動採点タスクにおいても最も攻略が困難な「最難関項目」になると位置づけている．

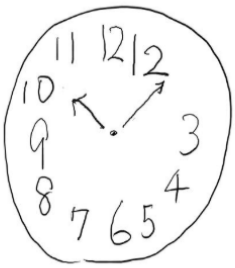

図 3　採点項目 4 において不適な例

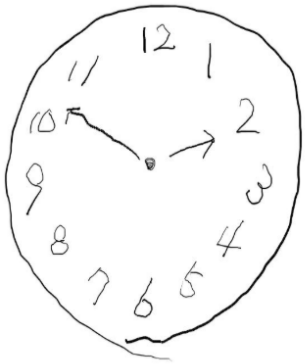

図 4　採点項目 7 において不適な例

2.4　先行研究：EVIDENT, EVIDENT 2.0

タブレットなどの ICT 端末を用いて描画検査を実施し，そのデータに機械学習を適用して回帰・分類を行う研究は，特に時計描画検査の領域で精力的に展開されてきた．一方で，機械学習型の自動採点には大規模データの確保が不可欠であり，さらに解釈可能性と精度の間にトレードオフがあることが指摘されている [12]．また，描画過程と診断結果の対応づけが困難なケースもあり，これらが実運用を阻む要因となっていた．

この課題認識のもと，我々の研究グループは，描画の順序やストローク速度といった過程情報を直観的に可視化し，医療従事者に提供する Web アプリケーション EVIDENT（Extraction and Visualization Interface of Drawing Execution in Neuropsychological Tests）を設計・実装した [4], [5]．EVIDENT により，描画特性に基づく新しい採点基準や診断プロトコルの検討が可能となる．また，EVIDENT はタブレット端末とインターネット環境があれば場所を選ばず検査を実施でき，記録された描画過程データを医療従事者が Web 上で随時確認できる．検査画面には手順ガイダンスも備えており，立会いなしでも検査を運用しやすく，オンライン診療等における診断業務の補完にも資する．さらに，認知症の早期発見には定期的な評価が重要である一方，各家庭での個別運用や医療機関での常時体制整備には負

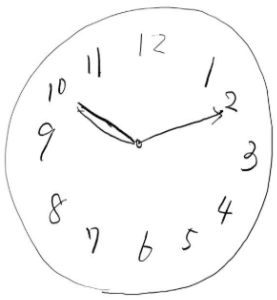

図 5　採点が人によって異なってしまうものの例

担が大きい．このため，定期評価は集団検査として実施するのが実務的であり，その目的に特化した枠組みも提案されている．

以上を踏まえ，我々は EVIDENT を集団実施に最適化した EVIDENT 2.0 へ拡張した [6]．本システムは，被験者が描画検査を行う EVIDENT EXAM，医療従事者が被験者情報や描画データの閲覧・分析を担う EVIDENT ADMIN，そして管理者が受検者・医療従事者の管理や検査準備を行う EVIDENT CONF の 3 モジュールから構成され，役割に応じた機能分離により集団運用の効率化を図っている．EVIDENT 2.0 では集団応用を行うために機能分担を行ったが，収集するデータに関しては変わっていない．そのため，以下ではまとめて EVIDENT と呼称する．

2.5　本研究で注目する課題

2.3 節で示した通り，時計描画検査の採点基準は「数字の配置」「針の指示時刻」「円の閉合」など多岐にわたり，その判定には専門的な知識と経験が求められる．また，2.4 節で述べた EVIDENT により描画検査データの収集基盤は整いつつあるものの，これらの複雑な項目を評価するプロセス自体は依然として課題が残る．本研究では，以下の 2 点を解決すべき課題として設定した．

P1:　採点者間における評価基準のばらつきの解消

現状の採点は評価者の目視に依存しているが，2.3 節で挙げたような細かな採点項目においては，基準に対する解釈が採点者によって異なる場合がある．実際に図 5 のようなものでは人によって数字の間隔が均一ととらえ人によっては不均一ととらえられてしまう．そのため，同一の描画データであっても採点者によって結果にばらつきが生じ，評価の客観性が損なわれる恐れがある．信頼性の高いスクリーニングを実現するためには，採点を行う主体や判断ロジックをシステム側で統一し，常に一定の基準で評価を行う必要がある．

P2:　採点およびフィードバック生成の即時化

専門家による人手での採点は，複数の採点項目を一つ一つ確認し点数化するために時間を要する．特に，単なる点数だけでなく，なぜその点数になったのかというフィードバック（評価根拠）を言語化して返却する場合，その作成コストはさらに増大する．検査結果を遅滞なく被験者や医療関係者に提供するためには，自動採点によって瞬時に評価とフィードバックを行い，タイムラグを解消する仕組みの実装が必要である．

3.　評 価 実 験

3.1　目的と概要

本研究の目的は，2.5 節で述べた課題を解決するため，EVIDENT を通じて収集された描画データに対し，LLM を用いた自動採点の有用性を評価することである．特に本研究では，LLM が人間の専門家と同様の基準で描画を評価し，その採点結果をどの程度正確に再現できるかという「判断の整合性」の調査に焦点を当てる．そのために，本研究では以下の 3 つのリサーチクエスチョンを設定する．

RQ1:　LLM は人間による採点結果をどの程度正確に再現できるか

人間が下した採点判定（適・不適の判断）に対し，LLM が同様の判定を出力できる程度，およびその限界（得意・不得意な項目）を検証する．これにより，LLM が人間の評価基準をどの程度模倣できているかを明らかにする．

RQ2:　LLM のモデル差によって採点の再現性にどのような違いが生じるか

モデルの規模や系統，特性の違いが，判定の傾向や精度にどのような影響を与えるかを調査する．これにより，タスクに適したモデル選定の指針を得る．

RQ3:　LLM による自動採点は実用的な代替手段としてどのような位置付けとなり得るか

RQ1 および RQ2 の結果を踏まえ，現状の技術水準において LLM が人間の専門家による採点業務を完全に代替可能か，あるいはあくまで補助的なツールとしての位置付けに留まるかを評価する．これにより，臨床現場への導入に向けた現実的な期待値を明らかにする．

本研究のキーアイデアは，人間による採点結果を正解と定義し，LLM の出力がそれにどれだけ近づけるかを定量的に比較することである．上記のリサーチクエスチョンに基づき複数の LLM を用いた人の手で行った採点との比較実験を行った．

3.2　実 験 準 備

本節では，3.1 節で設定したリサーチクエスチョンを検証するために行う，複数の LLM を用いた比較実験の具体的な手順について述べる．本実験は，EVIDENT によって取得された描画データに対し，複数の LLM を用いて自動採点を行い，既存の専門家による採点結果との整合性を検証するものである．

3.2.1　データセットと正解ラベルの定義

評価対象となる描画データには，先行研究において開発されたシステム（EVIDENT）を用いて収集された画像データを使用する．比較対象となる正解データには，先行して実施された専門家による確定済みの採点結果を使用する．この正解データは全 10 項目の採点基準に基づき，各項目について「基準を満たしている（適）」か「満たしていない（不適）」かの 2 値判定が付与されている．本実験で使用する採点項目の定義は図 1 で示されているものを利用する．

本実験では，複数の環境下で取得された広範なデータセットの中から，LLM の誤り検知能力を検証するために適した画像を抽出して使用した．具体的な抽出基準として，図 1 に示した

全 10 項目の採点基準それぞれについて，人間による採点で「不適」と判定された症例を重点的に選定した．各項目につき約 5 枚ずつの不適事例画像を抽出して評価用データセットを構成することで，特定の誤りに対する LLM の再現率を効率的に検証可能なデータセットとした．なお，本実験におけるデータ数が限定的である理由は，評価対象となる項目の誤りを明確に含みつつ，他の要因が判定に干渉しにくい症例を厳選したためである．これにより，複合的な要因によるノイズを排除し，各項目に対する純粋な検知能力を評価する探索的な実験設定とした．

3.2.2 使用した LLM と使用理由

自動採点を行うモデルとして，以下の 3 種類のマルチモーダル LLM を選定し，各社の API を通じて推論を行った．

- GPT-5-mini (OpenAI)
- GPT-4o (OpenAI)
- Qwen3-VL:235b-cloud (Alibaba Cloud)

商用 VLM はいずれも内部構造が公開されていないため，本研究では内部アーキテクチャに基づく比較ではなく，公式に示されているモデルの系統，世代，派生の違いが，実際の挙動にどのように現れるかという観点から比較対象を選定した．GPT-5-mini は，GPT 系モデルの中でも比較的新しい世代に属するモデルとして提供されており，軽量かつ高速な「mini」として区別された派生モデルである．GPT-4o は，GPT-4 世代に属するモデルであり，画像入力を含む汎用的なマルチモーダルモデルとして広く提供されている．Qwen3-VL:235b-cloud は，Vision–Language タスクを主な対象として提供されている大規模モデルであり，文書や図表などの視覚的構造を扱う用途が公式情報において例示されている．本実験では，GPT-5-mini と GPT-4o の比較により同一系統内における世代差および派生差の影響を，GPT 系と Qwen3-VL:235b-cloud の比較によりモデル系統の違いが結果に与える影響を分析できるようにした．

なお，本実験における Qwen3-VL:235b-cloud については，実験環境のマシンスペックの都合上，クラウド API 経由で推論を行っているが，これは十分なスペックを持つローカル環境で動作させた場合と同等の性能評価を意図したものである．

3.2.3 実験用 Web システムの構築と評価手順

各 LLM の API を用いた推論実験を効率的に行うため，専用の Web サービスを構築した．本システムはバックエンドに FastAPI，フロントエンドに HTML を用いて実装されており，描画画像のアップロードから各 LLM へのプロンプト送信，および回答の取得を一元管理するものである．構築したサービスの画面を図 6 に示す．

実験の手順は以下の通りである．

(1) 推論の実行　取得した描画画像と採点基準を本システム経由で各 LLM に入力する．プロンプトには「描画画像」と「採点項目の定義」を含め，画像として与えられた時計の描画内容（数字の配置や針の角度など）を視覚的に解析させる．出力の透明性を確保するため，単なる適・不適の判定だけでなく，その判定に至った「根拠」も記述させる．

(2) 採点精度の評価　各 LLM が出力した採点結果と，3.2.1 項で定義した正解ラベルを突合し，その一致度を評価する．評価

図 6　自動採点のサービス画面

にあたっては，全体の正答率に加え，「円」「数字」「針」といった描画要素ごとに，各モデルが「不適」事例をどの程度正確に検出できたかを個別に検証し，各モデルの画像認識能力の特性を明らかにする．

3.3 実験手法

まずは，各 LLM が時計描画検査における視覚的な特徴を正しく捉え，採点基準に基づいた判断ができるかを検証するための準備を行った．3.2.1 項で抽出した各画像データに対し，採点基準を適用するためのプロンプトを構築した．この際，検証対象となる特定の項目のみをプロンプトに含めるのではなく，図 4 に示した全採点項目を網羅的に記述する形式を採用した．これは，単一の項目のみを提示した場合と，全体の基準を提示した場合とで LLM の推論挙動や解釈精度に差異が生じる可能性を排除するためである．これにより，項目間の相互関係や全体的な文脈を考慮させた上で，実運用と同等の条件下での採点を行わせた．具体的には，LLM に対して「専門家の検査官」という役割を与え，全項目について「適」か「不適」かを判定させると同時に，その判定根拠を記述させる設計とした．

次に，構築したプロンプトと評価用画像を用いて，各 LLM による推論を実行した．4.1.2 項で選定した 3 つのモデル（GPT-5-mini，GPT-4o，Qwen3-VL:235b-cloud）に対し，Web インターフェース経由で採点を行わせた．入力においては，手動操作による人為的なミスやプロンプトの揺らぎを防ぐため，画像選択と同時に全項目を含む評価プロンプトが自動的に入力される定型入力の仕組みを用いた．これにより，すべての試行において完全に同一かつ厳密な条件下で推論を行わせたことを保証した．本実験の目的は，現在のプロンプト設計において LLM がどの程度自律的に判定可能かを検証することにあるため，出力結果に対する人手による修正や解釈の介在は行わず，モデルが出力した判定結果をそのまま評価対象として集計した．

推論完了後は，得られた全項目の出力の中から，各画像において検証対象として設定していた項目の判定結果のみを抽出し，人間による正解ラベルとの比較評価を行った．ここで，本実験における評価指標の定義について述べる．一般に，正常例を正しく判定する割合は「特異度」と呼ばれるが，本研究のようなスクリーニング検査においては，異常の兆候を見逃さないこと（偽陰性の低減）が最重要である．そのため，本実験では「不適」の事例を陽性クラスと定義した．その上で，本来「不適」である項目を，LLM が正しく「不適」であると検出できた割合

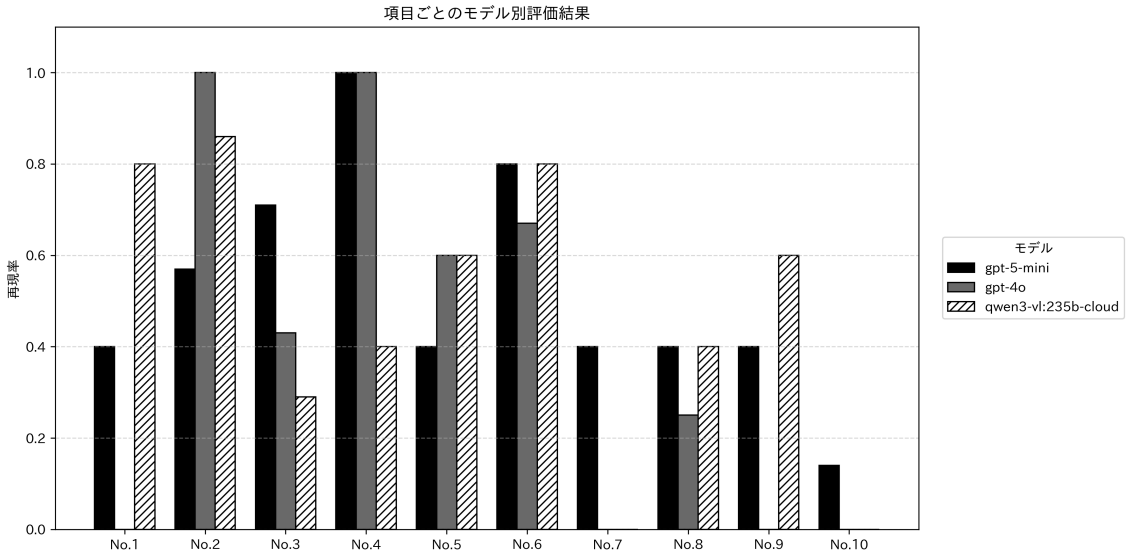

図 7　時計描画の採点項目における各 LLM の再現率

を算出することとし，これを不適クラスに対する「再現率」として評価指標に採用した．最終的に，各採点項目における各モデルのスコアを算出し，モデルごとの精度の違いを明らかにするため，結果を棒グラフとして可視化した．

3.4　実 験 結 果

図 7 に，各採点項目における各 LLM の再現率を示す．ここでは，人間による採点で「不適」と判定された画像に対し，LLM が正しく「不適」であると判定できた割合を評価した．

まず，項目 No.2「数字がすべて円内にあるか」について確認すると，GPT-4o および Qwen3-VL:235b-cloud が高い再現率を示した．この項目は，円という境界線と数字の位置関係を把握するタスクであるが，これらのモデルは高い精度で空間的な包含関係を認識できている結果となった．

次に，項目 No.4「数字の間隔は均等か」については，全体的に高い傾向が見られた．特に GPT-5-mini と GPT-4o においては再現率 100% を達成しており，数字の配置バランスの偏りを，不適事例として見落としなく検出することに成功した．しかしながら「適」として採点されないといけないものまで「不適」と判断される結果になっており特異度が現状高くないという結果になっていた．

一方で，最も難易度が高かった項目 No.10「時針より分針の方が長いか」については，モデル間で顕著な差と限界が見られた．GPT-4o と Qwen3-VL:235b-cloud の再現率が 0% であったのに対し，GPT-5-mini のみが約 14% の再現率を示し，わずかながら不適事例を検出できた．しかし，いずれのモデルにおいても数値は極めて低く，針の長さの比較判定が現状のマルチモーダル LLM にとって困難な課題であることを示している．

3.5　追加実験：プロンプトエンジニアリングによる精度改善の検証

3.4 節の結果より，LLM の自動採点には項目ごとに性質の異なる課題が存在することが明らかとなった．そこで本節では，プロンプトの記述内容を変更することで，これらの精度課題が改善されるかを検証した．なお，追加実験の対象モデルとしては，3.4 節の実験において全体的に最も高い再現率を示した GPT-5-mini を選定し，項目 No.4 および No.10 について再評価を行った．

3.5.1　項目 No.4 における特異度の改善

項目 No.4「数字の間隔は均等か」では，再現率が 100% であった一方で，本来は「配置が均等」である画像に対しても厳しく判定し，「不適」と誤分類してしまう傾向が確認された．そこで本実験では，「適」および「不適」を含む計 44 枚の画像データを用意し，プロンプトを改良することで，本来「適」であるものを正しく「適」と判定できるかを検証した．同時に，プロンプトを緩めることで本来「不適」なものを見逃してしまわないかについても確認を行った．具体的な変更点として，プロンプトに「被験者は高齢者であり，手書きであることを考慮して，多少のずれは許容して採点を行ってください」という文言を追加した．これにより，厳密さよりも，手書き文字特有のゆらぎを許容する柔軟な評価が可能になるかを調査した．

3.5.2　項目 No.10 における再現率の向上

項目 No.10「時針より分針の方が長いか」では，現状のプロンプトにおいて著しく低い再現率が課題となった．そこで本実験でも同様に，「適」および「不適」を含む計 44 枚の画像データを用いて再評価を行った．本項目におけるプロンプトの変更点として，LLM に対して推論のステップを強制する指示を追加した．具体的には，「被験者が描画した円の中心部分（丸の部分）を起点として測定し，左右の位置関係を明記した上でどちらの針が長いかを比較してから採点を行ってください」という指示を組み込んだ．これにより，単に全体を眺めるのではなく，中心点の特定，個々の針の長さ測定，そして相互比較という一連のプロセスを明示的に行わせることで，不適事例（針の長さが逆転，または同等）を正しく検出できるようになるかを検証した．

3.6　追加実験の結果

それぞれの実験結果を混同行列として図 8 および図 9 に示す．図 8 より項目 4「数字の間隔」の採点項目ではプロンプトの変更前と後で特異度が約 24.1 %から約 34.5 %に約 10 %の改善は見られたが依然として過半数の適事例を誤って判断しており，プロンプトの調整のみによる改善効果は限定的であった．また図 9 より項目 10「針の長さ」の採点項目では実際に針の長さを測って採点してもらうようにしたプロンプトの前と後では再現度の部分が約 38.1 %から約 47.6 %の改善傾向が見られたが改善幅も小さく，実運用において再現度を十分に活用できる水準には達しておらず，有用性の観点では課題が残る結果となった．

以上の結果より，プロンプトエンジニアリングによる一定の精度改善は確認されたものの，現時点では人間の専門家を完全に代替できる実用水準には達しておらず，特に幾何学的な厳密さを要する項目においては依然として課題が残ることが示唆された．

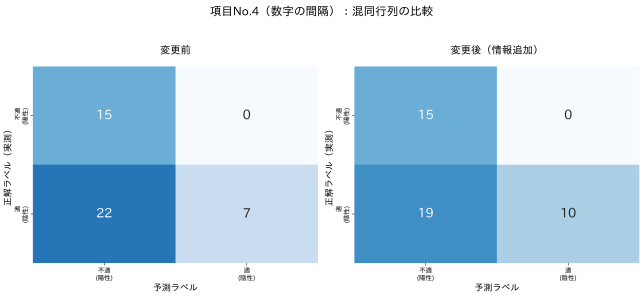

図 8　数字の間隔に関してのプロンプト変更前と変更後の混同行列

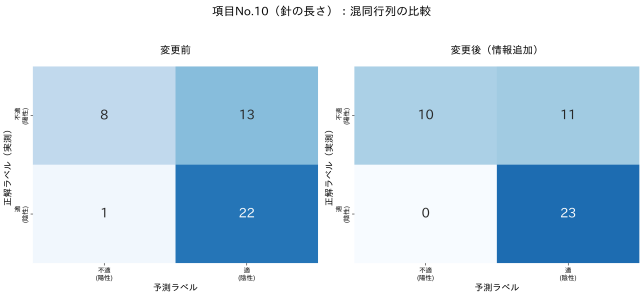

図 9　針に関してのプロンプトの変更前と変更後の混同行列

4.　考　　察

4.1　LLM による自動採点の全体傾向

本実験の結果，LLM を用いた時計描画検査（CDT）の自動採点には，タスクの性質による明確な得意・不得意が明らかとなった．項目 No.2「数字がすべて円内にあるか」のように，円と数字の包含関係や空間的な配置を問う項目では，比較的高い精度での採点が可能であった．これは，現在の LLM がオブジェクト認識やマクロな位置関係を把握する能力に長けていることを示唆している．一方で，項目 No.10（時針より分針の方が長いか）のような針の長短比較では，著しく低い精度となった．これは，手書き描画における線の長さを定量的に計測・比較するという，ミクロな幾何学的処理が，現状の LLM には依然として困難であることを示している．

4.2　モデルごとの特性と傾向

各モデルには独自の傾向が見られた．

4.2.1　GPT-5-mini

GPT-5-mini は，数字や針の「存在個数」の認識に関しては，比較した 3 モデルの中で最も高い精度を示した．また，針の認識において，単純な長さの計測ではなく，周囲の数字との位置関係（相対的な距離感）を手がかりとして推論している挙動が観察された．しかし，項目 No.4「数字の間隔」に関しては，判定基準が厳格すぎる傾向が見られた．人間の目には「概ね均等」と映る配置であっても，わずかなズレを検知して「不均衡（不適）」と判定するケース（偽陽性）が多発した．3.5 節の追加実験において，許容範囲を広げるプロンプトを与えたことで多少の改善は見られたものの，依然として厳格な判定傾向は残存しており，人間の感覚に近い「曖昧さの許容」には課題を残した．

4.2.2　GPT-4o

GPT-4o の最大の特徴は，その推論速度にある．他の 2 モデルが 1 枚あたり約 50 秒を要したのに対し，GPT-4o は約 20 秒で採点を完了しており，実運用における即時性の観点では優位性がある．

一方で，精度の評価に関しては慎重な解釈が必要である．図 7 の結果を一見すると，GPT-4o は他のモデルと比較して高い再現率を示しているように見える．しかし，詳細な事例分析を行うと，描画の形が大きく崩れている特定の症例画像において，視覚的な認識ができず採点不能となったり，時計としての構成要素不足を理由に採点を放棄したりするケースが確認された．すなわち，GPT-4o は難易度の高い「崩れた描画」を採点対象から事実上除外してしまい，相対的に採点が容易な（形状が保たれた）画像のみを処理した結果として，見かけ上の数値が高くなっている可能性が考えられる．このことから，図 9 で示された高い数値は必ずしもモデルの頑強性を示すものではなく，「採点しやすい画像だけを選別して回答した結果」であるという側面が強く，重度の認知機能低下に伴う崩れた描画に対する対応力には依然として課題があると言える．

4.2.3　Qwen3-VL:235b-cloud

Qwen は，円と数字の境界判定において極めて高い精度を示した．特に，数字が円の線上にかかっているような際どい事例においても，厳密に「円の外（または線上）」であると認識し，正確な判定を行うことができた．反面，数字同士の位置関係（間隔）に関しては，GPT-5-mini とは対照的に採点が甘くなる傾向が見られた．明らかにバランスが崩れている配置であっても「均等である」と判定するケースが散見された．また，円の分割線に関する項目においては，描画されている分割線を正しく認識できず，本数が不足していると誤判定するなど，微細な要素の採点能力は低いと言える

4.3　プロンプトエンジニアリングにおける課題

追加実験の結果，プロンプトの微修正のみでは解決できない根本的な課題が浮き彫りとなった．「数字の間隔」では「甘く見ろ」と指示しても幾何学的厳格さが勝り，「針の長さ」では「円の中心」という指示を「画像の中心座標」と解釈するよう

な LLM の認識のズレが確認された．ここで認知的観点から考察すると，人間は針同士が物理的に接触している「結合点」を瞬時に特定し，その点を起点として各直線の長さを追跡・比較するという構造的な視覚処理を行っている．対して LLM は，手書き特有の曖昧な結合点を正確な始点として認識できておらず，結果として絶対座標への依存や誤認識が生じていると考えられる．これにより，人間と AI の認識ギャップを埋めるには，根本的なプロンプト再設計が必要であることが示唆された．

4.4 リサーチクエスチョンに対する考察

本研究で設定した 3 つのリサーチクエスチョンに対し，得られた実験結果に基づき以下の通り考察する．

RQ1: LLM は人間による採点結果をどの程度正確に再現できるか

円や数字の配置といった大域的な空間認識においては，LLM は人間と同等の高い再現性を示した．一方で，針の長さの比較などの微細な幾何学的測定においては，プロンプトによる指示の最適化を行ってもなお精度が著しく低下することが明らかとなった．

RQ2: LLM のモデル差によって採点の再現性にどのような違いが生じるか

モデルごとに推論速度や採点傾向（厳格さや頑強性）に明確な差異が確認された．例えば，GPT-5-mini は相対的な関係性の把握に優れるが厳格すぎ，GPT-4o は高速だが崩れた描画に弱く，Qwen3-VL は境界判定には強いが微細な要素を見落とすといった特性が見られた．

RQ3: LLM による自動採点は実用的な代替手段としてどのような位置付けとなり得るか

現状の単一モデルのみでは，精度の安定性と幾何学的理解の不足により，専門家を完全に代替することは困難である．しかし，モデルごとの特性差を活かした補助ツールや，複数のモデルを組み合わせることで実用性を高められる可能性がある．

4.5 今後の課題と展望

課題として LLM の出力の不安定性と，日本語プロンプトによる言語的な制約が挙げられる．再現性の担保には多数決の導入や，英語プロンプトの活用が有効と考えられる．結論として，各モデルには明確な特性差が存在するため，単一の万能モデルに依存するのではなく，項目ごとの特性に応じてモデルを使い分けるアンサンブルアプローチの導入が，実用的な自動採点システムの実現には不可欠である．

5. ま と め

本研究では，マルチモーダル LLM を用いた時計描画検査の自動採点手法の構築とその評価を行った．研究の目的として，LLM の再現能力，モデル差による影響，および実用性の可否を検証するため，クラウドモデルおよびプライバシー保護を考慮したローカルモデル（GPT-4o，GPT-5-mini，Qwen3-VL）を用いた比較実験を実施し，さらにプロンプトエンジニアリングによる精度改善の可能性を検証した．

実験の結果，LLM は空間的な配置などの大域的特徴の認識には優れる一方，幾何学的な厳密さを要する測定タスクには課題が残ることが明らかとなった．これは，人間が描画の構造的特徴を把握して判断しているのに対し，LLM は座標ベースの処理や言語的な文脈に依存しているという認知的ギャップに起因すると考えられる．また，モデルごとの特性差が明確になったことから，現状の単一モデルのみでは専門家の完全な代替として実用化するには至らなかった．

しかし，この結果は必ずしも LLM の有用性を否定するものではない．本研究を通じて，LLM は単なる人間の「採点者」の代替ではなく，人間特有の評価の揺らぎを排し，「採点基準」を客観的かつ安定的に運用するための装置として有望であるという，新たな知見が得られた．

今後の展望として，各モデルの得意分野を活かして判定を行うアンサンブルアプローチの導入や，英語プロンプトによる指示の最適化が挙げられる．さらに，静止画のみの解析における限界を突破するため，EVIDENT の特性を活かし，筆順や速度といった「描画過程データ」をマルチモーダル入力として統合する新たな解析手法の構築を進めていく．

謝辞 本研究の一部は JSPS 科研費 JP25H01167, JP25K02946, JP25K24389, JP24K02765, JP24K02774, JP23K17006, JP23K28091, JP23K28383 の研究助成を受けて行われている．

文 献

[1] Agrell, B. and Dehlin, O.: The clock-drawing test, Age and ageing, Vol. 27, No. 3, pp. 399–403 (1998).

[2] S, M., A, O., S, M., K, O., T, S., I, K. and E., S.: Cube Copying Test(CCT) 採点法の信頼性・妥当性に関する臨床的検討，Japanese Journal of Comprehensive Rehabilitation Science, pp. 5–102 (2014).

[3] Masami, N., Masahiro, N., Eiji, I., Kanami, O., Masako, K., Akiko, H., Yasuyo, M. and Shinnichi, W.: MMSE and scoring of clock drawing test increase the accuracy of diagnosis of dementia, 医学検査, Vol. 68, No. 3, pp. 424–429 (2019).

[4] Ryukichi, S., Sachio, S., Masahide, N., Naoki, K. and Atsushi, S.: 神経心理学的描画検査における描画過程の可視化インターフェース EVIDENT の実装，電子情報通信学会技術研究報告，Vol. 120, No. 232 SC2020-31, pp. 63–69 (2020).

[5] Ryukichi, S., Sachio, S., Masahide, N., Naoki, K. and Atsushi, S.: 描画過程に基づく認知機能検査のデジタル化に向けたプラットフォームの作成，電子情報通信学会技術研究報告，Vol. 121, No. 416, pp. 151–156 (2022).

[6] Keisuke, Y., Sachio, S., Naoki, K., Atsushi, S., Sinan, C. and Masahide, N.: 描画過程に基づく認知機能検査アプリケーションの集団検査に向けた改良，電子情報通信学会技術研究報告， Vol. 123, No. 429, LOIS2023-60, pp. 070–077 (2024).

[7] Keisuke, Y., Shun, H., Sachio, S., Atsushi, S., Naoki, K. and Masahide, N.: 描画検査アプリ EVIDENT を用いた集団検査実験の実施と探索的データ分析，電子情報通信学会技術研究報告，Vol. 124, No. 245 SC2024-40, pp. 108–115 (2024).

[8] 日本看護協会：2019 年 病院看護実態調査，https://www.nurse.or.jp/up_pdf/20200330151534_f.pdf. (Accessed on 07/25/2025).

[9] 内閣府：2024 年 高齢社会白書，https://www8.cao.go.jp/kourei/whitepaper/w-2024/html/zenbun/s1_2_2.html. (Accessed on 07/29/2025).

[10] Koichi, T.: 神経心理学評価ハンドブック (2004).

[11] “How the clock-drawing test screens for dementia,” https://www.verywellhealth.com/the-clock-drawing-test-98619. (Accessed on 1/25/2024).

[12] W. Souillard-Mandar, R. Davis, C. Rudin, R. Au, and D. Penney, “Interpretable machine learning models for the digital clock drawing test,” 2016.