城西国際大学紀要　第34巻　第３号　2026年３月　15-25

〈研究論文〉

# SAFE-Home（Screening with Aggregated multi-run for Fall-hazard Evaluation in Home environments）：マルチモーダル大規模言語モデルによる家庭内転倒ハザードの安定化スクリーニング
## ―探索的パイロットスタディ―

桑　江　　豊

**【要旨】**

本研究の目的は、マルチモーダル大規模言語モデルであるGPT-4oを用いて、家庭内の静止画から転倒リスク要因を検出する手法の診断精度を、理学療法士の専門的判断を基準として検証することであった。在宅高齢者の生活空間を模擬した環境で撮影された140枚の室内写真を対象に、診断精度研究として計画した。指標テストには、同一画像に対し3回推論を行い多数決で判定を安定化させる「SAFE-Home（Screening with Aggregated multi-run for Fall-hazard Evaluation in Home environments）プロトコル」を適用したGPT-4oを用い、参照基準には理学療法士2名のコンセンサス判断を用いた。結果として、本手法は高い感度（0.917）を示したものの、特異度（0.603）と適合率（0.710）はそれに及ばず、偽陽性を多く生む傾向が明らかになった。偽陽性の多くは些細な特徴の過大評価に起因していた。結論として、GPT-4oは潜在的リスクの見逃しが少ないスクリーニングツールとして専門家を補助する可能性があるが、過剰検知の課題から単独での使用は時期尚早である。今後は専門家がAIの判断を監督・修正するHuman-in-the-loopの体制が重要になると示唆された。

**キーワード**：転倒リスク、家庭環境、GPT-4o、スクリーニング、診断精度

## １．はじめに

高齢者の転倒は、身体的傷害や自立の喪失、医療費の増大に繋がる、世界的に重要な公衆衛生上の課題である[1, 2]。これらの転倒の多くは、日常生活を送る家庭環境内に存在する様々な物理的ハザード（例：段差、滑りやすい床、不十分な照明）に起因する[1, 2]。そのため、家庭内のハザードを専門家が評価し、適切に修正・改善することが、科学的根拠のある効果的な転倒予防戦略として広く認識されている[2]。

しかし、理学療法士などの専門家による直接の家庭訪問は有効である一方、時間や移動コス

ト、人材確保の面で労働集約的であり、スケールアップや遠隔地への提供が難しいという根本的な課題を抱えている。この課題に対し、デジタル写真を用いて遠隔で家庭環境を評価するアプローチが登場し、専門家以外による評価でもある程度の妥当性を持つことが示唆されている[(3)]。このパラダイムは、画像とテキストといった複数のデータを統合して診断性能を向上させる、医療分野におけるマルチモーダル人工知能（AI）活用の潮流とも合致する[(4)]。しかし、この相乗効果の可能性があるにもかかわらず、これら高度なAIモデルを環境的な転倒リスクのスクリーニングに効果的に適用できるかどうかは未知数であり、患者安全技術における重要かつ未評価の領域となっている。

したがって、本探索的パイロットスタディの主目的は、マルチモーダル大規模言語モデル（multimodal Large Language Model; LLM）であるGPT-4oの診断精度を、静止画からの家庭内転倒リスク要因識別の観点から評価することである。筆者は、指標テストとしてのモデルの性能を、2名の理学療法士のコンセンサスによって確立された参照基準と比較した。さらに、生成AI固有の変動性に対処するため、筆者は新規の「SAFE-Home（Screening with Aggregated multi-run for Fall-hazard Evaluation in Home environments）」プロトコルを実装・評価した。このプロトコルでは、同一画像に対する3回の独立した推論からの多数決によって最終的な判断が下される。

本研究の貢献は以下の3点に要約される。（1）家庭内静止画を対象とした転倒ハザードの二値検出タスクを、理学療法士のコンセンサスを参照基準とする診断精度研究として形式化した。（2）LLMの3回推論を多数決で安定化させるSAFE-Homeプロトコルと、臨床的解釈が可能な誤差類型化コードブックを提示した。（3）研究の再現性と透明性を担保するため、中核となる標準化プロンプトの全文と、全画像の判定結果を含む匿名化データ表を開示した。

## ２．方法

### ２．１　研究デザインと報告ガイドライン

本研究は、探索的パイロットスタディとして計画された診断精度研究である。研究のデザインおよび報告は、診断精度研究報告のためのSTARD 2015声明に準拠した[(5)]。

### ２．２　画像データセットと対象者

データセットは、筆者の自宅（築20年、木造2階建て）の様々な室内環境を撮影した140枚のデジタルカラー写真で構成される。これは、本パイロットスタディの実現可能性を考慮し、在宅高齢者の生活空間を模擬した環境として設定した。写真は、玄関、廊下、浴室、トイレ、リビング、居間、寝室といった転倒が起こりやすい典型的な生活空間を対象とし、iPhone 14を用いて撮影された。全ての画像は分析前に匿名化処理を施した。

撮影機種：iPhone 14、撮影時期：2024年5月、時間帯：10:00～14:00、天候：晴、室内照明

は全点灯で統一した。構図は各室の動線全体が画面内に収まるように撮影した。

### ２．３　参照基準：理学療法士による評価

各画像の転倒リスク要因の有無に関する参照基準（Reference Standard）は、2 名の経験豊富な理学療法士（PT-A（臨床 22 年：急性期／回復期 7 年・退院前家屋調査多数、介護保険 8 年・通所／訪問・制度に基づく住宅評価を日常実施）／ PT-B（臨床 17 年：通所・デイ中心、介護保険利用前の住宅評価を多数経験））の独立判定とコンセンサスで確立した。両名は事前に本研究で用いる転倒リスク要因の定義と判定基準を協議し、試行評価で基準の共通理解を形成したのち、全画像に対して相互に盲検で独立評価を行った。不一致事例のみ協議し最終合意に至り、この合意判断を参照基準とした。

### ２．４　指標テスト：GPT-4o による評価と SAFE-Home プロトコル

指標テスト（Index Test）には、本研究で提案する SAFE-Home プロトコルを適用した。本プロトコルは、LLM である GPT-4o（OpenAI 社，2024 年 5 月版）の画像解析機能を利用し、その出力の安定性を確保することを目的とする。具体的には、評価には同モデルの Web インターフェースを使用し temperature・top_p・seed は UI 上で制御不可であるため、出力変動に対応して、各画像に対して同一の標準化プロンプト（prompt_v1）を用いて独立した推論を 3 回実行した（図 1）。最終的な分類は、これら 3 回の推論結果の多数決（3 回中 2 回以上が同一判断）によって決定した。使用した標準化プロトコルの全文を以下に示す。

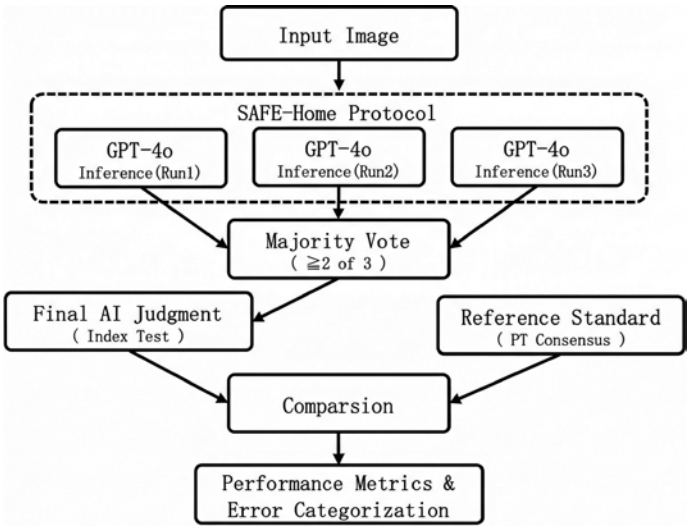


図 1　SAFE-Home プロトコル

入力画像から GPT-4o による 3 回推論、多数決による最終判定、参照基準との比較による混同行列の作成、そして誤差の類型化に至るまでの一連の評価プロセスを示す。出典：筆者作成

【プロンプト全文（prompt_v1）】

「送信された画像を解析して、転倒リスクの有無を判断し、リスクに関連する要素を詳しく説明します。床面の状態（滑りやすい箇所や段差）、家具や物品の配置による通行の障害、照明の不備や視認性の悪さ、水回りの安全性などに注目し、ユーザーがリスクを理解できるように詳細に説明します。想定する対象者は在宅で生活している健常高齢者とします。回答の順序として、まずは転倒リスクを「ある」「なし」と端的に述べ、その後に詳細な評価を行います。」

### ２．５　データ解析と統計手法

SAFE-Homeプロトコルの診断精度は、2 × 2 の混同行列（真陽性［True Positive; TP］、偽陽性［False Positive; FP］、偽陰性［False Negative; FN］、真陰性［True Negative; TN］）を作成して評価した。FPはA（些細な特徴の過大評価）／D（可能性の過剰指摘）、FNはF（微妙さの見落とし）／H（出力不安定）、TPはK1（主要因一致）／M（偶然一致）に分類した。この行列から、感度（再現率）、特異度、適合率、正解率、F1 スコア、陰性的中率（NPV）、バランスの取れた正解率（Balanced Accuracy）、マシューズ相関係数（MCC）、Youden’s J指数、および陽性・陰性尤度比（LR+/LR−）を算出した。単一比率の95％信頼区間（CI）はWilsonスコア法を用いて算出した。医療AIの評価においては単一の指標ではモデルの特性を捉えきれないため、複数の指標を報告することが推奨されている[(6)]。特に本研究のように陽性・陰性のサンプルサイズが不均衡な場合でも安定した評価が可能とされるマシューズ相関係数（MCC）も算出した[(7)]。F1 スコアの95％ CIは画像単位の非パラメトリック・ブートストラップ（B＝5000）で推定した。3 回推論の安定性は、一致率およびFleissの $\kappa$ 係数を用いて評価した。統計解析にはPython（ver. 3.10）および関連ライブラリ（pandas, scikit-learn, stats models）を用いた。

### ２．６　倫理的配慮

本研究は研究者自宅の室内写真のみを対象とし、人物の撮影・個人情報・介入・試料の取得を伴わない。在宅環境の安全評価手法の検討であり、当学科倫理委員に事前相談のうえ、学内規程に照らして「人を対象とする研究」には該当しないことを確認したため、正式な審査の対象外とした。ただし、研究データの取り扱いにおいては、個人情報の保護に関する法律を遵守した。画像は匿名化し、パスワード保護下で管理し、本研究以外の目的に用いない。

## ３．結果

### ３．１　参照基準評価の信頼性

理学療法士2名による140枚の画像に対する初期独立評価について、評価者間信頼性をCohenの $\kappa$ 係数で算出した結果、$\kappa = 0.899$（95％ CI: 0.827 – 0.972）となり、「ほぼ完璧な一致」が示された。

### ３．２　SAFE-Homeプロトコルの安定性

GPT-4oによる3回推論の結果、140枚中119枚の画像（85.0％）において3回の判断が完全に一致した。評価者3名（GPTの各推論を評価者と見なす）による評定の一致度を示すFleissのκ係数は0.688であり、「実質的な一致」を示した。

### ３．３　診断精度の全体成績

参照基準とSAFE-Homeプロトコルによる最終判断を比較した混同行列を表1に、そこから算出された各性能指標を表2に示す。GPT-4oは感度が0.917と高い値を示した一方で、適合率は0.710、特異度は0.603であった。

表1　混合行列（N＝140）

| | GPT-4o：リスクあり | GPT-4o：リスクなし | 行合計 |
|---|---|---|---|
| 参照基準：リスクあり | TP ＝ 66 | FN ＝ 6 | 72 |
| 参照基準：リスクなし | FP ＝ 27 | TN ＝ 41 | 68 |
| 列合計 | 93 | 47 | 140 |

出典：筆者作成

表2　GPT-4oの性能指標一覧

| 評価指標 | 値 | 95％信頼区間 |
|---|---|---|
| 感度（Recall） | 0.917 | 0.825 _ 0.962 |
| 特異度（Specificity） | 0.603 | 0.484 _ 0.711 |
| 適合率（Precision） | 0.710 | 0.609 _ 0.796 |
| 陰性的中率（NPV） | 0.872 | 0.748 _ 0.94 |
| 正解率（Accuracy） | 0.764 | 0.686 _ 0.829 |
| F1スコア | 0.8 | 0.720 _ 0.864* |
| バランスの取れた正解率 | 0.76 | _ |
| マシューズ相関係数（MCC） | 0.55 | _ |
| Youden’s J index | 0.52 | _ |
| 陽性尤度比（LR+） | 2.31 | 1.71 _ 3.12 |
| 陰性尤度比（LR−） | 0.138 | 0.063 _ 0.304 |

説明：F1スコアの95％信頼区間はブートストラップ法（5,000回）により算出。

出典：筆者作成

### ３．４　誤差要因分析

偽陽性（FP）および偽陰性（FN）となった事例の質的分析に基づき、誤差を類型化した（図2）。FP 27件の内訳は、「些細な特徴の過大評価（カテゴリA）」が74.1％（20件）、「可能性の過剰指摘（カテゴリD）」が25.9％（7件）であった。FPの代表例（図3（A））において、理学療法士は「移動スペースも十分にある」と判断し『リスクなし』としたが、GPT-4oは「ジョイントマットの継ぎ目に段差がある箇所が確認できます」と、マットの極めてわずかな不揃い

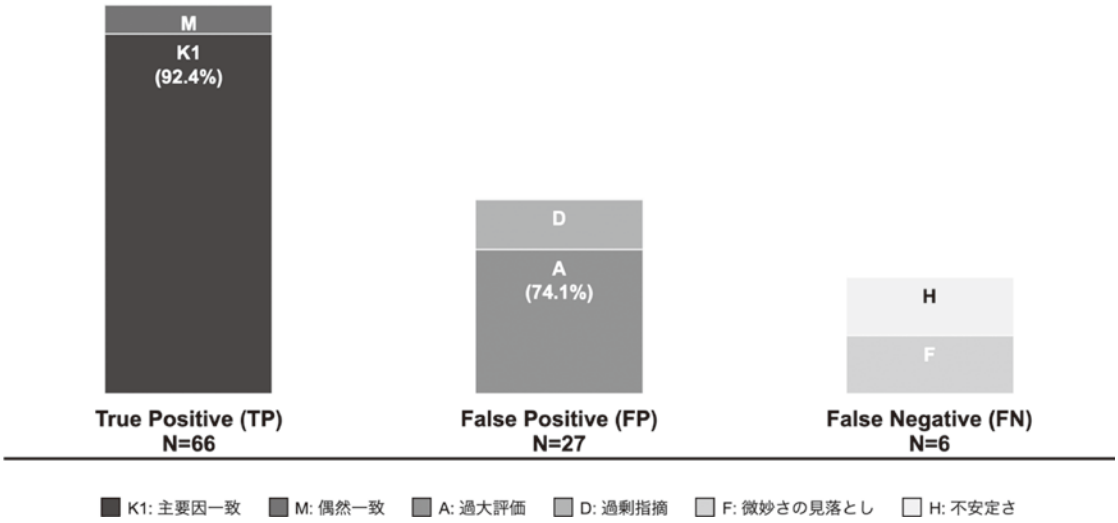


図 2　誤差の類型分布

真陽性（TP）、偽陽性（FP）、偽陰性（FN）の各事例における誤差・一致の類型（A、D、F、H、K1、M）の分布を示す。FPはカテゴリA（些細な特徴の過大評価）が、TPはK1（主要因一致）が大多数を占めることがわかる。出典：筆者作成

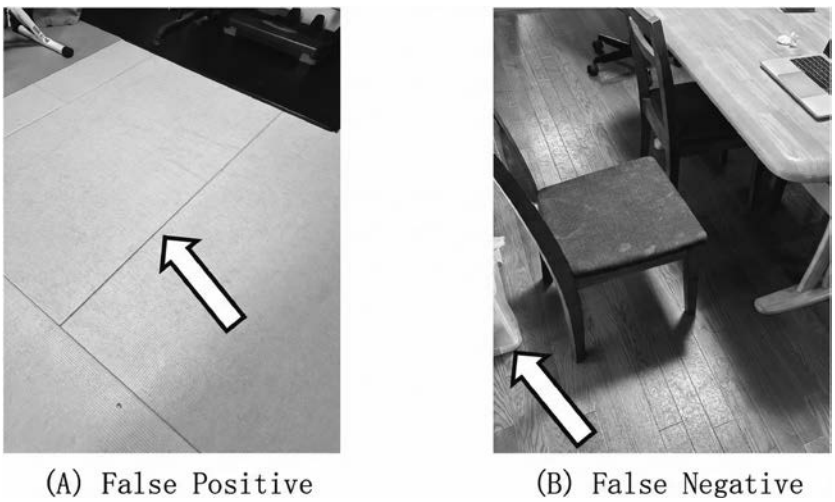


図 3　偽陽性（FP）および偽陰性（FN）の代表例

（A）偽陽性の代表例。理学療法士は「リスクなし」と判断したが、GPT-4oはジョイントマットの極めてわずかな継ぎ目を「些細な特徴の過大評価（カテゴリA）」として誤ってリスクと判定した。（B）偽陰性の代表例。理学療法士は「椅子の脚が通路にわずかにはみ出している」点をリスクと判断したが、GPT-4oは「微妙な障害の見落とし（カテゴリF）」によりリスクを認識できなかった。出典：筆者作成

を指摘し『リスクあり』と判断した。FN 6件の内訳は、「リスクの微妙さによる見落とし（カテゴリF）」と「出力の不安定さによる見落とし（カテゴリH）」がそれぞれ50％（3件）ずつであった。FNの代表例（図3（B））において、理学療法士は「動線に椅子が出ており、通行の

妨げになっている」点をリスクと指摘したが、GPT-4oは「床に障害物なし、通行スペース確保。椅子も適切な位置」と、リスクを認識できなかった。一方、真陽性（TP）66件のうち、92.4％（61件）は理学療法士の指摘と主要なリスク要因が一致しており（カテゴリK1）、説明も妥当であった。

### ３．５　部屋タイプ別の傾向

探索的分析として部屋タイプ別の誤差傾向を確認したところ、偽陽性（FP）は特に居間（20枚中10件）とリビング（20枚中７件）で多く観察された。

## ４．考察

本研究は、最新の大規模言語モデルGPT-4oが、住宅内の写真からどの程度転倒リスク要因を検出できるかを、理学療法士の専門的評価を基準として検証した探索的パイロットスタディである。主要な知見として、GPT-4oは高い再現率（感度：0.917）を示す一方で、適合率（0.710）と特異度（0.603）はそれに及ばず、偽陽性が多く生じる傾向が明らかになった。

GPT-4oの高い感度は、専門家がリスクと判断した状況の９割以上を見逃さなかったことを意味し、潜在的な危険を網羅的に拾い上げるスクリーニングツールとしての可能性を示唆する。これは、リスクの見逃しが重大な結果を招きかねない転倒予防の観点から、重要な特性である。

一方で、低い適合率の原因となった偽陽性の多さは、GPT-4oの実用化における主要な課題である。誤差要因分析の結果、FPの大部分が「些細な特徴の過大評価」や「可能性の過剰指摘」に起因していた。これは、モデルが高い正答率を示した場合でも、その判断根拠となる説明が論理的に破綻していることがある、という先行研究の指摘と合致する[8]。Jinらの研究では、GPT-4Vが正解を導きながらも、その約35.5％で欠陥のある推論を行っていたことが報告されている。本研究のFP事例は、まさにこの「もっともらしいが不正確な推論」の現れであり、LLMが自信を持って誤った情報を生成する「ハルシネーション」の一種と捉えることができる[9]。この特性は利用者にとって「アラート疲れ」やシステムの信頼性低下を引き起こす可能性がある。

偽陰性は少数であったが、その原因はリスクの微妙さをAIが認識できなかったケースや、画像の解釈が不安定だったケースに大別された。これは、人間の専門家が持つ空間認識能力や文脈理解能力に、現行のLLMがまだ及ばない部分があることを示す。事実、Kaczmarczykらの The New England Journal of Medicine画像チャレンジを用いた研究では、GPT-4Vが困難な質問を拒否したり、誤答したりする選択的な応答傾向を持つことが示されている[10]。本研究のFN事例は小物体の視覚的手掛かりの希薄さ[11]と部分的遮蔽[12]、狭隘空間での距離・左右関係の推論[13]、および視点依存[14]が複合して生じたと解釈できる。小物体と遮蔽は従来研究で

も検出困難な条件として繰り返し指摘されており、LLMは空間関係の理解でも人に対してギャップを残す。さらに、GPT-4Vは画像の提示様式や順序への脆弱性が報告されており、同一場面でも出力が揺れうる。このことが、LLMが不得手なケースを処理できなかった結果と解釈できる。

本研究の独自性は、「問題の定義と方法論の新規性」の組み合わせにある。筆者は、LLMのハザード検出能力を「診断精度研究」として形式化し、その評価枠組みを提示した。さらに、LLMの不安定性という課題に対し、3回推論の多数決による「SAFE-Home」プロトコルという具体的解決策を提案・評価した。TPの9割以上で専門家と主要なリスク認識が一致していた事実は、明確なリスク要因に対するGPT-4oの検出能力が非常に高いことを示しており、この点においては有用な補助ツールとなり得るだろう。

本研究の結果は、マルチモーダルLLMの画像理解が高感度な一方で、説明妥当性や出力安定性に課題が残ることを再確認した[8, 10]。したがって、本技術の安全な実装には、専門家がAIの判断を監督・修正するHuman-in-the-loopの枠組みが不可欠である。Festorらのシミュレーション研究では、臨床医はAIによる不安全な推奨の92％を拒否できたが、残りの8％は受け入れてしまったことが報告されており、人間の監督が重要であると同時に、その監督も完全ではないことが示唆されている[15]。AIを臨床判断の代替ではなく、あくまで増強ツールとして用いるべきであり、利用する専門家はその限界を理解する必要がある[16]。

本研究の限界は大きい。第一に、単一の住宅で撮影された画像のみを用いており、結果の一般化可能性は極めて低い。例えば、対象とした築20年の木造家屋では経年による床の微細なたわみ等が存在し、専門家は許容範囲と見なす一方、AIはこれをリスクとして過剰検出（偽陽性）した可能性が考えられる。第二に、ゴールドスタンダードは高い信頼性を確保したが、評価者の専門性によるバイアスは否定できない。さらに、本研究で提案したSAFE-Homeプロトコルは主に出力の安定性向上を目的としたが、診断精度自体への直接的な影響については、単一推論の結果との比較など、さらなる検証が必要である。第三に、GPT-4oのバージョン依存性や、その判断プロセスがブラックボックスである問題がある。特に、本研究のFP事例で示されたように、LLMの推論の信頼性は自明ではなく、正解を導き出す能力と、その理由を正しく説明する能力は別問題として検証されなければならない[8]。さらに、家庭内の写真を扱う本手法は、プライバシー、同意、バイアスといった倫理的課題を慎重に考慮する必要がある[16]。

今後の展望として、多様な住宅環境から収集した大規模なデータセットを用いた多施設共同研究が不可欠である。本研究で見られた過剰検出（偽陽性）と見逃し（偽陰性）は、AIのリスク認識閾値の問題として捉えることができる。静止画だけでなく動画データを用いることに加え、プロンプト設計の工夫によりこの閾値を調整し、偽陽性を削減できる可能性がある。例えば、「生命に危険を及ぼすような明らかな段差や障害物のみを指摘し、数ミリ程度の軽微な不揃いはハザードと見なさないでください」といった指示をプロンプトに加えることで、より

臨床的に有用な特異度を持つツールへと改善が期待できる。最終的には、AIをリスクの初期スクリーニングツールとして位置づけ、専門家が最終判断を下す協働モデルを構築することが、現実的かつ安全な応用への道筋となるだろう。

## ５．結論

本探索的パイロットスタディは、限られた条件下において、GPT-4oが住宅内の写真から転倒リスク要因を検出する能力を検証した。GPT-4oは、理学療法士の専門的評価と比較して高い感度（0.917）を示し、リスクの見逃しが少ない可能性が示唆された。特に、明確なリスク要因に対する検出能力は高く、その９割以上で専門家と主要なリスク認識が一致していた。しかし、専門家が許容範囲とするような些細な特徴や、潜在的な可能性を過剰に指摘する傾向があり、偽陽性が多く発生し適合率（0.710）が低下する課題も明らかになった。現状のGPT-4oを単独で正確な転倒リスク評価ツールとして使用するのは時期尚早であるが、その高い検出感度を活かし、専門家による評価の補助や、居住者自身が潜在的な危険に気づくための初期スクリーニングツールとしての将来的な可能性は示された。

## 【参考文献】


(1) Lee S. Falls associated with indoor and outdoor environmental hazards among community-dwelling older adults between men and women. BMC Geriatr. 2021;21: 547.

(2) Rani I, Zaheer S, Nasim S, Shah N, Hydrie MZI. Assessing hazards and associated fall risks among elderly population: a cross-sectional study of different residential settings in Karachi, Pakistan. BMC Public Health. 2025;25: 850.

(3) Ritchey KC, Meyer D, Ice GH. Non-therapist identification of falling hazards in older adult homes using digital photography. Prev Med Rep. 2015;2: 794-797.

(4) Jandoubi B, Akhloufi MA. Multimodal artificial intelligence in medical diagnostics. Information（Basel）. 2025;16: 591.

(5) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351: h5527.

(6) Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022;12: 5979.

(7) Chicco D, Jurman G. The advantages of the Matthews correlation coefficient（MCC）over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21: 6.

(8) Jin Q, Chen F, Zhou Y, Xu Z, Cheung JM, Chen R, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. NPJ Digit Med. 2024;7: 190.

(9) Aljamaan F, Temsah M-H, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: Development and usability study. JMIR Med Inform. 2024;12: e54345.

(10) Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. NPJ Digit Med. 2024;7: 205.

(11) Miri Rekavandi A, Rashidi S, Boussaid F, Hoefs S, Akbas E, Bennamoun M. Transformers in Small Object Detection: A benchmark and survey of state-of-the-art. ACM Comput Surv. 2026;58: 1-33.

(12) Ruan J, Cui H, Huang Y, Li T, Wu C, Zhang K. A review of occluded objects detection in real complex scenarios for autonomous driving. Green Energy and Intelligent Transportation. 2023;2: 100092.

(13) Shiri F, Guo X-Y, Far MG, Yu X, Haf R, Li Y-F. An empirical analysis on spatial reasoning capabilities of large multimodal models. In: Al-Onaizan Y, Bansal M, Chen Y-N, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2024. 21440-21455.

(14) GPT-4V(ision) System Card. 2023. Available: https://cdn.openai.com/papers/GPTV_System_Card.pdf

(15) Festor P, Nagendran M, Gordon AC, Faisal AA, Komorowski M. Safety of human-AI cooperative decision-making within intensive care: A physical simulation study. PLOS Digit Health. 2025;4: e0000726.

(16) Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med. 2024;7: 183.

# SAFE-Home (Screening with Aggregated multi-run for Fall-hazard Evaluation in Home environments): Stabilized Screening of In-Home Fall Hazards by a Multimodal LLM: An Exploratory Pilot Study

Yutaka Kuwae

Abstract

The objective of this study was to evaluate the diagnostic accuracy of a multimodal large language model (LLM), GPT-4o, for detecting in-home fall-risk hazards from static photographs, using the consensus judgment of physical therapists as the reference standard. This diagnostic accuracy study was conducted on 140 indoor photographs simulating the home environments of older adults. The index test was GPT-4o with the "SAFE-Home" protocol applied, which stabilizes the output through a majority vote from three independent inferences on the same image. The reference standard was the consensus judgment of two physical therapists. The method demonstrated high sensitivity (0.917) but moderate specificity (0.603) and precision (0.710), with a tendency to produce numerous false positives. Most false positives were attributed to the over-reading of minor features. In conclusion, while GPT-4o shows potential as a high-sensitivity screening tool to augment expert assessment, its standalone use is premature due to the challenge of over-detection. A human-in-the-loop framework is suggested as crucial for practical application.

Keywords: Fall risk, home environment, GPT-4o, screening, diagnostic accuracy