論文とETIC 15

覚えているAIは、いつ声をかけるべきか

記憶・タイミング・支援の強さ



Daiki Morimoto / Unknown Lab
ORCID: 0009-0009-6444-750X
Version 1.0 / 公開日 2026-10-28
DOI: 10.5281/zenodo.23717057
未査読 / 専門職としての著者確認済み
独立専門レビュー未実施 / AI支援執筆
ライセンス: CC BY 4.0

30秒で読む

生活の記録を長く覚えるAIは、尋ねられる前に助言できる。生活で役立てるには、記憶した内容に加え、話す時機と支援の強さも考える必要がある。本稿は二つの研究を並べる。MemoProAgent [1]は2026年の未査読プレプリントで、一人称映像と約一年分の合成履歴から根拠を探し、介入ゲートで話すべき事柄を絞る。COACH [2]は2008年の小規模試験で、認知症のある高齢者の手洗いを、確率的な状態推定と段階的な声かけで支えた。前者は記憶の検索と時機の難しさを、後者は本人の反応と支援の段階づけの難しさを具体的に示す。今回の二論文は両者の組合せやETIC 2.0 [4]を直接検証していない。Unknown Labは、新しいモデルを作る前に、両研究の既存部品を同じ条件の標準対照と比べ、「解決済み・取消・拒否」を明示した状態管理だけで足りるかを確かめることを提案する。この提案は未検証である。

仮定の場面と問い

以下は説明のための仮定である。ある人が先週、家族に「来週、牛乳を買う」と話した。今朝、冷蔵庫を開けたとき、眼鏡型のAIがその記録を思い出し、「牛乳を買う予定でしたね」と言う。役に立つ一言に見える。しかしその人は昨日もう買っていたかもしれないし、乳製品をやめると決めたかもしれない。聞き逃して、もう一度言ってほしいと思っているかもしれない。

この場面で、記憶そのものは正しい。問題は、その記憶が今の予定や望みを表しているか、そして話すならどの強さで、何回まで話すかである。本稿の問いは「生活を覚えるAIは、いつ声をかけ、どの程度の支援を出すべきか」である。新しいモデルの発明は目的にせず、既存研究の部品で何ができ、どこが未検証かを整理する。

MemoProAgent：雑多な履歴から根拠を探す

MemoProAgent [1]は、人から質問される前に、一人称映像と長期記憶から話す時機と内容を選ぶエージェントである。これまでの能動支援の課題では、短いプロフィールや目標をモデルに直接渡すことが多かった。この研究は、雑多な個人履歴の中から関連する根拠を探す課題に置き換えた点に特徴がある。記憶の保存と検索には、既存の記憶グラフ（出来事や選好を節点と関係で表す構造）や検索技術を使う。

データはEgo4D由来の動画と合成した個人履歴を組み合わせたもので、構成は表1のとおりである。履歴は平均362日分あるが、実在の1,804人を一年間追跡したものとは性質が異なる。注釈者が確かめたのは発話を認める時間窓と根拠の妥当性で、本人がその声かけを望むかどうかは調べていない。

項目	値
組数（訓練/開発/評価）	1,804（904/100/800）
履歴の長さ	平均362日（330〜365日）、約618Kトークン
関連する根拠の割合	履歴全体の0.19%
生成候補→1人以上がレビュー	2,168→2,153
注釈者と判断数	訓練済み32名、5,982判断（1,889標本は3名でレビュー）
採択	1,804（レビュー対象の83.8%）

表1。MemoProAgentのデータと注釈 [1]

処理は四段階で進む。場面に映る物体を認識し、出来事の記憶と持続的な選好・制約を含むグラフから候補を検索する。次に介入ゲートが「今の行動につながる未解決の事柄」を選び、最後に生成器が話すか黙るかを決める。生成器には過去の発話履歴も渡し、同じことの繰り返しを抑える。記憶には時刻と出典が付いており、後の履歴で解決済みと明確に読み取れる事柄はゲートで除かれる。検索で似た記録が見つかっただけでは話さない

設計である。ただし、履歴に現れない解決や希望の変更をどれだけ見落とすかは、現実の場面で検証されていない。

「一度だけ適切に話す」という採点

評価では、各正例クリップに発話を認める窓を一つ置く。窓は対象物が最初に映った時点を基準に作り、人が妥当性を確認した。8秒以下の短い視線外れは同じ出現として扱い、窓の終わりには2秒の許容幅を設けている。

主指標のPerfectは、根拠に沿う適切な発話を窓の中でちょうど一度だけ行い、ほかには何も話さなかったクリップの割合である。Hitは適切な発話が一度でもあった割合を指す。Over-triggerは一つのクリップで複数回話した割合で、警告の誤りを数える指標とは性質が異なり、発話の多さを測っている。発話内容が適切か（Good）の判定はGPT-4oが行う。窓と根拠は注釈者が検証したが、出力内容の主な採点はLLMに任されている。

ベンチマークとして、この設計は筋が通っている。正解の窓が一つなら、一度で言い当てて黙るのが最良になるからである。ただし「一度」はこの評価のために置いた条件で、ケア全体で最適な回数を示すものとは言えない。この点は論評の節で扱う。

結果：根拠探しと時機選び、二つの難所

主評価は、Qwen3.8-27Bを基礎モデルとし、直近4フレームを見て2秒ごとに判断する設定（windowed）で、評価用の800例を使った。標準RAG（検索した記録を生成器に渡す一般的な方式）については、検索件数k=1、2、5、10、30、50のうち最良の成績が載っている。

条件	Perfect	Hit	Over-trigger	発話数/クリップ
標準RAG	7.0%	42.5%	67.5%	3.58
harness（追加学習なし）	37.5%	54.6%	22.3%	1.06
harness + GRPO	42.4%	54.0%	13.5%	0.90

表2。windowed主評価800例の主な結果 [1]

harnessは、記憶検索、介入ゲート、発話履歴をまとめた枠組みである。追加学習なしでもPerfectを大きく引き上げた。そこにGRPO（複数の出力を相対的に比べて方策を更新する強化学習の一手法）を加えると、Hitはほぼ変わらず、Over-triggerが下がった。成績の伸びは、言い当てる回数が増えたことより、言いすぎが減ったことによるところが大きい。設定の異なるストリーミング評価でも、同じ3条件のPerfectは1.0%、33.3%、40.6%の順だった。

150例で部品を外した実験では、harness全体の37.3%に対し、ゲートを外すと24.0%、発話履歴を外すと8.0%に下がった。反復を抑える発話履歴の寄与が特に大きい。

難所を切り分けた診断実験が表3である。GPT-5.3-Chatに、根拠と時点の与え方を変えて解かせた。

与えた根拠	与えた時点	Perfect
正解の根拠	正しい窓の中だけ	96.0%
正解の根拠	毎時点（時機の判断は任せる）	38.0%
通常の検索結果	窓の中だけ	15.3%
通常の検索結果	毎時点	8.0%

表3。診断実験150例（GPT-5.3-Chat）[1]

正しい根拠を正しい時点に渡せば96.0%に達する。ただしこれは課題を分解するために作った計算上の条件で、通常の運用でこの成績が出るという意味ではない。正しい根拠があっても時機の判断を任せると38.0%まで落ち、検索結果が不正確でも大きく落ちる。根拠探しと時機選びは、それぞれ単独で成績を大きく左右している。

根拠となる記憶を除いた負例150クリップでは、何か話してしまった割合がRAG 77.3%、harness 15.3%、GRPO 14.0%だった。分母は正例の800例ではなく、この負例150クリップである。別の動画データEPIC-KITCHENS（100例）でもharness 41%、GRPO 49%と報告されているが、実際に暮らす人の生活で確かめたものではない。

学習の報酬と内容の採点には、同じGPT-4oが使われている。著者は、Hitが横ばいのまま反復だけが減ったことから、採点器への迎合は起きていないと解釈する。筋の通った解釈だが、独立した採点で迎合を排除したわけではない。訓練には904クリップから作った3,465時点を使い、H100を7基で14.6時間かけた。公開コード [3]（HEAD 8563e58、2026-10-22）は推論と評価が中心で、学習済みのチェックポイントや全再現一式は含まない。ライセンスはCC-BY-NC-4.0で非商用に限られ、Ego4D動画には別の許諾が要る。履歴データfacebook/MemProBenchは確認した時点で非公開で、利用には権限が必要だった。本稿の筆者はコードを実行しておらず、数値も再現していない。実運用での通信遅延、総費用、長く使ったときの受け入れられ方も評価されていない。

著者の結論は次のとおりである。記憶に基づく能動支援の難所は、まばらな根拠を探すことと話す時機を選ぶことにあり、介入ゲートと発話履歴を含む枠組み、さらに追加学習によって両方を改善できる。

COACH：見えない状態を確率で推定し、手助けを段階化する

COACH [2]は、認知症のある高齢者の手洗いを音声と動画の声かけで支えるシステムである。2008年の研究で古いが、声をかける時機と強さを確率的に選ぶ既存モデルの有力な先例である。

手洗いは五つの必須段階に分けられている。天井のカメラが手とタオルの位置を追い、その情報から今どの段階にいるかを確率で推定する。行動の選択にはPOMDP（部分観測マルコフ決定過程：直接見えない状態を推定し、行動の結果を見込んで次の行動を選ぶ既存の意思決定法）を使う。主な選択肢は、待機、短い音声、詳しい音声、音声と動画による促し、介護者の呼出しである。課題の進み具合や本人の反応のしやすさも確率で扱う。ただし前の版から視覚追跡、POMDP、動画などが同時に変わっており、この試験ではPOMDP単体の効果を切り分けられない。認知症の程度を診断する能力も確立していない。

試験はトロントの長期ケア施設で行われた。職員が候補を挙げ、65歳超、英語を話す、通常の会話が聞き取れる、重度の運動障害がないなどの条件を満たす8名を募集した。健康上の問題と攻撃的な行動で2名が離脱し、6名（中等度5名、重度1名）を解析した。代諾同意と倫理審査を経ている。8週間、平日に1日1回、各相10試行、1人あたり計40試行を計画し、A1-B1-A2-B2の固定順で実施した。Aは人による通常の支援、BはCOACHによる支援である。B相では研究者がCOACHを起動し、介護者は廊下から見守って、呼ばれたときや安全上必要なときに介入した。参加者は毎回車椅子で決まった場所へ誘導されており、評価は人がすぐに介入できる施設内の環境で行われた。実施年月は報告されていない。

評価指標は三つある。一つ目は、人の手助けなしに完了した段階の数（0〜5）で、機械の声かけで達成した段階も「独立」に数える。二つ目のFASは、手洗いの各段階を0〜7で採点し、合計35点とする尺度で、FIM（日常生活の自立度の標準的な評価）の全体を使ったものとは異なる。三つ目の人とのやり取りの回数は、介護負担の代わりの指標として使われた。心理的な負担の尺度や総労働時間は直接測っていない。熟練した研究者が

動画を採点し、全体の20%を第二の評価者が採点した。評価者間の一致度κは、独立0.96、やり取り0.69、FAS 0.63だった。解析は各相の時系列を視覚的に比較する方法で、正式な質的テーマ分析やインタビュー分析は行っていない。

COACHの結果と数字の照合

参加者	程度	独立した段階数	人とのやり取り
S1	重度	0.38→1.33	15.19→15.56
S3	中等度	4.95→5.00	0.11→0
S4	中等度	4.10→5.00	1.95→0
S5	中等度	3.60→4.53	4.10→1.33
S6	中等度	5.00→5.00	0.15→0
S8	中等度	4.57→5.00	2.05→1.94

表4。参加者別の変化（A→B、表6・7より）[2]

中等度の5名では、独立した段階数の平均が4.4から4.9に上がり、著者は11%の増加と報告した。重度のS1の独立工程数は0.38から1.33へ増えたが、人とのやり取りは減っていない。S8は表の上では5.00に達している。しかし本文には、石鹸の声かけを拒み、人の促しにだけ応じ、介護者の介入も残ったと書かれている。著者は「4人が完全に独立した」と述べるが、本稿はこれをそのまま繰り返さず、指標と記述の不整合として残す。

声かけの判断は750の観察条件に分類されている。同じ人の同じ試行に複数の記録が含まれるため、この数は人数や独立した事例の数を表していない。

分類	件数	その後の反応
Hit（必要なときに促した）	147	完了30（20.4%）、無視117
Miss（必要なのに促さなかった）	18	–
False Alarm（不要なのに促した）	152	反応16（未完了の試みを含む）、無視136
Correct Reject（正しく黙っていた）	433	–

表5。声かけ判断の分類（表9・図7） [2]

妥当な促しと正しい沈黙を合わせると580/750=77.3%になる。一方で、必要なときの促しのうち完了まで至ったのは20.4%にとどまった。判断の妥当さと、本人がその声で動けるかどうかの間には大きな開きがある。

報告箇所によって数値が食い違うところがあるため、本稿での扱いを表6にまとめた。

項目	論文内の記載	本稿の扱い	理由
正しい沈黙	表9・図7は433、本文p15は443	433を採用	表と図が一致し、4分類の合計が750になる
妥当な判断の割合	要約78%、本文77%	77.3%（580/750）	表の値から再計算
人とのやり取りの減少	要約60%、本文・表5は66%	一つに統一しない	表7の中等度5名の単純平均では約60.9%減。集計の過程と原データが不明
石鹸の誤った促し	107件を全条件の8%と記載	主要な数値に使わない	107/750は14.3%で、分母を確認できない

表6。COACHの数値の照合 [2]

減少率の解釈は保留し、表4の参加者別結果を判断材料とする。

観察の記述にも重要な示唆がある。不要な促しは、いらだちや混乱を招くことがあった。済んだ動作をやり直してしまい、追加の援助が必要になった場面もある。S8は促しを理解したうえで断った。著者は、MMSE（認知機能の簡易検査）の点数だけでは反応の違いを説明できないとし、注意、身体能力、気分、好みなどが影響した可能性を挙げている。ただしこれは少人数の観察から出た推測で、一般的な因果関係として扱うことはできない。

技術的な不具合も具体的に記録されている。完了した段階の推定が時間とともに「未完了」へ戻ってしまう推定のずれ（drift）が起き、手を乾かしている最中に石鹸を再び促すことがあった。石鹸の誤った促し107件のうち59件は、その段階が終わった後に起きている。著者は終了判定を改善する予定だと述べているが、改善の効果は示されていない。待つ時間についても、本人が自分で再開する助けになった場合と、待ちすぎて修正が間に合わなかった場合の両方が報告されている。

限界も多い。固定順の小規模な試験で、支援者の位置や行動も相ごとに変わった。慣れや日々の体調の変動も影響しうる。人をすぐ呼べる研究環境で、天井カメラ、21インチ画面、PC、専任の介護者、研究者を必要とした。導入費、保守費、続けられるかどうかの比較もない。著者は、中等度の認知症のある人と介護者を支える可能性を示したと結論し、個別化の必要性と、家庭での大規模な試験を次の課題に挙げた。一般的な介護負担の軽減や人員削減の効果までは示していない。

批判的論評：覚えていることと、いま声をかけてよいこと

第一に、評価の到達点を先へ延ばす必要がある。MemoProAgentでの成功は、注釈された窓の中で根拠に沿って一度話すことであり、本人がその声で行動できたか、それを望んだかは測っていない。COACHでは、必要なときに出した促しでも完了につながったのは約2割だった。COACHには、不要な繰り返しが混乱を招いた場面が記録されている。一方、聞き逃した人への言い直しは、有益な支援となる可能性がある。これは本稿が追加で検証すべきと考える場面である。このような場面へ評価を広げる際には、必要な言い直しまで一律に減点しない設計が求められる。

第二に、記憶が正しくても、今の予定や望みは変わる。MemoProAgentは、後の履歴で解決が明確な事柄を除くが、取消や拒否がはっきり記録されない場合の振る舞いは確かめていない。COACHで起きた、済んだ段階を再び促す不具合は、状態管理の失敗がそのまま本人の混乱とやり直しにつながることを示す具体的な警告である。必要なのは新しい推論機構よりも、「解決済み」「取消」「拒否」を記録して以後の判断に反映する終了状態の管理かもしれない。既存の方法で足りる可能性を先に確かめるべきである。

第三に、本人が理解したうえで断る反応を評価に残す必要がある。S8の観察を踏まえると、達成した工程数だけでは本人の意思を尊重できたかが分からない。評価を始める前に、拒否を尊重する条件、重大な危険があるときに人へ相談する条件、声かけや見守りの負担を誰が担うかを定義しておく必要がある。これは評価の設計についての提案であり、臨床の手順として勧めるものではない。

既存部品で組めるか

両研究には、組み合わせの候補になる部品がある。ただし、二つの論文のどちらもこの組み合わせを試していない。

部品	入力→出力	使えそうな点	証拠と制約	未検証の点
出典付き記憶（MemoPro）	履歴→時刻・出典付きの記録	根拠をたどれる	合成履歴、CC-BY-NC、データの利用に権限が必要	実生活の記録の誤りや欠け
介入ゲート（MemoPro）	候補→未解決で関連する事柄	似ているだけでは話さない	外すと37.3%→24.0%	明示されない取消や拒否
発話履歴（MemoPro）	過去の発話→反復の抑制	言いすぎを減らせる	外すと8.0%	言い直しが必要な場面
確率的な状態推定（COACH）	映像→各段階の確率	不確かさを扱える	小規模な試験、推定のずれ	長期記憶との接続
段階的支援（COACH）	状態→待つから動画まで	支援の強さを調整できる	POMDP単体の効果は不明	拒否されたときの振る舞い
人への接続（COACH）	状態→介護者を呼ぶ	必要時に介護者を呼べる	研究者と専任介護者がいる環境	家庭での負担の所在

表7。既存部品の候補と制約

Unknown Labの提案は、いきなり独自のモデルを増やさないことである。まず既存の方法を、同じ情報、同じ計算資源、同じ相手の反応条件のもとで比べる。それでも足りない部分が残ったときにだけ、小さな追加を試す。

ETIC 2.0への含意

ETIC 2.0 [4]は、生活の中の活動、人、環境、関係から技術の価値を捉える枠組みである。今回の二つの論文はETICを直接検証していない。以下は、論文の知見をETICの観点から読み替えた含意である。

考え方と重なる点

COACHは手洗いを段階に分け、本人の反応や介護者の位置まで含めて支援を設計した。技術の価値を活動と関係の中で捉えるという点で、ETICの考え方と重なる。MemoProAgentの介入ゲートも、記憶の量より、今の行動につながるかどうかで話す価値を決めており、同じ方向を向いている。

答えるべき問い

ETICで技術の価値を論じるとき、声かけの回数と強さをどの単位で評価するのか。本人の拒否を、活動ができなかったこととして数えるのか、意思の表明として数えるのか。支援者の負担を、やり取りの回数のような代わりの指標で十分とみなすのか。これらに答える必要がある。

具体的な見直し案

評価項目に、解決済み・取消・拒否の扱い、言い直しを認める条件、人へつなぐ条件、負担を誰が担うかを明記することを提案する。この見直し案も未検証であり、次の節の検証で必要性が示されなければ採用しない。

検証案と、提案を縮小・撤回する条件

次の検証は、公開データや合成した場面から始める。「買う予定を撤回した」「支援を断った」「声かけを聞き逃した」「すでに完了した」「危険を伴う作業が未完了」という五つの状況を、段階的に難しくして与える。開発に使う場面と最終評価に使う場面は分け、最終評価の結果を見て設計を調整しない。対照には、同じ基礎モデル、同じ利用可能な情報、同じ予算で組んだ、よく整えた標準方式（検索と発話履歴を備えたもの）を置く。

見る指標は、声かけの回数、課題の完了、不要なやり直し、拒否の尊重、見逃し、本人と支援者の負担である。これらは一つの合成点にまとめない。人を対象とする研究はまだ行っていないため、本人と支援者の負担は当面、場面の設計に組み込んだ代わりの指標でしか見られない。この制約が、結果をどこまで解釈できるかの範囲を決める。

提案を撤回するのは、終了状態を明示しても、標準対照より不要な声かけややり直しが減らない場合である。縮小するのは、終了状態の管理だけで同等の成績になり、確率的な状態推定や段階的支援を加えても差が出ない場合で、そのときは追加した要素を削る。変更が不要になるのは、よく整えた標準対照がすべての状況で同等に振る舞う場合で、その場合は追加機構の導入を見送る。拒否や負担の記録は、標準方式の妥当性を確かめるためにも維持する。

結論と次に読む論文

MemoProAgentの著者は、能動支援の難所を根拠探しと時機選びに分け、介入ゲートと発話履歴によって言いすぎを大きく減らせることを示した。COACHの著者は、確率的な状態推定と段階的な声かけが中等度の認知症のある人の手洗いを支えうると結論し、個別化と家庭での試験を今後の課題とした。二つを並べると、「いま声をかけてよいか」「断られたらどうするか」が、記憶の活用に伴う具体的な研究課題として見えてくる。Unknown Labは、既存の部品と状態管理で足りるかを先に確かめることを提案するが、これはまだ検証していない。

次に読む候補はLiら [5]である。今回は書誌情報と抄録の確認にとどまり、数値の根拠には使っていない。対話の好みの記憶とその変化を扱う研究であり、言い直してよい条件や、話しかけ方が変わったときの扱いという、本稿に残った問いにつながる可能性がある。

参考文献

[1] Zeng S, Yang X, Ma T, et al. MemoProAgent: Toward Memory-Driven Proactive Assistance. arXiv:2610.22671v2 [preprint, 未査読]. v1: 2026-10-15; v2: 2026-10-21. https://doi.org/10.48550/arXiv.2610.22671（HTML版: https://arxiv.org/html/2610.22671v2）

[2] Mihailidis A, Boger JN, Craig T, Hoey J. The COACH prompting system to assist older adults with dementia through handwashing: An efficacy study. BMC Geriatrics. 2008;8:28. https://doi.org/10.1186/1471-2318-8-28

[3] facebookresearch. Memo-Pro-Agent [source code]. GitHub. Commit 8563e58d3106c8e0b3108bff64c04d00e1c8f095 (2026-10-22). CC-BY-NC-4.0. https://github.com/facebookresearch/Memo-Pro-Agent

[4] Morimoto D. エコテクノロジカル・インテグラティブ・ケア(ETIC)2.0, Version 2.1改訂版: AI時代の作業療法とケア生態系設計. Unknown Lab; 2026. https://doi.org/10.5281/zenodo.21848456

[5] Li J, Chen Z, Luo H, Salam H. Why Bother Me Again? Benchmarking Personal Agent Memory for Interaction Preferences. arXiv:2610.30485v1 [preprint]. 2026-10-21. https://doi.org/10.48550/arXiv.2610.30485

注：本稿はAIの支援を受けて作成した未査読の批判的論評であり、独立した専門家によるレビューは経ていない。個別の臨床利用を推奨するものではない。