

HDTwin: Building a Human Digital Twin using Large Language Models for Cognitive Diagnosis

Gina Sprint¹, PhD; Maureen Schmitter-Edgecombe², PhD; Diane J. Cook², PhD

¹Gonzaga University, Spokane, WA, USA

²Washington State University, Pullman, WA, USA

Corresponding Author:

Diane J Cook, PhD

School of Electrical Engineering & Computer Science

Washington State University

Pullman, WA 99164-2752, USA

djcook@wsu.edu

Abstract

Background: Human digital twins have the potential to change the practice of personalizing cognitive health diagnosis because these systems can integrate multiple sources of health information and influences into a unified model. Cognitive health is multifaceted, yet researchers and clinical professionals struggle to align diverse sources of information into a single model.

Objective: In this paper, we introduce a method called HDTwin, for unifying heterogeneous data using large language models. HDTwin is designed to predict cognitive diagnoses and offer explanations for its inferences.

Methods: HDTwin integrates cognitive health data from multiple sources, including demographic, behavioral, EMA, n-back test, speech, and baseline experimenter testing session markers. Data are converted into text prompts for a large language model. The system then combines these inputs with relevant external knowledge from the scientific literature to construct a predictive model. The model's performance is validated using data from three studies involving n=124 participants, comparing its diagnostic accuracy with baseline machine learning classifiers. The source code for HDTwin is available at <https://github.com/WSU-CASAS/HDTwin>.

Results: HDTwin achieves a peak accuracy of 0.81 based on automated selection of markers, significantly outperforming baseline classifiers. On average, HDTwin yielded accuracy=0.77, precision=0.88, recall=0.63, and Matthews correlation coefficient (mcc)=0.57. In comparison, the baseline classifiers yielded average accuracy=0.65, precision=0.86, recall=0.35, and mcc=0.36. The experiments also reveal that HDTwin yields superior predictive accuracy when information sources are fused compared to single sources. HDTwin's chatbot interface provides

interactive dialogues, aiding in diagnosis interpretation and allowing further exploration of patient data.

Conclusions: HDTwin integrates diverse cognitive health data, enhancing the accuracy and explainability of cognitive diagnoses. This approach outperforms traditional models and provides an interactive interface for navigating patient information. The approach shows promise for improving early detection and intervention strategies in cognitive health.

Keywords: digital twin, smartwatch, digital behavior markers, large language models, text analysis, machine learning

Introduction

Mild cognitive impairment (MCI) is a transition state between healthy aging and dementia. Digital health technologies can enhance early detection and improve the ecological validity of traditional MCI diagnostic assessments. Current technology-assisted approaches often focus on a small set of data sources, such as speech and text ^{1,2}, mobile tests ^{3,4}, self-reported in-the-moment states ⁵, and digital behavior markers ⁶⁻⁹. Each contributes valuable insights, but they are fragmented.

We propose the construction of a human digital twin that utilizes these diverse pieces of digital information to form a more comprehensive model of an individual. This digital twin integrates data from multiple sources, recorded at different times and in real-world settings. The result offers a holistic view that enhances early diagnosis of cognitive impairment and facilitates timely interventions to slow progression, in line with the quest for precision health.

Designing digital twins faces the challenges associated with merging data that differ in acquisition times, devices, formats, and fidelity. Clinicians also need help navigating these types of information with traditional dashboards ¹⁰. To overcome these challenges, we propose a system called HDTwin that uses large language models (LLMs) to create a cohesive digital twin from heterogeneous data sources. In this paper, we detail the design of HDTwin and evaluate the system in the context of automating cognitive health diagnosis for 124 participants from three studies. Specifically, we highlight the system's ability to perform information retrieval, data fusion, and inference explanations. We demonstrate that HDTwin successfully integrates traditional machine learning models, numeric reasoning resources, scientific literature, audio recordings, and Ecological Momentary Assessment (EMA) responses to generate diagnosis predictions that are comparable or superior to traditional ensemble classifiers. We further investigate the ability of HDTwin to process information and articulate clear diagnostic explanations interactively, providing a bidirectional flow of information between a clinician and the computational model.

Related Work

Recent advances in language models have tremendously impacted health question answering and information retrieval. These prior works focus on leveraging specialized corpora for enhanced performance. Language models are trained on biomedical texts to refine their capabilities in summarizing documents and answering complex health-related questions^{11,12}. Models have been further enhanced by efforts like KeBioSum¹³, which integrates medical knowledge into model training to improve response accuracy.

Language models also support health prediction. As an example, AD-BERT¹⁴ processes electronic health record (EHR) notes with pretrained models to forecast a patient's progression from mild cognitive impairment to Alzheimer's disease. Research by Asgari et al.¹⁵ leverages text markers from recorded speech to predict mild cognitive impairment. Jiang et al.¹⁶ train an LLM on medical language to predict hospital readmission, and Kim et al.¹⁷ evaluate prompting strategies for LLMs on a variety of health prediction tasks. Because LLMs can inherently predict future state, Xue and Salim¹⁸ explore pretrained LLMs to predict future temperature, electricity consumption, and movement trajectories. Similarly, Sprint et al.¹⁹ demonstrated that LLMs could anticipate future health states based on past EMA reports and sensor-based behavior data.

Recently researchers have extended LLM capabilities to interpret non-textual data inputs. Yu et al.²⁰ direct LLMs to diagnose sleep apnea and cardiac conditions by leveraging large databases. Jin et al.²¹ convert time-series data into text to forecast electricity usage. Partnering principal component analysis (PCA) with text reports, as explored by de Zarza et al.²², enhances predictive accuracy for forecasting weather and traffic volume.

The next step in the evolution of LLMs for health diagnostics involves fusing diverse information sources. Girdhar et al.²³ aligned video, text, and audio by creating unified image embeddings. Xu et al.²⁴ paired images with radiology reports. While these prior efforts illustrate the potential of LLMs to synthesize information across modalities, Cascella et al.²⁵ caution that LLMs still face challenges in aligning personal and general information sources effectively.

This paper aims to contribute to the evolving landscape by exploring the use of LLMs to create a human digital twin from diverse information sources. First, we consider a novel integration of digital behavior markers into the language model that are collected from continuous sensor data. Second, we enhance the language model for cognitive health domains by incorporating self-report, traditional clinical assessment, and automated performance scores. Third, we investigate whether LLMs can offer an effective mechanism for creating a digital twin from these varied components that enhances the accuracy of cognitive health diagnosis and the explainability of system inferences.

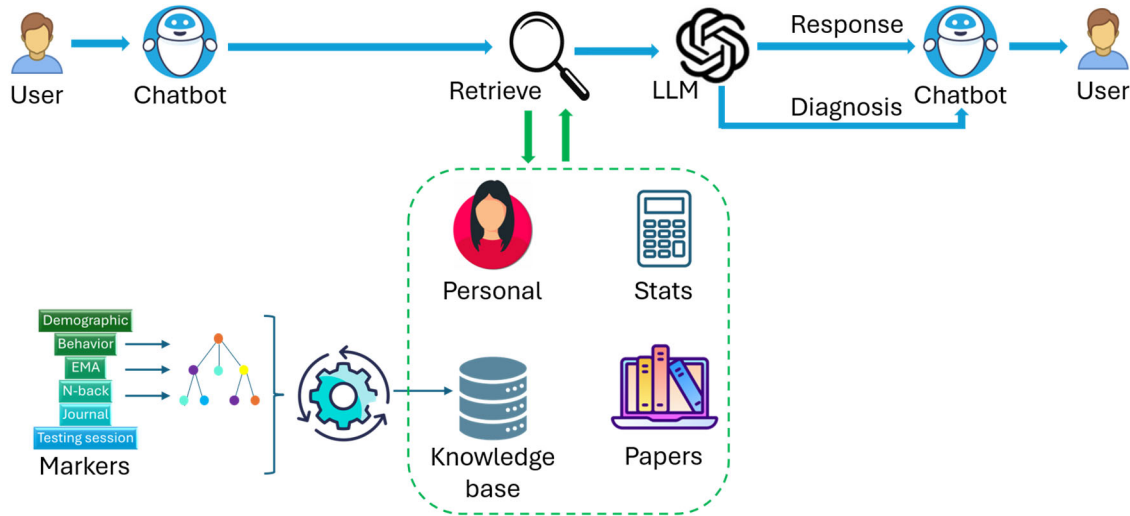


Figure 1. HDTwin information processing pipeline. A user interacts with the LLM interface to request summary information about a person or a suggested diagnosis. Based on the query, HDTwin retrieves personalized markers together with paper abstracts and data from a knowledge base that informs a response. The query response is presented to the user, supporting an ongoing conversation about the person or explanation of the query response.

Methods

HDTwin LLM

We built HDTwin using langchain and OpenAI’s GPT-3.5-turbo-0125 language model. As with other GPTs, the model is based on a transformer architecture and uses a self-attention mechanism to aid in capturing dependencies and context within the text. HDTwin’s processing pipeline is illustrated in Figure 1. Using custom langchain tools, HDTwin retrieves information from personal data, statistical summaries, paper abstracts, and a knowledge base to input as prompts to a large language model, which then generates output for the user in response to a query or diagnosis request. The knowledge base incorporates diverse personal markers which are prompt engineered for input to the language model.

We validate HDTwin using data from three studies. Participants in these studies were 124 independent-living older adults (mean age=70.51; 72.4% female). Participant recruitment and screening were similar across the studies. Recruitment included community health and wellness fairs, TrialMatch, advertisements on social media, physician referrals, and online posts. Inclusion criteria were age 50+ and have the ability to speak English; exclusion criteria included current psychoactive substance use; significant auditory visual, or cognitive impairment; presence of a psychiatric, neurologic or medical condition that greatly attributed to cognitive complaints; and Telephone Interview for Cognitive Status (TICS) ²⁶ score < 26. Participants provided informed consent, and the studies were approved by the

Washington State University Institutional Review Board. Source code for HDTwin and a video demonstration of the chatbot interface are available online ^{27,28}.

Each participant was assigned a fictitious name, sampled from a repository ²⁹. This step was performed to anonymize references to the names that appeared in processed text. Participants were categorized as cognitively healthy older adults (HOA, n=75) or older adults with mild cognitive impairment (MCI, n=49). To perform these categorizations, at study baseline interviews were conducted, questionnaires were completed (e.g., Patient-Reported Outcomes Measurement Information System ³⁰), and standardized neuropsychological tests evaluating the cognitive domains of memory, language, executive functioning, and attention (three scores per domain) were administered. These included the Wechsler Adult Intelligence Scale – Fourth Edition ³¹ Digit Span Forward and Backward subtests, the Delis-Kaplan Executive Function System ³² Category Switching test, the Five Point Test ³³, the California Verbal Learning Test ³⁴, and self-reported measures from the Patient-Reported Outcomes Measurement Information System ³⁰. Jak/Bondi criteria ³⁵ were followed to classify individuals as MCI. These participants were primarily single domain (80%), and most met the criteria for amnesic MCI (78%).

Ethical Considerations

These studies were reviewed and approved by the Institutional Review Board at Washington State University. To participate in any of the studies, participants needed to sign an informed consent; each person received compensation between \$55 and \$120 for their participation, consistent with the time demands of the study. All data were anonymized before performing analyses.

Data

Participant data stem from multiple numeric and text-based marker sources. Language models currently struggle with numeric reasoning for real-valued variables. HDTwin includes agents to summarize and learn models from raw numeric data. To handle the cases where prompts are fed directly to the LLM, however, we transform each real-valued marker to a 0-10 integer scale.

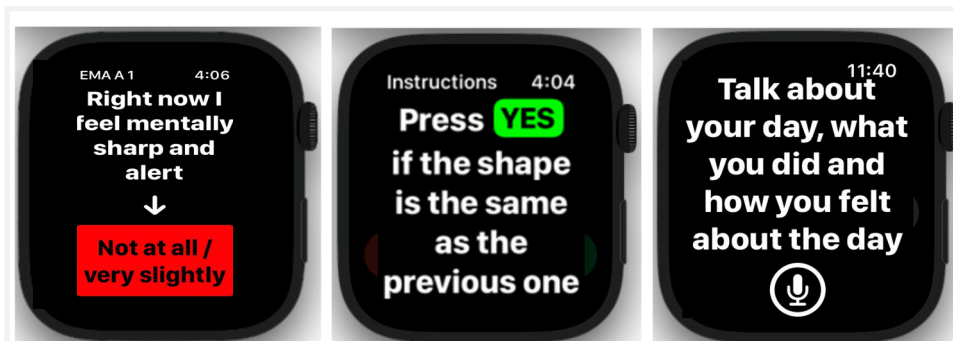


Figure 2. In addition to collecting sensor data, the smartwatch app queries the user for their current state, includes an n-back shape test, and collects daily audio data.

Demographic markers

Each participant's age, sex, and number of education years were included as markers.

Behavior markers

All participants wore a smartwatch (Apple Watch) daily for a minimum of two weeks. The watches continuously collected acceleration, rotation, and location data at 10Hz. From the location coordinates, we defined the participant's home as the most-frequent visited location among the first 300 readings each day. From these data, we extracted activity level (estimated as total acceleration) and distance from home. These values were aggregated by day, then we calculated mean and variance over the entire data collection period. The missing data rate was 14%, and missing entries were not included in the calculations. Figure 2 shows screenshots of the smartwatch app.

EMA markers

Ecological momentary assessment (EMA) responses were collected 4x/day at random points within specified time windows. During this session, each participant responded to the prompt "Right now I feel mentally sharp" on a Likert scale of 1 ("Not at all") to 5 ("Extremely"). For each person, we extract the mean and variance for the EMA response value and response rate / compliance.

N-back markers

Prior research indicated that the *n*-back task, delivered via a mobile tablet, can capture cognitive capacity for older adults as it is influenced by fatigue, mental sharpness, and the environment⁹. We adapted this *n*-back shape test to the smartwatch. Three shapes (circle, square, and diamond) were displayed on the watch screen, and participants indicated whether the current shape was the same as the prior. We computed accuracy for each 45-second task.

Earlier studies reveal the importance of considering *n*-back performance in terms of the learning phase (when scores start low but increase sharply) as well as characteristics of performance over the entire sampled period. After the learning phase, daily performance varies with fatigue, mood, mental sharpness, cognitive changes, and environmental factors^{36,37}. Applying linear regression to the sequence of daily scores, we extracted the slope for the first 6 scores (the learning rate). We also computed the overall score mean and standard deviation.

Speech markers

Each day, participants provided a verbal description of their day in response to the prompt, "Talk about your day, what you did, and how you felt." Participant responses were collected by the smartwatch. The audio files were then converted to text and fed verbatim to the language models. A total of 2,995 audio files were provided by 85 participants across the three studies. All descriptions from a single participant were aggregated into one text entry per person.

Testing session markers

From the administered cognitive assessment tests that were delivered orally at study baseline, we converted the recorded testing session audio files to text. Based on the experimenter script, we input experimenter questions as prompts to the LLM. The purpose of this step was to have the language model quantitatively evaluate participant responses based solely on the recorded interaction. This contrasts with the traditional method of using a predefined set of rules to determine the score for each assessment task. There were 23 types of participant responses, generated from tests for phonemic and semantic fluency, list learning and delayed memory, cognitive screening, and health literacy.

As shown in Table 1, we asked the LLM to extract and rate participant responses on a scale of 1 (poor) to 5 (excellent) or missing. While some prompts included a list of correct answers (e.g., “where is our lab located?”), many did not (e.g., “list all of the animals you can think of”). In the second case, response rating consequently relied on the model’s reasoning. Of the 124 baseline testing session transcripts, 116 had at least 3 non-missing responses and were input to the LLM.

Table 1. Example prompt for processing a baseline testing session data. The user provides the system message and prompt; HDTwin generates the corresponding response.

System message	You are a helpful assistant that extracts and rates participant responses to interview questions.
Example prompt	<p>The following is an unlabeled transcript of an interview between an interviewer and a participant: "all right so the recordings on again for the next task I'm going to"</p> <p>There should be a part of the transcript where the interviewer asks the participant to list as many words that start with the letter S as they can think of. Using only exact quotes from the transcript, provide the participant's response to the interviewer's question. Then, using a scale of 1 (poor) – 5 (excellent), rate the participant’s ability to think of words that start with S. Briefly explain your rating. If you cannot find the participant’s response, respond only with: Could not find the participant’s response.</p>
Example response	<p>Participant's Answer: Sun, stairs, startled, stars, stilt, stern, south, song</p> <p>Rating: 5</p> <p>Explanation: The participant was able to list multiple words that start with the letter S, demonstrating a strong ability to recall words.</p>

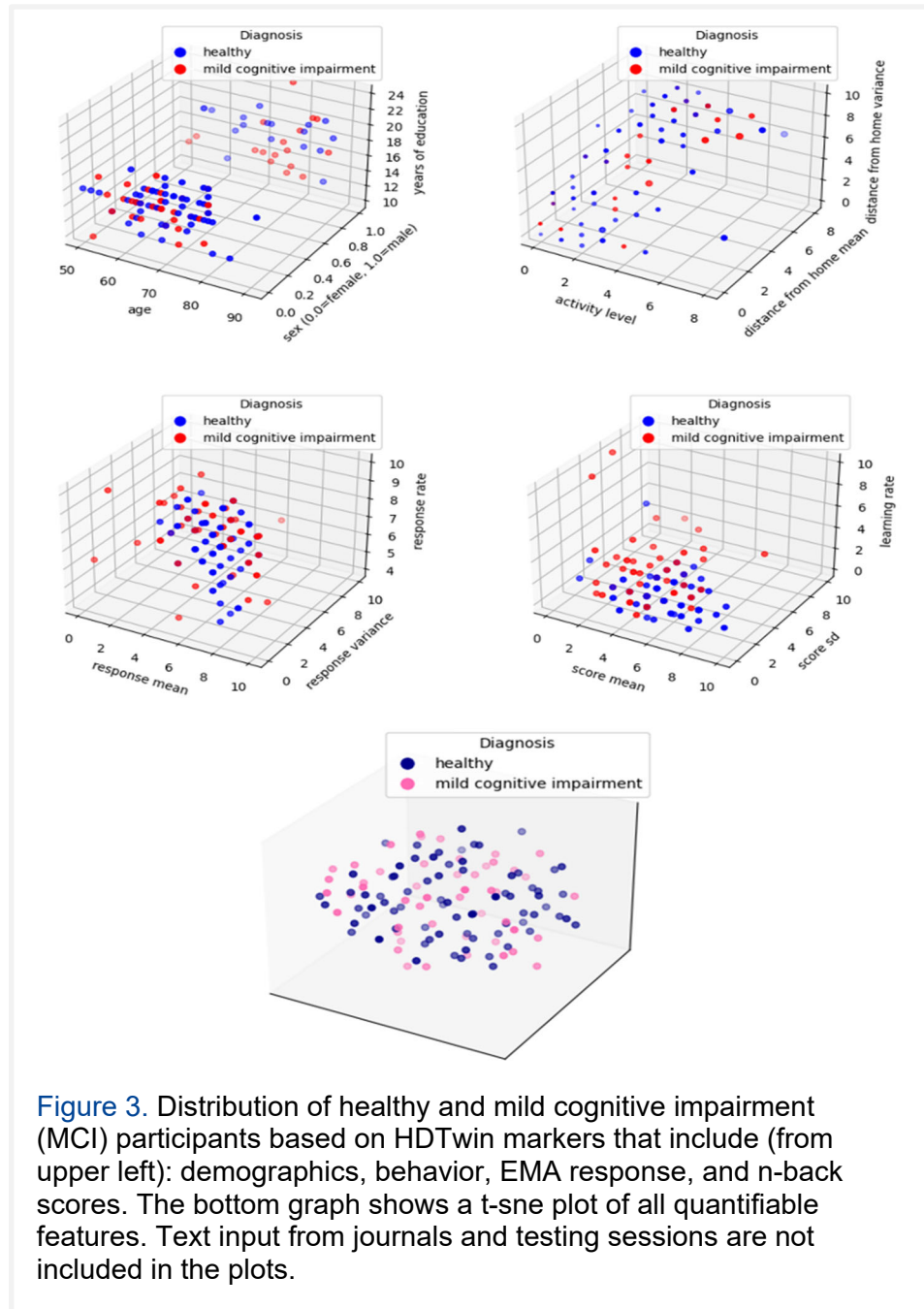


Figure 3. Distribution of healthy and mild cognitive impairment (MCI) participants based on HDTwin markers that include (from upper left): demographics, behavior, EMA response, and *n*-back scores. The bottom graph shows a t-sne plot of all quantifiable features. Text input from journals and testing sessions are not included in the plots.

Figure 3 illustrates the distribution of participants between the healthy and mild cognitive impairment categories based on each type of quantifiable marker (i.e., demographics, behavior, EMA response, and *n*-back score). As can be seen from the plots, the categories are highly intermixed for each category of marker as well as the t-sne combination of these markers. The high degree of intermixing between categories indicates that there are no clear, simple class boundaries separating the groups, making diagnosis prediction more complex. This complexity necessitates leveraging all available information to improve prediction accuracy, which requires

integrating information with diverse modalities, granularities, and availability. This challenge motivates the adoption of an LLM-based method for this task, which is well-suited to handling and synthesizing such heterogeneous data.

Knowledge Base Construction

To predict cognitive diagnoses with LLMs, we constructed a prompt template that combines two primary sources of context to be combined by the LLM:

1. Insights gleaned from external knowledge about the field, and
2. Insights gathered from personalized data markers.

We utilize these information sources by transforming them into text prompts for the LLM. To integrate insights from the field, each prompt is framed based on a finding from a relevant publication. Language models have demonstrated their ability to summarize medical research articles ³⁸. In contrast, we use such research articles as just one piece of the digital twin to provide more comprehensive reasoning about a person’s state. In this context, we introduce a novel use of such articles, leveraging them as sources of external neuropsychology insights. Table 2 provides a summary of the relevant literature sources we included for HDTwin.

Table 2. Rules found in external information sources. These rules represent summaries of statements found in the included research articles.

More hours spent outside the home and greater distance from home were associated with better cognitive function. ⁶
Physical exercise demonstrated a protective factor for MCI; was associated with higher mean semantic memory and executive function. ^{8,39}
Association was found between participant sentence complexity and levels of cognitive impairment. ²
Individuals with MCI were observed to perform worse on shape test than healthy older adults. ⁴
Response / compliance rate was greater for healthy older adults compared to mild cognitive impairment. ⁵
Detected changes in behavior patterns was early indicator of cognitive decline. ⁴⁰

Three types of LLM prompts are created. First, prompts that process text directly (i.e., journal entries, testing session transcripts) are formalized based on a statement found in the corresponding research article (i.e., external knowledge). HDTwin contains 11 of these rules, such as “if journal_text has a small vocabulary, short sentences, and/or low sentence complexity then more likely mild cognitive impairment” ⁴¹ (the full set of rules is provided in Table 6). The LLM must interpret

the meaning of the rule in the context of the input text. Second, as described earlier, the LLM is directed to generate numeric ratings of the testing session participant responses.

Third, a decision tree algorithm processes numeric markers from the training set, learning a concept distinguishing HOA from MCI. The purpose of these models is to provide a mechanism for learning from real-valued numeric data and demonstrate the ability of HDTwin to harness heterogeneous types of information. Decision trees are advantageous because the learned model is easily interpretable. In particular, these trees are automatically converted to if-then rules that are provided to the LLM in a prompt. From the large set of possible rules, we include those that support $\geq 10\%$ of the training data and the probability of rule occurrence in the tree is $> 70\%$. In the case of single-term rules resulting from decision stumps, the rule inversion is also added. This process resulted in 20 rules, such as “if `distance_traveled_from_home` > 4.5 then more likely healthy” and “if `physical_activity_level` ≤ 1.5 and `EMA_compliance` ≤ 9.5 then more likely mild cognitive impairment.” The full set of decision tree-generated rules is provided in Table 7.

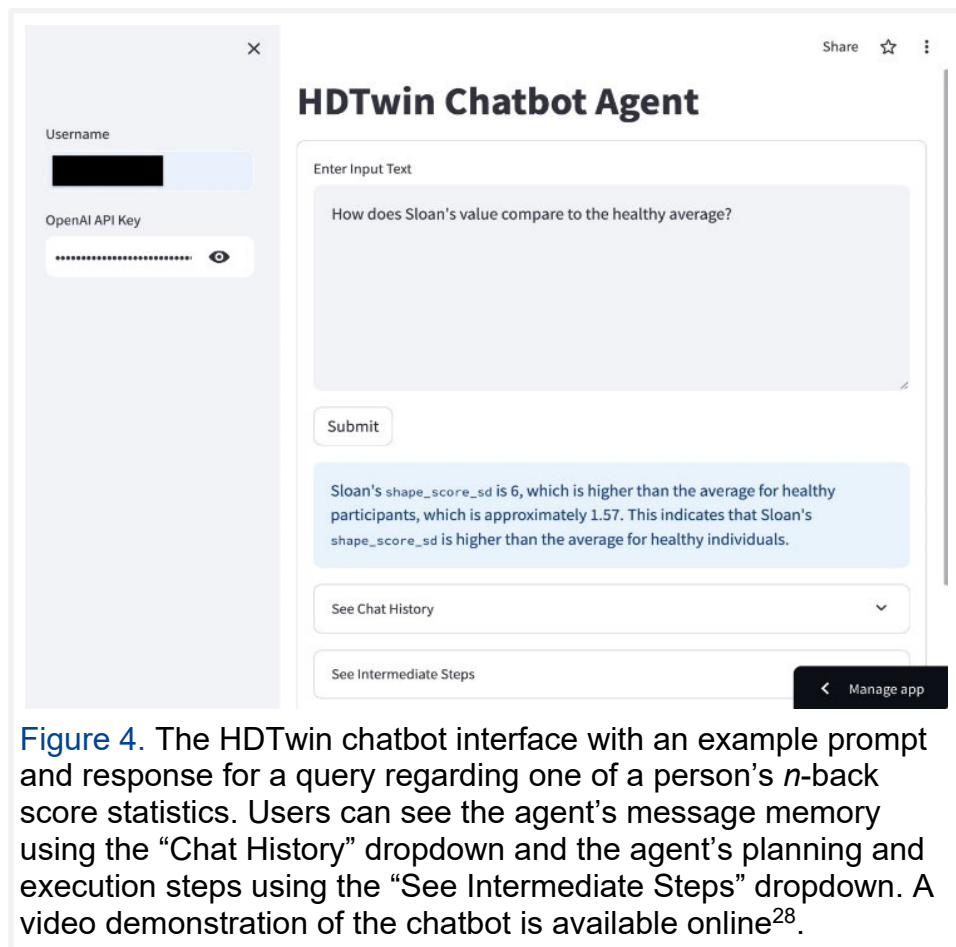
Chatbot Agent Tools

The chatbot agent is designed as a Python langchain OpenAI tools agent. We created custom tools that allow the chatbot agent to plan and take actions to solve specialized tasks. By integrating these tools, HDTwin links the LLM with additional sources of information and additional functionality, yielding a more complete digital twin. These agent tools include:

1. Participant data retriever tool. We embedded and stored the test dataset in a FAISS vector database. We then created a langchain retriever tool for the database that the agent can query to find markers for a participant.
2. Reference data calculation tool. We wrapped a langchain pandas agent in a custom tool that loads the reference data into a pandas DataFrame. The agent can query this calculation tool to get on-the-fly summary statistics for HOA and/or MCI training participants, such as grouping data by diagnosis then calculating a mean for a marker of interest. This tool allows HDTwin to interactively answer questions regarding a particular person, including their scores on each assessment component and how the scores compare with the participant cohort.
3. Knowledge base retriever tool. We embedded and stored the knowledge base in a FAISS vector database. Similar to the participant data retriever tool, the agent can query a langchain retriever tool to find knowledge in the database that may interest the user. For example, HDTwin may quote one of the findings from the literature as part of the reason it predicted a particular diagnosis category for a participant.
4. Paper abstract retriever tool: Using the langchain PubMed application programming interface, we created a tool that searches PubMed abstracts for information that may be relevant to the query tasks. Text from the abstracts is

included verbatim as input. The tool can also be leveraged to include literature from other public domain sources.

5. Diagnosis classification tool: Using the same classification prompting strategy described in the previous section, we created a langchain structured tool that uses a custom chat chain to classify a participant as HOA or MCI. The chain constructs a prompt using information from the knowledge base and personalized markers to request that the language model generate predictions (see the Diagnosis Prediction section for more details about diagnosis prompt construction).



We designed the agent with an interface built using the Python streamlit library. Figure 4 provides a screenshot of this interface. As the figure demonstrates, HDTwin leverages the participant’s personal markers (demographics, behavior, EMA response, n -back scores, testing session data), rules generated from the training cohort, and literature from the field to respond to user queries. Upon request, the agent lists the steps that were performed by generating a response, pulling from the message memory as needed (i.e., interaction history), as well as the information sources.

Table 3. Example cognitive diagnosis prompt. The user input the system message and prompt, HDTwin generated the corresponding response. In this case, the classification of diagnosis from HDTwin is correct.

System Message	You are a knowledgeable health assistant designed to diagnose someone as "healthy" or "mild cognitive impairment" using diagnosis rules and relevant personal information.
Example Prompt	<p>Diagnose Sloan as healthy or mild cognitive impairment by applying the following rules to Sloan's information.</p> <p>Diagnosis Rules: Rule #1: if shape_score_sd > 2.5 then more likely mild cognitive impairment Rule #2: if shape_score_sd <= 2.5 then more likely healthy Rule #3: if shape_score_sd > 2.5 and shape_learning_rate > 1.5 then more likely mild cognitive impairment Rule #4: if journal_text has a large vocabulary, long sentences, and/or high sentence complexity then more likely healthy Rule #5: if some interview_assessment ratings are <= 3 then more likely mild cognitive impairment</p> <p>Sloan's Information: shape_learning_rate: 9 shape_score_sd: 6 journal_text: made a list and went to the store and got everything on it and didn't alter from it interview_assessment: Category Fluency (Animals): Participant's Answer: "turkey" Rating: 2 <text removed for brevity></p> <p>Answer using the format: Diagnosis: ... Explanation: ...</p>
Example Response	<p>Diagnosis: Mild Cognitive Impairment</p> <p>Explanation: Based on the provided information, Sloan's shape_score_sd is 6, which is greater than 2.5, and the shape_learning_rate is 9, indicating a high learning rate. According to Rule #3, when shape_score_sd is greater than 2.5 and shape_learning_rate is higher than 1.5, the individual is more likely to have mild cognitive impairment. Additionally, some of Sloan's interview_assessment ratings are <= 3, such as the ratings for Category Fluency (Animals) and Letter Fluency (F), further supporting the diagnosis of mild cognitive impairment.</p>

Diagnosis Prediction

The long-term goal of HDTwin is to build a digital twin representing the cognitive health state of a physical human. The digital twin can respond to queries about the person's behavior, task performance, and predicted cognitive health diagnosis. Unlike traditional machine learning tools, the system will interactively explain its reasoning and ingest additional information that is provided on the fly.

In this paper, we consider the role of HDTwin in performing cognitive diagnosis prediction. Specifically, we seek to validate the ability of HDTwin to perform cognitive diagnosis from a fusion of heterogeneous information. Using the LLM model, we constructed a classification pipeline that we performed for each participant in the test set. The pipeline begins by constructing the context needed to form a prompt. Table 3 provides an example of this context, which primes the language model to perform reasoning and generate a response. As shown in the example, the prompt includes instructions, the knowledge base rules that it utilizes, and participant markers. The LLM is prompted to return a response with the diagnosis, which we extract using a regular expression pattern search. Upon request, the LLM will also provide an explanation for the diagnosis. By default, HDTwin provides a short (i.e., few-sentence) explanation. The full explanation can be extracted upon request. The latter not only provides insight into the LLM's reasoning but can help with response accuracy using chain-of-thought reasoning⁴².

We measure prediction performance using accuracy, MCI precision and recall, and Matthew's correlation coefficient (mcc). To provide a baseline for comparison, we trained three traditional classifiers using the numeric markers. We trained the models on the same participants used to build the LLM knowledge base and evaluated performance using the same test set. The baseline classifiers are random forest, gradient boosting, and XGBoost.

Reference, training, and test sets

Traditional supervised machine learning uses data from a training set to train a model. In contrast, HDTwin performs retrieval-augmented generation by combining information retrieval with a generative model⁴³. In HDTwin, therefore, training data are used to build a knowledge base of text, data, and rule prompts for the language model to retrieve. Following literature guidelines for evaluating large language models⁴⁴, we created three sets - reference, training, and test - from the participant group. Data from the *reference* set support the interactive chatbot functions. A subset of the reference group is selected as the *training* set. In our experiments, the training set is used to build a knowledge base of text input, numeric data, and decision tree rules for HDTwin to perform diagnosis prediction. Participants not in the reference (and training) set are reserved to form a *test* set. Prediction performance is determined based on these held-out data.

Participants included in the training and test sets were required to have complete behavior, EMA, and *n*-back markers, as these markers were essential for conducting

an ablation analysis of their contribution to diagnosis prediction. To ensure a uniform class distribution, 24 participants were randomly selected from each diagnosis group (MCI, HOA) to form a test set of 48 participants. Of these participants, 51 were labeled healthy and 25 were labeled MCI. For the training set, we randomly selected 25 reference set participants from each diagnosis category (51 labeled healthy, 25 labeled MCI), resulting in an equal split.

Table 4. Example chatbot interaction. User-inputted text is highlighted with italic font. Each HDTwin response is listed with the chatbot agent tools that were employed to generate the corresponding response.

Prompt	Tools Used
<i>What do you know about Sloan?</i>	
<p>Sloan is a 57-year-old female with 14 years of education. She has a Shape Score Standard Deviation of 6 and a Shape Learning Rate of 9. Sloan's EMA compliance is rated at 8.0, and her physical activity level is 4.0. She has a mental sharpness mean of 6.0 and a mental sharpness variance of 5.0.</p> <p>Sloan's journal entry mentions activities such as making a list, going grocery shopping, using the EMMA app to remind her to take pills, and engaging in social and leisure activities.</p> <p>If you would like more specific information about Sloan or her cognitive health status, please let me know.</p>	Participant data retriever
<i>Would you diagnose Sloan as mild cognitive impairment?</i>	
<p>Sloan has been diagnosed with mild cognitive impairment based on her cognitive assessment results. If you have any further questions or need more information, feel free to ask.</p>	Diagnosis
<i>Can you explain why you made that diagnosis?</i>	
<p>Sloan has been diagnosed with mild cognitive impairment based on her cognitive assessment results. The diagnosis is supported by the fact that Sloan's shape_score_sd is greater than 2.5 and her shape_learning_rate is greater than 1.5, meeting the criteria for mild cognitive impairment.</p>	Message memory, Diagnosis
<i>How does Sloan's shape_score_sd compare to the average?</i>	
<p>Sloan's shape_score_sd value is 6, which is higher than the average value of approximately 1.57 for healthy participants. This indicates that Sloan's shape_score_sd value is significantly higher compared to the average value of healthy participants.</p>	Message memory, Training calculation

Results

Table 4 demonstrates an interaction with the HDTwin chatbot agent in which the agent executes custom tools and accesses a memory cache for conversation history. The generated rules were used to form two classification prompt variants, one with a requested explanation that aids in understanding the LLM’s decision-making process. Experiment results showed no substantial difference between these variations, so results are reported for the smaller, diagnosis-only prompt. Lastly, HDTwin performs a wrapper-based selection of rules to include in the diagnoses. In the wrapper method, a greedy search is employed to identify the final set of features, or rules, to include ⁴⁵.

Table 5 summarizes the diagnostic accuracy of HDTwin compared to baseline classifiers. These include results for the LLM using wrapper selection, a single type of information source, or all available rules. Due to LLM non-determinism, each variation was executed 30 times; the mean and standard deviation were reported. To quantify the improvement gained by the wrapper approach, we computed a one-tailed unpaired t-test, comparing the LLM wrapper results against the best single marker set (*n*-back; $p < 0.001$) and the best traditional classifier (XGBoost; $p < 0.001$).

Table 5. Prediction performance. Performance is reported in terms of prediction accuracy, precision and recall for the MCI category, and Matthews correlation coefficient (mcc). The best performer for each metric is highlighted in bold.

Classification	Accuracy	MCI Precision	MCI Recall	MCC
LLM / demographics	0.48 ± 0.00	0.47 ± 0.00	0.33 ± 0.00	-0.04 ± 0.00
LLM / behavior	0.43 ± 0.02	0.43 ± 0.02	0.43 ± 0.04	-0.14 ± 0.05
LLM / <i>n</i> -back	0.75 ± 0.00	0.93 ± 0.00	0.54 ± 0.00	0.55 ± 0.00
LLM / EMA	0.65 ± 0.00	0.89 ± 0.00	0.33 ± 0.00	0.37 ± 0.00
LLM / journal	0.57 ± 0.04	0.56 ± 0.03	0.63 ± 0.05	0.13 ± 0.07
LLM / test session	0.61 ± 0.01	0.78 ± 0.01	0.30 ± 0.02	0.28 ± 0.02
LLM / all	0.56 ± 0.01	0.64 ± 0.02	0.29 ± 0.02	0.15 ± 0.02
LLM / wrapper	0.77 ± 0.02	0.88 ± 0.05	0.63 ± 0.03	0.57 ± 0.05
Random forest	0.62 ± 0.04	0.86 ± 0.03	0.27 ± 0.00	0.31 ± 0.07
Gradient boosting	0.65 ± 0.01	0.89 ± 0.01	0.35 ± 0.03	0.39 ± 0.02
XGBoost	0.67 ± 0.00	0.83 ± 0.00	0.42 ± 0.00	0.38 ± 0.00

Discussion

Principal Results

The goal of this study was to explore the use of large language models as a mechanism to fuse multi-modal information relevant to understanding and predicting the cognitive health diagnosis for an individual. As demonstrated in Table 4, the HDTwin chatbot agent conversationally provides answers related to the

dataset and cognitive diagnosis. When prompted for a diagnosis for a particular participant, it succinctly responds with a predicted class label. When asked to explain the diagnosis, the agent correctly cites a rule from the knowledge base, though the chatbot may offer only a subset of rules that were used for the inference.

We note that when the chatbot is requested to compare a participant to the training set, the LLM calls the training calculation tool, which generates and executes the code to select, filter, and summarize the underlying DataFrame (e.g., generating `{'query': "df[df['diagnosis'] == 'healthy'] ['shape_score_sd'].mean()"}`). In this case, the training calculation tool produces the correct value (1.57). Because LLMs are non-deterministic systems, the chatbot is not guaranteed to return the same response each time.

Non-deterministic behavior also affected the classification results. This behavior is evidenced by the non-zero performance standard deviations that are listed in Table 5. We further observed that longer prompts (e.g., journal and testing session) generally led to less consistent performance (e.g., higher standard deviation in Table 5). As the prompt text increases in length, there is a higher risk of divergence due to the model latching onto different parts of the prompt in different ways across each run, creating more variability in the output.

Another factor influencing classification performance is the type of information that is utilized and the way the information is incorporated into the LLM prompt. While HDTwin can use all information sources, Table 5 illustrates that not all information was equally effective at discriminating between diagnosis classes. Of the six marker sets, *n*-back offers the most predictive rules (0.75 accuracy). While the wrapper method occasionally selected the text-based journal and testing session markers, this marker set did not perform the best. The top-performing case, with 0.81 accuracy, was a run utilizing the combination of *n*-back and behavior markers. The worst-performing cases combined demographic markers with behavior markers. This is not surprising considering only two rules utilizing demographic markers were supported with a high probability of occurrence in the training set. Our results show that the LLM wrapper method significantly outperformed the best traditional classifier as well as the best individual marker set (*n*-back). Results for the wrapper method are comparable to reported results, which use MRI and cerebrospinal fluid to perform a similar task, yielding an accuracy of 76.4%⁴⁶. These findings provide evidence supporting in-the-home data collection and the design of LLMs technologies for improved mild cognitive impairment diagnosis.

Limitations

This study faced several limitations, including missing data across several of the marker sets, reliance on an LLM to extract and label participant responses to testing session questions, and a small sample size of training participants (n=50) whose data formed the knowledge base. Additionally, LLM non-determinism affects the reproducibility of the results. In the future, we plan to explore other LLMs for the HDTwin chatbot agent and diagnosis classification task, as well as improve

prediction accuracy with more in-depth prompt engineering and an expansion of the knowledge base.

Many of HDTwin's data sources are continually updated. These include the behavior markers, n -back scores, and journal entries. The rules can be periodically updated, allowing HDTwin to adapt to this new information dynamically. Our approach does not fine-tune the model but provides new context. Future work will investigate how to direct the LLM to model the historical evolution of the prompts and forecast the future state of the individual. In this setting, the digital twin can provide a tool for testing scenarios and predicting outcomes of various behavior and other changes. Such a tool can help optimize potential treatment decisions for each person before they are administered.

Conclusions

In this paper, we built a custom agent called HDTwin for interactively exploring a multimodal, in-the-wild health dataset ($n=124$), supporting the creation of a digital twin. The HDTwin digital twin contains an agent that supports diagnosis of mild cognitive impairment, guiding more informed and actionable cognitive assessment. To build this classifier with an LLM, we explored the predictive capability of diverse individual and fused health markers. A fusion of knowledge and participant data from different marker sets yielded the strongest performance. Our findings indicate that HDTwin significantly outperforms traditional classifiers in diagnosing mild cognitive impairment. Integrating diverse data sources through LLMs provides a comprehensive view of cognitive health, enhancing diagnosis and intervention strategies. Future studies can continue to explore approaches for increasing the accuracy of LLMs to improve the accuracy of custom agents like HDTwin for aiding clinicians with healthcare diagnosis and prediction of outcome.

Acknowledgements

This work is supported in part by NIH grants R01AG065218 and R35AG071451. The work was also supported in part by DOD grant AZ 190055.

Data Availability

The source code for HDTwin is available at <https://github.com/WSU-CASAS/HDTwin>. The code for the smartwatch n -back task is also available online at <https://github.com/WSU-CASAS/shapetest-watch-demo>. The data sets analyzed for this study are available from the authors upon reasonable request.

Table 6. HDTwin text processing rules. These results are used to process information found in text including journal entries and interview assessments.

Marker Set	Rule
Speech	if journal_text is empty the more likely mild cognitive impairment
	if journal_text has a large vocabulary, long sentences, and/or high sentence complexity then more likely healthy
	if journal_text has a small vocabulary, short sentences, and/or low sentence complexity then more likely mild cognitive impairment
	if journal_text uses positive emotion words then more likely healthy
	if journal_text uses negative emotion words then more likely mild cognitive impairment
	if journal_text appears to have more than one entry then more likely healthy
	if journal_text appears to only have one entry then more likely mild cognitive impairment
Baseline Testing Session	if some interview_assessment ratings are ≤ 3 then more likely mild cognitive impairment
	if most interview_assessment ratings are ≥ 4 then more likely healthy
	if the interview_assessment explanations suggest the participant confidently answered the questions correctly then more likely healthy
	if the interview_assessment explanations suggest the participant struggled to answer the questions correctly then more likely mild cognitive impairment

Table 7. HDTwin numeric processing rules. These results are created from the trained decision trees and are listed with the corresponding probability of the rule occurring in the decision tree (Pr) and number participants supporting the rule (Pa).

Rule	Pr	Pa
if distance_traveled_from_home > 4.5 then more likely healthy	100.00	6
if distance_traveled_variance > 8.0 and physical_activity_level <= 5.5 then more likely healthy	91.67	12
if physical_activity_level <= 1.5 and EMA_compliance <= 9.5 then more likely mild cognitive impairment	85.71	7
if mental_sharpness_variance > 2.5 and mental_sharpness_mean > 4.5 then more likely mild cognitive impairment	83.33	6
if shape_score_sd <= 2.5 and mental_sharpness_variance <= 2.5 then more likely healthy	81.82	22
if distance_traveled_variance > 8.0 then more likely healthy	80.00	15
if mental_sharpness_variance <= 2.5 and mental_sharpness_mean <= 4.5 then more likely mild cognitive impairment	80.00	5
if mental_sharpness_mean > 5.5 and physical_activity_variance <= 8.0 then more likely healthy	79.17	24
if sex = male and age <= 79.5 then more likely mild cognitive impairment	77.78	9
if mental_sharpness_variance > 2.5 then more likely mild cognitive impairment	77.78	9
if sex = female and shape_learning_rate <= 4.5 then more likely healthy	76.47	17
if shape_score_sd > 2.5 and shape_learning_rate > 1.5 then more likely mild cognitive impairment	75.00	16
if physical_activity_variance <= 0.5 then more likely mild cognitive impairment	75.00	8
if mental_sharpness_mean <= 4.5 then more likely mild cognitive impairment	75.00	8
if shape_score_sd > 2.5 then more likely mild cognitive impairment	70.59	17
if shape_score_sd <= 2.5 then more likely healthy	60.61	33
if distance_traveled_variance <= 8.0 then more likely mild cognitive impairment	60.61	33
if physical_activity_variance > 0.5 then more likely healthy	57.50	40
if mental_sharpness_variance <= 2.5 then more likely healthy	57.50	40
if mental_sharpness_mean > 4.5 then more likely healthy	56.10	41

References

1. O'Malley RPD, Mirheidari B, Harkness K, et al. Fully automated cognitive screening tool based on assessment of speech and language. *BMJ Neurol Neurosurg Psychiatry*. 2021;92:12-15.
2. Aronsson FS, Kuhlmann M, Jelic V, Ostberg P. Is cognitive impairment associated with reduced syntactic complexity in writing? Evidence from automated text analysis. *Aphasiology*. 2021;35(7):900-913.
3. Nicosia J, Aschenbrenner AJ, Balota DA, et al. Unsupervised high-frequency smartphone-based cognitive assessments are reliable, valid, and feasible in older adults at risk for Alzheimer's disease. *J Int Neuropsychol Soc*. 2023;29(5):459-471.
4. Schmitter-Edgecombe M, Luna C, Beech B, Dai S, Cook DJ. Capturing cognitive capacity in the everyday environment across a continuum of cognitive decline using a smartwatch n-back test and ecological momentary assessment. *Neuropsychology*.
5. Cook DJ, Walker A, Minor B, et al. A cross-study analysis of mobile EMA in monitoring behavior and well-being: Insights to refine EMA methods. *JMIR mHealth uHealth*. Published online 2024.
6. Petersen J, Austin D, Mattek N, Kaye J. Time out-of-home and cognitive, physical, and emotional wellbeing of older adults: A longitudinal mixed effects model. *PLoS One*. 2015;10(10):e0139643.
7. Cook D, Schmitter-Edgecombe M. Fusing ambient and mobile sensor features into a behaviorome for predicting clinical health scores. *IEEE Access*. 2021;2:65033-65043.
8. Almeida ML, Pederson AM, Zimmerman SC, et al. The association between physical activity and cognition in a racially/ethnically diverse cohort of older adults: Results from the Kaiser healthy aging and diverse life experiences study. *Alzheimer's Dis Assoc Disord*. Published online 2024.
9. Schmitter-Edgecombe M, Sumida CA, Cook DJ. Bridging the gap between performance-based assessment and self-reported everyday functioning: An ecological momentary assessment approach. *Clin Neuropsychol*. 2020;34(4):678-699.
10. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *npj Digit Med*. 2020;3:136.
11. Wang B, Xie Q, Pei J, et al. Pre-trained language models in biomedical domain: A systematic survey. *ACM Comput Surv*. 2023;56(3):55.
12. Luo R, Sun L, Xia Y, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. 2022;23(8). doi:10.1093/bib/bbac409
13. Xie Q, Bishop JA, Tiwari P, Ananiadou S. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Syst*. 2022;252:109460.
14. Mao C, Xu J, Rasmussen L, et al. AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's

- disease. *J Biomed Inform.* 2023;144:104442. doi:10.1016/j.jbi.2023.104442
15. Asgari M, Kaye J, Dodge H. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's Dement.* 2017;3(2):219-228.
 16. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature.* 2023;619:357-362.
 17. Kim Y, Xu X, McDuff D, Breazeal C, Park HW. Health-LLM: Large language models for health prediction via wearable sensor data. *arXiv:240106866.* Published online 2024.
 18. Xue H, Salim FD. PromptCast: A new prompt-based learning paradigm for time series forecasting. *IEEE Trans Knowl Data Eng.* Published online 2024:1-14.
 19. Sprint G, Schmitter-Edgecombe M, Weaver R, Wiese L, Cook DJ. CogProg: Utilizing large language models to forecast in-the-moment health assessment. *JMIR Mhealth Uhealth.* Published online 2024.
 20. Yu H, Guo P, Sano A. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. In: *Machine Learning for Health Symposium.* ; 2023:650-663.
 21. Jin M, Wang S, Ma L, et al. Time-LLM: Time series forecasting by reprogramming large language models. In: *International Conference on Learning Representations.* ; 2024:1-24.
 22. de Zarza I, De Curto J, Roig G, Calafate CT. LLM multimodal traffic accident forecasting. *Sensors.* 2023;23:22. doi:<https://doi.org/10.3390/s23229225>
 23. Girdhar R, El-Nouby A, Liu Z, et al. Imagebind: One embedding space to bind them all. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* ; 2023:15180-15190.
 24. Xu S, Yang L, Kelly C, et al. ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv:230801317.* Published online 2023.
 25. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. *J Med Syst.* 2024;48(1):22.
 26. Brandt J, Folstein M. *Telephone Interview for Cognitive Status.* Psychological Assessment Resources, Inc.; 2003.
 27. CASAS. HDTwin. Published 2024. <https://github.com/WSU-CASAS/HDTwin>
 28. Sprint G. hdtwin1. youtube. <https://youtu.be/F1aRgUk9nls>
 29. Nonbinary. Neutral names starting with A. https://nonbinary.wiki/wiki/Neutral_names_starting_with_A
 30. Cella D, Riley W, Stone A, et al. Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005-2008. *J Clin Epidemiol.* 2010;63(11):1179-1194.
 31. Wechsler D. *Wechsler Memory Scale - Fourth Edition.* Pearson; 2009.
 32. Delis DC, Kaplan E, Kramer JH. *Delis-Kaplan Executive Function System: Examiner's Manual.* The Psychological Corporation; 2001.
 33. Regard M, Strauss E, Knapp P. Children's production on verbal and non-verbal fluency tasks. *Percept Mot Skills.* 1982;55(3):838-844.
 34. Delis DC, Kramer JH, Kaplan E, Ober BA. *The California Verbal Learning Test -*

Second Edition.; 2000.

35. Jak AJ, Bondi MW, Delano-Wood L, et al. Quantification of five neuropsychological approaches to defining mild cognitive impairment. *Am J Geriatr Psychiatry*. 2009;17:368-375.
36. Dai S, Kehinde OJ, Schmitter-Edgecombe M, French B. Modeling daily fluctuations in everyday cognition and health behaviors at general and person-specific levels: a GIMME Analysis. *Behaviormetrika*. 2023;50:563-583.
37. Cormack F, McCue M, Skirrow C, et al. Characterising longitudinal patterns in cognition, mood and activity in depression with 6-week high-frequency wearable assessment: an observational study. *JMIR Ment Heal*. Published online 2024.
38. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. *npj Digit Med*. 2023;6:158.
39. Liu X, Jiang Y, Peng W, et al. Association between physical activity and mild cognitive impairment in community-dwelling older adults: Depression as a mediator. *Front Aging Neurosci*. 2022;14:964886.
40. Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans Speech Audio Process*. 1995;3(1):72-83.
41. Weyerman JJ, Rose C, Norton MC. Personal journal keeping and linguistic complexity predict late-life dementia risk: The Cache County journal pilot study. *Journals Gerontol Ser B, Psychol Sci Soc Sci*. 2017;72(6):991-995.
42. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903.
43. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *International Conference on Neural Information Processing Systems*. ; 2020:9459-9474.
44. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. *ACM Trans Intell Syst Technol*. 2024;15(3):39.
45. El Aboudi N, Benhlima L. Review on wrapper feature selection approaches. In: *International Conference on Engineering & MIS*. ; 2016.
46. Zhang D, Wang Y, Zhou L, Yuan H, Shen D. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*. 2011;55(3):856-867.