DEEPQUEST: FROM KNOWLEDGE GAPS TO SCIENTIFIC DISCOVERY IN ALZHEIMER'S DISEASE

Anonymous authors
Paper under double-blind review

ABSTRACT

Automating experiments does not necessarily advance scientific understanding. Research agents must identify meaningful knowledge gaps, determine what available data can resolve, and assess whether their findings advance existing knowledge. We introduce DeepQuest, a knowledge-gap-driven AI scientist framework for Alzheimer's disease research. DeepQuest transforms unresolved questions from domain reviews into traceable, testable hypotheses and prioritizes them through data feasibility screening and groupwise Elo ranking. It couples iterative experimentation with two-level auditing that separates experimental reliability from knowledge contribution relative to a frozen literature corpus. On OASIS-3, DeepQuest produces 42 feasible hypotheses and outperforms the strongest baselines in 14 of 15 evaluator–dimension comparisons across three language-model reviewers. The full framework achieves an audit-assessed scientific finding rate of 95.24%, versus 78.57% for Codex and 76.19% for Claude Code on the same hypotheses. Representative findings concern neuropsychiatric symptoms and dementia risk, longitudinal amyloid spatial stability, and neocortical tau-associated clinical impairment, with key associations supported in independent ADNI and AIBL cohorts. These results show how aligning knowledge gaps, data feasibility, and evidence auditing can turn automated research into traceable, evidence-supported scientific discovery. Our source code is available at `https://anonymous.4open.science/r/DeepQuest-3066/`.

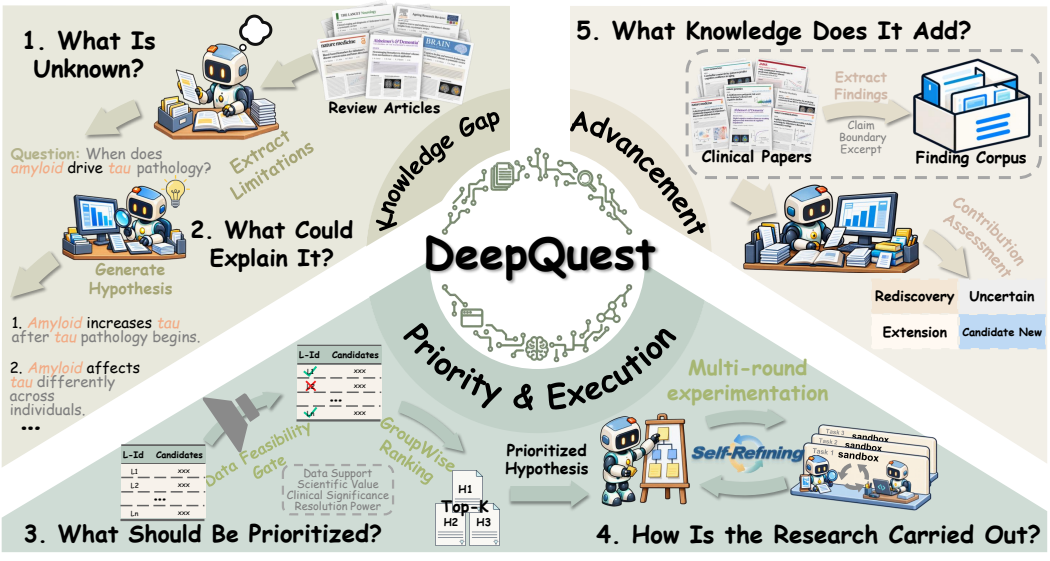

Figure 1: Overview of the DeepQuest framework for automated AD research. The framework integrates knowledge-gap-driven hypothesis formulation, data-constrained task selection, iterative experimentation, and literature-based knowledge contribution assessment.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

A central challenge in Alzheimer's disease (AD) research is to translate the growing body of heterogeneous evidence into testable scientific questions that advance understanding of the disease. The complex pathology of AD spans multiple biological levels (Jack Jr et al., 2024; Zhao et al., 2025; Shwab et al., 2025). Large scale cohorts and data sharing initiatives have assembled multimodal data encompassing clinical assessments, cognitive measures, genetic information, biofluid biomarkers, magnetic resonance imaging (MRI), and positron emission tomography (PET), providing a foundation for investigating disease states, progression, and pathological associations (Travaglia & Hoffmann, 2024; Toga et al., 2024; Veitch et al., 2026; Weiner et al., 2025). DiaMond learns cross-modal disease representations from MRI and PET (Li et al., 2025b). CARE-AD extracts information on disease evolution from longitudinal clinical records (Li et al., 2025a). The multitask model proposed by Wang et al. (2025) jointly analyzes brain age, cognition, and amyloid pathology. These advances strengthen data analysis while placing greater demands on evidence organization, research question formulation, and the evaluation of existing disease understanding.

Research agents have extended automation across multiple stages of the scientific workflow, yet completing the workflow does not by itself ensure scientifically valuable discoveries. General-purpose AI Scientist systems such as DeepScientist can generate candidate hypotheses, plan experiments, and iteratively refine their exploration based on feedback (Weng et al., 2026; Yamada et al., 2025; Tang et al., 2025). Domain-specific systems, such as Medical AI Scientist, AMID and SPARK, further integrate domain knowledge, specialized tools, and quality control mechanisms to support complex analytical workflows in medicine and neuroscience (Wu et al., 2026; Liu et al., 2026; Han et al., 2026; Trost et al., 2026). However, generating hypotheses and executing experiments for a predefined objective still requires assessing whether they address important knowledge gaps, are testable with available data, and advance existing understanding.

Autonomous scientific discovery in AD requires three interdependent decisions: problem formulation, task selection, and knowledge contribution assessment. First, limitations identified in the literature must be translated into experimentally testable hypotheses with explicit evidential provenance, thereby establishing a traceable link between existing knowledge and newly formulated research questions. Second, scientifically meaningful questions are not necessarily answerable with available medical data. Task selection must therefore balance scientific value against execution cost under constraints imposed by data feasibility and experimental resources. Third, experimental reliability and scientific contribution should be assessed separately. Improvements in predictive performance or statistical significance may support specific experimental conclusions, but neither alone establishes that a study has produced new understanding of the disease. Integrating these stages enables research agents to pursue questions according to their potential knowledge gain.

To this end, we propose DeepQuest, an AI scientist framework for AD research designed to advance existing knowledge by integrating knowledge gap identification, data-constrained task selection, experimental validation, and knowledge contribution assessment as shown in Fig. 1. DeepQuest first extracts structured research limitations from domain reviews and transforms them into testable hypotheses. It then applies multi-stage screening and relative priority ranking under data feasibility constraints to select research tasks that are both scientifically valuable and feasible to execute. During experimentation, DeepQuest combines iterative execution with a two-level audit of workflow reliability and the contribution of conclusions beyond existing knowledge. This integrates problem formulation, task selection, and conclusion assessment into a unified research process.

Experimental results support the effectiveness of DeepQuest at three levels: hypothesis quality, knowledge contribution assessment, and validation on independent cohorts. Across all three evaluator models, mean evaluation scores for the 42 executable hypotheses generated by DeepQuest were better than those for five strong generative-model baselines. The evidence audit classified the resulting experimental claims from 40 of 42 hypotheses (95.24%) as Candidate New Discovery or Extension. All ten top-ranked hypotheses were identified as having the potential to advance existing knowledge. Three representative case studies on amyloid spatial topology, neuropsychiatric symptoms and dementia risk, and the relationship between tau pathology and clinical impairment further demonstrate DeepQuest's ability to carry research from problem formulation through evidence validation. The corresponding findings were broadly supported in independent external cohorts.

Overall, our contributions are as follows:

- We introduce a review-driven approach to formulating scientific questions in AD research. The method transforms research limitations identified in AD domain reviews into structured representations and testable hypotheses, establishing a traceable mapping from literature evidence and knowledge gaps to research questions.

- We introduce a data-constrained task selection and optimization approach. It formulates task selection as an optimization problem over the intersection between the literature-derived research space and the data feasibility region. Through multi-stage screening and priority ranking, it balances scientific value with execution feasibility.

- We develop a two-level auditing mechanism that separates experimental reliability from knowledge contribution. DeepQuest independently examines whether experiments are reliable and conclusions extend beyond existing knowledge, providing traceable evidence for assessing candidate scientific discoveries and their contributions.

2 Related Work

General Autonomous Scientific Research Systems. General-purpose AI research systems automate multiple research stages by connecting hypothesis generation, experimental search, and feedback (Lu et al., 2024; Schmidgall et al., 2025; Mitchener et al., 2025). AI Scientist-v2 (Yamada et al., 2025) organizes experimental exploration through agent-based tree search and performs result analysis and manuscript generation. AI-Researcher (Tang et al., 2025) integrates literature investigation, hypothesis formulation, and algorithm implementation through multiagent collaboration. DeepScientist (Weng et al., 2026) formulates scientific discovery as a Bayesian optimization problem and combines discovery memory with hierarchical validation to iteratively select and explore promising candidate directions. Collectively, these systems demonstrate the capacity for open-ended scientific ideation and autonomous research direction selection, with representative evaluations conducted primarily on machine learning tasks. Applying these mechanisms to AD cohort studies requires aligning hypotheses with target populations, observable variables, follow-up windows, and confounding-control requirements, so empirical evidence constrains question selection.

Biomedical Research Agents. Biomedical research agents advance hypothesis exploration for medical problems by incorporating domain knowledge and specialized validation procedures (Huang et al., 2026). Co-Scientist (Gottweis et al., 2026) iteratively refines research directions through hypothesis generation, critique, and ranking, and conducts experimental validation in biomedical settings. Robin (Ghareeb et al., 2026) combines literature-based reasoning, candidate selection, and experimental feedback to support the discovery of therapeutic candidates. Medical AI Scientist (Wu et al., 2026) provides three modes for paper-based reproduction, literature-inspired innovation, and task-driven exploration, using coordinated medical and engineering reasoning to identify research gaps, formulate hypotheses, and execute experiments. These systems integrate literature-grounded reasoning and question exploration into scientific workflows, providing important points of reference for developing autonomous biomedical research.

Research agents for medical imaging and clinical cohort studies further incorporate data adaptation, analytical reliability, and hypothesis validation into the scientific workflow. AMID (Liu et al., 2026) plans candidate methods according to the characteristics and constraints of available data and integrates experimental auditing into model optimization. NEXUS (Han et al., 2026) uses quality feedback to revise neuroimaging analysis pipelines. NeuroAgent (Zhong et al., 2026) connects multimodal preprocessing, predictive modeling, and interactive analysis within a unified workflow. More closely related to our setting, the Alzheimer's Disease and Related Dementias (ADRD) hypothesis-testing system proposed by Saichandran et al. (2025) maps literature-derived hypotheses to available variables in the NACC cohort and uses iterative analysis and methodological review to assess testability and conduct statistical tests. Its evaluation focuses primarily on replicating previously reported associations. Our work extends the focus to the formulation, selection, and contribution assessment of candidate research questions in AD. DeepQuest constructs structured domain-specific knowledge gaps from domain reviews, prioritizes research tasks under data constraints, and separately audits experimental reliability and the incremental contribution of conclusions beyond existing knowledge.

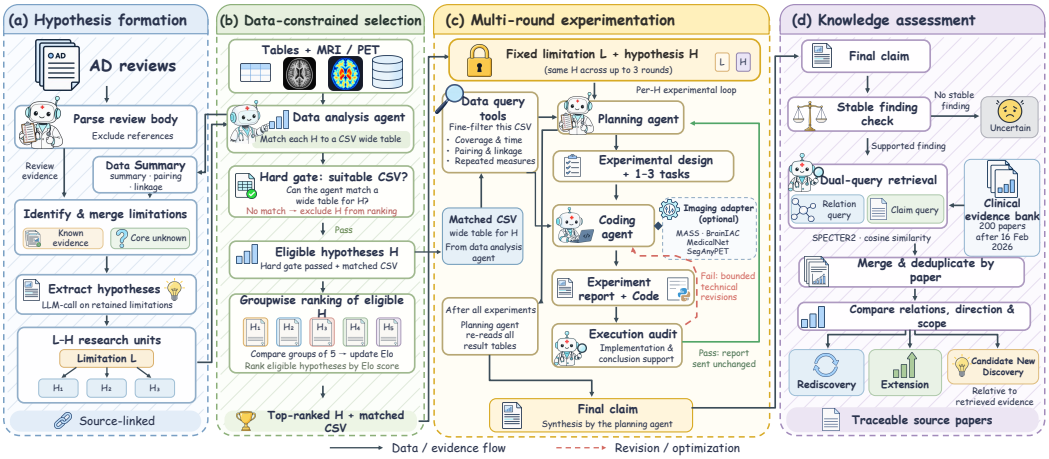

Figure 2: Architecture of DeepQuest. The framework consists of four stages: (a) knowledge-gap-driven hypothesis formulation, (b) data-constrained task selection, (c) iterative experimentation with execution auditing, and (d) literature-based knowledge contribution assessment.

3 METHOD

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

We formulate automated AD research as a knowledge-gap-driven process of task selection under data constraints and evidence assessment. Given a collection of domain reviews R and medical data D, DeepQuest constructs candidate research directions $u_i = (l_i, H_i)$, where l_i denotes a structured research limitation and H_i denotes the corresponding set of hypotheses. It then proceeds through knowledge gap extraction, data feasibility assessment, research priority ranking, iterative experimentation, and knowledge contribution assessment, as illustrated in Fig. 2.

3.2 REVIEW-DRIVEN KNOWLEDGE GAP AND HYPOTHESIS CONSTRUCTION

Knowledge gap extraction organizes established knowledge and key unknowns from domain reviews into traceable research limitations. DeepQuest extracts findings and unresolved relationships concerning AD pathology, biomarkers, imaging, and clinical research. Each limitation records known facts, unresolved phenomena, and supporting source evidence. Semantically equivalent unknowns across reviews are consolidated into a unified limitation l_i.

Hypothesis construction transforms research limitations into scientific propositions testable against observable outcomes. For each l_i, DeepQuest generates a hypothesis set $H_i = \{h_{ij}\}_{j=1}^{n_i}$ and specifies empirical predictions that distinguish competing explanations. Each knowledge gap and its hypothesis set constitute a candidate research direction. This organization separates unresolved questions from explanations while preserving traceability to the original problem.

3.3 DATA FEASIBILITY ASSESSMENT AND RESEARCH ASSET PREPARATION

The data summary provides an initial assessment of whether candidate research questions can be examined with the available data. A data analysis agent examines the available variables, subject structure, imaging records, and multimodal longitudinal follow-up relationships, and constructs the following comprehensive structured representation:

$$z_D = A_{\text{data}}(D) = (z_{\text{summary}}, z_{\text{pairing}}, z_{\text{linkage}}), \tag{1}$$

where z_{summary} summarizes variable semantics and measurement information. z_{pairing} characterizes joint observation patterns across subjects, visits, and temporal dimensions. z_{linkage} records entity identifiers, temporal keys, and cross-table linkage relationships. Based on this structured summary,

DeepQuest applies soft gating to remove research directions that are clearly unsupported by the available data, according to variable semantics and observational coverage.

Hypothesis-level assessment determines whether available data meet the requirements for empirical validation. A data preparation agent examines each hypothesis alongside the structured summary and raw data, checking correspondence among variables, samples, temporal information, and imaging modalities. The feasible hypothesis set for each knowledge gap is denoted by H_i^{feas}, and the retained individual hypotheses form U_D. This stage establishes testability and specific analytical strategies are determined during experimentation.

Research asset preparation organizes the validated data into reusable inputs for downstream experiments. DeepQuest registers verified assets according to data granularity, variable type, and scope of applicability, and prioritizes retrieval and reuse of existing results. For data that have not yet been curated, the system extracts and integrates the required information according to the variables and observation levels specified by each hypothesis. Imaging preprocessing is performed in a modality-specific manner. Structured data and preprocessed images are available for later experiments.

3.4 Groupwise Comparative Research Prioritization

Research ranking compares the value of exploring different knowledge gaps among data feasible candidates under the available cohort and modality constraints. DeepQuest further filters hypotheses that address core unresolved questions in AD, offer potential knowledge gain, and admit empirically discriminative predictions, yielding the candidate set $U_C \subseteq U_D$.

Global research priorities are established through repeated comparisons within groups of five candidates. Inspired by Co-Scientist's pairwise ranking strategy (Gottweis et al., 2026), DeepQuest evaluates each group using anonymized identifiers and randomized presentation order. Let G_t denote the set of candidate indices in round t , with $|G_t| = 5$. Let $s_i^{(t)}$ denote the current score of candidate i, and let $\rho_i^{(t)} \in \{1, \ldots, 5\}$ denote its rank within the group. For any two candidates i, j in G_t, we compute the expected pairwise probability that candidate i is preferred over candidate j using the Elo rating model (Elo, 1978):

$$p_{ij}^{(t)} = \frac{1}{1 + 10^{\left(s_j^{(t)} - s_i^{(t)}\right)/400}}. \tag{2}$$

The scores are synchronously updated according to the within-group rankings as follows:

$$s_i^{(t+1)} = s_i^{(t)} + \eta \left[\frac{5 - \rho_i^{(t)}}{4} - \frac{1}{4} \sum_{j \in G_t \setminus \{i\}} p_{ij}^{(t)} \right], \quad i \in G_t, \tag{3}$$

where η denotes the update step size. The first term represents the proportion of the other four candidates that a given candidate outranks, while the second term is the average expected pairwise preference probability computed from the scores before the update. After T_{rank} comparison rounds, DeepQuest prioritizes individual hypotheses by their final scores, which indicate relative research priority only.

3.5 Iterative Experimentation and Execution Auditing

Experimental planning translates a selected hypothesis into an analytical task with explicit comparators and a defined evidence pathway. For a hypothesis h_k entering the experimental stage, the planning agent fixes its associated research limitation and core proposition, invokes data tools to verify the data structures required for the current round, and constructs the shared dataset D_t. Each round consists of an experimental design P_t and an execution task T_t, which specify the analysis target, variable configuration, statistical method, and expected evidence. Given P_t, T_t and the available dataset D_t, the result of round t is expressed as

$$R_t = f_{\text{exec}}(P_t, T_t, D_t), \tag{4}$$

where f_{exec} denotes the automated experimental execution procedure and R_t denotes the resulting evidence from the current experimental round.

Experimental execution and technical auditing check whether the implementation follows the current design and whether the results support the reported conclusions. A coding agent performs data loading, sample construction, variable processing, and statistical or machine learning analysis in an isolated sandbox. For imaging tasks, representations are extracted from predeployed models through a unified adapter. The audit agent inspects the code implementation, core comparisons, and the correspondence between evidence and conclusions under the current design, and provides technical revision feedback when necessary. The experimental design remains fixed throughout revision. If the analysis still fails to pass the audit after reaching the maximum allowed number of revisions, the run is marked as an irrecoverable execution failure.

Cross-round optimization updates the experimental plan to address unresolved evidence gaps while keeping the core hypothesis unchanged. At the end of each round, the planning agent identifies remaining evidence gaps from the current result R_t and determines whether to revise the comparator, data organization, or analytical design based on audit feedback:

$$P_{t+1} = f_{\text{update}}(h_k, P_t, R_t), \tag{5}$$

where f_{update} denotes the cross-round experimental design update process. DeepQuest conducts at most three rounds, then summarizes the conclusions and their overall scope of applicability under the data constraints. Within-round auditing checks implementation–design consistency, while cross-round planning develops the evidence around the original hypothesis.

3.6 Literature-Based Assessment of Knowledge Contribution

The literature evidence repository provides a traceable reference for comparing experimental conclusions against prior research. We decompose 200 clinical studies on AD and related topics published after February 16, 2026 into individual finding entries to construct the evidence repository K where each entry $e \in K$ records the core claim, boundary and excerpt. This repository defines the literature scope against which subsequent knowledge contribution is assessed.

Evidence sufficiency is assessed before literature comparison to distinguish interpretable conclusions from insufficiently supported results. Results are labeled as Uncertain when valid evidence is lacking because of inadequate statistical power, unstable estimates, nonestimable parameters, or incomplete analysis. Conclusions supported by evidence, including directional associations, conditional differences, or null effects, proceed to literature comparison. Statistical nonsignificance alone is not treated as evidence of a null effect in isolation.

Literature-grounded scientific synthesis and novelty assessment increasingly combine broad retrieval, semantic reranking, citation verification, and structured comparison against prior work (Lin et al., 2025; Shahid et al., 2025). Dual-route retrieval matches both the research relationship and the specific claim to obtain complementary literature evidence. For a claim c , DeepQuest constructs a relation query $q_{\text{rel}}(c)$ and a claim query $q_{\text{claim}}(c)$. SPECTER2 (Singh et al., 2023) is used to encode both queries and literature entries into vector representations, which are ranked by cosine similarity:

$$\text{sim}(q, e) = \frac{f_Q(q)^\top f_E(e)}{\|f_Q(q)\|_2 \, \|f_E(e)\|_2}. \tag{6}$$

Each route's results are deduplicated at the paper level. The top 16 papers from each route are retained and merged into a candidate evidence set covering at most 32 unique papers.

Knowledge relation auditing assesses each experimental claim against the retrieved evidence based on relation structure, claim direction, and scope. Claims consistent with existing evidence and introducing no new proposition are labeled Rediscovery. Those revealing additional temporal, interaction, nonlinear, spatial, or population-specific patterns are labeled Extension. Claims are labeled Candidate New Discovery when the retrieved evidence contains no direct precedent for the same underlying relationship or only broadly related precedents. DeepQuest retains the supporting rationale and source literature. Candidate New Discovery denotes potential knowledge gain relative to the retrieved evidence from the frozen repository, without establishing the absence of precedents in the broader literature. Execution auditing assesses procedural consistency, whereas knowledge relation auditing assesses literature level contribution.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP AND EVALUATION PROTOCOL

We evaluate DeepQuest at three levels: hypothesis quality, research output, and medical discovery. OASIS-3 (LaMontagne et al., 2019) serves as the primary cohort. Three case studies are selected to provide complementary clinical and scientific coverage, and their analyses are further validated on independent ADNI (Mueller et al., 2005) or AIBL (Ellis et al., 2009) cohorts to assess transferability.

We compare 42 DeepQuest hypotheses against five language model baselines. All methods generate hypotheses from the same review set. Three independent language model based reviewers evaluate each hypothesis along five dimensions: novelty, scientific advancement, clinical significance, empirical discriminative power, and feasibility.

To disentangle the contributions of hypothesis generation, execution, and the full pipeline, we compare hypothesis sets from different sources under a unified execution environment and classify final outcomes using a frozen literature evidence repository. We define the Scientific Finding Rate (SFR) as the proportion of evaluated tasks classified by the audit as Candidate New Discovery, Rediscovery, or Extension. Tasks that fail to produce a result because of execution failure, timeout, or reaching the maximum number of audit revisions are labeled Skip. Further implementation details are provided in Appendix A.

4.2 HYPOTHESIS QUALITY AND RESEARCH PRIORITIZATION

DeepQuest achieves the highest overall mean across all three evaluators, with its strongest gains concentrated in dimensions related to scientific question formulation and empirical discriminative power. Overall improvements over each evaluator's best baseline are 0.27, 0.24, and 0.21 points, as shown in Fig. 3. DeepQuest outperforms the best baseline in 14 of 15 evaluator–dimension comparisons. The largest gains are 0.35 in empirical discriminative power under GPT-5.6-sol, 0.34 in novelty under Claude Opus 4.8, and 0.26 in scientific advancement under DeepSeek-v4-pro-0813. Clinical significance shows relatively smaller gains than other evaluated dimensions: under GPT-5.6-sol, DeepQuest scores 3.43 versus the best baseline's 3.45.

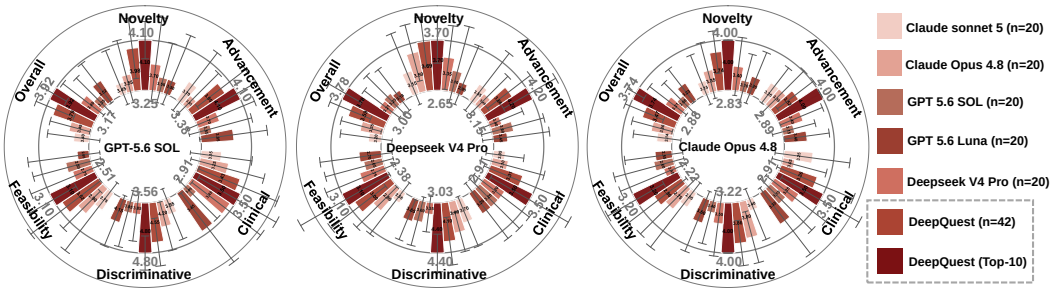

Figure 3: Hypothesis quality evaluation. Bars show mean scores from three independent language model reviewers across multiple evaluation dimensions for baseline methods, DeepQuest, and DeepQuest Top-10, with standard deviations indicated in gray.

Across all three evaluator models, the Top-10 subset scores higher than the full candidate set and exceeds the best corresponding baseline on all five dimensions, indicating alignment between ranking and quality assessments. Because Top-10 is nested within the full set, this provides only descriptive evidence of ranking effectiveness rather than an independent validation of ranking performance.

4.3 FROM HYPOTHESES TO SCIENTIFIC CLAIMS

The full DeepQuest pipeline achieves a higher rate of auditable claims on the same hypothesis set. Of 42 tasks, 40 yield Candidate New Discovery or Extension, giving an SFR of 95.24%, compared with 78.57% (33/42) for Codex(OpenAI, 2025) and 76.19% (32/42) for Claude Code(Anthropic, 2025), as shown in Table 1. Compared with Claude Code, DeepQuest yields more outcomes classified as

Extension (20 vs. 12) and fewer classified as Skip (1 vs. 8). Compared with Codex, it yields more classified as Candidate New Discovery (20 vs. 14) and fewer classified as Uncertain (1 vs. 9). These gains reflect the complete system and cannot be attributed to a single module.

Table 1: Discovery outcome comparison across methods. Candidate New denotes Candidate New Discovery. Scientific Finding Rate (SFR) represents the proportion of tasks classified as Candidate New Discovery, Rediscovery, or Extension. Top-10 SFR is computed among the highest-ranked DeepQuest-H tasks. N/A indicates not applicable.

	Discovery outcome					Finding rate	
Method	**Candidate New**	**Rediscovery**	**Extension**	**Uncertain**	**Skip**	**SFR**	**Top-10 SFR**
GPT-H ($n = 29$)							
Codex	13 (44.83%)	5 (17.24%)	4 (13.79%)	7 (24.14%)	0 (0.00%)	22 (75.86%)	N/A
Claude Code	7 (24.14%)	2 (6.90%)	11 (37.93%)	2 (6.90%)	7 (24.14%)	20 (68.97%)	N/A
Claude-H ($n = 30$)							
Codex	9 (30.00%)	0 (0.00%)	11 (36.67%)	8 (26.67%)	2 (6.67%)	20 (66.67%)	N/A
Claude Code	10 (33.33%)	0 (0.00%)	12 (40.00%)	2 (6.67%)	6 (20.00%)	22 (73.33%)	N/A
DeepQuest-H ($n = 42$)							
Codex	14 (33.33%)	2 (4.76%)	17 (40.48%)	9 (21.43%)	0 (0.00%)	33 (78.57%)	9 (90.00%)
Claude Code	20 (47.62%)	0 (0.00%)	12 (28.57%)	2 (4.76%)	8 (19.05%)	32 (76.19%)	8 (80.00%)
DeepQuest	20 (47.62%)	0 (0.00%)	20 (47.62%)	1 (2.38%)	1 (2.38%)	40 (95.24%)	10 (100.00%)

Under either independent executor, DeepQuest-H achieves the highest SFR among the three hypothesis sets. For DeepQuest-H, GPT-H, and Claude-H, SFRs are 78.57%, 75.86%, and 66.67% with Codex, and 76.19%, 68.97%, and 73.33% with Claude Code, respectively. Because the sets differ in size and composition, these comparisons reflect combined generation and selection performance rather than an isolated effect of hypothesis quality.

Top-10 SFRs were 90%, 80%, and 100% under Codex, Claude Code, and DeepQuest, respectively, exceeding their full-set rates. All ten DeepQuest outcomes were Candidate New Discovery or Extension. Given the small subset and lack of an additional ranking baseline, these results provide only descriptive evidence of ranking effectiveness.

4.4 Representative Medical Findings and External Replication

4.4.1 Neuropsychiatric Symptoms and Subsequent Dementia Risk

First observed neuropsychiatric symptom positivity was associated with higher incident dementia risk in the two-year lag analysis, as shown in Fig. 4. Using the 10 Neuropsychiatric Inventory Questionnaire (NPI-Q) symptoms, DeepQuest constructed time-varying risk intervals for 0, 365, and 730-day lags and fitted time-varying Cox models to symptom-history states. At the 730-day lag, first observed positivity had a hazard ratio (HR) of 2.496 (95% CI: 1.528–4.077) in OASIS-3 and 3.326 (1.443–7.668) in ADNI, relative to no prior or current positive symptom record.

In OASIS-3, HR point estimates were 1.847 for prior positivity with current negativity, 2.496 for first observed positivity, and 3.598 for positivity both previously and at the current visit. ADNI did not show the same ordering of point estimates, but first observed symptom positivity was likewise associated with a higher risk of incident dementia under the 730-day lag.

4.4.2 Longitudinal Preservation of Relative Regional Amyloid Distribution

DeepQuest investigated whether regional amyloid distribution provides a stable imaging phenotype beyond global amyloid burden, as shown in Fig. 4. Using the first and last eligible PET scans, it applied a centered log-ratio transformation to derive relative amyloid burden measures for the frontal, parietal, and occipital composite regions. Longitudinal preservation was then evaluated after adjusting for global cortical amyloid burden and relevant covariates.

In OASIS-3 (n=284), the within-region mapping coefficients were 0.7276 (95% CI:0.5587–0.8965) for the frontal region and 0.9471(0.7881–1.1061) for the parietal region. Corresponding estimates in AIBL (n = 159) were 0.6243 (0.2257–1.0229) and 0.5779 (0.0648–1.0910). All four estimates

support within-region preservation across visits. The partial R^2 values for baseline regional configuration were 0.7638 in OASIS-3 and 0.7193 in AIBL. These results support the longitudinal stability of amyloid distribution and its potential as an imaging phenotype beyond global burden.

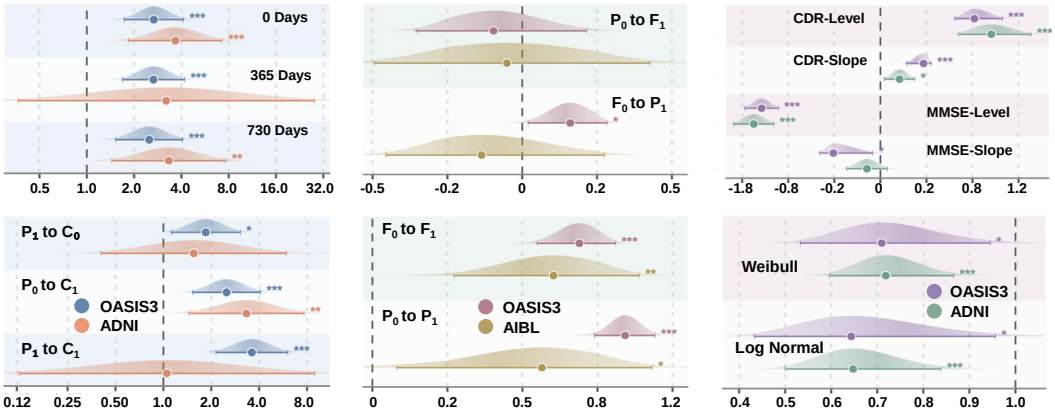

Figure 4: Representative case studies produced by DeepQuest. Results are validated in OASIS-3 with external evaluation in ADNI or AIBL. Left: NPS–dementia associations across lags (upper) and symptom-history states at 730-day lag (lower). Mid: cross- and within-region longitudinal amyloid associations. Right: neocortical tau associations with cognitive levels/slopes (upper) and AFT time ratios for clinical impairment (lower). Dots and bars indicate estimates and 95% CIs.

4.4.3 Neocortical Tau and Clinical Impairment Trajectories and Timing

Higher neocortical tau was associated with greater clinical impairment and faster Clinical Dementia Rating Sum of Boxes (CDR-SB) increases, whereas associations with Mini-Mental State Examination (MMSE) change differed across cohorts, as shown in Fig. 4. Using standardized neocortical tau burden from the first tau PET scan, DeepQuest fitted linear mixed-effects models with subject-specific random intercepts and tau-by-time interactions. Both cohorts showed higher CDR-SB and lower MMSE levels with greater neocortical tau. CDR-SB slope associations were 0.1877 (95% CI:0.1134–0.2619) in OASIS-3 and 0.0831(0.0171–0.1491) in ADNI, supporting an association between higher neocortical tau and faster clinical progression. MMSE slope associations coefficients were -0.2113 (-0.3890, -0.0335) in OASIS-3 and -0.0583 (-0.1470, 0.0304) in ADNI, indicating longitudinal association replication depends on the clinical measure.

Survival models incorporating censoring further linked higher neocortical tau to earlier recorded clinical impairment across the available clinical history. DeepQuest defined the first occurrence of global CDR (> 0) as the event and fitted accelerated failure time (AFT) models to 405 OASIS-3 and 578 ADNI participants. Weibull time ratios were 0.7097 (95% CI: 0.5327–0.9454) in OASIS-3 and 0.7185 (0.5964–0.8657) in ADNI. Log-normal models showed effects in the same direction. These results support associations of neocortical tau with clinical progression and earlier impairment. Further implementation details are provided in Appendix B.

5 Conclusion

In this work, we introduced DeepQuest, a knowledge-gap-driven AI scientist framework for Alzheimer's disease research. It integrates hypothesis formulation, data-constrained prioritization, and iterative experimentation with two-level auditing that distinguishes experimental reliability from knowledge contribution. Each candidate contribution is linked to experimental evidence and assessed against a frozen literature corpus, making its evidential basis explicit. Evaluation on OASIS-3 demonstrates improved hypothesis quality and audit-assessed scientific finding rates over the evaluated baselines, with key associations supported in independent ADNI and AIBL cohorts. These results highlight a path beyond automated experimentation: selecting meaningful, answerable questions and assessing what the resulting evidence adds to existing knowledge.

AI USE STATEMENT

We used generative AI tools solely for manuscript refinement, including language translation, grammar correction, and improving clarity and readability. AI tools were not used for research ideation, hypothesis generation, experimental design, data analysis, result generation, dataset creation, mathematical derivations, proofs, or scientific claims. All AI-assisted suggestions were reviewed and verified by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

Because AI-generated scientific outputs may contain unsupported interpretations or overlook important limitations, DeepQuest is designed with safeguards to promote responsible AI-assisted scientific discovery. All analyses in this work are conducted using publicly available datasets with established data governance and privacy protection procedures. The system does not access or collect additional personally identifiable information.

The findings generated by DeepQuest should be regarded as evidence-supported research hypotheses or candidate scientific insights, rather than final scientific conclusions or clinical recommendations. AI-generated results require independent validation, expert interpretation, and further investigation before being considered for scientific publication or practical application. DeepQuest is intended to augment human scientific reasoning rather than replace expert judgment. The responsibility for experimental design, result interpretation, and final scientific claims remains with human researchers.

REPRODUCIBILITY STATEMENT

Source code and evaluation artifacts are available at `https://anonymous.4open.science/r/DeepQuest-3066/`

REFERENCES

Anthropic. Claude code, 2025. URL `https://claude.ai/claude-code`.

Kathryn A Ellis, Ashley I Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T Lautenschlager, Nat Lenzo, Ralph N Martins, Paul Maruff, et al. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International psychogeriatrics*, 21(4):672–687, 2009.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York, 1978.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Dmytro Shved, Gavin J Gyimesi, Jon M Laurent, Samantha M Wright, Muhammed T Razzak, et al. A multi-agent system for automating scientific discovery. *Nature*, pp. 1–3, 2026.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Keqi Han, Songlin Zhao, Yao Su, Xiang Li, Yixuan Yuan, Lifang He, and Carl Yang. Towards a virtual neuroscientist: Autonomous neuroimaging analysis via multi-agent collaboration. *arXiv preprint arXiv:2605.09366*, 2026.

Shimaa A Heikal, Eman M Khedr, Gharib Fawi, Mai Othman, Nesma G Elsheikh, Heba M Tawfik, Hany I Hassanin, Sara ElFarrash, Samer Salama, Eman M Ali, et al. Outcome-specific risk factors challenge universal dementia prevention priorities: Findings from the egyptian dementia network registry. *Alzheimer's & Dementia*, 22(6):e71527, 2026.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Ryan Li, Yusuf Roohani, Lin Qiu, Shiyi Cao, Gavin Li, et al. Autonomous biomedical research with an artificial intelligence agent. *Science*, pp. eadz4351, 2026.

Clifford R Jack Jr, J Scott Andrews, Thomas G Beach, Teresa Buracchio, Billy Dunn, Ana Graf, Oskar Hansson, Carole Ho, William Jagust, Eric McDade, et al. Revised criteria for diagnosis and staging of alzheimer's disease: Alzheimer's association workgroup. *Alzheimer's & Dementia*, 20(8):5143–5169, 2024.

Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv*, pp. 2019–12, 2019.

Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541, 2025a.

Yitong Li, Morteza Ghahremani, Youssef Wally, and Christian Wachinger. Diamond: Dementia diagnosis with multi-modal vision transformers using mri and pet. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 107–116. IEEE, 2025b.

Ethan Lin, Zhiyuan Peng, and Yi Fang. Evaluating and enhancing large language models for novelty assessment in scholarly publications. In *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, pp. 46–57, 2025.

Shengyuan Liu, Jia-Xuan Jiang, Boyun Zheng, Cheng Wang, Zipei Wang, Wentao Pan, Hongtao Wu, Houwen Peng, Yu Gu, Lichao Sun, et al. Towards autonomous and auditable medical imaging model development. *arXiv preprint arXiv:2607.10522*, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C Landsness, Daniel L Barabasi, Siddharth Narayanan, Nicky Evans, et al. Kosmos: An ai scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

OpenAI. Codex, 2025. URL `https://openai.com/codex`.

Ketan S Saichandran, Karim Elzokm, Osman B Guney, and Vijaya B Kolachalama. Agentic ai for automated hypothesis testing in alzheimer's disease and related dementias. *medRxiv*, 2025.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, 2025.

Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S Weld, and Tom Hope. Literature-grounded novelty assessment of scientific ideas. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pp. 96–113, 2025.

Atsushi Shimizu, Yu Iwabuchi, Shogyoku Bun, Memi Watanabe, Masahito Kubota, Sho Shimohama, Toshiki Tezuka, Keisuke Takahata, Hajime Tabuchi, Morinobu Seki, et al. Florzolotau (18f) retention is linked to neuropsychological performance in tauopathy. *Alzheimer's & Dementia*, 22(4):e71319, 2026.

E Keats Shwab, Gita A Pathak, Joshua Harvey, Michael E Belloy, Corinne E Fischer, Michael W Lutz, Sonja W Scholz, Noah Cook, Danielle M Reid, Jingchun Chen, et al. Leveraging multiomic approaches to elucidate mechanisms of heterogeneity in alzheimer's disease: Neuropsychiatric symptoms, co-pathologies, and sex differences. *Alzheimer's & Dementia*, 21(8):e70549, 2025.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 5548–5566, 2023.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *Advances in Neural Information Processing Systems*, 38:9481–9520, 2025.

Arthur W Toga, Scott Neu, Sidney Taiko Sheehan, Karen Crawford, and Alzheimer's Disease Neuroimaging Initiative. The informatics of adni. *Alzheimer's & Dementia*, 20(10):7320–7330, 2024.

Alessio Travaglia and Steve Hoffmann. Data sharing as the foundation of discovery: Adni and breakthroughs in alzheimer's disease. *Alzheimer's & Dementia*, 20(10):7395–7398, 2024.

Florian Trost, Bide Zhang, Ines Aring, Marcus Bauer, Lennert Glamann, Michael Wessolly, Kyra Johnson, Heike Göbel, Tristan Lerbs, Taban Sangenne, et al. An agentic framework for autonomous scientific discovery in cancer pathology. *Nature Medicine*, pp. 1–13, 2026.

Dallas P Veitch, Melanie J Miller, Shaveta Kanoria, Paul S Aisen, Laurel A Beckett, Robert C Green, Danielle J Harvey, Clifford R Jack Jr, William Jagust, Edward B Lee, et al. Contributions of the alzheimer's disease neuroimaging initiative to advancing ad research: a targeted review of recent publications. *Alzheimer's & Dementia*, 22(6):e71566, 2026.

Chenxi Wang, Weiwei Zhang, Ming Ni, Qiong Wang, Chang Liu, Linbin Dai, Mengguo Zhang, Yong Shen, and Feng Gao. Deep-learning based multi-modal models for brain age, cognition and amyloid pathology prediction. *Alzheimer's Research & Therapy*, 17(1):126, 2025.

Michael W Weiner, Shaveta Kanoria, Melanie J Miller, Paul S Aisen, Laurel A Beckett, Catherine Conti, Adam Diaz, Derek Flenniken, Robert C Green, Danielle J Harvey, et al. Overview of alzheimer's disease neuroimaging initiative and future clinical trials. *Alzheimer's & Dementia*, 21(1):e14321, 2025.

Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. Deepscientist: Advancing frontier-pushing scientific findings progressively. In *International Conference on Learning Representations*, volume 2026, pp. 47981–48037, 2026.

Hongtao Wu, Boyun Zheng, Dingjie Song, Yu Jiang, Jianfeng Gao, Lei Xing, Lichao Sun, and Yixuan Yuan. Towards a medical ai scientist. *arXiv preprint arXiv:2603.28589*, 2026.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Kun Zhao, Pindong Chen, Dong Wang, Rongshen Zhou, Guolin Ma, and Yong Liu. A multiform heterogeneity framework for alzheimer's disease based on multimodal neuroimaging. *Biological Psychiatry*, 97(11):1022–1033, 2025.

Lujia Zhong, Yihao Xia, Jianwei Zhang, Jiaxin Yue, Mingyang Xia, Yonggang Shi, et al. Neuroagent: Llm agents for multimodal neuroimaging analysis and research. *arXiv preprint arXiv:2605.06584*, 2026.

A Evaluation of Hypothesis Quality and Scientific Discovery

A.1 Protocol for Hypothesis Quality Evaluation

Experimental Setup. DeepQuest included 42 candidate hypotheses that passed data availability screening. For comparison, we selected five language models as baselines: GPT-5.6-luna, GPT-5.6-sol, Claude Sonnet 5, Claude Opus 4.8, and DeepSeek-v4-pro-0813. Each baseline model generated 20 hypotheses. All baseline models received the same 10 Alzheimer's disease review articles used by DeepQuest as literature input. During question refinement, DeepQuest used automatically generated data summaries to identify research questions supported by the available data. To reduce differences in data context, the official OASIS-3 overview was also provided to all baseline models.

The comparison therefore evaluates the respective hypothesis-generation and selection workflows, including differences in data context and feasibility screening.

Evaluation Protocol. GPT-5.6-sol, Claude Opus 4.8, and DeepSeek-v4-pro-0813 served as independent evaluators. All three models received the same candidate hypotheses and standardized data description, and evaluated each hypothesis across five dimensions: Novelty, Scientific Advancement, Clinical Significance, Empirical Discriminative Power, and Feasibility. Novelty measures the extent to which a hypothesis introduces new insights beyond the provided reviews and known research questions. Scientific Advancement assesses its potential to resolve scientific uncertainty or advance the investigation of unknown relationships. Clinical Significance evaluates its potential value for clinical interpretation, stratification, monitoring, or understanding of outcomes. Empirical Discriminative Power measures whether the available data and analyses can distinguish among alternative explanations or competing expectations. Feasibility assesses whether the required variables, samples, and analytical conditions are attainable within the given data context. Each dimension was rated on a 1 to 5 scale. Overall was calculated as the equally weighted mean of the five dimensions. Results are reported as hypothesis level means and sample standard deviations.

A.2 Evaluation of Scientific Finding Rate

Hypothesis Set Construction. The execution evaluation used hypothesis sets from three sources. DeepQuest-H comprised all 42 DeepQuest hypotheses that passed data availability screening. GPT-H combined hypotheses generated by GPT-5.6-luna and GPT-5.6-sol. Researchers verified the required variables and analytical conditions against the available data assets, retaining 29 hypotheses. Claude-H was constructed analogously by combining hypotheses generated by Claude Opus 4.8 and Claude Sonnet 5, with 30 hypotheses retained. DeepQuest Top-10 consisted of the highest-priority tasks selected from the 42 DeepQuest-H hypotheses according to groupwise ranking. Independent execution environments included Codex powered by GPT-5.6-luna and Claude Code powered by Claude Sonnet 5, both of which were compared with the complete DeepQuest pipeline.

On DeepQuest Top-10, Codex produced auditable conclusions for 9 of 10 tasks, yielding an SFR of 90%. Across all 10 tasks, 5 were classified as Candidate New Discovery, 3 as Extension, 1 as Rediscovery, and 1 as Uncertain. Claude Code produced auditable conclusions for 8 of 10 tasks, yielding an SFR of 80%. Its 10 task outcomes comprised 4 Candidate New Discovery, 4 Extension, 1 Uncertain, and 1 Skip. DeepQuest produced auditable conclusions for all 10 tasks, yielding an SFR of 100%. Of these, 6 were classified as Candidate New Discovery and 4 as Extension, with no Rediscovery, Uncertain, or Skip outcomes.

B Additional Details on Representative Medical Findings

B.1 Neuropsychiatric Symptoms and Subsequent Dementia Risk

B.1.1 Symptom Definition and Sample Inclusion

Neuropsychiatric symptoms were defined using the ten symptom items of the NPI-Q. A visit was used to update exposure status only when all ten items had valid records, the visit was marked as a complete NPI-Q assessment, and the participant had not developed dementia by that visit. NPI-Q records obtained after dementia onset were not used to define exposure. Based on current and prior records, symptom status was classified into four categories: negative both previously and currently, previously positive but currently negative, previously negative but currently positive, and positive both previously and currently.

The OASIS-3 input data comprised 7,172 records from 1,071 participants. Among them, 1,052 participants contributed 5,947 candidate visits that met the criteria for dementia-free status and complete NPI-Q assessment. After applying lags of 0, 365, and 730 days, the three risk sets included 880 participants with 190 events, 868 participants with 178 events, and 837 participants with 149 events, respectively. The raw ADNI data comprised 4,982 records from 593 participants. Among them, 425 participants had complete NPI-Q assessments, and 117 incident dementia events were identified.

B.1.2 Risk Intervals and Statistical Models

Lag windows of 0, 365, and 730 days were evaluated separately. For each window, counting-process risk intervals were reconstructed. Each risk interval began on the date of the current complete NPI-Q visit plus the corresponding lag and ended at the earliest of the next complete NPI-Q visit plus that lag, the participant's last observation date, or the first dementia date. An event was recorded when the interval ended at the first dementia date. Only intervals beginning before dementia onset contributed risk time. Consequently, the three lag windows yielded distinct risk sets rather than simple temporal shifts of exposure dates within an identical sample.

The OASIS-3 analyses at lags of 0, 365, and 730 days included 880 participants with 190 events, 868 participants with 178 events, and 837 participants with 149 events, respectively. The corresponding ADNI analyses included 256 participants with 35 events, 226 participants with 33 events, and 200 participants with 29 events.

Symptom history was analyzed using a time-varying Cox proportional hazards model, with participants who were negative both previously and currently as the reference group. Models adjusted for baseline age, sex, years of education, and cumulative vascular status, with participant-clustered robust standard errors. Cumulative vascular status was derived from medical history recorded up to the corresponding NPI-Q exposure visit. Intervals with unknown vascular status were excluded.

B.2 Longitudinal Preservation of Relative Regional Amyloid Distribution

B.2.1 Regional Measures and Inclusion of Repeated PET Scans

The analysis summarized amyloid burden across the frontal, parietal, and prespecified occipital composite regions. The occipital composite comprised the cuneus, lateral occipital cortex, and lingual gyrus. For each PET visit, the mean standardized uptake value ratio (SUVR) was first calculated for each of the three composite regions, followed by a centered log-ratio transformation:

$$\mathrm{clr}(r_k) = \log(r_k) - \frac{1}{3}\sum_{j=1}^{3}\log(r_j), \tag{7}$$

where r_k denotes the SUVR of the k-th composite region, and $clr(r_k)$ denotes its centered log-ratio transformed value. Because the three clr coordinates sum to zero by construction, only the frontal and parietal coordinates were retained as two nonredundant dimensions.

OASIS-3 included participants with at least two eligible amyloid PET scans and complete regional and global cortical SUVR measurements. The first and last eligible PET scans were paired, yielding a final sample of 284 participants. For AIBL, the two PET scans were required to come from the same participant, use the same tracer, and fall within the prespecified intervisit interval. The PiB-only analysis ultimately included 159 participants.

B.2.2 Multivariate Topographic Model

The follow-up frontal and parietal coordinates were modeled jointly as outcomes, with the baseline frontal and parietal coordinates as the primary predictors. The model additionally adjusted for sex, scan interval, and global cortical amyloid burden at both baseline and follow-up. Tracer type was further adjusted for when multiple tracers were present. The model can be written as:

$$\begin{pmatrix} F_1 \\ P_1 \end{pmatrix} = \mathbf{B}\begin{pmatrix} F_0 \\ P_0 \end{pmatrix} + \mathbf{\Gamma}C + \boldsymbol{\varepsilon}, \tag{8}$$

where F_0 and P_0 denote the baseline frontal and parietal clr-transformed coordinates, F_1 and P_1 denote the corresponding follow-up coordinates, and C denotes the covariates.

In OASIS-3, the cross-regional coefficient for $F_1 \leftarrow P_0$ was -0.0767 (95% CI: -0.3256–0.1722). The coefficient for $P_1 \leftarrow F_0$ was 0.1282 (0.0164–0.2404). The joint test of the four baseline topographic coefficients yielded $\chi^2(4) = 509.94$, $P = 4.74 \times 10^{-109}$.

In AIBL, the coefficient for $F_1 \leftarrow P_0$ was -0.0409 (95% CI: -0.4938–0.4121). The coefficient for $P_1 \leftarrow F_0$ was -0.1086 (-0.4465–0.2292). The joint test yielded $\chi^2(4) = 51.440$, $P = 1.81 \times 10^{-10}$.

B.3 Neocortical Tau and Clinical Impairment Trajectories and Timing

B.3.1 Linear Mixed-Effects Models of Cognitive Trajectories

Across both cohorts, time was defined in years relative to the first eligible tau PET scan (date difference divided by 365.25), and only post-scan cognitive assessments were included. CDR-SB and MMSE were modeled in their original score units.

In OASIS-3, the first eligible tau PET scan for each participant was defined as baseline and matched to the nearest quantitative amyloid PET scan within 365 days. Medial temporal and neocortical tau were defined as mean SUVR across the entorhinal cortex and hippocampus, and across the inferior parietal cortex, precuneus, and temporal pole, respectively. Tau measures were standardized across participants with imaging data, whereas amyloid burden was standardized within tracer. The complete-case CDR-SB and MMSE models included 981 observations from 374 participants and 577 observations from 301 participants, respectively, and were fitted using linear mixed-effects models with participant-specific random intercepts.

In ADNI, the earliest tau PET scan that passed quality control was used for each participant. The CDR-SB model included 1,697 observations from 561 participants, and the MMSE model included 1,564 observations from 558 participants. Models adjusted for medial temporal tau, age, sex, and years of education. Because a matched quantitative amyloid measure was unavailable, amyloid burden, APOE status, imaging-pair interval, and amyloid tracer were not included as covariates. Accordingly, ADNI served as an external replication under a reduced covariate specification and should not be described as using a model identical to that of OASIS-3.

B.3.2 Interval-Censored Accelerated Failure Time Model

Event-time bounds were measured in days relative to the first eligible tau PET scan. Before fitting, all finite bounds were shifted by a cohort-specific constant (5,609 days in OASIS-3 and 5,231 days in ADNI) to ensure positivity. Reported time ratios therefore refer to the shifted time scale.

The first occurrence of global CDR greater than 0 was defined as the clinical impairment event. If a CDR of 0 had been recorded before the first positive assessment, the event time was considered to lie between the last normal assessment and the first positive assessment and was treated as interval-censored. Participants whose first valid assessment was already positive were treated as left-censored. Participants who remained normal throughout follow-up were treated as right-censored at their last normal assessment.

OASIS-3 included 405 participants, comprising 57 left-censored, 41 interval-censored, and 307 right-censored cases. The base model included standardized amyloid burden, medial temporal tau, the date difference between the paired amyloid and tau PET scans, and amyloid tracer. The full model additionally included neocortical tau:

$$M_0: \quad \mathrm{Surv}(L, R) \sim \text{amyloid} + \text{MTL Tau} + \text{pairing interval} + \text{tracer}, \tag{9}$$

$$M_1: \quad M_0 + \text{neoTau}, \tag{10}$$

where L and R denote the lower and upper bounds of the impairment event time. M_0 is the reduced model, and M_1 is the full model including neocortical tau. ADNI included 578 participants, comprising 228 left-censored, 106 interval-censored, and 244 right-censored cases. The full model included neocortical tau, medial temporal tau, age, sex, and years of education. The reduced model excluded neocortical tau.

Both cohorts were modeled using Weibull and log-normal distributions, with the distribution selected within each cohort according to the AIC of the full model. In OASIS-3, the AIC values were 629.95 for the Weibull model and 646.11 for the log-normal model. In ADNI, the corresponding values were 1463.64 and 1497.68. The Weibull model yielded the lower AIC in both cohorts.

C FINDING REPRODUCTION

C.1 EXPERIMENTAL SETUP

We use published studies as external references and evaluate DeepQuest through finding reproduction, focusing on its understanding of scientific questions, experimental reasoning, and analytical execution. We select Shimizu et al. (2026), which studied spatial associations between regional tau deposition and cognition, and Heikal et al. (2026), which examined differences between clinical diagnosis and continuous cognitive outcomes. The goal is to test whether DeepQuest can independently recover key scientific relationships under available data constraints.

We use GPT-5.6 to abstract each study into a high level Limitation–Hypothesis (L–H) pair and provide an initial experiment design. The L–H pair preserves the original knowledge gap and target relationship, while the initial design specifies only the analysis objective and comparison direction. DeepQuest then refines the experiment autonomously. Because it receives an abstract scientific question rather than the original analysis pipeline, finding reproduction evaluates whether the system can recover the key scientific relationship through an independently designed experiment, rather than replicate the original procedure step by step. DeepQuest subsequently operationalizes the hypothesis, designs and executes the analysis using OASIS-3, and compares the resulting findings with the core conclusions of the original study.

C.2 REGIONAL TAU DEPOSITION AND COGNITIVE FUNCTION

Shimizu et al. (2026) used Florzolotau (18F) PET to examine spatial associations between regional tau deposition and neuropsychological performance across the AD continuum. Higher tau retention was generally associated with poorer cognition, with prominent signals in posterior cortical regions including the temporal and parietal lobes. The study performed voxelwise analyses in SPM12, normalized tau PET to Montreal Neurological Institute (MNI) space, and computed standardized uptake value ratios (SUVRs). Voxelwise regressions adjusted for age and sex were combined with cluster level family wise error (FWE) correction. Memory measures showed the largest tau associated clusters, with 28,425 voxels for Logical Memory delayed recall (LM-delay) and 28,210 for Logical Memory immediate recall (LM-immediate). Attention and executive measures yielded 21,149 voxels for Trail Making Test Part B (TMT-B) and 5,662 for Part A (TMT-A). In item level analyses, ADAS-Cog word recall and word recognition were mainly associated with left temporal tau retention. MMSE registration and delayed recall were primarily associated with bilateral temporal tau, whereas calculation was mainly associated with the left temporal region.

DeepQuest operationalized this question in OASIS-3 as an association analysis between regional tau burden and cognitive performance. Using AV-1451 PET, the system selected five regions of interest (ROIs): entorhinal cortex, hippocampus, temporal pole, inferior parietal cortex, and precuneus. Memory was assessed with Selective Reminding Test–Free Recall (SRTFree), while attention and executive function were assessed with TMT-A. Models adjusted for age, sex, and education, used HC3 robust standard errors, and applied Benjamini–Hochberg correction across all primary region-by-measure comparisons.

The memory results broadly reproduced the spatial pattern reported in the original study. Higher tau burden in all five regions was associated with poorer free recall and remained significant after FDR correction. The strongest association was observed in the entorhinal cortex, with $\beta = -0.3485$, followed by the temporal pole, inferior parietal cortex, precuneus, and hippocampus, with $\beta = -0.2968$, -0.2945, -0.2590, and -0.1973, respectively. These results support the reported association between temporal tau and poorer memory and suggest that the relationship may extend beyond the medial temporal lobe. Attention and executive function showed a more posterior cortical pattern. TMT-A was most strongly associated with the inferior parietal cortex, with $\beta = 0.4611$, and precuneus, with $\beta = 0.4774$. Associations in the entorhinal cortex, temporal pole, and hippocampus were not significant after FDR correction. This spatial pattern is consistent with the posterior cortical distribution of TMT related tau associations reported in the original study.

DeepQuest reproduced the core finding that higher regional tau burden is associated with poorer cognitive performance. Across different data and analytical settings, it preserved the spatial pattern of stronger temporal associations for memory and stronger posterior association cortex involvement for attention and executive function. The original study used Florzolotau (18F) PET and voxelwise

Table 2: Regional tau associations with cognitive performance. β denotes regression coefficients with 95% confidence intervals, p and q denote unadjusted and FDR-adjusted p-values, respectively. BH correction was applied across the full family of region-by-measure comparisons, including comparisons not displayed in this table.

Brain region	Test	β [95% CI]	p	q
Entorhinal cortex	SRTFree	-0.3485 [$-0.4525, -0.2445$]	5.14×10^{-11}	1.03×10^{-9}
	TMT-A	$+0.1656$ [0.0124, 0.3188]	3.415×10^{-2}	7.588×10^{-2}
Hippocampus	SRTFree	-0.1973 [$-0.3187, -0.0760$]	1.435×10^{-3}	4.100×10^{-3}
	TMT-A	$+0.0768$ [-0.0204, 0.1741]	1.213×10^{-1}	2.425×10^{-1}
Temporal pole	SRTFree	-0.2968 [$-0.4010, -0.1925$]	2.41×10^{-8}	2.41×10^{-7}
	TMT-A	$+0.1610$ [0.0204, 0.3016]	2.485×10^{-2}	6.212×10^{-2}
Inferior parietal	SRTFree	-0.2945 [$-0.4573, -0.1316$]	3.93×10^{-4}	1.573×10^{-3}
	TMT-A	$+0.4611$ [0.2115, 0.7107]	2.93×10^{-4}	1.467×10^{-3}
Precuneus	SRTFree	-0.2590 [$-0.3806, -0.1375$]	2.96×10^{-5}	1.97×10^{-4}
	TMT-A	$+0.4774$ [0.2026, 0.7522]	6.61×10^{-4}	2.205×10^{-3}

spatial analysis, whereas DeepQuest evaluated the corresponding relationship using AV-1451 PET, predefined ROIs, and cognitive measures available in OASIS-3 that approximate the original assessments. Therefore, the results are better interpreted as cross-setting validation of the reported finding across different datasets, tau tracers, and analytical settings, rather than as an exact reproduction of the original experimental pipeline.

C.3 Clinical Impairment Status and Outcome Specificity of Continuous Cognitive Performance

Heikal et al. (2026) showed that clinical dementia diagnosis and continuous cognitive performance are related but capture distinct dimensions of disease phenotype. The same factor may influence the risk of crossing a clinical diagnostic threshold, while others may affect only continuous cognitive performance. Using 660 participants with mild cognitive impairment (MCI) or dementia from the Egyptian Dementia Network, the study applied logistic regression to dementia diagnosis and linear regression to continuous MMSE scores. Education and social engagement were associated with both outcomes, whereas several vascular, metabolic, and genetic factors were associated only with MMSE performance. These findings indicate distinct association patterns for binary clinical status and continuous cognitive function.

DeepQuest operationalized this question in OASIS-3 as a paired comparison of clinical impairment status and continuous cognitive performance within the same visit. The analysis included 1,376 participants with valid CDR and MMSE records. Clinical impairment was defined as CDRTOT > 0, and MMSE represented continuous cognitive performance. Logistic regression was used for the CDR outcome, with odds ratios (OR) and marginal risk differences reported, while MMSE was modeled using censored regression to account for the 30 point ceiling. All models adjusted for age, sex, education, socioeconomic status, and study time.

DeepQuest first identified factors associated with both outcomes. A one standard deviation increase in age was associated with higher clinical impairment risk, with an OR of 2.120, and a 0.849 point decrease in MMSE. A one-standard-deviation increase in education was associated with lower odds of impairment (OR = 0.646) and higher MMSE scores (β = 0.406). APOE $\varepsilon 4$ homozygosity was also associated with higher impairment risk, with an OR of 6.224, and a 1.584 point decrease in MMSE. These results indicate that the two outcomes share several risk factors.

Several factors showed different patterns of statistical significance across outcomes. Race and socioeconomic status were primarily associated with MMSE, with effects of -0.426 and -1.825, respectively, but showed no significant association with CDR defined clinical impairment. Anxiety and elation were primarily associated with clinical impairment, with ORs of 3.077 and 4.078, but

Table 3: Outcome-specific association patterns for clinical and cognitive measures. OR and β denote odds ratios and regression coefficients for the respective outcomes with 95% confidence intervals. Age and education are expressed per standard deviation. Comparisons are APOE $\varepsilon 4$: $\varepsilon 4/\varepsilon 4$ vs $\varepsilon 3/\varepsilon 3$. Race: Black vs White. SES denotes socioeconomic status. ANXIETY and OCD denote a diagnosis or history of anxiety and obsessive-compulsive disorder, respectively. ELAT denotes elation or elevated mood. FOCLDEF denotes focal neurological deficits. q denotes the BH-adjusted p-value across all tested comparisons within each factor group and outcome.

	CDR			**MMSE**		
Factor	**OR [95% CI]**	p	q	β **[95% CI]**	p	q
Age	2.120 [1.823, 2.464]	1.33×10^{-22}	6.64×10^{-22}	-0.849 [-0.979, -0.692]	3.95×10^{-31}	1.97×10^{-30}
Education	0.646 [0.533, 0.783]	8.09×10^{-6}	4.05×10^{-5}	$+0.406$ [0.275, 0.574]	1.07×10^{-7}	5.35×10^{-7}
APOE $\varepsilon 4$	6.224 [3.623, 10.690]	3.48×10^{-11}	1.74×10^{-10}	-1.584 [-2.562, -0.844]	3.03×10^{-4}	7.57×10^{-4}
Race	0.737 [0.506, 1.073]	1.12×10^{-1}	1.86×10^{-1}	-0.426 [-0.735, -0.089]	9.82×10^{-3}	1.64×10^{-2}
SES	0.933 [0.290, 3.009]	9.08×10^{-1}	9.08×10^{-1}	-1.825 [-3.426, -0.559]	1.26×10^{-2}	3.15×10^{-2}
ANXIETY	3.077 [1.438, 6.585]	3.78×10^{-3}	7.55×10^{-3}	-0.246 [-0.740, 0.176]	2.92×10^{-1}	3.69×10^{-1}
OCD	0.595 [0.065, 5.461]	6.46×10^{-1}	8.62×10^{-1}	$+0.894$ [0.215, 1.638]	1.38×10^{-2}	2.55×10^{-2}
ELAT	4.078 [1.688, 9.855]	1.79×10^{-3}	4.30×10^{-3}	-0.108 [-1.258, 0.651]	8.24×10^{-1}	8.24×10^{-1}
FOCLDEF	3.185 [1.121, 9.047]	2.96×10^{-2}	5.08×10^{-2}	-1.934 [-3.714, -0.822]	8.77×10^{-3}	1.75×10^{-2}

not with MMSE. In contrast, obsessive-compulsive disorder was significantly associated with higher MMSE, whereas focal neurological deficits were primarily associated with lower MMSE.

DeepQuest reproduced the finding that binary cognitive impairment status and continuous cognitive performance share some associated factors while also exhibiting distinct association patterns, with some factors showing opposite point estimate directions. The original study used clinical diagnoses of dementia and MCI as the binary outcome, whereas DeepQuest used CDRTOT > 0 and censored regression for MMSE. These results therefore provide cross setting validation across different populations, outcome definitions, and analytical settings, showing that DeepQuest can recover both shared and outcome specific patterns in cognitive phenotypes.