

A Scoping Review of Foundation Models in Dementia Research

Matilda Beinat

Queen Mary University of London

Eamonn Walsh

King's College London

Charles Marshall

Queen Mary University of London

Venet Osmani

Queen Mary University of London

Petroula Proitsi

p.proitsi@qmul.ac.uk

Queen Mary University of London

Article

Keywords: Artificial Intelligence, Foundation Models, Pre-trained Transformer, Large Language Models, Dementia, Alzheimer's Disease, Diagnosis, Detection, Challenges, Recommendations

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11129492/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Foundation models (FMs) are increasingly applied to dementia for multiple purposes, but their main scopes and limitations remain unclear. We examined how FMs are used for dementia detection, diagnosis, prognosis and risk prediction. Eligible studies applied a pre-trained, fine-tuned or commercial LLM FM to these tasks in people with or at risk of dementia, in English, January 2022 to August 2026. Searches of PubMed, Europe PMC, arXiv and Google Scholar yielded 474 unique records, screened to 50 analysed papers, charted by taxonomy, scope, dataset, performance and validation. Of these, 8 used pre-trained FMs, 26 fine-tuned FMs and 16 commercial LLMs (detection and diagnosis predominated). Performance was not comparable across classes (tasks, datasets and metrics differed). Demographic reporting was poor: race and ethnicity in 4/50 studies, age in 24/50, studies concentrated in the USA. The literature remains at proof of concept, requiring larger, diverse datasets and rigorous validation.

1 Introduction

Dementia poses a significant social care and health challenge costing the UK economy £42 billion per year (1), and by 2050 it is expected to cost the world economy \$15 trillion (2). It is the leading cause of death in the UK, with the expectation that 1 in 3 people born today will develop dementia, and of these, 2 in 3 people will develop Alzheimer's disease (AD)(3, 4). As life expectancy increases, cases of dementia are expected to rise correspondingly. New treatments, such as Lecanemab and Donanemab show promise in early disease stages (5), yet they have critical limitations and are not reimbursed in many countries due to concerns about cost effectiveness (6). Therefore, the urgency to improve early detection to enable the identification of effective disease-modifying treatments has never been greater.

Artificial Intelligence (AI) has shown promise within multiple areas of research, including dementia. Foundation models (FMs) are AI models trained on extensive and highly diverse data, covering multiple domains, generally with self-supervision, that can be adapted to a wide range of downstream tasks (7). Traditional supervised learning requires task-specific feature engineering and produces fixed, closed-form outputs, such as binary classification. FMs instead learn generalisable representations from large-scale, heterogeneous data and generate flexible, high-dimensional outputs, better suited to the complexity and multimodality of clinical data (8, 9). Literature nonetheless uses the term inconsistently. Studies involving FMs may be using Large Language Models (LLMs), Vision Transformers (ViTs), self-supervised learning, transfer learning, or conventional neural networks with pre-trained components. Instead of basing ourselves on a term, we defined an explicit Population, Concept, Context (PCC) and taxonomy (Tables 1–2). FMs are advancing rapidly in industry: learning from current applications, alongside challenges researchers may have faced, will accelerate progress in the field.

To gain firmer footing in this evolving area, we conducted a scoping review on the usage of FM approaches applied to dementia care research. This review therefore 1) examined how models are being used across different stages of the dementia care pathway by systematically categorising studies into their respective scopes and research objectives, 2) assessed their performance to identify patterns in

model performance, 3) identified research gaps that require further investigation, to realise the clinical potential of AI-driven approaches in dementia healthcare.

FMs may offer advantages where dementia research requires representation of heterogeneous information, transfer across tasks, or integration of multiple modalities. These properties matter because dementia data are siloed, expensive to annotate, and inconsistently available across sites. However, these theoretical advantages do not eliminate the requirement for sufficiently large, representative and accessible clinical datasets. In many of the articles reviewed here, smaller task-specific models remained competitive.

The evidence showed in this review should therefore be read as demonstrating a research opportunity as the state of FMs in dementia (strengths, limitations, and optimal methods of usage) is still not understood. The question therefore is to understand under which data, task and validation conditions their additional complexity provides a benefit that task-specific models do not. Specifically, we address the following questions: i) how FMs are built or applied in dementia research, exploring scopes, architectures and methods, ii) what the current evidence on their clinical performance is by understanding the performance ranges for each architecture, and iii) delving into limitations of each research paper to understand what the critical knowledge gaps are.

2 Methods

2.1 Design

This review adheres to PRISMA-ScR (10) guidelines and the five-stage scoping review framework of Arksey & O'Malley (11), updated by JBI (12). Eligibility was structured with the Population Concept Context (PCC) framework found in Table 1. This review analysed published literature only, no human participants were involved, and ethical approval was not required.

2.2 Eligibility Criteria

Concepts following the taxonomy set out in Table 2, applied to detection, diagnosis, prognosis or risk prediction. Window: January 2022 - August 2026, starting the year after the FM paradigm was formalised (Bommasani et al., first posted August 2021) (7).

Table 1
Population, Concept, Context (PCC) criteria for inclusion and exclusion of papers.

Element	Include	Exclude
Population	Studies discussing people with or at risk of dementia (AD, vascular, Lewy body, frontotemporal), MCI and pre-diagnostic cohorts	Do not include papers discussing Parkinson's disease without dementia outcomes. Do not include non-human studies. Do not include any form of review papers.
Concept	Models in taxonomy classes 1–3 (Table 2) for detection, diagnosis, prognosis/progression or risk prediction, traditional ML admissible as comparator arm only within studies	Traditional ML or statistical models as the index model, technical metrics without clinical outcomes. Do not include convolutional neural networks when they are the index model.
Context	Any geography or setting, peer-reviewed and preprint full papers, English language, 2022–2026	Do not include conference papers. Do not include non-English publications. Do not include papers that were published before 2022.

Table 2
Model taxonomy, where all models are foundation models, each falling under one classification.

Class	Definition
1. Foundation Model (pre-trained)	A model trained from randomly initialised weights, ground-up, on a large-scale general corpus, optimising for an objective. All weights, across all layers, are learned from scratch. Continued pretraining of an existing checkpoint on a large-scale corpus is also included (7).
2. Foundation Model (fine-tuned)	A model that starts from pre-trained weights (already initialised and not random) and performs further gradient updates on smaller, task specific, or domain specific, datasets. Does not rebuild the model. This can be done with full finetuning: every parameter in the model is updated, or parameter-efficient finetuning (PEFT) such as LoRA, adapters, prefix-tuning (base weights are frozen and only a small set of additional or low rank parameters are trained (8).
3. Commercial LLMs	Large language models used with frozen weights (no gradient updates) through prompting, retrieval or in-context learning. Most are commercial (proprietary weights, architecture not visible), open-weight models used without weight updates are also included. Control is limited to input (prompt, context, attached documents) and configuration (max tokens, temperature, system instructions, tools) (9).

There are multiple research papers that utilise multiple models, e.g. one paper can use two models: one that is used to translate speech audio to speech transcripts, and the other uses the speech transcripts to fine-tune the FM. For such cases, the model that determines the results discussed in the research papers is classed as the index model. The index model is the model that falls under the 3 taxonomy classes.

2.3 Search Strategy

Figure 1 depicts the search pipeline in detail. A systematic search of research studies was conducted following a combined manual approach and LLM research tools. The manual search was conducted

through Google Scholar and Pubmed using the following Boolean search strategy: (“foundation models” OR “fm” OR “large language model” OR “llm” OR “pre-trained transformer” OR “self-supervised learning” OR “ssl”) AND (“alzheimer*” OR “dementia”), with the window 2022–2026. This search returned 1,369 records.

Manual search results were screened using the same exact-match and deduplication procedure applied to the LLM deep research output (below). Of 1,369 combined records, 336 satisfied the Boolean string exactly in title, abstract, or keywords. Deduplication reduced this to 309, and 95 were excluded at screening (conference/workshop papers, books/theses, patents, and records where terms matched only as homonyms (FM as fibromyalgia or authors initials in reference list)), leaving 214 included records.

For the LLM-assisted portion of the search, we used a human-in-the-loop deep research approach. Prompts for the LLM deep research were generated using Claude Sonnet 5 (full input and output prompts in Supplementary Data Prompt 1, August 2026). The same search string used in the manual portion of the search was incorporated into the LLM deep research prompt.

We instructed Claude Sonnet 5 to search five databases: Scopus, IEEE Xplore, Embase, Google Scholar, and Pubmed. As several of these were not directly searchable by the LLM, Claude instead searched PubMed, Europe PMC, and arXiv. All outputs were manually checked to confirm that titles and DOIs existed and were not hallucinated (i.e., not fabricated by the model). This stage resulted in 2,216 records (PubMed 348, Europe PMC 281, arXiv 1,587). Each was screened to satisfy the Boolean string exactly in the title, abstract, or indexed keywords. This returned 715 records (PubMed 288, Europe PMC 248, arXiv 179), which were deduplicated (DOI as primary key, then title + first author + year) to 480. These were screened for publication type, homonym matches, and AI scope, excluding 42, leaving 438 included.

The 438 records from the LLM deep research stage and 214 from the manual stage (N = 652) were merged using the same DOI-primary, then title + first author + year key (a title-only fall back resolved three cases where a preprint (deep research stage) and its later published version (manual stage) carried different DOIs and years, with both DOIs retained on the merged record). This removed 178 duplicates, leaving 474 unique records. The manual PubMed search contributed no records beyond those found by Claude deep research.

A final stage of full text screening was conducted to confirm the papers met the PCC criteria. Full text was obtained via a combination of automated open-access retrieval (Unpaywall, CORE, and PubMed Central) and manual retrieval through institutional library access for papers not available through these sources. Of the 474 papers, 133 were excluded prior to full text screening as they were inaccessible through institutional access or judged, on the basis of title and abstract, to fall outside of the scope of the review and two duplicate papers were dropped at this stage as well, returning 339 papers to screen.

The remaining 339 papers were screened at full text using a purpose-built agent (Claude Sonnet 5, accessed via Anthropic Batches API) with a metaprompt (Supplementary Data Prompt 2 and 3, Code available at: <https://github.com/MatildaBeinat/Scoping-Review-Screening-System/tree/main>)

incorporating the PCC criteria (Table 1). For each paper, the script evaluated the PCC criteria independently, returning a structured decision for each meets / fails / insufficient information decision, supported by a verbatim quotation located in the supplied text. When a quote could not be found, the criterion was coded as insufficient information rather than inferred. An overall decision of Include, Exclude, or Maybe was derived from the three sub-decisions (PCC), with a reviewer not summarising the decision made based on each subdivision: 113 Include, 223 exclude, and 3 Maybe.

To assess the reliability of the automated full-text screening stage, a random stratified sample of 30 papers (seed = 42) was drawn from the 339 full-text-screened records, stratified by the agent's overall decision (Include, Exclude, Maybe) to ensure all three categories were represented: all 3 Maybe decisions were retained, and the remaining 27 papers were allocated proportionally between Include (n = 9) and Exclude (n = 18), reflecting their relative representation in the full set. These 30 papers were screened independently by the lead author (M.B.), blind to the agent's decisions, against the same PCC criteria (Table 1), and Cohen's kappa was computed between the human and agent decisions.

Agreement was 76.7% (23/30), $\kappa = 0.57$, indicating moderate agreement (Supplementary Data Table 1: confusion matrix). Disagreement was concentrated around the Maybe category: 3 of 7 disagreements involved a Maybe decision on one side, whereas agreement on the binary Include/Exclude distinction was considerably higher (84%, 21/25 non-Maybe agent decisions).

Across the full set, the final full-text screening stage returned a valid decision for all 339 papers: 113 Include, 223 Exclude, and 3 Maybe. Given the volume of literature and resource constraints, as a pragmatic decision, the present analysis was restricted to the subset of Include papers receiving a high-confidence decision (N = 50), the remaining Include papers were rated medium confidence. Confidence was self-assessed by the screening agent as part of its structured output rather than governed by an explicit rubric and should therefore be interpreted as suggestive rather than a calibrated measure. An exploratory review of reviewer notes (created by the agent, gives formal reasoning for each decision made) associated with the medium-confidence Include decisions showed recurring PCC classification ambiguity, warranting further analysis in future work. A summary of the research corpus of included papers is listed in Table 3, with the complete table listed in Supplementary Data Table 2.

Table 3

Summary of research corpus (n = 50) data grouped by taxonomy class: Number of studies, typical approach, data type, main datasets, sample size, metric range, validation, USA-involved author teams.

Studies	FM pre-trained	FM fine-tuned	Commercial LLM	All
	8 (16%) (13–20)	26 (52%) (21–46)	16 (32%) (47–62)	50
Typical approach	Ground up transformers, Vision Transformers, Mamba, EHR foundation models	BERT/RoBERTa, speech encoders, Qwen/LLaMA with LoRA, QLoRA, Vision Transformer with LoRA	Prompted or RAG, GPT4/4o, Claude, Bard, frozen open-weight LLMs	
Main datasets	Multi-cohort PSG, ADNI, OASIS-3, UK Biobank, NYU Langone	ADReSS/ADReSSo/Pitt, ADNI, NACC, LEADS, PREPARE	ADReSS/ADReSSo/Pitt, in house clinical cohorts, MGB, OHSU	
Sample size, median (IQR)	155-1,288,333, median 18,169.5, IQR 3,655 – 47,182	156 – 54,535, median 1,057, IQR 403-2,043	1-384,925, median 196.5, IQR 67 – 1,049	1–1,288,333, median 843, IQR 244-3,655
Metric range	Most frequent band 0.65–0.79 (50%)	Most frequent band 0.80–0.90 (38.5%)	Most > 0.90	
External validation	6/8	10/26	2/16	18
Statistical testing	4/8	8/26	5/16	17
No validation	0	0	4	4
USA involved authors	7/8	18/26	10/16	35

2.4 Data Extraction

Relevant information from each paper was extracted: publication year, paper geolocation, taxonomy class, model used, datasets used, demographics (including female/male splits, and number of participants), scope, results, validation method, limitations stated by research papers.

This data was extracted manually and analysed using Microsoft Excel (Version 16.113.1).

Descriptive quantitative and qualitative analysis was conducted: frequency counts, median sample sizes and parameter ranges, proportion of papers falling into metric bands, themes, narrative comparisons,

and validation cross tabulating. There was no statistical comparison between taxonomy classes conducted as the research papers are heterogeneous, metrics reported diverge and scopes for methods are too small to compare and reach a robust conclusion.

Agreement between the screening agent and the human rater was quantified as percentage agreement and unweighted Cohen's κ ($n = 30$). High and medium-confidence Include papers were compared on scope (detection versus diagnosis; $n = 84$ scope assignments, dual-scope papers counted in each) with Pearson's χ^2 test of independence (two-sided, $\alpha = 0.05$, no continuity correction, all expected counts > 5), exact P values are reported. Class distributions were compared descriptively. No statistical comparisons were made between taxonomy classes.

3 Results

3.1 Study selection and characteristics

The search returned 50 records: two were published in 2022, two in 2023, seven in 2024, 23 in 2025, and 16 in 2026. Thirty-one papers were peer reviewed and 19 were preprints. By taxonomy class, eight were FMs pre-trained (16%), 26 were FMs fine-tuned (52%), and 16 were Commercial LLMs (32%). Total study sample size, including pretraining data where authors performed pretraining, was highly right skewed (median 843, interquartile range 244-3,655, range 1 to 1,288,333). The median sample size for FMs pre-trained was 18,169.5, for FMs fine-tuned was 1,057, and for Commercial LLMs was 196.5. The model parameter sizes per taxonomy: Commercial LLMs have the largest parameters (estimated at about 1.7T for GPT4 and 200B for GPT4o (63)), FMs fine-tuned range from 13M to 30B, and FMs pre-trained 44K to 1.2B.

To describe how the analysed set relates to the remaining eligible studies, the 63 medium-confidence Include papers were compared descriptively with the 50 high-confidence papers on taxonomy class and scope (Supplementary Data Tables 3–5). This was a descriptive check, not a test of selection bias, and the medium-confidence papers were not analysed further.

The rank order of classes was identical in both sets (fine-tuned FMs $>$ commercial LLMs $>$ pre-trained FMs). Fine-tuned FMs made up 52% versus 41% of the high and medium-confidence sets, commercial LLMs 32% versus 32%, and pre-trained FMs 16% versus 27%.

A chi-square test of independence examined the association between scope (detection vs. diagnosis) and confidence classification (high vs. medium) across the 84 scope assignments falling under these two scopes (Supplementary Data Table 6). High-confidence studies were more frequently detection-focused (28/46, 60.9%) than diagnosis-focused (18/46, 39.1%), while medium-confidence studies showed the reverse (detection: 16/38, 42.1%; diagnosis: 22/38, 57.9%). This association did not reach statistical significance ($\chi^2(1, N = 84) = 2.94, p = 0.087$), although the comparison is descriptive and the samples are small. The divergence in scope distribution, together with scopes in the medium set that fall outside the four analysed (screening, staging, monitoring, biomarker discovery), indicates the medium-

confidence set is not a direct extension of the high-confidence set's composition and should be interpreted with this difference in mind.

3.2 Scopes

Four scopes were identified: detection (28), diagnosis (18), risk prediction (7), prognosis (3). Six studies addressed dual scopes: diagnosis & risk prediction (1), detection & risk prediction (1), detection & diagnosis (3), and diagnosis & prognosis (1). Figure 2 shows the distinct counts of papers for each category.

3.3 Geography, datasets, and data modalities

Author affiliations spanned 16 unique countries. The majority of papers had a single country author team ($n = 41$). The most common country was the United States of America (USA) ($n = 35$): followed by China ($n = 5$) (Fig. 3).

There were 72 unique datasets used, with 30 papers using a single dataset. The most common data set used was ADNI ($n = 11$), followed by the ADReSS challenge dataset ($n = 8$), DementiaBank Pitt Corpus ($n = 8$), and the ADReSSo challenge dataset ($n = 8$, updated version of the ADReSS challenge) (Fig. 4).

Speech transcripts were the predominant modality ($n = 22$), followed by speech audio ($n = 11$), clinical notes ($n = 8$), structural MRI ($n = 7$), structured EHR ($n = 5$), functional MRI (3). Figure 5 shows the full paper count vs modality per component. Out of the 50 papers, 34 were unimodal and 16 were multimodal. The modality most used in unimodal papers were speech transcripts ($n = 12$) and clinical notes ($n = 6$). For multimodal papers it was speech audio & speech transcripts ($n = 8$). Table 4 shows complete analyses.

Table 4
Unimodal and multimodal classifications.

Unimodal	Paper Count	Multimodal	Paper Count
speech transcripts / language	12	speech audio + speech transcripts / language	8
clinical notes / free text	6	speech audio + image + speech transcripts / language	1
structural MRI	3	structural MRI + PET imaging	1
structured EHR / clinical tabular data	3	clinical notes / free text + structured neuropsychological/language test scores + structural MRI	1
structured clinical tabular data (non-EHR)	2	structured EHR / clinical tabular data + clinical notes / free text	1
functional MRI (fMRI)	2	functional MRI (fMRI) + demographic / phenotypic data	1
speech audio	1	structural MRI + structured EHR / clinical tabular data	1
polysomnography signals	1	structural MRI + genomic / genetic data	1
digital pathology (whole-slide imaging)	1	speech audio + speech transcripts / language + demographic / phenotypic data	1
retinal imaging	1	Total	16
wearable / sensor data	1		
multi-omic / genomic data	1		
Total	34		

3.4 Demographic reporting and analysis

Only 4/50 papers reported race/ethnicity with a breakdown, of which the majority participants were white. Twenty-six papers did not report the age of participants. Out of the 50 papers, only 20 reported female and male demographics separately, with a higher proportion of female participants (49.4% female – 58.96% female). Only 2 of the 50 papers reported sex differences in their models (TRADE: model performance broken out by gender and SpeechCARE: includes a fairness audit the reports equality of opportunity scores for males and females). However, no papers built models specifically for each sex.

3.5 Performance of corpus

Performance is summarised descriptively: tasks, modalities, datasets and metrics differ between studies, so values are not comparable across classes and no ranking is implied.

Commercial LLMs (n = 16):

Widest range of reported metrics of any class. Some results match or beat fine-tuned models, these were all bigger models, with larger parameter sizes (RAG GPT-4o F1 0.933 (47), GPT-4 + attention-based deep neural network + XGBoost ensemble F1 92.1% (53), linguistic-feature classifier Acc 97.88%, specificity 0.99 and precision 0.98 (61)). Others sit near chance (Bard/GPT-4/GPT-3.5 chatbot prompting, overall accuracy 38–54% and F1 29–71% despite specific/precision 96%/89% - with BARD performing better than GPT models (59), LLaMA2 pipeline F1 0.34 (51), however this was due to lack of demographics in input data). A tabular in-context FM was comparable to tuned XGBoost (AUC 0.879 vs 0.887, $p > 0.05$ (55) (55)). Limitations for the highest-performing models in this class followed a pattern: small, single site, or narrow sourced samples that generalise poorly. Reference standards (gold standard to compare to) are incomplete proxies for real diagnosis, and external/multi-site validation is mostly missing. Studies with the weakest results reported limitations in prompting. This class had the most research papers with metrics above 0.90 (43.75%).

Foundation Model fine-tuned (n = 26):

Most results sit in the 0.80–0.99 range on accuracy/F1/AUC (CD-Tron F1 0.929 (26), RoBERTa AUC 0.988 (45), DeepSeek-R1 SLM Acc 91.67% and F1 91.71% (42), mixture of dataset R-GenIMA Acc 99.01% and F1 98.50% (39), T5-large Acc 0.95 (46). Weak outliers exist (GRPO-tuned LLaMA-3.1 BACC 67.33% (22), LUNAR MCI F1 0.53 (29)) but these are harder multi-class or MCI-specific tasks, not a class-wide weakness. This class focused on LLM/FM pipelines built on pre-trained weights that were further fine-tuned, including parameter-efficient methods. This taxonomy class leaned towards explainability, using SHAP, attention, or synthetic rationale generation/reasoning chains, as a first-class design goal, majority with the scope of detection and/or diagnosis (n = 21). This class had the most research papers with metrics between 0.80–0.90 (38.46%), with 30.77% reaching 0.90+.

Foundation Model pre-trained (n = 8):

Reported values were more tightly grouped than in the commercial class and rarely reached the 0.90 + band. Representative AUC/accuracy values were 0.73–0.86 (TRADE AUROC 0.735 (18), mosGraphGPT Acc 75.09% (17), ViT-B AUC 0.8619 (20)). Two results are exceptions at the high end (ADFound AUC 0.967 (14), SleepFM C-index up to 0.99 for risk prediction of senile dementia with AUROC 0.87/C-index 0.85 for dementia risk prediction (13)). Notably, in at least one paper, a foundation-model backbone did not beat a simpler CNN baseline (DenseNet121 CNN AUC 0.8865 vs best ViT-B PEFT variant 0.8619 (20)). FMs show promise but did not always outperform simpler baselines. This class had the most research papers with metrics between 0.65–0.79 (50%), with 12.5% reaching 0.90+.

3.6 Validation

Validation varied. Thirteen of the 50 papers used internal train/test/validate split only. External validation was reported by 18 papers: seven alongside cross validation and 11 without it. Statistical testing (analysing whether there is significant difference between test and validation performance) was reported by 17 papers. Of the 18 papers with external validation, nine reported no statistical testing. Of the 17 papers with statistical testing, eight had no external validation. Only five papers combined cross & external validation and statistical testing. Four papers used none of these approaches because their design did not allow it: zero-shot prompting with no training step, a same session comparison against a clinical reference standard, a single case study, and a diagnostic accuracy study without a data split. Two papers used postmortem neuropathology as the reference standard.

3.7 Limitations of corpus

Limitations reported by the research corpus followed 7 recurring themes.

1. Sample bias and single source data, with most papers using single datasets, making results difficult to transfer to other populations.
2. Demographic and linguistic narrowness, with cohorts skewed towards English speakers, white race/ethnicity or highly educated populations.
3. Black box interpretability, with papers acknowledging that LLMs, cannot be explained mechanistically.
4. LLM specific instability, suffering from hallucinations, sensitivity to prompt phrasing and rapid model obsolescence.
5. Many papers relying on single modality data.
6. Research to clinic gap, with authors stating that the tools they developed were not clinically deployable.
7. Absent or vague limitations (missing entirely) and are inferred 'from future' work statements.

4 Discussion

4.1 Principal findings

This review analysed 50 research papers applying FMs to detect, diagnose, predict the progression, and risk of dementia. Most papers focused on detection (n = 28) or diagnosis (n = 18). Performance depended on type of model used. Fine-tuned models (n = 26) performed best and most reliably, and many were designed to explain their reasoning. Commercial LLMs (n = 16) showed the widest range of results: some outperformed fine-tuned models, while others barely performed better than random guessing. This inconsistency stemmed from small sample sizes, weak prompt design, and incomplete patient data. Pre-trained models (n = 8) performed moderately and more consistently than commercial LLMs but rarely reached top performance.

The central finding reveals a trade-off between model size and reliability. Commercial LLMs, despite having most parameters (internal settings that determine a model's capabilities) and never being

adjusted for this specific task, produced the highest individual results. Fine-tuned models, with fewer parameters, delivered the most dependably strong performance (scores range 0.80–0.90). Pre-trained models, with the fewest parameters, showed the lowest typical performance of this cohort (scores range 0.65–0.79) (Fig. 6).

4.2 Equity

Reporting of demographic and geographic characteristics was inconsistent across the studies reviewed (section 3.3–3.4). Race/ethnicity were rarely reported recorded, and where the population was described, it was mostly white. Age was often left unreported, and while sex was reported more frequently, results were never broken down by sex in the analysis. Authorship was also heavily concentrated in the USA. These gaps limit what can be concluded about how these models perform across people who differ in ethnicity, language, age, or sex. The body of research is not yet diverse enough to support claims that the findings generalise to these groups, and there is a real risk that results do not hold up for populations that were underrepresented in the studies.

4.3 Corpus results

Distributions across model type and research scope are reported in sections 3.1–3.3. The concentration of studies in detection and diagnosis, whether considered together or separately, suggests these are either the main research priorities for foundation models in this field, or simply the task best suited to what these models can currently do. Speech-based approaches were the most common (3.3), which fits with FMs' particular strength in processing large volumes of text and audio, compared with structured data or brain scans.

Performance also appeared to track a plausible trade-off with computing cost, though no study in this review directly reported how much computing power or money was spent on training. Pre-trained models, built from scratch, are by definition the most resource intensive approach, yet had the lowest typical performance: half of these studies scored between 0.65 and 0.79, one failed to outperform a simpler, older type of model (20), and another cited high computing costs as a barrier to further progress (15). Fine-tuned models, which adjust an existing model rather than training one from scratch, were often described using cheaper techniques (PEFT/LoRA), and produced the most consistent results (0.80–0.99, where 38.46% falling between 0.80–0.90). Commercial LLMs required no training computing power, since their internal settings are fixed, yet some of these studies produced the highest scores in the entire review (43.75% above 0.90), consistently by relying on larger underlying models (47, 53, 61).

On the measures reported here, the two cheaper approaches outperformed the most expensive one. This is a small sample of studies, none of which test cost against performance directly, so this is a pattern, not a controlled comparison. Even so, the current evidence does not show that training a FM from scratch delivers a performance advantage that justified its resource cost, comparing with fine-tuning an existing model or simply using a commercial AI system.

4.4 Validation

Validation methods varied considerably across the studies. The most common approach was internal validation, splitting a single dataset into training, testing, and validation subsets (13/50). External validation, testing a model on entirely separate data, without additional cross validation was the single most frequent method (11/50), while 7 papers combined external validation with cross validation. Statistical testing to determine whether differences between test and validation results were meaningful was carried out in only 17/50 papers. Among the papers that used external validation, 9 did not perform statistical testing, among the papers that did perform statistical testing, 8 relied only on internal validation.

Just 5 papers combined all three approaches. Four papers used none of them, in each case because the study's design ruled them out. Only 2 papers validated their findings against postmortem brain tissue analysis, the gold standard for confirming a dementia diagnosis.

This pattern indicates that rigorous validation, particularly the combination of external data with statistical testing, remains limited, and this held true regardless of model type. These findings should be treated cautiously given the size of the review, but they point to a clear need for validation against gold-standard methods wherever possible. Notably, only 2 papers used postmortem validation. Postmortem data will not be feasible for every dataset, since it depends on brain donation (a practice whose frequency itself is understudied (64)). Yet to establish real-world clinical usability, establishing a gold-standard is encouraged. As things stand, this research cannot progress toward clinical application and will remain at the proof of concept stage until validation practices improve.

4.5 From research to clinical utility

This body of literature sits at the pilot stage. Recurring limitations reported by the studies include data drawn from a single research site, instability specific to LLMs (such as sensitivity to how questions are phrased, model quickly becoming outdated, and generating plausible but false information), reliance on a single type of data, and explicit statements from authors that their tool was not ready for clinical use. Less commonly, some papers lacked a formal limitations section altogether, instead folding their limitations into a more optimistically framed “future work” section.

The performance figures in this review therefore show what these models can do on specific, sometimes narrow, groups of patients and against the particular reference standards each study chose, not what they would achieve in a broader, representative clinical population. This matter because, with one exception (48), none of the studies validated their results against postmortem brain tissue analysis, even though this remains the gold standard for a definitive dementia diagnosis (65). This distinction is especially important for commercial LLMs: studies reporting the strongest results attributed them to careful prompt design and complete input data, while studies with weaker results cited shortcomings in exactly these areas. This suggests these models are highly sensitive to how they are implemented, which may make results difficult to reproduce elsewhere. The limitations found across this body of research are what reveal the shaky foundations underlying current FMs in dementia research.

Understanding why these tools are not ready yet for clinical use is essential for guiding future research toward genuinely usable tools in everyday dementia care.

4.6 Strengths and limitations of this review

This review's main methodological contribution is a three-category classification system, developed to address inconsistent use of the term "foundation model" in the existing literature. The review demonstrates that distinguishing between these three categories, rather than grouping them under one broad label, is itself a meaningful finding, and this distinction shapes how the results are structured throughout.

The search and screening process combined a manual keyword-based search with AI-assisted research to identify relevant papers. Only one reviewer assessed each paper (rather than a more rigorous approach of two independent reviewers), but agreement between the AI system and the human reviewer on which papers to include or exclude was high (84%), lending some confidence to the AI's screening decisions, though the human reviewer was stricter than the AI on three occasions. The statistical measure of agreement used here (Cohen's kappa) reflects agreement between the AI and the person, this is not the same as an independent check. Future work should include a second, independent reviewer to confirm these results.

The final analysis was included 50 papers that the AI screening system rated as high confidence for inclusion after reading the full text. A further 63 papers rated medium confidence were excluded from analysis, and this confidence rating was the AI's own self-assessment, made without explicit criteria, so it should be treated as indicative rather than a rigorously calibrated measure. The model type and research scope of these 63 excluded papers were nonetheless counted (Supplementary Data Tables 3–4) and compared descriptively against the 50 included papers. The ranking of model types was the same in both groups, but pre-trained models appeared proportionally more often in the excluded group (27% versus 16%) and detection studies appeared less often (42% versus 61% of detection and diagnosis papers combined), this difference in scope was not statistically significant ($p = 0.087$). This comparison did not extend to performance, validation methods, or demographic reporting, and cannot rule out the possibility that the excluded papers differ from the included ones in other important ways.

4.7 Recommendations

Based on the findings of this review, 5 main recommendations are made.

1. Prioritise larger, multimodal, and demographically diverse datasets (where possible). Single source, single country, and single modality studies dominate the literature corpus, and this limits generalisability more than any modelling choice.
2. Report race/ethnicity, age, and sex as a standard. With only 4/50 reporting ethnicity, and 26/50 not reporting age, the corpus does not support conclusions about how these models perform across demographic groups.

3. Conduct sex stratified performance analyses, rather than just reporting sex differences. This uncovers sex specific biological patterns, and ensures risks are accurately understood for both sexes. This is more important for dementia research as sex differences are yet to be understood fully, and evidence shows there are real sex differences (66–73).
4. Expand external validation and statistical comparison together. Only 5/50 combined cross validation, external validation, and statistical comparison. Validation types should be selected to match claims made in a study, rather than defaulting to internal splits.
5. Treat FMs as a research trajectory under construction rather than an established performance advantage. The evidence in this review shows continued investment in FM approaches due to the heterogeneity of data, but it does not yet support claims that they outperform task-specific or classical models at scale.

5 Conclusion

This review found evidence of technical promise across all three types of FMs to dementia research. Most studies focused on models fine-tuned specifically on dementia-related data, while commercial LLMs produced widely varying results and general pre-trained models produced moderate, more consistent results.

At present, this research resets on a narrow evidence base: it is dominated by USA based studies, typically drawn from a single research site, with demographic characteristics often poorly reported, inconsistent or incomplete validation practices, and a recurring but varied pattern of limitations. As a result, this body of research cannot yet demonstrate how these models perform across populations differing in ethnicity, language, age, or sex.

Moving these tools toward real clinical use will require larger, more diverse datasets that combine multiple types of data, external validation on independent datasets, and testing in real-world clinical settings.

Declarations

AI disclosures

As part of our methodology, as disclosed in the methods section, Claude Sonnet 5 was used as part of our LLM-assisted screening protocol.

Competing Interests:

The authors declare no conflicts of interest.

Supplemental Data

Supplementary data can be found separately to this scoping review.

Author Contribution

MB, PP, VO conceived the study. MB conducted the literature search, built the study design, performed manual title/abstract screening and full-text screening, developed the analysis agent, conducted formal analysis, curated and interpreted the data, and created all figures and tables. MB wrote the original manuscript draft and all subsequent drafts. PP and VO provided supervision, validated the methodology and results, and contributed to data interpretation, manuscript writing and editing. CM and EW contributed to manuscript review and editing. PP and VO directly accessed and verified all underlying data reported in the manuscript. All authors accept responsibility for the integrity of the work and approve the final manuscript for submission.

Acknowledgements:

This paper receiving no funding.

Data Availability

The data underlying this scoping review are the published studies included in the analysis, cited in Table 3 and Supplementary Data Table 2, together with the list of 63 medium-confidence papers excluded from full analysis, provided in the Supplementary Data. Code for the full-text screening pipeline is publicly available at <https://github.com/MatildaBeinat/Scoping-Review-Screening-System/tree/main>. The full-text analysis dataset (CSV) and records of the 339 screening decisions made by the AI-assisted agent are available from the corresponding author (M.B.) upon reasonable request.

References

1. Alzheimer's Society. People living with undiagnosed dementia are three times more likely to visit A&E | Alzheimer's Society [Internet]. 2024 [cited 2025 Nov 13]. Available from: <https://www.alzheimers.org.uk/news/2024-09-16/living-with-undiagnosed-dementia-three-times-more-likely-visit-ae>
2. Chen S, Cao Z, Nandi A, Counts N, Jiao L, Prettnner K, et al. The global macroeconomic burden of Alzheimer's disease and other dementias: estimates and projections for 152 countries or territories. *Lancet Glob Health*. 2024;12(9):e1534-43. doi:10.1016/S2214-109X(24)00264-X
3. Alzheimer's Society. Alzheimer's Society [Internet]. 2025 [cited 2025 Nov 11]. Dementia is the UK's leading cause of death | Alzheimer's Society. Available from: <https://www.alzheimers.org.uk/about-us/dementia-UK-leading-cause-of-death>

4. Alzheimer's Society. What is Alzheimer's disease? | Alzheimer's Society [Internet]. 2025 [cited 2025 Nov 11]. Available from: <https://www.alzheimers.org.uk/about-dementia/types-dementia/alzheimers-disease>
5. Volloch V, Rits-Volloch S. Effect of Lecanemab and Donanemab in Early Alzheimer's Disease: Mechanistic Interpretation in the Amyloid Cascade Hypothesis 2.0 Perspective. *J Alzheimers Dis.* 2023;93(4):1277–84. doi:10.3233/JAD-230164 PubMed PMID: 37212119; PubMed Central PMCID: PMC10357217.
6. Mahase E. NICE rejects Alzheimer's drug donanemab owing to cost and "significant health risks." *BMJ.* 2024;387:q2342. doi:10.1136/bmj.q2342 PubMed PMID: 39442962.
7. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, Arx S von, et al. On the Opportunities and Risks of Foundation Models [Internet]. arXiv; 2022 [cited 2026 Aug 24]. Available from: <http://arxiv.org/abs/2108.07258> doi:10.48550/arXiv.2108.07258
8. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature.* 2023;616(7956):259–65. doi:10.1038/s41586-023-05881-4
9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172–80. doi:10.1038/s41586-023-06291-2
10. PRISMA statement [Internet]. [cited 2026 Jan 26]. Scoping. Available from: <https://www.prisma-statement.org/scoping>
11. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol.* 2005;8(1):19–32. doi:10.1080/1364557032000119616
12. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. *JB I Evid Synth.* 2020;18(10):2119–26. doi:10.11124/JBIES-20-00167 PubMed PMID: 33038124.
13. Thapa R, Kjaer MR, He B, Covert I, Moore Iv H, Hanif U, et al. A multimodal sleep foundation model for disease prediction. *Nat Med.* 2026;32(2):752–62. doi:10.1038/s41591-025-04133-4
14. Yang G, Du K, Yang Z, Du Y, Cheung EYW, Zheng Y, et al. ADFound: A Foundation Model for Diagnosis and Prognosis of Alzheimer's Disease. *IEEE J Biomed Health Inform.* 2025;29(11):8395–408. doi:10.1109/JBHI.2025.3576436
15. Verma R, Kandoi S, Afzal R, Chen S, Jegminat J, Karlovich MW, et al. Domain-Specific Foundation Model Improves AI-Based Analysis of Neuropathology [Internet]. arXiv; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2512.05993> doi:10.48550/ARXIV.2512.05993
16. Wei Z, Dan T, Wu G. Large Connectome Model: An fMRI Foundation Model of Brain Connectomes Empowered by Brain-Environment Interaction in Multitask Learning Landscape [Internet]. arXiv; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2510.18910> doi:10.48550/ARXIV.2510.18910
17. Zhang H, Huang D, Chen E, Cao D, Xu T, Dizdar B, et al. mosGraphGPT: a foundation model for multi-omic signaling graphs using generative AI [Internet]. *Bioinformatics*; 2024 [cited 2026 Aug 19].

Available from: <http://biorxiv.org/lookup/doi/10.1101/2024.08.01.606222>
doi:10.1101/2024.08.01.606222

18. Zhu W, Tang H, Zhang H, Rajamohan HR, Huang SL, Ma X, et al. Predicting Risk of Alzheimer's Diseases and Related Dementias with AI Foundation Model on Electronic Health Records [Internet]. *Neurology*; 2024 [cited 2026 Aug 19]. Available from: <http://medrxiv.org/lookup/doi/10.1101/2024.04.26.24306180> doi:10.1101/2024.04.26.24306180
19. Bijlani N, Gustavo, Nilforooshan R, Group C, Barnaghi P, Kouchaki S. Remote Monitoring in Dementia Care - Lightweight, Explainable AI Validated for Early Warning of Health Events in the Home [Internet]. In Review; 2025 [cited 2026 Aug 19]. Available from: <https://www.researchsquare.com/article/rs-6294980/v1> doi:10.21203/rs.3.rs-6294980/v1
20. Dhinagar NJ, Ozarkar SS, Buwa KU, Thomopoulos SI, Owens-Walton C, Laltoo E, et al. Parameter Efficient Fine-tuning of Transformer-based Masked Autoencoder Enhances Resource Constrained Neuroimage Analysis [Internet]. *Neuroscience*; 2025 [cited 2026 Aug 19]. Available from: <http://biorxiv.org/lookup/doi/10.1101/2025.02.15.638442> doi:10.1101/2025.02.15.638442
21. Park C, Kim C. A Novel Chain-of-Thought Reasoning Approach for Alzheimer's Disease Detection Using Large Language and Vision-Language Models. *IEEE Trans Neural Syst Rehabil Eng*. 2025;33:4386–95. doi:10.1109/TNSRE.2025.3624762
22. Zamai A, Fijalkow N, Mansencal B, Simon L, Navet E, Coupe P. An Explainable Diagnostic Framework for Neurodegenerative Dementias via Reinforcement-Optimized LLM Reasoning [Internet]. *arXiv*; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2505.19954> doi:10.48550/ARXIV.2505.19954
23. Agbavor F, Liang H. Artificial Intelligence-Enabled End-To-End Detection and Assessment of Alzheimer's Disease Using Voice. *Brain Sci*. 2022;13(1):28. doi:10.3390/brainsci13010028
24. Mekulu K, Aqlan F, Yang H. Automated Detection of Early-Stage Dementia Using Large Language Models: A Comparative Study on Narrative Speech [Internet]. *Health Informatics*; 2025 [cited 2026 Aug 19]. Available from: <http://medrxiv.org/lookup/doi/10.1101/2025.06.06.25329081> doi:10.1101/2025.06.06.25329081
25. Gupta RD, Morol MK, Fahad N, Hosain MT, Choya SBZ, Hossen MJ. BRAINS: A Retrieval-Augmented System for Alzheimer's Detection and Monitoring [Internet]. *arXiv*; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2511.02490> doi:10.48550/ARXIV.2511.02490
26. Guan H, Novoa-Laurentiev J, Zhou L. CD-Tron: Leveraging large clinical language model for early detection of cognitive decline from electronic health records. *J Biomed Inform*. 2025;166:104830. doi:10.1016/j.jbi.2025.104830
27. Mehdoui SS, Bouzid A, Sierra-Sosa D, Elmaghraby A. Dementia Insights: A Context-Based MultiModal Approach [Internet]. *arXiv*; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2503.01226> doi:10.48550/ARXIV.2503.01226
28. Wang L, Rezaii N, Dickerson BC, Glass J. Do Multimodal Large Language Models Need Reasoning to Classify Dementia from Speech? [Internet]. *arXiv*; 2026 [cited 2026 Aug 19]. Available from:

<https://arxiv.org/abs/2607.00260> doi:10.48550/ARXIV.2607.00260

29. Kowshik SS, Jasodanand VH, Bellitti M, Puducheri S, Xu L, Liu Y, et al. Domain-adapted language model using reinforcement learning for various dementias [Internet]. *Neurology*; 2026 [cited 2026 Aug 19]. Available from: <http://medrxiv.org/lookup/doi/10.64898/2026.03.17.26348154> doi:10.64898/2026.03.17.26348154
30. Li J, Mao L, Mao X, Wang H, Wang Z, Ni XS. Early Detection of Alzheimer's Disease and Related Dementias From Spontaneous Speech Using Foundation Speech and Language Models: Comparative Evaluation. *JMIR Form Res*. 2026;10:e79411. doi:10.2196/79411
31. Han C, Kim J, Lee H, Joo H, Chang J, Park SM. Explainable foundation model for dementia screening and risk stratification using retinal fundus images. *Npj Digit Med*. 2026 Jul 4. doi:10.1038/s41746-026-02968-w
32. Dong Z, Guo H, Xu W, Zhao H, Zhang Z. Foundation Model-based Evaluation of Neuropsychiatric Disorders: A Lifespan-Inclusive, Multi-Modal, and Multi-Lingual Study [Internet]. *arXiv*; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2512.20948> doi:10.48550/ARXIV.2512.20948
33. Cheng Y, Malekar M, He Y, Bommareddy A, Magdamo C, Singh A, et al. High-Throughput Phenotyping of the Symptoms of Alzheimer Disease and Related Dementias Using Large Language Models: Cross-Sectional Study. *JMIR AI*. 2025;4:e66926. doi:10.2196/66926
34. Makarov N, Bordukova M, Quengdaeng P, Garger D, Rodriguez-Esteban R, Schmich F, et al. Large language models forecast patient health trajectories enabling digital twins. *Npj Digit Med*. 2025;8(1):588. doi:10.1038/s41746-025-02004-3
35. Songdechakraiwut T. Leveraging LLMs for Early Alzheimer's Prediction [Internet]. *arXiv*; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2510.23946> doi:10.48550/ARXIV.2510.23946
36. Zolnour A, Azadmaleki H, Haghbin Y, Taherinezhad F, Nezhad MJM, Rashidi S, et al. LLMCARE: early detection of cognitive impairment via transformer models enhanced by LLM-generated synthetic data. *Front Artif Intell*. 2025;8:1669896. doi:10.3389/frai.2025.1669896
37. Chen Q, Hu X, Wang Z, Hong Y. MedBLIP: Bootstrapping Language-Image Pre-training from 3D Medical Images and Texts [Internet]. *arXiv*; 2023 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2305.10799> doi:10.48550/ARXIV.2305.10799
38. Ilias L, Askounis D. Multimodal Deep Learning Models for Detecting Dementia From Speech and Transcripts. *Front Aging Neurosci*. 2022;14:830943. doi:10.3389/fnagi.2022.830943
39. Zhao K, Dai S, Zhang Y, Liu G, Gu P, Lin C, et al. R-GenIMA: Integrating Neuroimaging and Genetics with Interpretable Multimodal AI for Alzheimer's Disease Progression [Internet]. *arXiv*; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2512.18986> doi:10.48550/ARXIV.2512.18986
40. Zhou X, Ai M, Adeli E, Zhang Y, Liu YM, Vankee-Lin F. Resting-state fMRI foundation models enable robust and generalizable latent neural target discovery in cognitive aging interventions [Internet]. *Bioengineering*; 2026 [cited 2026 Aug 19]. Available from: <http://biorxiv.org/lookup/doi/10.64898/2025.12.30.697042> doi:10.64898/2025.12.30.697042

41. Li VO, Han Y, Lam JC, Cheung LY. Reverse-Speech-Finder: A Neural Network Backtracking Architecture for Generating Alzheimer's Disease Speech Samples and Improving Diagnosis Performance [Internet]. arXiv; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2505.17477> doi:10.48550/ARXIV.2505.17477
42. Addepalli VR, Abdalnabi N, Kummerfeld E, Hembroff G, Kiselica A, Rao P, et al. Small-Sized Reasoning Language Models for Linguistic Screening of Alzheimer's Disease [Internet]. Health Informatics; 2026 [cited 2026 Aug 19]. Available from: <http://medrxiv.org/lookup/doi/10.64898/2025.12.24.25342972> doi:10.64898/2025.12.24.25342972
43. Azadmaleki H, Haghbin Y, Rashidi S, Momeni Nezhad MJ, Zolnour A, Zolnoori M. SpeechCARE: dynamic multimodal modeling for cognitive screening in diverse linguistic and speech task contexts. Npj Digit Med. 2025;8(1):677. doi:10.1038/s41746-025-02026-x
44. Kearney S, Yang S, Wen Z, Lyu W, Hou B, Duong-Tran D, et al. Tabular LLMs for Interpretable Few-Shot Alzheimer's Disease Prediction with Multimodal Biomedical Data [Internet]. arXiv; 2026 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2603.17191> doi:10.48550/ARXIV.2603.17191
45. Rezaii N, Wong B, Aisen P, Beckett L, Dage JL, Eloyan A, et al. Voiceprints of cognitive impairment: analyzing digital voice for early detection of Alzheimer's and related dementias. Npj Dement. 2025;1(1):35. doi:10.1038/s44400-025-00040-0
46. Jiang L, Zhou Y, Parde N. What Do LLMs Know About Alzheimer's Disease? Multi-loss Fine-Tuning and Probing for AD Detection [Internet]. arXiv; 2026 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2602.11177> doi:10.48550/ARXIV.2602.11177
47. Wang L, Liu B, Yang R, Chuang YW, Estiri H, Murphy S, et al. A retrieval-augmented generation large language model framework for accurate dementia identification from electronic health records [Internet]. Health Informatics; 2026 [cited 2026 Aug 19]. Available from: <http://medrxiv.org/lookup/doi/10.64898/2026.01.24.26344477> doi:10.64898/2026.01.24.26344477
48. Galligani C, Miller ZA, Mandelli ML, Rosen HJ, Ezzes Z, Lin M, et al. Agentic Generative Artificial Intelligence System for Classification of Pathology-Confirmed Primary Progressive Aphasia Variants [Internet]. Neurology; 2025 [cited 2026 Aug 19]. Available from: <http://medrxiv.org/lookup/doi/10.1101/2025.10.28.25338977> doi:10.1101/2025.10.28.25338977
49. Xu L, Kadambi P, Goldinger S, Berisha V, Mueller KD, Liss J. AI-based Cognitive-linguistic Features for Dementia Assessment in Picture Description [Internet]. arXiv; 2026 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2606.18054> doi:10.48550/ARXIV.2606.18054
50. Bang J, Han S, Kang B. Alzheimer's disease recognition from spontaneous speech using large language models. ETRI J. 2024;46(1):96–105. doi:10.4218/etrij.2023-0356
51. Wang J, Ahn S, Dalal T, Zhang X, Pan W, Zhang Q, et al. Augmented Risk Prediction for the Onset of Alzheimer's Disease from Electronic Health Records with Large Language Models [Internet]. arXiv; 2024 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2405.16413> doi:10.48550/ARXIV.2405.16413

52. Rabaya YJ, Qarariya SA, Abualhaija TM, Salah AA, Ashqar HI. Automated MoCA scoring for Arabic speakers using hybrid AI of multimodal speech, vision, and LLM integration. *Front Psychol.* 2026;17:1833118. doi:10.3389/fpsyg.2026.1833118
53. Du X, Novoa-Laurentiev J, Plasek JM, Chuang YW, Wang L, Marshall GA, et al. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *eBioMedicine.* 2024;109:105401. doi:10.1016/j.ebiom.2024.105401
54. Kleiman MJ. Evaluating large language model performance and reliability in scoring picture description tasks for neuropsychological assessment. Lu HHS, editor. *PLOS Digit Health.* 2026;5(4):e0001385. doi:10.1371/journal.pdig.0001385
55. Ye B, Soykan B, Koksalmis GH, Huang HH, Brattain LJ. Evaluating TabPFN for Mild Cognitive Impairment to Alzheimer's Disease Conversion in Data Limited Settings [Internet]. *arXiv*; 2026 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2604.27195> doi:10.48550/ARXIV.2604.27195
56. Heitz J, Schneider G, Langer N. Linguistic Features Extracted by GPT-4 Improve Alzheimer's Disease Detection based on Spontaneous Speech [Internet]. *arXiv*; 2024 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2412.15772> doi:10.48550/ARXIV.2412.15772
57. Jung OJ, Park J, Oh M. Listening Between the Lines: Joint Learning of ASR Embeddings and LLM-Augmented Linguistics for Dementia Detection [Internet]. *arXiv*; 2026 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2606.30675> doi:10.48550/ARXIV.2606.30675
58. Sweidan J, El-Yacoubi MA, Semmar N. MMSE-Calibrated Few-Shot Prompting for Alzheimer's Detection [Internet]. *arXiv*; 2025 [cited 2026 Aug 19]. Available from: <https://arxiv.org/abs/2509.19926> doi:10.48550/ARXIV.2509.19926
59. B.T B, Chen JM. Performance Assessment of ChatGPT versus Bard in Detecting Alzheimer's Dementia. *Diagnostics.* 2024;14(8):817. doi:10.3390/diagnostics14080817
60. McCoy TH, Perlis RH. Predicting hospice eligibility among dementia patients using language models. *Alzheimers Dement.* 2025;21(11):e70878. doi:10.1002/alz.70878
61. Rezaii N, Quimby M, Wong B, Hochberg D, Brickhouse M, Touroutoglou A, et al. Using Generative Artificial Intelligence to Classify Primary Progressive Aphasia from Connected Speech [Internet]. *Neurology*; 2023 [cited 2026 Aug 19]. Available from: <http://medrxiv.org/lookup/doi/10.1101/2023.12.22.23300470> doi:10.1101/2023.12.22.23300470
62. Navarra AN, Ikramuddin S, Greenberg SM, Doraiswamy PM. Utility of the Medical Knowledge AI Copilot, OpenEvidence: A Patient With 100 Cerebral Microhemorrhages. *Cureus.* 2026 May 1. doi:10.7759/cureus.108098
63. Abacha AB, Yim W wai, Fu Y, Sun Z, Yetisgen M, Xia F, et al. MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes [Internet]. *arXiv*; 2025 [cited 2026 Sep 21]. Available from: <http://arxiv.org/abs/2412.19260> doi:10.48550/arXiv.2412.19260
64. Sweeney A, Rush A, Stevens J, Sutherland GT. A Survey of Community Perceptions on Brain Donation for Research. *Biopreservation Biobanking.* 2025;23(1):31–8. doi:10.1089/bio.2023.0158 PubMed PMID: 38457650; PubMed Central PMCID: PMC12344117.

65. Suemoto CK, Leite REP. Autopsy studies are key to identifying dementia cause. *Lancet Healthy Longev.* 2023;4(3):e94-5. doi:10.1016/S2666-7568(23)00022-3 PubMed PMID: 36870340.
66. Stites SD, Cao H, James R, Harkins K, Coykendall C, Flatt JD. A systematic review of measures of gender and biological sex: Exploring candidates for Alzheimer's disease and related dementias (AD/ADRD) research. *Alzheimers Dement Diagn Assess Dis Monit.* 2023;15(1):e12359. doi:10.1002/dad2.12359
67. Castro-Aldrete L, Einsiedler M, Novakova Martinkova J, Depypere H, Alvin Ang TF, Mielke MM, et al. Alzheimer disease seen through the lens of sex and gender. *Nat Rev Neurol.* 2025;21(5):235–49. doi:10.1038/s41582-025-01071-0
68. Pinho-Gomes AC, Gong J, Harris K, Woodward M, Carcel C. Dementia clinical trials over the past decade: are women fairly represented? *BMJ Neurol Open.* 2022;4(2):e000261. doi:10.1136/bmjno-2021-000261 PubMed PMID: 36110923; PubMed Central PMCID: PMC9445798.
69. Wretling A, Xu J, Chen W, Velayudhan L, Ashton NJ, Zetterberg H, et al. Lipid profiling reveals unsaturated lipid reduction in women with Alzheimer's disease. *Alzheimers Dement.* 2025;21(8):e70512. doi:10.1002/alz.70512
70. Lopez-Lee C, Torres ERS, Carling G, Gan L. Mechanisms of sex differences in Alzheimer's disease. *Neuron.* 2024;112(8):1208–21. doi:10.1016/j.neuron.2024.01.024 PubMed PMID: 38402606.
71. Sex and Gender Differences in Alzheimer's Disease: Genetic, Hormonal, and Inflammation Impacts [Internet]. 2025 [cited 2025 Oct 28]. Available from: <https://www.mdpi.com/1422-0067/25/15/8485>
72. Arenaza-Urquijo EM, Boyle R, Casaletto K, Anstey KJ, Vila-Castelar C, Colverson A, et al. Sex and gender differences in cognitive resilience to aging and Alzheimer's disease. *Alzheimers Dement.* 2024;20(8):5695–719. doi:10.1002/alz.13844
73. Aggarwal NT, Mielke MM. Sex Differences in Alzheimer's Disease. *Neurol Clin.* 2023;41(2):343–58. doi:10.1016/j.ncl.2023.01.001 PubMed PMID: 37030962; PubMed Central PMCID: PMC10321561.

Figures

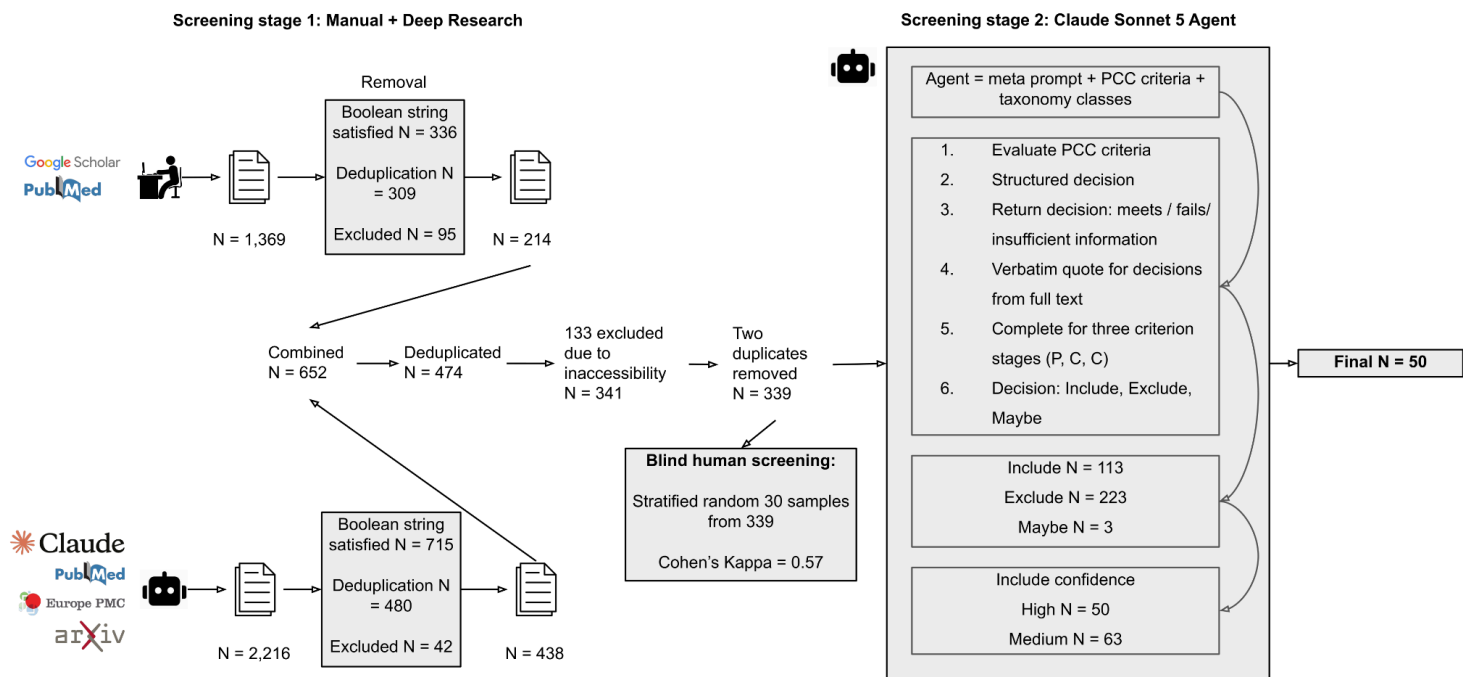


Figure 1

Search Pipeline in visual format.

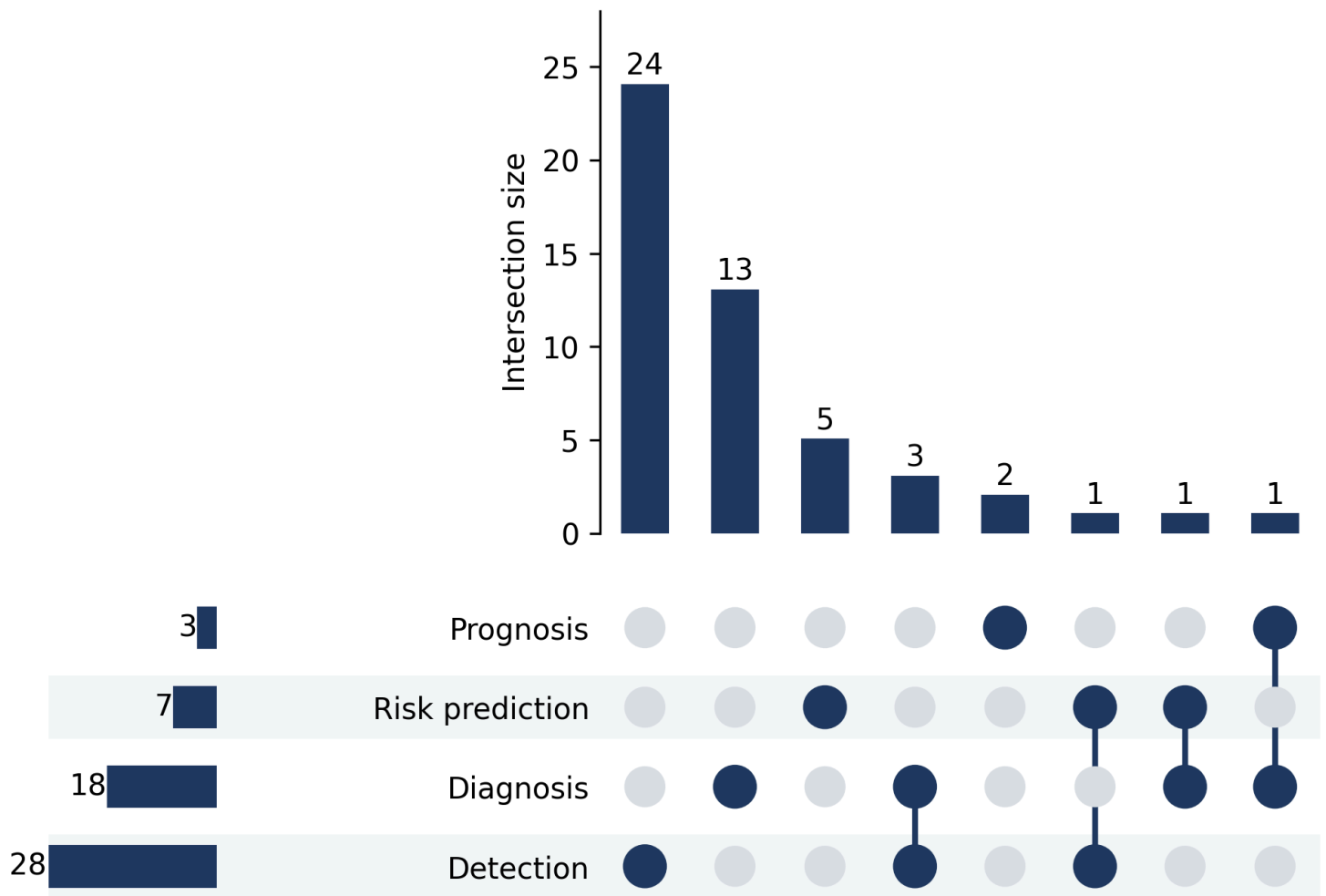


Figure 2

Upset Plot for spread of scopes in the research corpus analysed. Some papers had dual scopes, seen by the lines linking each scope.

Single-country and Multi-country

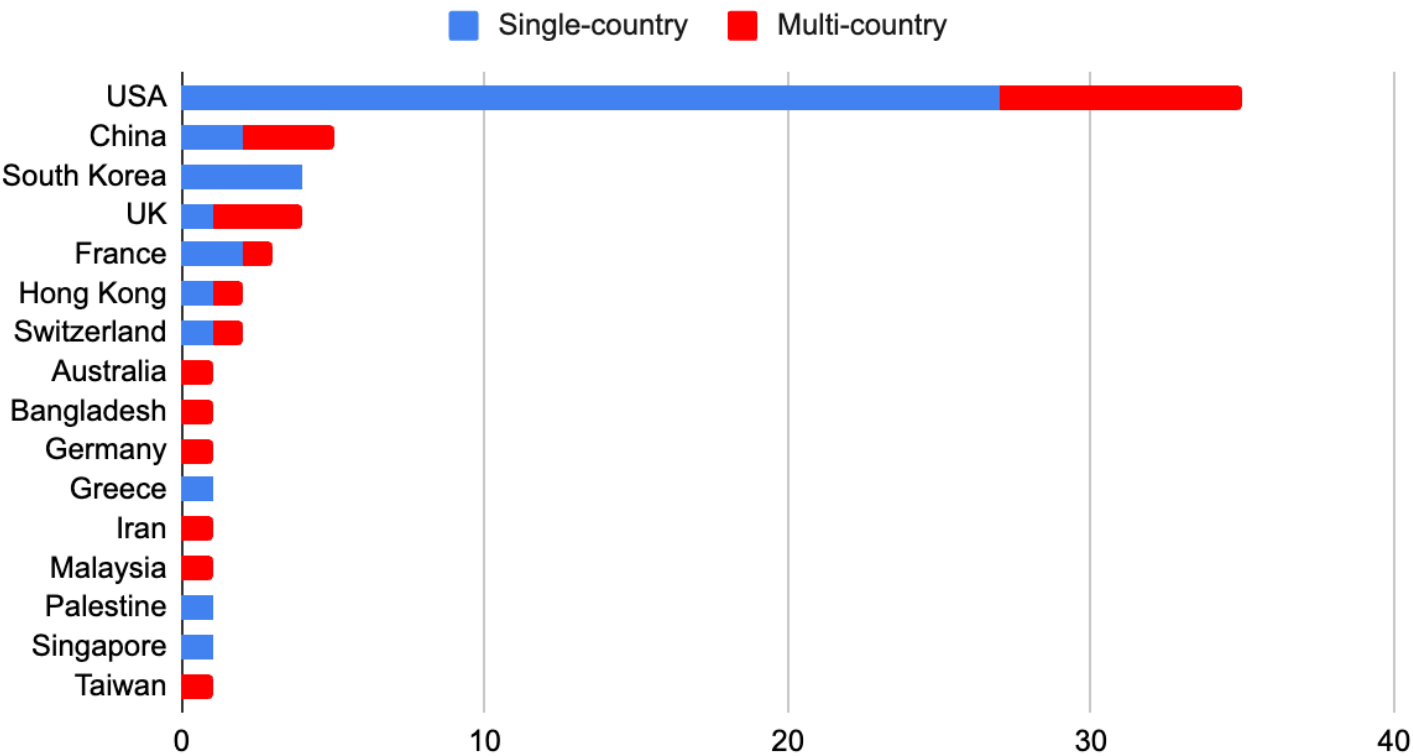


Figure 3

Country of author distribution per paper.

Dataset Usage by Individual Paper

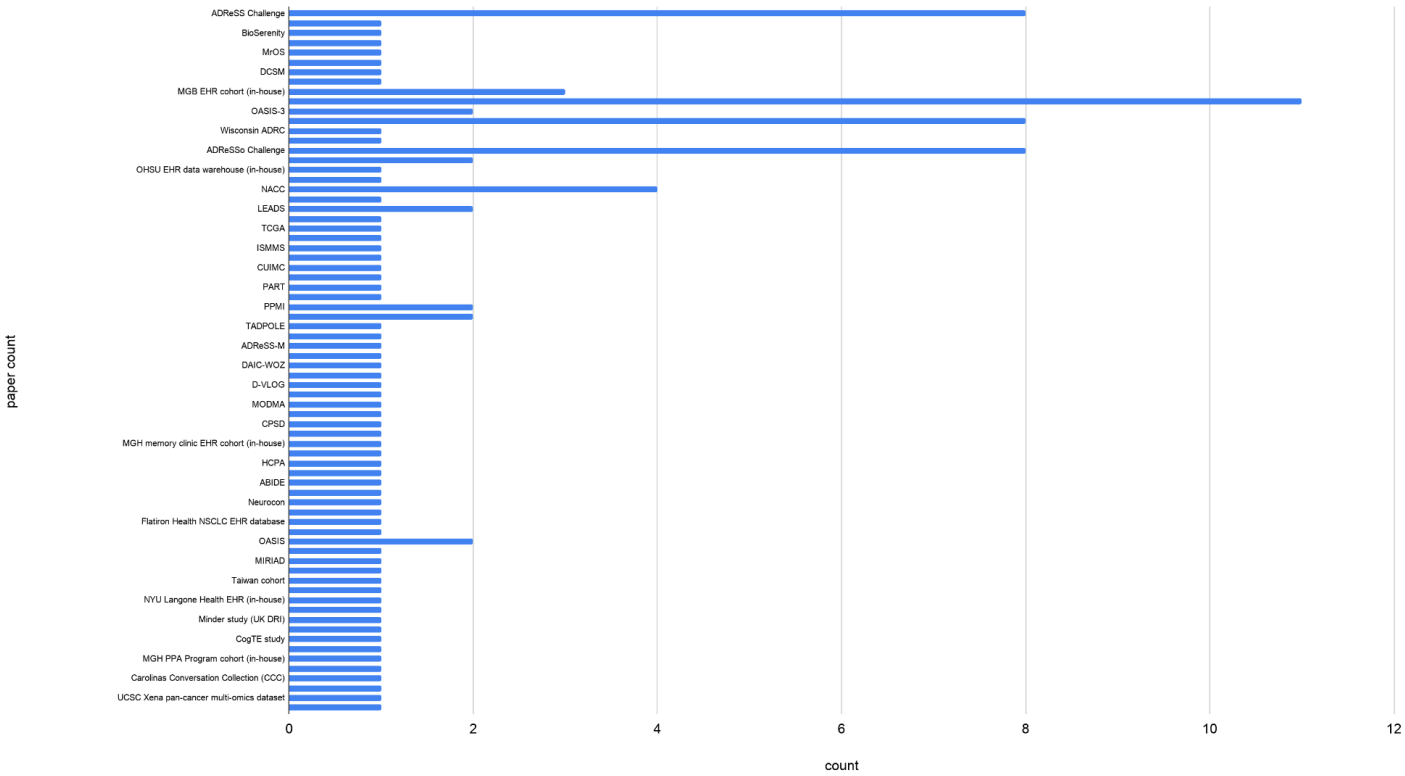


Figure 4
Count of datasets used by individual paper.

Paper count vs Modality (component)

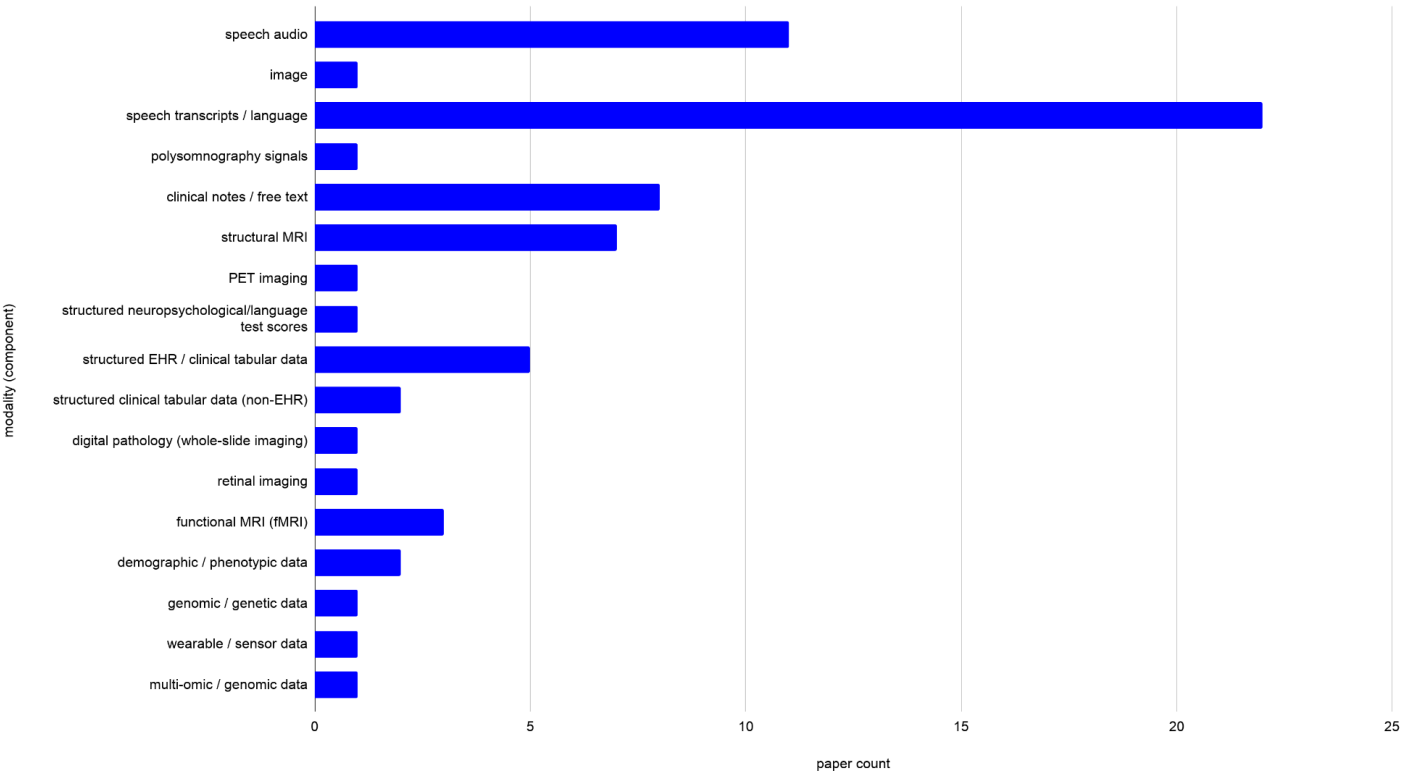


Figure 5

Paper count vs individual modality of input data.

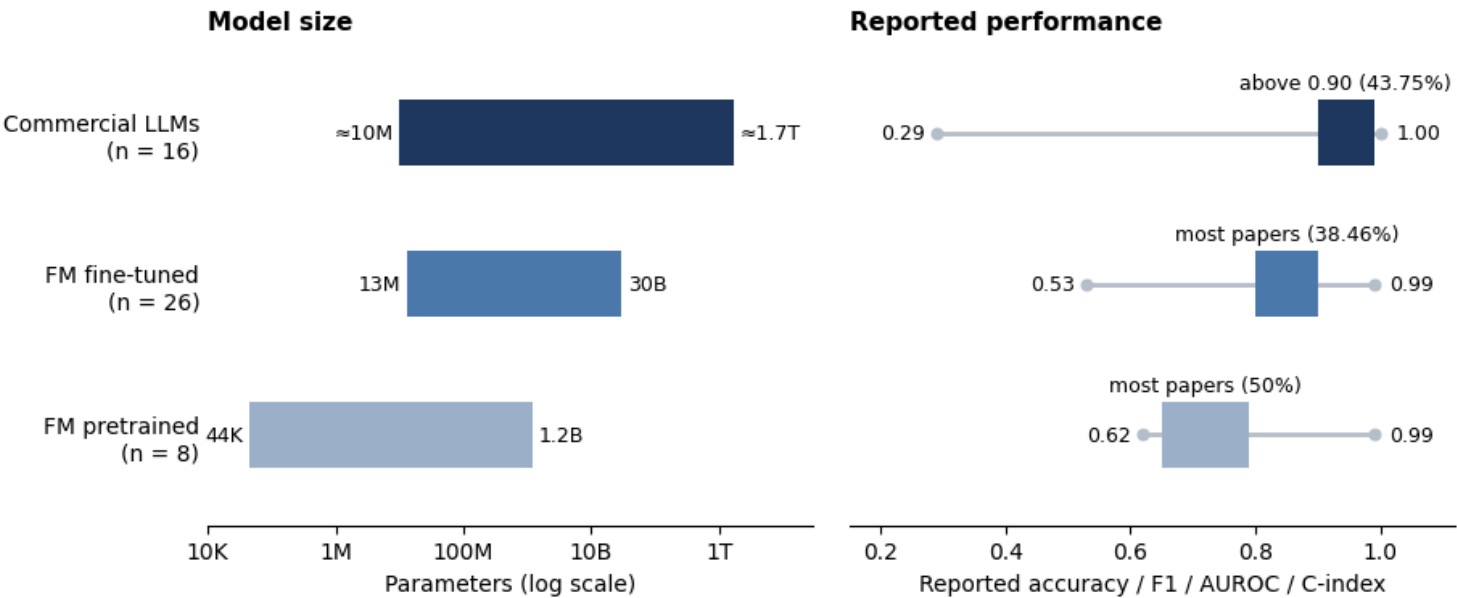


Figure 6

Parameter counts and reported performance across commercial LLMs, fine-tuned FMs, and pre-trained FMs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [NPJDementiaSupplementaryData.docx](#)