Data Management and Sharing Plan

Scientific data from this project will be managed, preserved, and shared based on the below plan.

Data type: (1) *Data that will be generated.* The project will collect conversation video/audio recording, conversation transcripts with names and direct identifiers removed, derived language and quantitative measures, survey responses, and system use logs from 20 MCI and 20 normal cognition participants. (2) *Data that will be preserved and shared.* Participant-level data described in 1 will be de-identified before submitting to repository and preserved through deposition of the data in a controlled access public repository. The data collection system will be designed to minimize the collection of personally identifiable information (PII). Raw audio/video will not be shared. (3) *Meta data and documents.* Submitted documentation will include study protocols, a table that explains each variable, and survey instruments.

Related tools, software, and code: Data processing and analysis Python code will be shared in a public GitHub code repository, with clear instructions (including parameters) to reproduce reported results. Tools used in the project will include open-source python software and libraries, and any custom tools will be shared with documentation in the code repository.

Standards: (1) Tabular data from demorgraphic and acceptability survey and other measurable outcomes (adherence rate, recruitment rate and engagement level) will be provided in non-proprietary formats such as comma-separated values (CSV). Variable names, definitions, and allowed values will be documented in a table that explains each variable. Where applicable, we will use common measures and reporting formats used in aging and cognitive health research. (2) Audio data will be stored in WAV format, and video data in MP4 format. Transcripts will be stored in jsonl format.

Data preservation, access, and timelines: De-identified datasets and documentation will be deposited in a public repository that provides permanent identifiers and long-term preservation, such as AD Data Initiative Repository. At least, the study website will create a permanent link for the data request form along with the contact personnel information. The identifier will be included in the final publication. A data use agreement will be required to access the data. Data will be shared no later than the time of an associated publication, and will remain available for at least five years after release.

Access, distribution, or reuse considerations: We will maximize appropriate sharing while protecting participant privacy and confidentiality. (1) *Factors affecting subsequent access, distribution, or reuse of scientific data*: The data will be collected with the informed consent: The dataset will be only used for research purposes and does not include the study of population origins or ancestry. (2) *Access Control*. Data will be reviewed for the risk that someone could match the data back to a person. A Data Use Agreement (DUA) will be required to executed before accessing the data. (3) *Privacy Protection*: Study data will be stored and processed within Massachusetts General Brigham (MGB) systems in accordance with the Health Insurance Portability and Accountability Act (HIPAA), encrypted in transit and at rest, and protected by secure authentication and periodic security review.

Oversight of data management and sharing: The Principal Investigator for the project, Dr. Junyuan Hong, together with Dr. Liu Chen (data manager), Ms. Cathrine Young (the program manager), and a designated study coordinator, will oversee compliance with this plan, review data sharing readiness before release, and conduct compliance review annually. The progress toward the plan's DMS activities will be included in reports submitted to the data repository officer. At the project conclusion, the final progress report will summarize how the DMS objectives were fulfilled and provide links to the shared dataset(s).