AGENTIC COGNITIVE PROFILING: REALIGNING AUTOMATED ALZHEIMER'S DISEASE DETECTION WITH CLINICAL CONSTRUCT VALIDITY

Jiawen Kang, Kun Li, Dongrui Han, Jinchao Li, Junan Li, Lingwei Meng, Xixin Wu, Helen Meng
The Chinese University of Hong Kong
jwkang@se.cuhk.edu.hk

ABSTRACT

Automated Alzheimer's Disease (AD) screening has predominantly followed the inductive paradigm of pattern recognition, which directly maps the input signal to the outcome label. This paradigm sacrifices construct validity of clinical protocol for statistical shortcuts. This paper proposes an Agentic Cognitive Assessment Framework that realigns automated screening with clinical protocol logic across multiple cognitive domains. Rather than learning opaque mappings from transcripts to labels, the framework decomposes standardized assessments into atomic cognitive tasks and orchestrates specialized LLM agents to extract verifiable scoring primitives. Central to our design is decoupling semantic understanding from measurement by delegating all quantification to deterministic function calling, thereby mitigating hallucination and restoring construct validity. Unlike popular datasets that typically comprise around a hundred participants under a single task, we evaluate on a clinically-annotated corpus of 402 participants across eight structured cognitive tasks spanning multiple cognitive domains. The framework achieves 90.5% score match rate in task examination and 85.3% accuracy in AD prediction, surpassing popular baselines while generating interpretable cognitive profiles grounded in behavioral evidence. This work demonstrates that construct validity and predictive performance need not be traded off, charting a path toward AD screening systems that explain rather than merely predict.

1 INTRODUCTION

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder characterized by the deterioration of specific cognitive faculties, such as memory and executive function. Since neuropathological changes are often clinically silent in early stages, clinicians rely on standardized cognitive tests—such as the Montreal Cognitive Assessment (MoCA) (Hobson, 2015)—as causal probes to detect underlying deficits (Harvey, 2012). Unlike passive observation, these tests are engineered protocols designed to isolate and quantify distinct cognitive domains. For instance, list learning tasks explicitly stress memory encoding and retrieval (Delis et al., 2000), while visual naming tasks probe the integrity of semantic knowledge and visual perception (Giles et al., 1996). Consequently, this paradigm of structured cognitive profiling yields verifiable diagnostic evidence, establishing a reliable cornerstone for clinical decision-making.

In contrast to this *deductive* clinical logic, the field of automated AD screening has predominantly followed the paradigm of *inductive* pattern recognition. Early approaches relied on handcrafted features (e.g., lexical diversity, syntactic complexity) (Fraser et al., 2016; Weiner et al., 2019), while recent works leverage the Pre-trained Language Models (PLMs) (Balagopalan et al., 2020; Yuan et al., 2020) to map input transcripts directly to diagnostic labels. Although PLM-based methods have significantly improved benchmark performance (e.g., on ADReSS (Luz et al., 2020)), they fundamentally treat AD detection as a *"de-contextualized"* classification task, which identifies arbitrarily possible statistical correlations between featurized inputs and diagnostic labels. These "black-box" models face inherent reliability challenges. Notable studies have demonstrated the "Clever Hans" effect (Liu et al., 2024; Sahidullah et al., 2025; Kang et al., 2025), where models could

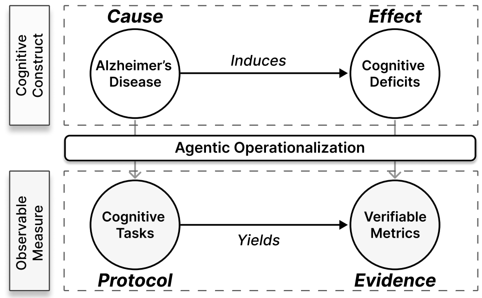


Figure 1: The Conceptual Framework. We align AD detection with clinical construct validity by operationalizing the causal chain from Alzheimer's Disease to Cognitive Deficits into an agentic workflow comprising Cognitive Tasks and Verifiable Metrics.

predict correct labels based on non-pathological shortcuts. Besides, the field observes inconsistent findings regarding feature robustness on speech recognition errors (Kang et al., 2024; Li et al., 2024a). Fundamentally, these issues are stem from their intrinsic paradigm: *pattern recognition models predict outcomes rather than extract causal evidence, as in clinical protocol.*

From the lens of construct validity (Strauss & Smith, 2009), the divergence between these two paradigms reveals a fundamental measurement gap (Van der Wal et al., 2024). The clinical protocol largely operationalizes the target neurocognitive constructs: performance decrements are directly attributable to specific failures in cognitive domains (e.g., memory deficits). Conversely, predominant data-driven approaches often sacrifice construct validity for predictive validity, thereby encoding more construct-irrelevant variance—confounding proxy features (e.g., dialect, or acoustic shortcuts) with genuine pathological signals.

This paper aims to realign automated AD screening with clinical grounding. To this end, we propose an Agentic Cognitive Assessment Framework that shifts from inductive pattern recognition to deductive clinical profiling across multiple cognitive domains. The key insight is to operationalize clinical scoring logic through a multi-agent workflow, decoupling semantic understanding from measurement by delegating all quantification to deterministic functions, ensuring verifiable and construct-valid scoring. We make the following contributions:

- **Multi-faceted Assessment**: Existing public benchmarks for automated AD detection (e.g., ADReSS, N=156) are typically built upon a single elicitation task, capturing a limited facet of cognitive manifestation. We evaluate on a clinically-annotated corpus of 402 participants across eight structured tasks spanning multiple cognitive domains, providing the first empirical evidence for the feasibility and effectiveness of automated cognitive scoring and AD screening under a multi-domain structured assessment setting.
- **Framework**: We propose an Agentic Cognitive Assessment Framework that, for the first time, realigns automated AD screening with clinical construct validity. Rather than learning opaque mappings from transcripts to labels, the framework decomposes standardized assessments into atomic cognitive tasks and orchestrates specialized agents to extract verifiable scoring primitives, supporting both zero-shot and supervised screening with interpretable cognitive profiles.
- **Performance**: The framework achieves 90.5% score match rate in task examination and 85.3% accuracy in AD screening, surpassing PLM-based baselines in both zero-shot and supervised settings while maintaining full interpretability.

2 Related Work

2.1 Clinical Cognitive Assessment

Clinical cognitive assessments are standardized protocols designed to isolate and quantify neurocognitive deficits (Harvey, 2012). These instruments generally fall into three categories based on granularity: *(a) Rapid screening tests*, such as the AD8 (Galvin et al., 2005), typically question daily functioning (e.g., troubles with making decisions) and detect initial functional shifts before formal testing; *(b) Domain-specific tests* target distinct faculties. In language function, the "Cookie Theft" task (from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001)) underpins bench-

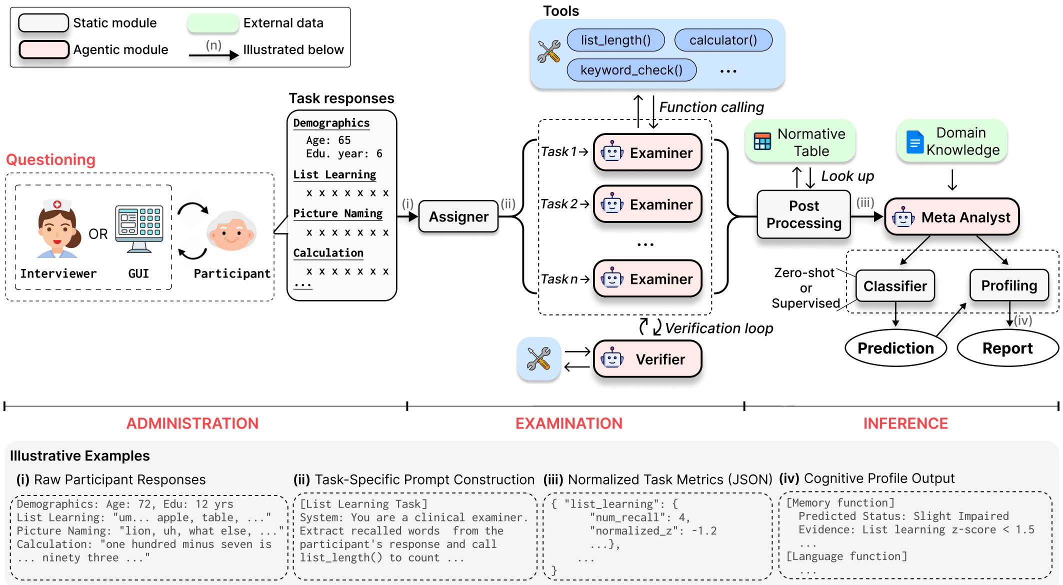


Figure 2: Overview of the Agentic Cognitive Assessment Framework. The workflow comprises three stages: (1) Administration: Collection of standardized task responses; (2) Examination: Multi-agent Workflow with Deterministic Function Calling and Verification Loop; (3) Inference: Aggregation of verified metrics for classification and explainable reporting.

marks like ADReSS (Luz et al., 2020; 2021). In memory function, list-learning tasks (Delis et al., 2000; Au et al., 2003) serve as the gold standard for episodic memory; *(c) Comprehensive screening tests* profile impairment across multiple domains. The Mini-Mental Status Examination (Cockrell & Folstein, 2002) is a widely used baseline, but it suffers from ceiling effects in early-stage pathology. Consequently, the Montreal Cognitive Assessment (MoCA) (Hobson, 2015; Yeung et al., 2014) has emerged as the standard for detecting Mild Cognitive Impairment (MCI) due to its sensitivity to executive dysfunction. We adopted MoCA and the Hong Kong List Learning Test (HKLLT) (Au et al., 2003) as the clinical foundation for this work.

2.2 Automatic AD detection

Dominant research in natural language-based AD detection has been driven by the exploration of effective features to improve discrimination. Early efforts utilized handcrafted acoustic and linguistic features (Fraser et al., 2016; Weiner et al., 2019), such as decreased jitter and lexical diversity. The advent of pre-trained models has shifted the paradigm toward deep representation learning, achieving state-of-the-art performance across text (Balagopalan et al., 2020; Meng et al., 2023; Yuan et al., 2020; Wang et al., 2022), speech (Haulcy & Glass, 2021; Li et al., 2023; Zhu et al., 2021), and multi-modal settings (Koo et al., 2020; Li et al., 2025a; Syed et al., 2021) by capturing rich semantic dependencies.

More recently, distinct approaches have explored incorporating certain clinical knowledge as priors in model design. Li et al. (2024b); Park et al. (2025) assess information retrieval in the Cookie Theft picture description task, and Li et al. (2025a) quantifies image-narrative alignments in image-based storytelling tasks. These methods provide a degree of explainability yet lack explicit modeling of clinical decision making, thus limiting both transparency and coverage of cognitive profiling.

Beyond algorithmic advances, a body of research has also validated the deployment practices of automated AD detection systems, including Ding et al. (2022); An et al. (2025); Breithaupt et al. (2025). These efforts adopt a human-computer interaction perspective, exploring the use of graphical user interfaces (GUIs) to administer test items to participants. Such work has progressively narrowed the gap between GUI-based interaction and human-administered assessment.

Table 1: Overview of Cognitive Tasks and Scoring Primitives. Scoring primitives define measurable performance units and their maximum scores.

Task	**Scoring Primitives**	**Score**
MoCA-SL (Assess language, attention, executive functions)		
Picture Naming	Per-item correctness (×3)	3
Digit Span	Forward/backward correctness	2
Serial 7 Subtraction	# correct subtractions	3
Sentence Repetition	Per-sentence correctness (×2)	2
Animal Fluency	# valid animals	1
Abstraction	Per-pair correctness (×2)	2
	Total:	13
HKLLT (Assess learning and memory functions)		
Trial-4	# recalled words (10-min delay)	16
Trial-5	# recalled words (30-min delay)	16

2.3 LLM-based Agentic Workflow

Agentic workflows enable LLMs with the capacity to navigate complex problem spaces through iterative cycles of perception, planning, and execution. This shift turns opaque generation into transparent, verifiable actions. Agents have shown great potential in multiple domains, including coding automation (Yang et al., 2024), scientific discovery (M. Bran et al., 2024), and open-world exploration (Wang et al., 2024).

In the context of clinical cognitive assessment, very few works have explored this area. (Bazgir et al., 2025) employs LLM agents for AD disease management. (Li et al., 2025b) and (Hou et al., 2026) develop agents for data preprocessing and model selection, respectively, while still relying on external neural networks for classification.

3 Methodology

3.1 Task Definition and Cognitive Constructs

We formulate the automatic AD detection process as a cognitive profiling task. Formally, given a session transcript X, our objective is to extract a set of interpretable *Scoring Primitives*, $\mathcal{S} = \{s_1, s_2, ..., s_k\}$, representing the atomic units of clinical evidence (e.g., "successful naming a depicted animal", or "successful recall of 'velvet'"). These primitives serve as a grounded representation to derive both a human-readable screening report and a binary detection label.

Structured Input: Unlike open-ended conversations where cognitive signals are sparse and entangled, we ground our framework in standardized clinical protocols. Based on MoCA and HKLLT, we adopt a MoCA-SL (Spoken Language subset of MoCA) and Trial $4/5$ in the HKLLT as our profiling tasks, to ensure construct validity and reliable cognitive profiling. The corresponding cognitive domains and scoring primitives are summarized in Table 1, and we provide further details in Appendix A. We assume structured input: participant demographics (age, years of education) paired with verbal responses to each cognitive task. This input format directly mirrors clinical practice, and it is scalable for large-scale deployment, where a growing body of HCI research (Ding et al., 2022; An et al., 2025; Breithaupt et al., 2025) has demonstrated that GUI-based interfaces can reliably administer these standardized tests.

Explainable Output. Beyond binary prediction ($y \in \{\text{AD}, \text{HC}\}$), our framework generates a structured cognitive profile that details *why* a screening conclusion is reached. We argue that an interpretable screening system should not be limited to binary prediction. In clinical practice, assessment of cognitive impairment is never a simple yes-or-no decision, it is equally crucial to understand how the assessment leads to a particular conclusion. Such information is essential for patient communication and treatment planning. Our framework addresses this by generating cognitive reports that detail task-level evidence alongside clinical interpretations (e.g., failure in delayed recall).

3.2 Agentic Examination

The Examination stage employs a multi-agent workflow to transform verbal responses into structured scoring primitives (see Figure 2). We describe each component below.

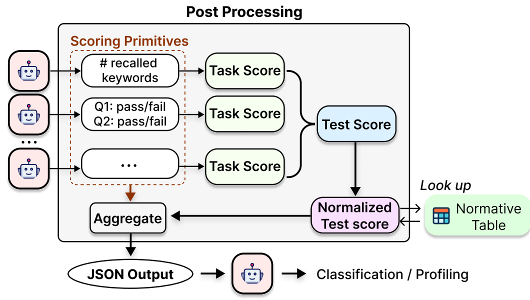


Figure 3: Cognitive profile inference pipeline. Verified scoring primitives from each task are aggregated, normalized against demographic norms, and used for classification and report generation.

3.2.1 Task-Specific Examiner Agents

The Assigner routes each task response to a dedicated Examiner agent equipped with task-specific prompts. Task prompts are structured, comprising four components: *(a) Task Introduction*: clinical role and assessment objective; *(b) Guidelines*: processing rules and edge cases; *(c) Output Format*: response structure; and *(d) Examples*: demonstrations for consistent behavior. And each Examiner therefore applies task-specific rules to process transcription, outputting task-specific scoring primitives for downstream processing. Prompt template and examples are provided in Appendix E.1.

This modular design enables flexible task-specific processing. For instance, the Animal Fluency Examiner requires semantic understanding to parse responses, identify valid animals, and deduplicate lexical variants before counting. Others could act as simple bridges—the Sentence Repetition Examiner directly passes the target sentence and transcription to `keyword_check()` without intermediate processing. While more autonomous planning architectures are feasible, our preliminary experiments indicate that fixed, template-based guidelines yield more *stable* and *input-robust* outputs.

3.2.2 Deterministic Function Calling

A key design principle is the decoupling of semantic understanding from measurement. While LLMs interpret natural language content, such as eliminating disfluencies, understanding dialectal variations, and semantics. All quantification is delegated to deterministic functions, mitigating hallucination (i.e., fabrication of unsupported outputs) in numerical outputs. Specifically, Examiner agents have access to a library of scoring tools. Some functions are shared across tasks, while others are task-specific. For instance, `keyword_check()` function checks if certain keywords appear in a list, and could serve both Sentence Repetition and Digit Span tasks. `parse_hkllt()` function extracts HKLLT-specific metrics such as semantic clustering.

3.2.3 Verification Loop

We observe that Examiner agents exhibit hallucination when processing transcripts—LLMs occasionally fabricate false evidence to justify incorrect scores. This issue is particularly pronounced when deploying smaller, locally-hosted models. To ensure reliability, each Examiner's result undergoes validation by a *verifier* agent. The verifier receives the original transcript alongside the Examiner's output and evaluates correctness. If discrepancies are detected—such as hallucination or incorrectly parsed responses— verifier agent provides feedback specifying the error, and the Examiner regenerates its output. This loop continues for up to 3 iterations, after which the result is accepted regardless of verification status.

3.3 Cognitive Profile Inference

The Inference stage aggregates verified scoring primitives, normalizes them against population norms, and generates both a prediction and an interpretable clinical report (see Figure 3).

3.3.1 Score Normalization

In this stage, scoring primitives from all examiners are collected into a unified JSON structure. Task scores are computed by aggregating primitives according to standard clinical protocols, for example, summing per-item correctness flags in Picture Naming task. These task scores then aggregate into test scores: MoCA-SL (maximum 13 points) and HKLLT trial scores (maximum 16 words each).

To obtain norm-referenced scores, we consult age-education stratified normative tables. For HKLLT, published Hong Kong norms are directly available. For MoCA-SL, however, existing normative data are based on the full 30-point assessment Wong et al. (2015). We address this by linearly rescaling the full MoCA norms proportionally to the MoCA-SL score range, keeping comparable distributional properties. Alternative estimation methods are compared in Appendix C.

3.3.2 Prediction

We implement two prediction approaches: zero-shot and supervised. The zero-shot method applies established clinical thresholds directly without requiring any training data: a participant is flagged as AD if the MoCA-SL score falls below the 16th percentile, or if either the HKLLT delayed recall score (10-min or 30-min) falls below -1.0 SD. These cutoffs correspond to standard clinical practice for identifying mild cognitive impairment, ensuring full transparency in decision logic.

We also find that behaviorally-grounded scoring primitives serve as effective features for supervised classification. Using all primitives as input, we train an SVM classifier that achieves strong performance. Unlike black-box approaches that operate on abstract learned representations, this classifier relies exclusively on interpretable behavioral evidence extracted from task performance.

3.3.3 Cognitive Profiling

Beyond binary prediction, the Meta Analyst agent explains why a particular outcome is reached, grounding each conclusion in task-level evidence. The agent receives the participant's scoring primitives, norm-referenced scores, and a domain knowledge document specifying each task's clinical significance, normal ranges, and interpretation guidelines.

The output comprises two steps: (a) a risk analysis regarding multiple cognitive domains (e.g., memory, executive function), with each domain containing a status indicator, supporting evidence, and clinical interpretation; and (b) a narrative statement summarizing the participant's overall cognitive profile. The cognitive profile and final prediction are derived from identical evidence, ensuring that the screening outcome is not merely a label but a transparent, auditable clinical judgment.

4 Data

To our knowledge, no publicly available dataset provides structured responses from standardized cognitive assessments. Therefore, we evaluate our framework on an in-house Cantonese speech corpus collected for cognitive screening research. The corpus comprises recordings from 1,063 older adults, each participating in a clinician-guided assessment session lasting approximately 1.5 hours. Sessions cover multiple standardized cognitive tests, including AD8, MoCA, HKLLT, and the Modified Boston Naming Test (mBNT).

Inclusion Criteria. Participants were required to be (1) aged 60 years or above, (2) proficient in spoken Cantonese, and (3) capable of completing all tests with adequate vision and hearing (corrective aids permitted).

Annotation and Labels. Of the full corpus, 402 sessions have been manually transcribed with task-level segmentation. Each participant was assigned a binary label—AD-risk (AD) or healthy control (HC)—by professional clinical assessors. We use this annotated subset for all experiments in this work.

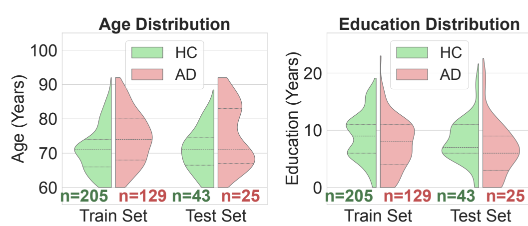


Figure 4: Demographic distribution of participants in datasets. Subplots depict age and years of education for Alzheimer's (AD) and healthy control (HC) groups.

Data Split. We partition the annotated data into training (N=334) and test (N=68) sets. The age and education distributions of participants are presented in Figure 4, which shows balanced demographic patterns between the AD and healthy control groups.

Task Coverage. As detailed in Table 1, we extract responses from two test batteries: (1) MoCA-SL, comprising six spoken-language (SL) tasks from the full MoCA protocol, and (2) HKLLT, a representative test assessing learning and memory ability.

Table 2: Task-level examination result. Score Match Rate (SMR) indicates the exact agreement percentage with manual scores; Mean Absolute Error (MAE) contextualizes error magnitude. Score range is annotated per task. *Func. Call*: Deterministic Function Calling.

	Full		**w/o Verifier**		**w/o Func. Call**	
Task	SMR	MAE	SMR	MAE	SMR	MAE
MoCA-SL						
Picture Naming (/3)	97.0%	0.03	97.0%	0.03	92.6%	0.07
Digit Span (/2)	98.5%	0.01	98.5%	0.01	77.3%	0.24
Serial 7 Sub. (/3)	82.4%	0.19	70.6%	0.32	63.2%	0.63
Sentence Rep. (/2)	89.7%	0.10	89.7%	0.10	86.8%	0.13
Animal Fluency (/1)	98.5%	0.01	97.1%	0.03	98.5%	0.02
Abstraction (/2)	82.3%	0.18	64.7%	0.38	67.6%	0.32
Score-Weighted Avg.	90.5%	0.10	85.5%	0.16	79.2%	0.27
HKLLT						
Trial-4 (/16)	94.1%	0.07	94.1%	0.07	27.9%	2.23
Trial-5 (/16)	92.6%	0.07	92.6%	0.07	23.5%	2.32

5 Experiments and Results

5.1 Experimental Setup

We implement all agents using Qwen3-8B deployed locally via vLLM, with temperature set to 0.3 for examiners and 0.1 for verifiers. We compare against three baseline categories: (1) traditional approaches using handcrafted linguistic features (13 features); (2) PLM-based methods applying BERT and RoBERTa on concatenated task transcripts; and (3) LLM-CoT, which prompts the same backbone model with concatenated transcripts in a Chain-of-Thought (CoT) manner (Park et al., 2025). For task examination, we report Score Match Rate (SMR) and Mean Absolute Error (MAE), where SMR indicates the exact agreement percentage with manual scores; MAE contextualizes error magnitude. For screening inference, we report Accuracy, F1, Precision, and Recall. Supervised classifiers (SVM with RBF kernel, MLP) are trained on extracted scoring primitives. All experiments are repeated 5 times, and we report the mean performance. Implementation details are provided in Appendix B.

5.2 Task Examination Evaluation

Examination Results. Table 2 presents per-task SMR and MAE. Low-inference tasks, such as Picture Naming, Digit Span, and Animal Fluency, achieve near-ceiling accuracy ($> 97\%$) as they involve straightforward extraction aligned with clinical scoring rules. High-inference task Serial 7 Subtraction and Abstraction show slightly lower but robust performance (82%), as these tasks require nuanced semantic judgment. HKLLT delayed recall tasks also achieve high accuracy ($> 92\%$). Note that Sentence Repetition exhibits lower SMR (89%) due to pronunciation-level ambiguities where clinical assessors and annotators often disagree; we assign lower weight to subsequent analyses. Overall, the framework achieves 90.5% weighted-average SMR with 0.10 MAE, demonstrating reliable examination across diverse cognitive tasks. We leave more results and discussions in Appendix D.

Ablation Study. Table 2 also quantifies the contribution of the verification loop and deterministic function calling. For low-inference tasks, removing either component yields negligible changes, as performance already approaches the ceiling. However, on high-inference tasks, the Verifier proves essential: its removal causes substantial drops in Abstraction and Serial 7 Subtraction, indicating it corrects hallucinated evidence in nuanced reasoning. Removing function calling causes more severe degradation on counting-intensive tasks (e.g., HKLLT Trial-4), confirming that LLM requires deterministic computation for reliable measurement. Together, these results demonstrate that both components are critical for challenging tasks.

Case Study. We conduct error analysis to investigate failure modes. Details can be found in Appendix F. For low-inference tasks, errors primarily stem from model hallucinations that the verifier fails to handle. As we implemented an 8B model, we expect larger models might be more stable, thus mitigating such issues. For high-inference tasks, errors typically occur at boundary cases due to

Table 3: Comparison of Alzheimer's disease screening performance (%) across baseline and proposed systems. Results exceeding 80% are highlighted in bold.

System	Classifier	Accuracy	F1	Precision	Recall
Baseline					
Handcrafted	MLP / SVM	68.2 / 70.6	55.7 / 67.8	58.3 / 56.8	54.4 / **84.0**
BERT	MLP / SVM	73.7 / 79.4	59.5 / 72.0	68.3 / 72.0	52.8 / 72.0
RoBERTa	MLP / SVM	70.4 / 75.0	53.9 / 66.7	63.1 / 65.4	47.2 / 68.0
LLM-CoT	-	70.6	66.7	57.1	**80.0**
Proposed					
Zero-shot	-	**82.4**	**80.2**	**81.4**	**80.2**
w/o Verifier	-	79.4	77.4	78.1	77.5
w/o Function	-	69.1	63.3	66.8	63.0
Supervised	MLP / SVM	**81.5 / 85.3**	66.9 / 78.3	**95.7 / 85.7**	52.0 / 72.0
w/o Verifier	MLP / SVM	73.2 / 77.9	44.5 / 61.5	**94.4 / 85.7**	29.6 / 48.0
w/o Function	MLP / SVM	74.1 / **82.4**	49.8 / 76.0	**86.6** / 76.0	35.2 / 76.0

response diversity. For instance, in the train-bicycle similarity task, the model accepted "both can move" as correct, while the assessor judged it too generic.

5.3 Screening Inference Evaluation

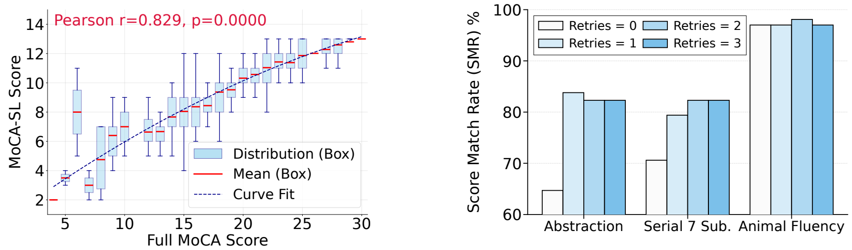
Prediction Results. Table 3 compares AD screening performance across baseline and proposed systems. The proposed agentic framework outperforms prior PLM-based methods in both zero-shot and supervised settings. Among baselines, BERT achieves the best performance (79.4% accuracy), consistent with prior findings on PLM-based AD detection. LLM-CoT attains 70.6% accuracy in a fully zero-shot manner, demonstrating the advantage of structured clinical inputs for reasoning with domain knowledge—yet its performance remains bounded by the absence of verifiable measurement. Our proposed framework addresses this limitation through grounded scoring primitives. The zero-shot approach surpasses the best baseline without any training data, achieving the highest F1 score (80.2%) across all systems. The supervised variant further improves accuracy to 85.3%, establishing a substantial margin over prior methods. We note that SVM consistently outperforms MLP across all configurations, likely due to class imbalance causing naive MLP to exhibit prediction bias toward the majority class. Ablation results align with section 5.2: removing the verifier or function calling degrades both zero-shot and supervised performance, confirming their critical role in reliable scoring primitive extraction.

Cognitive Profiling. Beyond binary prediction, our framework generates structured cognitive profiles that explain why a screening conclusion is reached. Each report is organized by cognitive domain—memory, executive function, attention, and language—with status indicators, supporting evidence, and clinical interpretations grounded in norm-referenced performance. This dual-output bridge automated screening with clinical workflow, ensuring that outcomes are transparent and auditable rather than opaque labels. We showcase several complete example reports in Appendix E.2.

5.4 Analysis

Validate MoCA-SL scores. MoCA-SL serves as the core protocol for assessing language, attention, and executive functions, directly influencing both screening outcomes and cognitive profiling. As a spoken-language subset, a natural question arises: Does MoCA-SL introduce measurement noise compared to the full MoCA gold standard? To validate MoCA-SL, we examine its alignment with full MoCA scores across all participants with both assessments available. Figure 5 plots MoCA-SL against full MoCA scores. Despite variance at low-score ranges due to sparse samples, the two measures exhibit a strong linear relationship (Pearson $r = 0.829$, $p < 0.001$). This correlation confirms that MoCA-SL preserves the discriminative validity of the full assessment.

Verifier Max Retries. The Verification Loop iteratively corrects Examiner hallucinations, but excessive retries may introduce latency without performance gains. We analyze how the maximum retry N_{max} limit affects examination accuracy. We vary N_{max} from 0 to 3 and measure SMR on high-inference tasks where verification is most impactful. Figure 5 (b) shows SMR results. Performance



(a) Score Alignment **(b) Retries Analysis**

Figure 5: (a) Alignment of the full MoCA score and the subset MoCA-SL score; (b) Performance comparison of the Score Match Rate (SMR) with varying maximum retry limits.

improves sharply from N_{max} = 0 to N_{max} = 1, with diminishing returns beyond N_{max} = 2. Based on this analysis, we set N_{max} = 3 to balance accuracy and efficiency.

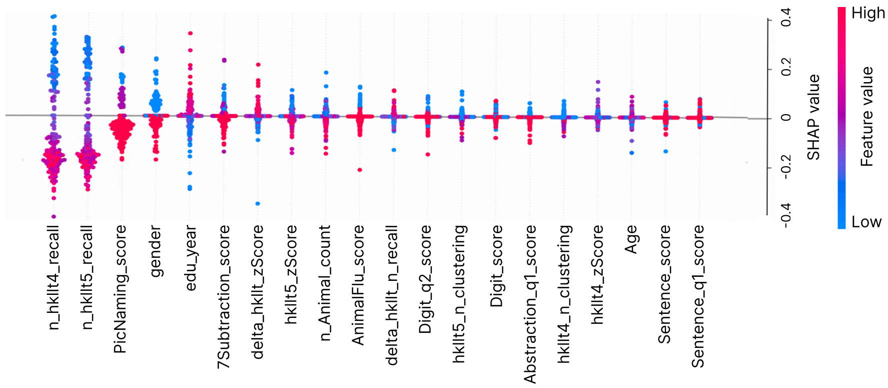


Figure 6: Feature importance analysis using SHAP values for the supervised SVM classifier. Top-20 features are depicted.

Feature Importance. We analyze SHAP values on the SVM classifier to interpret feature contributions. Figure 6 shows that memory-related HKLLT features (e.g., n_hkllt4_recall) are the most influential, consistent with clinical consensus that episodic memory is an effective early marker of AD risk. MoCA-SL task scores (e.g., PicNaming, 7Subtraction) follow, aligning with clinical expectations of language and attention as secondary indicators. Notably, edu_year shows a negative relationship—higher education is associated with reduced predicted AD risk, consistent with the cognitive reserve effect documented in clinical literature and the education-adjusted normative cutoffs used in clinical practice (see Appendix C). Among low-impact features, we hypothesize that the Sentence task contributes minimally likely due to ceiling effects, and that HKLLT semantic clustering metrics may be masked by the dominant n_recall signal.

6 Conclusion

We presented an agentic cognitive assessment framework that bridges the gap between clinical protocol logic and automated Alzheimer's Disease screening. By decomposing standardized assessments into atomic cognitive tasks and delegating all quantification to deterministic functions, the framework restores construct validity to automated screening without sacrificing predictive performance. Experiments on eight cognitive tasks and 402 participants confirm two key findings: (a) the zero-shot setting already surpasses PLM-based baselines trained on task transcripts, suggesting that clinical protocol logic itself carries sufficient discriminative signal without requiring data-driven pattern extraction; (b) SHAP analysis confirms that the most influential features—notably episodic memory recall—align with established clinical markers of early AD, indicating that the extracted scoring primitives capture genuine cognitive constructs rather than statistical shortcuts. Together, these results demonstrate that construct validity and predictive performance are complementary rather than competing objectives, offering a principled foundation for automated cognitive assessment systems that explain rather than merely predict.

LIMITATIONS

1. Dependence on Predefined Scoring Rules

The framework's operation aligns with standardized clinical protocols, and its validity is therefore contingent on the implementation of established scoring rules. While this ensures fidelity to clinical constructs, the assessment is inherently confined to the cognitive domains predefined by the tasks.

2. Reliance on LLM Semantic Capability

While deterministic functions ensure objective quantification, the agents' semantic parsing performance remains limited by the LLM's comprehension capabilities.

3. Restricted Dataset Accessibility

Due to Institutional Review Board constraints and ethical guidelines, the clinical corpus used in this work cannot be shared publicly, limiting direct reproducibility. However, the framework's architecture is not language-specific. Adapting it to other languages requires only re-authoring task-specific prompts and normative tables, while the overall pipeline structure remains transferable.

ETHICAL CONSIDERATIONS

We used LLMs (e.g., Claude and Gemini) to assist with language editing and clarification during the preparation of this paper. The use of LLMs was limited to improving readability and expression, and all technical content, ideas, and research contributions are solely those of the authors. We confirm compliance with ACL ethical guidelines.

REPRODUCIBILITY STATEMENT

All prompts and scoring criteria are provided in Appendices B–E. The assessment protocol and feature extraction procedures are detailed in Sections 3–4 and Appendix A. The clinical dataset cannot be publicly released due to IRB constraints on protected health information, but we provide sufficient methodological detail to enable replication with independently collected data.

ACKNOWLEDGMENTS

This work is supported by the HKSARG Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N), the CUHK Stanley Ho Big Data Decision Research Centre, and the Centre for Perceptual and Interactive Intelligence (CPII) Ltd, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

REFERENCES

Jiaxin An, Siqi Yi, Yao Lyu, Houjiang Liu, and Yan Zhang. Conversational agents for older adults' health: A systematic literature review, 2025. URL https://arxiv.org/abs/2503.23153.

Alma Au, Agnes S Chan, and Helen Chiu. Verbal learning in alzheimer's dementia. *Journal of the International Neuropsychological Society*, 9(3):363–375, 2003. doi: 10.1017/S1355617703930025.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. In *Interspeech 2020*, pp. 2167–2171, 2020. doi: 10.21437/Interspeech.2020-2557.

Adib Bazgir, Amir Habibdoust, Xing Song, and Yuwen Zhang. Agenticad: A specialized multiagent system framework for holistic alzheimer disease management, 2025. URL https://arxiv.org/abs/2510.08578.

Andrew G. Breithaupt, Nayoung Choi, James D. Finch, Jeanne M. Powell, Arin L. Nelson, Oz A. Alon, Howard J. Rosen, and Jinho D. Choi. Designing and evaluating a conversational agent for early detection of alzheimer's disease and related dementias, 2025. URL https://arxiv.org/abs/2509.11478.

Joseph R Cockrell and Marshal F Folstein. Mini-mental state examination. *Principles and practice of geriatric psychiatry*, pp. 140–141, 2002. doi: 10.1002/0470846410.ch27(ii).

Dean C Delis, Joel H Kramer, Edith Kaplan, and Beth A Ober. California verbal learning test–second edition. *Assessment*, 2000. doi: 10.1037/t15072-000.

Zijian Ding, Jiawen Kang, Tinky Oi Ting Ho, Ka Ho Wong, Helene H Fung, Helen Meng, and Xiaojuan Ma. Talktive: A conversational agent using backchannels to engage older adults in neurocognitive disorders screening. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–19, 2022. doi: 10.1145/3491102.3502005.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2016. doi: 10.3233/JAD-150520.

JE Galvin, CM Roe, KK Powlishta, MA Coats, SJ Muich, E Grant, JP Miller, M Storandt, and JC Morris. The ad8: a brief informant interview to detect dementia. *Neurology*, 65(4):559–564, 2005. doi: 10.1212/01.wnl.0000172958.95282.2a.

Elaine Giles, Karalyn Patterson, and John R Hodges. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information. *Aphasiology*, 10(4):395–408, 1996. doi: 10.1080/02687039608248419.

Harold Goodglass, Edith Kaplan, and Sandra Weintraub. *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001. URL https://utdallas.primo.exlibrisgroup.com/discovery/fulldisplay/alma9927692213901421/01UT_DALLAS:UTDALMA.

Philip D Harvey. Clinical applications of neuropsychological assessment. *Dialogues in clinical neuroscience*, 14(1):91–99, 2012. doi: 10.31887/DCNS.2012.14.1/pharvey.

R'mani Haulcy and James Glass. Classifying alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, 11:624137, 2021. doi: 10.3389/fpsyg.2020.624137.

John Hobson. The montreal cognitive assessment (moca). *Occupational Medicine*, 65(9):764–765, 2015. doi: 10.1093/occmed/kqv078.

Wenlong Hou, Guangqian Yang, Ye Du, Yeung Lau, Lihao Liu, Junjun He, Ling Long, and Shujun Wang. ADAgent: LLM agent for alzheimer's disease analysis with collaborative coordinator. In Jianing Qiu, Jinlin Wu, Curtis Langlotz, Baoru Huang, Zhen Lei, Honghan Wu, Hongbin Liu, and Weidi Xie (eds.), *AI for Clinical Applications*, pp. 23–32. Springer Nature Switzerland, 2026. ISBN 978-3-032-06004-4. URL https://doi.org/10.1007/978-3-032-06004-4_3.

Jiawen Kang, Junan Li, Jinchao Li, Xixin Wu, and Helen Meng. Not all errors are equal: Investigation of speech recognition errors in alzheimer's disease detection. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 254–258. IEEE, 2024. doi: 10.1109/ISCSLP63861.2024.10800483.

Jiawen Kang, Dongrui Han, Lingwei Meng, Jingyan Zhou, Jinchao Li, Xixin Wu, and Helen Meng. On the Within-class Variation Issue in Alzheimer's Disease Detection. In *Interspeech 2025*, pp. 5668–5672, 2025. doi: 10.21437/Interspeech.2025-2751.

Junghyun Koo, Jie Hwan Lee, Jaewoo Pyo, Yujin Jo, and Kyogu Lee. Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition. In *Interspeech 2020*, pp. 2217–2221, 2020. doi: 10.21437/Interspeech.2020-3153.

Changye Li, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. Useful blunders: Can automated speech recognition errors improve downstream dementia classification? *Journal of biomedical informatics*, 150:104598, 2024a. ISSN 1532-0464. doi: 10.1016/j.jbi.2024.104598.

Jinchao Li, Kaitao Song, Junan Li, Bo Zheng, Dongsheng Li, Xixin Wu, Xunying Liu, and Helen Meng. Leveraging pretrained representations with task-related keywords for alzheimer's disease detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096205.

Jinchao Li, Yuejiao Wang, Junan Li, Jiawen Kang, Bo Zheng, Ka Ho Wong, Brian Mak, Helene H Fung, Jean Woo, Man-Wai Mak, et al. Detecting neurocognitive disorders through analyses of topic evolution and cross-modal consistency in visual-stimulated narratives. *IEEE Journal of Selected Topics in Signal Processing*, 2025a. doi: 10.1109/JSTSP.2025.3622049.

Junan Li, Yunxiang Li, Yuren Wang, Xixin Wu, and Helen Meng. Devising a set of compact and explainable spoken language feature for screening alzheimer's disease. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 471–475. IEEE, 2024b. doi: 10.1109/ISCSLP63861.2024.10800414.

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, 2025b. doi: https://doi.org/10.1038/s41746-025-01940-4.

Yin-Long Liu, Rui Feng, Jia-Hong Yuan, and Zhen-Hua Ling. Clever hans effect found in automatic detection of alzheimer's disease through speech. In *Interspeech 2024*, pp. 2435–2439, 2024. doi: 10.21437/Interspeech.2024-1018.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Interspeech 2020*, pp. 2172–2176, 2020. doi: 10.21437/Interspeech.2020-2571.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Detecting cognitive decline using speech only: The adresso challenge. In *Interspeech 2021*, pp. 3780–3784, 2021. doi: 10.21437/Interspeech.2021-1220.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: https://doi.org/10.1038/s42256-024-00832-8.

Helen Meng, Brian Mak, Man-Wai Mak, Helene Fung, Xianmin Gong, Timothy Kwok, Xunying Liu, Vincent Mok, Patrick Wong, Jean Woo, Xixin Wu, Ka Ho Wong, Shensheng Xu, Naijun Zheng, Ranzo Huang, Jiawen Kang, Xiaoquan Ke, Junan Li, Jinchao Li, and Yi Wang. Integrated and enhanced pipeline system to support spoken language analytics for screening neurocognitive disorders. In *Interspeech 2023*, pp. 1713–1717, 2023. doi: 10.21437/Interspeech.2023-2249.

Chanwoo Park, Anna Seo Gyeong Choi, Sunghye Cho, and Chanwoo Kim. Reasoning-based approach with chain-of-thought for alzheimer's detection using speech and large language models. In *Interspeech 2025*, pp. 2185–2189, 2025. doi: 10.21437/Interspeech.2025-1226.

Md Sahidullah, Hye-jin Shim, Rosa Gonzalez Hautamäki, and Tomi H. Kinnunen. Shortcut learning in binary classifier black boxes: Applications to voice anti-spoofing and biometrics. *IEEE Journal of Selected Topics in Signal Processing*, 2025. doi: 10.1109/JSTSP.2025.3569430.

Milton E Strauss and Gregory T Smith. Construct validity: Advances in theory and methodology. *Annual review of clinical psychology*, 5:1–25, 2009. doi: 10.1146/annurev.clinpsy.032408.153639.

Zafi Sherhan Syed, Muhammad Shehram Shah Syed, Margaret Lech, and Elena Pirogova. Automated recognition of alzheimer's dementia using bag-of-deep-features and model ensembling. *IEEE Access*, 2021. doi: 10.1109/ACCESS.2021.3090321.

Oskar Van der Wal, Dominik Bachmann, Alina Leidinger, Leendert Van Maanen, Willem Zuidema, and Katrin Schulz. Undesirable biases in nlp: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*, 79:1–40, 2024. doi: 10.1613/jair.1.15195.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ehfRiF0R3a.

Yi Wang, Tianzi Wang, Zi Ye, Lingwei Meng, Shoukang Hu, Xixin Wu, Xunying Liu, and Helen Meng. Exploring linguistic feature and model combination for speech recognition based automatic ad detection. In *Interspeech 2022*, pp. 3328–3332, 2022. doi: 10.21437/Interspeech.2022-723.

Jochen Weiner, Claudia Frankenberg, Johannes Schröder, and Tanja Schultz. Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 674–681, 2019. doi: 10.1109/ASRU46091.2019.9003908.

Adrian Wong, Lorraine SN Law, Wenyan Liu, Zhaolu Wang, Eugene SK Lo, Alexander Lau, Lawrence KS Wong, and Vincent CT Mok. Montreal cognitive assessment: one cutoff never fits all. *Stroke*, 46(12):3547–3550, 2015. doi: 10.1161/STROKEAHA.115.011226.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik R Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://arxiv.org/abs/2405.15793`.

PY Yeung, LL Wong, CC Chan, Jess LM Leung, CY Yung, et al. A validation study of the hong kong version of montreal cognitive assessment (hk-moca) in chinese older adults in hong kong. *Hong Kong Medical Journal*, 2014. doi: 10.12809/hkmj144219.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Interspeech 2020*, pp. 2162–2166, 2020. doi: 10.21437/Interspeech.2020-2516.

Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A Batsis, and Robert M Roth. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. In *Interspeech 2021*, pp. 3790–3794, 2021. doi: 10.21437/Interspeech.2021-332.

A Detailed Description of Cognitive Assessments

A.1 Montreal Cognitive Assessment (MoCA)

Table 4 presents a summary of the Montreal Cognitive Assessment (MoCA), a widely used screening instrument designed to assess mild cognitive dysfunction by probing multiple cognitive domains, including attention, executive functions, memory, language, and conceptual thinking. In this work, we utilized a specific subset termed MoCA-SL (Spoken Language), which is highlighted with an asterisk (*) in the table. These tasks—including Picture Naming, Digit Span, Serial 7 Subtraction, Sentence Repetition, Verbal Fluency, and Abstraction—are based entirely on verbal interaction. Unlike visual-motor tasks (e.g., trail making), these spoken tasks could be readily recorded and processed by an automated pipeline.

Table 4: Detailed description and scoring criteria for the Montreal Cognitive Assessment (MoCA, Hong Kong version). Tasks marked with * belong to the spoken-language subset (MoCA-SL) used in this study

Task	Description	Cognitive Domain (DSM-5)	Scoring Primitives	Score
Alternating Trail Making	Draw a line to connect numbers and characters in alternating order	Exec. Function	Successful connection pattern	1
Copy Cube	Copy a 3-dimensional cube structure	Perceptual Motor	Correct 3D structure	1
Clock Drawing	Draw a clock face, numbers, and set time to 11:10	Perceptual Motor	Contour (1), Numbers (1), Hands (1)	3
*Picture Naming	Name three depicted animals (e.g., Lion, Rhino, Camel)	Language	Per-item correctness ($\times 3$)	3
*Digit Span	Repeat digits forward and backward	Complex Attention	Forward (1), Backward (1)	2
Vigilance (Tap)	Tap hand at each target digit (e.g., '1') in a sequence	Complex Attention	≤ 1 error allowed	1
*Serial 7 Subtraction	Subtract 7 from 100 consecutively	Complex Attention	4-5 correct (3), 2-3 correct (2), 1 correct (1)	3
*Sentence Repetition	Repeat two sentences verbatim	Language	Per-sentence correctness ($\times 2$)	2
*Animal Fluency	Generate animal words starting with a specific character/category in 1 min	Language, Execution Function	$\geq$ threshold (e.g., 11 words)	1
*Abstraction	Identify similarity between word pairs	Executive Function	Per-pair correctness ($\times 2$)	2
Delayed Recall	Recall 5 words learned earlier (after $\sim$5 min)	Learning and Memory	Per-word correctness ($\times 5$)	5
Orientation	Identify current date, month, year, day, place, and city	Perceptual Motor, Orientation	Per-item correctness ($\times 6$)	6
			MoCA Total:	30

A.2 Hong Kong List Learning

Table 5 presents a summary of the Hong Kong List Learning Test (HKLLT), a standardized neuropsychological assessment that mainly probes verbal learning and episodic memory through the recall of a 16-word list. In this work, we specifically focused on Trial 4 (10-minute delayed recall) and Trial 5 (30-minute delayed recall). These delayed recall measures are regarded as particularly effective for early diagnosis, as clinical research demonstrates that the rate of rapid forgetting over these intervals is the most sensitive discriminator between patients with mild Alzheimer's disease and normal elderly controls.

B Implementation Details

Model & Deployment. We use Qwen3-8B[1] as the backbone for all agents, deployed locally via vLLM with half precision and 0.85 GPU memory utilization. All experiments are conducted on 4 NVIDIA RTX 8000 (48GB).

[1]https://huggingface.co/Qwen/Qwen3-8B

Table 5: Detailed description and scoring criteria for the Hong Kong List Learning Test (HKLLT)

Trial / Task	Description	Cognitive Domain (DSM-5)	Scoring Primitives	Score
Acquisition & Learning (Form 1: Random Condition)				
Trial 1	Listen to a list of 16 words (4 categories) presented orally, followed by immediate free recall.	Immediate Memory	# correctly recalled words	16
Trials 2 & 3	Listen to the same list repeated (same order), followed by free recall each time.	Learning Rate	# correctly recalled words (per trial)	16
Delayed Recall (Retention)				
Trial 4 (10-min Delay)	Unannounced free recall of the 16 words after a 10-minute non-verbal interval.	Short-term Retention	# correctly recalled words	16
Trial 5 (30-min Delay)	Unannounced free recall of the 16 words after a 30-minute interval.	Long-term Retention	# correctly recalled words	16
Recognition				
Recognition	Identify the 16 target words mixed with 16 distractors (foils) presented orally (Yes/No).	Recognition Memory	Hits (True Positives) & False Alarms. *Discrim. Score* = $\frac{\text{Hits}-\text{False Alarms}}{16} \times 100\%$	16

Table 6: Handcraft linguistic features used in this work. "%": "ratio of", "#": "number of", "dur": "duration".

ID	Feature Name	Description
L1	# words	The number of words
L2	% Stop words	# stop words / # words
L3	% Filled pauses	# filler words / # words
L4	% Lexical filler	# lexical filler / # words
L5	% Backchannel	# backchannel words / # words
L6	% Repetition	# repeated words / # words
L7	% Adj	# adjectives / # words
L8	% Adv	# adverbs / # words
L9	% Noun	# nouns / # words
L10	% Pronoun	# pronouns / # words
L11	% Verb	# verb / # words
L12	% Func	# functional words / # words
L13	% CTTR	# types / rooted(2 * # tokens)

Agent Configuration. Key inference parameters include: max sequence length 4096, temperature 0.3 (Examiner) / 0.1 (Verifier), and top_p 0.9.

Baselines. Handcrafted linguistic features are detailed in Table 6. For PLM-based methods, we use BERT[2] and RoBERTa[3] with transcripts translated to Simplified Chinese. We do not finetune these models, as preliminary experiments showed no consistent improvement. Task responses are concatenated with "[<TASK>-<NAME>]" delimiters as input. For LLM-COT, the prompt template is provided in Appendix E.

Classifiers. SVM uses RBF kernel with C=1.0 and gamma=scale. MLP consists of two hidden layers (32, 16) with ReLU activation, trained for 20 epochs with dropout 0.2, batch size 16, and initial learning rate 0.001.

[2]https://huggingface.co/google-bert/bert-base-chinese
[3]https://huggingface.co/hfl/chinese-roberta-wwm-ext

Scoring Primitives. The complete list of scoring primitives used as classifier input is provided in Table 7 .

Table 7: Cognitive scoring primitives used in this work. "#": "number of", "z-score": "standardized score relative to norms".

ID	Feature Name	Description
HKLLT (Hong Kong List Learning Test)		
C1	**hkllt4_zScore**	Trial 4 z-score
C2	**hkllt5_zScore**	Trial 5 z-score
C3	**n_hkllt4_recall**	Trial 4 recall count
C4	**n_hkllt5_recall**	Trial 5 recall count
C5	**hkllt4_n_clustering**	Trial 4 semantic clustering count
C6	**hkllt5_n_clustering**	Trial 5 semantic clustering count
C7	**delta_hkllt_zScore**	Z-score difference between Trial 4 and Trial 5
C8	**delta_hkllt_n_recall**	Recall count difference between Trial 4 and Trial 5
MoCA-SL (Montreal Cognitive Assessment)		
M1	**n_Animal_count**	Animal fluency count
M2	**AnimalFlu_score**	Animal fluency score
M3	**7Subtraction_score**	Serial 7s subtraction score
M4	**Abstraction_q1_score**	Abstraction item 1 score
M5	**Abstraction_q2_score**	Abstraction item 2 score
M6	**Digit_fwd_score**	Digit span forward score
M7	**Digit_bwd_score**	Digit span backward score
M8	**Digit_score**	Digit span total score
M9	**PicNaming_score**	Picture naming score
M10	**Sentence_q1_score**	Sentence repetition item 1 score
M11	**Sentence_q2_score**	Sentence repetition item 2 score
M12	**Sentence_score**	Sentence repetition total score

C Population Norms for MoCA-SL Scores

Since MoCA-SL comprises only a subset of the full MoCA assessment, existing normative tables cannot be directly applied. To enable norm-referenced scoring and percentile calculation, we explored three approaches to estimate population norms for MoCA-SL.

Approach 1: Direct Empirical Estimation. The most straightforward approach involves computing MoCA-SL norms directly from healthy control participants in our corpus. We stratified 536 cognitively normal participants by age and education, then calculated mean and standard deviation for each stratum. While this method is intuitive, it carries the risk that our sample may not adequately represent the broader Hong Kong elderly population, particularly in undersampled demographic cells.

Approach 2: Proportional Rescaling. This method leverages the established MoCA normative table (Wong et al., 2015) by rescaling all values proportionally. Given that MoCA-SL covers 13 out of 30 total points, we apply a linear scaling factor of 13/30 to all norm values in the published table:

$$Z_{\text{norm}} = Y_{\text{norm}} \times \frac{13}{30}$$

where Y_{norm} denotes norm values from the full MoCA table and Z_{norm} represents the estimated MoCA-SL norms. This approach assumes that the subset score distribution maintains proportional relationships with the full score, ensuring full grounding in the clinically validated gold standard.

Approach 3: Regression-Based Estimation. We adopt a data-driven approach by fitting a linear regression model on our corpus to characterize the relationship between full MoCA scores (Y) and MoCA-SL scores (Z):

$$Z = \alpha + \beta \cdot Y$$

After estimating coefficients α and β from healthy control participants, we apply this mapping to the published MoCA normative values to derive corresponding MoCA-SL norms:

$$Z_{\text{norm}} = \alpha + \beta \cdot Y_{\text{norm}}$$

This method combines empirical data with established norms, theoretically providing accurate estimates anchored to the gold standard while accounting for any non-linear ceiling or floor effects in the subset.

Experiments We conduct experiments to compare AD screening performance using different approaches; the results are shown in Table 8. We finally adopt Approach 2 (Proportional Rescaling) as the primary method due to its simplicity and superior performance. The obtained normative table is shown in Table 9.

Table 8: Comparison of AD screening performance using different MoCA-SL normative estimation approaches.

Approach	Accuracy	Precision	Recall	F1
1	0.818	0.813	0.786	0.796
2	0.824	0.814	0.802	0.802
3	0.818	0.813	0.786	0.796

Table 9: Normative Table for MoCA-SL scores using proportional rescaling.

Age	Education	N	Median	IQR	Percentile		
(years)	(years)				16th	7th	2nd
65–69	0–3	64	9.1	1.7	7.4	6.1	3.9
	4–6	82	10.0	2.2	8.2	7.8	5.6
	7–9	74	10.4	2.2	9.1	8.2	6.9
	10–12	82	10.8	1.7	9.5	8.7	7.4
	>12	67	11.7	1.3	10.8	10.0	9.1
70–79	0–3	76	8.2	2.2	6.5	6.1	4.8
	4–6	82	9.5	1.7	7.8	6.5	4.3
	7–9	66	10.0	1.7	8.7	7.8	6.5
	10–12	76	10.4	1.7	9.5	8.2	7.8
	>12	67	10.8	2.2	9.5	8.7	6.9
≥80	0–6	37	7.8	2.6	5.6	5.6	4.3
	>6	21	8.7	2.2	7.4	6.5	5.6

D Task Examination Evaluation Details

Table 10 extends the main evaluation (Table 2) with a comprehensive breakdown of performance indicators, including exact score matching (SMR Exact), tolerance-based matching (SMR ± 1), and granular error metrics (MAE, RMSE) for both final scores and intermediate counts.

E Prompts and System Output

E.1 Prompt for Examiner Agent

All Examiner agents follow a standardized four-component template structure. Each prompt comprises: (1) Task Introduction defining the agent's clinical role and assessment objective, (2) Guidelines specifying processing rules and edge cases, (3) Output Format detailing the required response structure, and (4) Examples providing demonstrations for consistent behavior.

We provide two representative examples of Examiner prompts below. Figure 7 presents the prompt for the MoCA Abstraction task, which requires the agent to extract the participant's responses to two similarity questions and judge whether each response reflects abstract categorical reasoning (e.g., transportation" for train-bicycle) versus concrete associations (e.g., they have wheels"). Figure 8 presents the prompt for the MoCA Animal Naming Fluency task, which instructs the agent to extract all valid animal names from the participant's verbal response, handle Cantonese colloquial expressions, deduplicate semantically equivalent items, and invoke the `list_length()` function to obtain the final count.

E.2 META ANALYST

The Meta Analyst agent transforms verified scoring primitives into interpretable cognitive profile reports. Its instruction prompt comprises two components: (1) **Clinical Protocol Context** (Figure 9), which provides the evaluation framework including normal ranges, impairment thresholds, and clinical significance for each HKLLT and MoCA subtest; and (2) **Output Requirements** (Figure 10, upper panel), which specifies the report format—assessments organized by four cognitive domains (memory, executive function, attention & working memory, language), each containing status judgment, supporting evidence, and clinical interpretation.

Figure 10 presents an example input for a 75-year-old male with 6 years of education, showing HKLLT z-scores in the mild-to-moderate impairment range and mixed MoCA subtest performance. Figure 11 displays the corresponding model output, where the Meta Analyst identifies moderate memory impairment based on delayed recall z-scores (-1.5 to -2.0 range), mild executive and language dysfunction, while correctly recognizing preserved attention and working memory—culminating in a HIGH risk assessment with recommendation for further clinical evaluation.

Figure 12 illustrates another output for a cognitively normal participant. Despite one incorrect abstraction item, the agent appropriately concludes that overall executive function remains within normal range, demonstrating the framework's ability to contextualize isolated errors within the broader performance pattern rather than over-interpreting individual task failures.

E.3 PROMPT FOR LLM-COT BASELINE

The LLM-CoT baseline shares the same Clinical Protocol Context as the Meta Analyst (Figure 9). Figure 13 presents the output requirements, which specify a JSON format containing cognitive assessments per domain, three-step chain-of-thought reasoning, and a final diagnosis with confidence score. Unlike our framework, this baseline directly processes raw transcripts without intermediate scoring primitive extraction or verification.

F CASE STUDY

We analyze representative error cases to characterize the failure modes of our framework, particularly where the verification loop proves insufficient.

Hallucination (Figure 14) In the Serial 7 Subtraction task, the Examiner fabricates numbers not present in the transcript (e.g., 76, 69, 62). While the Verifier successfully detects these hallucinations and triggers re-generation, the Examiner responds by simply removing the flagged numbers rather than re-examining the transcript for valid responses. This reveals a limitation: the verification loop targets hallucination but cannot recover *missing* valid answers that the Examiner failed to extract initially.

Boundary Case (Figure 15) In the Abstraction task, the subject responds "use wheels to move" for the train-bicycle similarity question. The Examiner judges this as correct, but the Verifier overrides it as incorrect, reasoning that the response describes a concrete feature rather than an abstract category (transportation). However, this judgment conflicts with the human assessor's scoring, highlighting that boundary cases involving nuanced semantic distinctions remain challenging—even with verification, and verifier judgments may not always align with human assessors on ambiguous cases.

Prompt Example: MoCA Abstraction (Bilingual Contrast)

- Original Prompt -

[Instruction]:

任務介紹
你是一位醫學助理，評估MoCA抽象概唸任務。任務是從轉錄中提取受試者對兩個問題的回答（Q1:火車-單車相似處, Q2:手表-直尺相似處），並判斷是否正確。

注意事項
- Q1和Q2之間有<|question-change|> 分隔
- Q1正確答案：交通工具、運輸工具、代步工具等（錯誤：有輪子、會動、沒區別、唔相似）
- Q2正確答案：度量工具、測量工具、計量器具等（錯誤：有數字、可以看時間、不知道、諗唔到）
- 轉錄粵語口語

格式要求
Q1回答: <提取概唸> Q1判斷: <true/false>
Q2回答: <提取概唸> Q2判斷: <true/false>
{ "Q1": {"response": ["..."], "is_correct": <true/false>}, "Q2": {"response": ["..."], "is_correct": <true/false>} }

範例**1**
轉錄: "單車同火車, 都繫交通 工具, <|question-change|>, 手表, 間尺, 大家都繫, 計數目..."
Q1回答: 交通工具 Q1判斷: true
Q2回答: 計數目 Q2判斷: false
{ "Q1": {"response": ["交 通 工 具"], "is_correct": true}, "Q2": {"response": ["計數目"], "is_correct": false} }

[User Input]:
轉錄
請按格式要求輸出Q1、Q2回答和判斷，並輸出JSON結果。
轉錄: "誒交通工具囉, 運輸工具囉, <|question-change|>, 都繫量度一 囉, 一個時間一個長度"

[Model Output (Intermediate)]:
Q1回答: 交通工具 Q1判斷: true
Q2回答: 量度一 Q2判斷: false
{ "Q1": {"response": ["交 通 工 具"], "is_correct": true}, "Q2": {"response": ["量度一"], "is_correct": false} }

[Verifier]:
Q2判斷錯誤（假陰性）：回答"量度一" 包含度量概唸，應該判正確
- 模型判斷: is_correct = false
- 正確判斷: is_correct = true
修正建議：將Q2.is_correct 改成true

[Final Model Output]:
{ "Q1": {"response": ["交 通 工 具"], "is_correct": true}, "Q2": {"response": ["量度一"], "is_correct": true} }

[Verifier]:
所有回答都在轉錄中找到且判斷正確:
- Q1回答"交通工具" (is_correct=True)
- Q2回答"量度一" (is_correct=True)

- Translated English Prompt -

[Instruction]:

Task Description
You are a medical assistant evaluating the MoCA Abstraction task. Extract the subject"s answers to two questions (Q1: Train-Bicycle similarity, Q2: Watch-Ruler similarity) from the transcript and determine correctness.

Notes
- Q1 and Q2 are separated by <|question-change|>.
- Q1 Correct: Transportation, Vehicle, Commute tools, etc. (Wrong: have wheels, move, no difference, not similar).
- Q2 Correct: Measuring tools, Measurement instruments, Gauges, etc. (Wrong: have numbers, tells time, don"t know, can"t think of it).
- Transcript is in Cantonese colloquial.

Format Requirements
Q1 Resp: <Extracted Concept> Q1 Judge: <true/false>
Q2 Resp: <Extracted Concept> Q2 Judge: <true/false>
{ "Q1": {"response": ["..."], "is_correct": <true/false>}, "Q2": {"response": ["..."], "is_correct": <true/false>} }

Example 1
Transcript: "Bicycle and train, both are transport tools, <|question-change|>, watch, ruler, everyone is, counting numbers..."
Q1 Resp: Transport tools Q1 Judge: true
Q2 Resp: Counting numbers Q2 Judge: false
{ "Q1": {"response": ["Transport tools"], "is_correct": true}, "Q2": {"response": ["Counting numbers"], "is_correct": false} }

[User Input]:
Transcript
Please output Q1/Q2 response and judgment in JSON format.
Transcript: "Eh transport tools lo, transportation tools lo, <|question-change|>, both measure things lo, one time one length"

[Model Output (Intermediate)]:
Q1 Resp: transport tools Q1 Judge: true
Q2 Resp: measure things Q2 Judge: false
{ "Q1": {"response": ["transport tools"], "is_correct": true}, "Q2": {"response": ["measure things"], "is_correct": false} }

[Verifier]:
Q2 Judgment Error (False Negative): Response "measure things" contains measurement concept, should be correct
- Model: is_correct = false
- Ground Truth: is_correct = true
Correction: Change Q2.is_correct to true.

[Final Model Output]:
{ "Q1": {"response": ["transport tools"], "is_correct": true}, "Q2": {"response": ["measure things"], "is_correct": true} }

[Verifier]:
All responses found in transcript and correctly judged:
- Q1 response "transport tools" (is_correct=True)
- Q2 response "measure things" (is_correct=True)

Figure 7: An example of the Examiner's prompt with a verifier step for MoCA Abstraction task. The left column shows the original Cantonese prompt, and the right column shows the English translation.

Prompt Example: MoCA Animal Naming Fluency (Bilingual Contrast)

- Original Prompt -

[Instruction]:

任務介紹
你是一位醫學助理，評估MoCA動物命名流暢性任務。任務是從轉錄中提取受試者出的所有動物名稱，然後調用list_length()函數計數。

注意事項
- 提取所有動物名稱，按順序記錄
- 重複的動物隻計算一次
- 轉錄粵語口語，需理解粵語(如：馬騮/馬餾=猴子)

格式要求
轉錄: <原文片段>
動物列表: [動物1, 動物2, 動物3, ...]
```
<tool_call>
{"name": "list_length", "arguments":
{"list": ["動物1", "動物2", ...]}}
</tool_call>
```

範例
轉錄: "嗯, 獅子, 老虎, 豹, 駱駝, 大笨象, 熊, 熊同埋狐狸, 狼, 猴子, 誒野貓, 恐龍, 嗯"
動物列表: [獅子, 老虎, 豹, 駱駝, 大笨象, 熊, 狐狸, 狼, 猴子, 野貓, 恐龍]
```
<tool_call>
{"name": "list_length", "arguments":
{"list": ["獅 子", "老 虎", "豹", "駱 駝",
"大笨象", "熊", "狐狸", "狼", "猴子", "野貓",
"恐龍"]}}
</tool_call>
```

[User Input]:
轉錄
請按格式要求輸出動物列表並調用函數。
轉錄: "哦, 越多越好, 誒即繫即繫水上又得, 即繫繫咪, 哦嗷得啦得啦, 鼠老鼠牛誒牛貓狗, 誒羊, 駱, 我繫飛誒飛蛾, 長嗯長頸鹿, 大笨象, 啊馬, 羊山羊, 天誒隻誒, 有誒水水, 隻叫龜啊水龜, 龜同埋, 誒, 啊"

[Model Output]:
動物列表: [老鼠, 牛, 貓, 狗, 羊, 駱駝, 飛蛾, 長頸鹿, 大笨象, 馬, 山羊, 龜]
```
<tool_call>
{"name": "list_length", "arguments":
{"list": ["老 鼠", "牛", "貓", "狗", "羊",
"熊", "駱 駝", "飛 蛾", "長 頸 鹿", "大 笨 象",
"馬", "山羊", "龜"]}}
</tool_call>
```

[Verifier]:
Pass, 所有12 個動物都在轉錄中找到。

- Translated English Prompt -

[Instruction]:

Task Description
You are a medical assistant evaluating the MoCA Animal Naming Fluency task. Extract all animal names mentioned by the subject from the transcript, then call the list_length() function to count them.

Notes
- Extract all animal names in order
- Repeated animals are counted only once
- Transcript is in Cantonese colloquial, need to understand Cantonese (e.g., 馬騮/馬餾= monkey)

Format Requirements
Transcript: <original text>
Animal list: [Animal1, Animal2, Animal3, ...]
```
<tool_call>
{"name": "list_length", "arguments":
{"list": ["Animal1", "Animal2", ...]}}
</tool_call>
```

Example
Transcript: `"Um, lion, tiger, leopard, camel, elephant, bear, bear and fox, wolf, monkey, eh wild cat, dinosaur, um"`
Animal list: [Lion, Tiger, Leopard, Camel, Elephant, Bear, Fox, Wolf, Monkey, Wild cat, Dinosaur]
```
<tool_call>
{"name": "list_length", "arguments":
{"list": ["Lion", "Tiger", "Leopard",
"Camel", "Elephant", "Bear", "Fox",
"Wolf", "Monkey", "Wild cat",
"Dinosaur"]}}
</tool_call>
```

[User Input]:
Transcript
Please output the animal list and call the function as required.
Transcript: `"Oh, the more the better, eh that is on water also works, that is right, oh then ok ok, mouse rat cow eh cow cat dog, eh sheep, cam, I am fly eh moth, long um giraffe, elephant, ah horse, sheep goat, sky eh that one eh, have eh water water, that one called turtle ah water turtle, turtle and, eh, ah"`

[Model Output]:
Animal list: [Rat, Cow, Cat, Dog, Sheep, Camel, Moth, Giraffe, Elephant, Horse, Goat, Turtle]
```
<tool_call>
{"name": "list_length", "arguments":
{"list": ["Rat", "Cow", "Cat", "Dog",
"Sheep", "Bear", "Camel", "Moth",
"Giraffe", "Elephant", "Horse", "Goat",
"Turtle"]}}
</tool_call>
```

[Verifier]:
Pass, all 12 animals found in transcript.

Figure 8: An example of the Examiner's prompt with a verifier step for MoCA Animal Fluency task. The left column shows the original Cantonese prompt, and the right column shows the English translation.

Clinical Protocol Context (Bilingual Contrast)

- Original Version -

評估框架

核心評估指標及臨床意義

1. HKLLT記憶測試（**Hong Kong List Learning Test**）
正常範圍與損害分級：
- 正常範圍: z-score > -1.0
- 輕度損害: z-score -1.0 至-1.5
- 中度損害: z-score -1.5 至-2.0
- 重度損害: z-score < -2.0

臨床意義：
情景記憶是阿爾茨海默症最早受損的認知域。HKLLT z-score是核心診斷指標。
- Trial 4（10分鐘延遲回憶）: 評估短期記憶鞏固
- Trial 5（30分鐘延遲回憶）: 評估長期記憶保留
- 語義聚類次數: 反映記憶組織策略的有效性

2. MoCA認知評估子項

2.1 動物命名測試（**Semantic Fluency**）
- 正常: e11個動物名稱/分鐘
- 異常: <11個
- 臨床意義: 評估語義流暢性和執行功能，對額葉-顳葉功能敏感

2.2 連續減法（**Serial 7s**）
- 滿分: 3分（4-5個正確）
- 輕度損害: 2分（2-3個正確）
- 重度損害: 0-1分
- 臨床意義: 評估注意力、工作記憶和計算能力

2.3 命名測試（**Visual Naming**）
- 滿分: 3分（獅子、犀牛、駱駝）
- 臨床意義: 評估視覺命名能力，對語義記憶敏感

2.4 句子重複（**Sentence Repetition**）
- 滿分: 2分
- 臨床意義: 評估語言功能和工作記憶

2.5 數字廣度（**Digit Span**）
- 滿分: 2分（順向+逆向）
- 臨床意義: 評估注意力和工作記憶容量

2.6 抽象思維（**Abstraction**）
- 滿分: 2分
- 臨床意義: 評估執行功能和概唸推理能力

- Translated English Version -

 # Evaluation Framework

Core Assessment Indicators and Clinical Significance

1. HKLLT Memory Test (Hong Kong List Learning Test)

Normal Range and Impairment Levels:
- Normal range: z-score > -1.0
- Mild impairment: z-score -1.0 to -1.5
- Moderate impairment: z-score -1.5 to -2.0
- Severe impairment: z-score > -2.0

Clinical Significance:
Episodic memory is the earliest impaired cognitive domain in Alzheimer"s disease. HKLLT z-score is a core diagnostic indicator.
- Trial 4 (10-min delayed recall): Assesses short-term memory consolidation
- Trial 5 (30-min delayed recall): Assesses long-term memory retention
- Semantic clustering count: Reflects effectiveness of memory organization strategies

2. MoCA Cognitive Assessment Subtests

2.1 Animal Naming Test (Semantic Fluency)
- Normal: 11 animal names/minute
- Abnormal: ¡11
- Clinical significance: Assesses semantic fluency and executive function, sensitive to frontal-temporal lobe function

2.2 Serial Subtraction (Serial 7s)
- Full score: 3 points (4-5 correct)
- Mild impairment: 2 points (2-3 correct)
- Severe impairment: 0-1 points
- Clinical significance: Assesses attention, working memory, and calculation ability

2.3 Naming Test (Visual Naming)
- Full score: 3 points (Lion, Rhino, Camel)
- Clinical significance: Assesses visual naming ability, sensitive to semantic memory

2.4 Sentence Repetition
- Full score: 2 points
- Clinical significance: Assesses language function and working memory

2.5 Digit Span
- Full score: 2 points (Forward + Backward)
- Clinical significance: Assesses attention and working memory capacity

2.6 Abstraction
- Full score: 2 points
- Clinical significance: Assesses executive function and conceptual reasoning ability

Figure 9: Clinical Protocol Context is a section in Meta Analyst's instruction prompt, showing core assessment indicators and clinical significance. The left column shows the original Chinese version, and the right column shows the English translation.

Prompt Example: Cognitive Profile Report Generation (Bilingual Contrast)

- Original Prompt -
[Instruction]:

 # Clinical Protocol Context:
(..省略..)

輸出要求
請以自然語言陳述的形式，該受試者撰一份詳細的認知功能概況報告。報告應包含以下四個認知域的評估，每個域需包含：狀態判斷、支持証據、臨床解釋。

報告格式範例：
【認知功能概況報告】

1. 記憶功能（**Memory Function**）
狀態：[正常/輕度損害/中度損害/重度損害]
証據：例如，該受試者在HKLLT-4（10分鐘延遲回憶）中獲得z-scoreX，回憶了Y個詞匯，語義聚類Z次；在HKLLT-5（30分鐘延遲回憶）中獲得z-scoreX，回憶了Y個詞匯。
解釋：[根據z-score閾值和表現模式，明記憶功能的損害程度及其臨床意義...]

2. 執行功能（**Executive Function**）
狀態：[正常/損害]
証據：在動物命名測試中，該受試者在1分鐘內出了X個動物名稱（標準：11個正常）；在抽象思維測試中，Q1回答[正確/錯誤]，Q2回答[正確/錯誤]。
解釋：[明語義流暢性和概唸推理能力的表現]

3. 注意力與工作記憶
狀態：[正常/損害]
証據：在連續減法（Serial 7s）中獲得X/3分（Y/5個正確）；在數字廣度測試中，順向[通過/未通過]，逆向[通過/未通過]，總分X/2分。
解釋：[明注意力持續性和工作記憶容量的表現]

4. 語言功能（**Language Function**）
狀態：[正常/損害]
証據：在命名測試中獲得X/3分；在句子重複測試中獲得X/2分。
解釋：[明視覺命名和語言複述能力的表現]

綜合評述：
綜合以上各認知域的表現，該受試者的認知功能整體呈現[描述整體模式]...結合風險等級[LOW/MODERATE/HIGH/VERY_HIGH]，[提供整體臨床印象]。

[User Input]:
當前案例

基本信息
- 年齡: 75歲
- 性別: 男
- 教育年限: 6.0年

評估結果

HKLLT指標
- HKLLT-4 z-score: -0.71 (回憶4個詞，語義聚類2次)
- HKLLT-5 z-score: -0.83 (回憶3個詞，語義聚類1次)

MoCA認知評估子項
- 動物命名: 14個(通過標準)
- 連續減法: 2/3分(2/5個正確)
- 數字廣度: 1/2分(順向pass，逆向fail)
- 命名測試: 3/3分
- 句子重複: 2/2分
- 抽象思維: Q1 pass，Q2 fail

[Model Output]:
(..省略..)

- Translated English Prompt -
[Instruction]:

 # Clinical Protocol Context:
(..omitted..)

 # Output Requirements
Please write a detailed cognitive profile report for the subject in **natural language narrative** form. The report should include assessments of the following four cognitive domains, each containing: status judgment, supporting evidence, and clinical interpretation.

Report Format Example:
【`Cognitive Profile Report`】

1. Memory Function
Status: [Normal/Mild impairment/Moderate impairment/Severe impairment]
Evidence: For example, the subject obtained a z-score of X in HKLLT-4 (10-minute delayed recall), recalled Y words, with Z semantic clusters; obtained a z-score of X in HKLLT-5 (30-minute delayed recall), recalled Y words.
Interpretation: [Based on z-score thresholds and performance patterns, explain the degree of memory impairment and clinical significance...]

2. Executive Function
Status: [Normal/Impaired]
Evidence: In animal naming test, the subject named X animals within 1 minute (criterion: 11 for normal); in abstraction test, Q1 answered [correctly/incorrectly], Q2 answered [correctly/incorrectly].
Interpretation: [Explain semantic fluency and conceptual reasoning performance]

3. Attention & Working Memory
Status: [Normal/Impaired]
Evidence: In Serial 7s, scored X/3 points (Y/5 correct); in digit span test, forward [passed/failed], backward [passed/failed], total score X/2.
Interpretation: [Explain sustained attention and working memory capacity performance]

4. Language Function
Status: [Normal/Impaired]
Evidence: Scored X/3 in naming test; scored X/2 in sentence repetition test.
Interpretation: [Explain visual naming and verbal repetition abilities]

Overall Summary:
Integrating performance across all cognitive domains, the subject"s overall cognitive function presents [describe overall pattern]...Combined with risk level [LOW/MODERATE/HIGH/VERY_HIGH], [provide overall clinical impression].

[User Input]:
 # Current Case

Basic Information
- Age: 75 years
- Gender: Male
- Education: 6.0 years

Assessment Results

HKLLT Metrics
- HKLLT-4 z-score: -0.71 (recalled 4 words, 2 semantic clusters)
- HKLLT-5 z-score: -0.83 (recalled 3 words, 1 semantic cluster)

MoCA Cognitive Sub-items
- **Animal naming**: 14 (passed criterion)
- **Serial 7s**: 2/3 points (2/5 correct)
- **Digit span**: 1/2 points (forward pass, backward fail)
- **Naming test**: 3/3 points
- **Sentence repetition**: 2/2 points
- **Abstraction**: Q1 pass, Q2 fail

[Model Output]:
(..omitted..)

Figure 10: An example of the prompt for generating cognitive profile reports. The left column shows the original Cantonese prompt, and the right column shows the English translation.

Model Output Example: Cognitive Function Summary Report (Bilingual Contrast)

- Original Output -

【認知功能概況報告】

1. 記憶功能（**Memory Function**）
狀態：中度損害
証據：該受試者在HKLLT-4（10分鐘延遲回憶）中獲得z-score-1.65，回憶了2個詞匯，語義聚類1次；在HKLLT-5（30分鐘延遲回憶）中獲得z-score-1.7，回憶了1個詞匯，語義聚類1次。
解釋：該受試者在延遲回憶測試中表現明顯下降，z-score處於中度損害範圍（-1.5至-2.0）。回憶詞匯數量減少且語義聚類次數有限，提示情景記憶的鞏固和長期保留能力受損，符合阿爾茨海默症早期記憶功能損害的典型表現。

2. 執行功能（**Executive Function**）
狀態：輕度損害
証據：在動物命名測試中，該受試者在1分鐘內出了16個動物名稱（通過標準）；在抽象思維測試中，Q1回答錯誤，Q2回答錯誤。
解釋：動物命名測試表現正常，表明語義流暢性未受明顯影響。然而，抽象思維測試中兩題均錯誤，提示概唸推理和執行功能存在輕度損害，可能與額葉-顳葉功能異常有關。

3. 注意力與工作記憶（**Attention & Working Memory**）
狀態：正常
証據：在連續減法（Serial 7s）中獲得3/3分（4/5個正確）；在數字廣度測試中，順向通過，逆向通過，總分2/2分。
解釋：連續減法測試和數字廣度測試均表現良好，表明注意力持續性和工作記憶容量未受明顯影響，提示該受試者在執行任務時的注意力和信息處理能力保持正常。

4. 語言功能（**Language Function**）
狀態：輕度損害
証據：在命名測試中獲得1/3分；在句子重複測試中獲得2/2分。
解釋：命名測試中僅能正確命名1個詞匯，提示視覺命名能力受損，可能與語義記憶或語言處理中樞功能異常有關。然而，句子重複測試表現正常，明語言複述能力未受明顯影響。

綜合評述：
綜合以上各認知域的表現，該受試者的認知功能整體呈現記憶功能中度損害、執行功能輕度損害、語言功能輕度損害，而注意力與工作記憶保持正常。結合風險等級HIGH，提示存在較高的阿爾茨海默症可能性，需進一步進行影像學和臨床評估以明確診斷。

- Translated English Output -

[Cognitive Function Summary Report]

1. Memory Function
Status: Moderate impairment
Evidence: The subject obtained a z-score of -1.65 in HKLLT-4 (10-minute delayed recall), recalling 2 words with 1 semantic clustering; in HKLLT-5 (30-minute delayed recall), obtained a z-score of -1.7, recalling 1 word with 1 semantic clustering.
Interpretation: The subject showed significant decline in delayed recall tests, with z-scores in the moderate impairment range (-1.5 to -2.0). The reduced number of recalled words and limited semantic clustering suggest impaired consolidation and long-term retention of episodic memory, consistent with typical early-stage memory dysfunction in Alzheimer"s disease.

2. Executive Function
Status: Mild impairment
Evidence: In the animal naming test, the subject named 16 animals within 1 minute (passed criterion); in the abstraction test, Q1 was incorrect, Q2 was incorrect.
Interpretation: Normal performance in animal naming test indicates semantic fluency is not significantly affected. However, both questions in the abstraction test were incorrect, suggesting mild impairment in conceptual reasoning and executive function, possibly related to frontal-temporal lobe dysfunction.

3. Attention & Working Memory
Status: Normal
Evidence: Scored 3/3 in Serial 7s (4/5 correct); in digit span test, passed forward span, passed backward span, total score 2/2.
Interpretation: Good performance in both Serial 7s and digit span tests indicates sustained attention and working memory capacity are not significantly affected, suggesting the subject maintains normal attention and information processing abilities during task execution.

4. Language Function
Status: Mild impairment
Evidence: Scored 1/3 in naming test; scored 2/2 in sentence repetition test.
Interpretation: Only correctly naming 1 item in the naming test suggests impaired visual naming ability, possibly related to semantic memory or language processing center dysfunction. However, normal performance in sentence repetition indicates language repetition ability is not significantly affected.

Comprehensive Summary:
Integrating performance across cognitive domains, the subject"s overall cognitive function shows moderate memory impairment, mild executive function impairment, mild language impairment, while attention and working memory remain normal. Combined with HIGH risk level, this suggests a high probability of Alzheimer"s disease, requiring further imaging and clinical evaluation for definitive diagnosis.

Figure 11: An example of the model-generated cognitive function summary report. The left column shows the original Chinese output, and the right column shows the English translation.

Model Output Example: Cognitive Function Summary Report - Normal Case (Bilingual Contrast)

- Original Output -

【認知功能概況報告】

1. 記憶功能（**Memory Function**）
狀態：正常
証據：該受試者在HKLLT-4（10分鐘延遲回憶）中獲得z-score-0.18，回憶了7個詞匯，語義聚類2次；在HKLLT-5（30分鐘延遲回憶）中獲得z-score0.08，回憶了7個詞匯，語義聚類4次。
解釋：該受試者的記憶功能處於正常範圍，z-score均高於-1.0。雖然語義聚類次數在HKLLT-4中較低，但整體回憶表現穩定，未出現明顯的記憶鞏固或長期記憶保留障礙，不符合阿爾茨海默症早期快速遺忘的典型模式。

2. 執行功能（**Executive Function**）
狀態：正常
証據：在動物命名測試中，該受試者在1分鐘内出了24個動物名稱（標準：11個正常）；在抽象思維測試中，Q1回答正確，Q2回答錯誤。
解釋：該受試者在語義流暢性方麵表現優異，遠超正常閾值，提示其語義記憶和執行功能未受明顯影響。雖然抽象思維測試中Q2回答錯誤，但整體執行功能仍處於正常範圍，未提示額葉-顳葉功能受損。

3. 注意力與工作記憶（**Attention & Working Memory**）
狀態：正常
証據：在連續減法（Serial 7s）中獲得3/3分（5/5個正確）；在數字廣度測試中，順向通過，逆向通過，總分2/2分。
解釋：該受試者在連續減法和數字廣度測試中均表現優異，提示其注意力集中能力良好，工作記憶容量充足，未髮現注意力或工作記憶方麵的損害。

4. 語言功能（**Language Function**）
狀態：正常
証據：在命名測試中獲得3/3分；在句子重複測試中獲得2/2分。
解釋：該受試者在視覺命名和句子重複測試中均表現正常，提示其語言功能完整，無明顯的語言障礙或複述能力下降。

綜合評述：
綜合以上各認知域的表現，該受試者的認知功能整體呈現正常模式，各主要認知域均未髮現明顯損害。結合風險等級MODERATE，提示存在輕度認知變化的可能，但尚未達到阿爾茨海默症的診斷標準。建議進一步觀察和隨訪，排除其他潛在認知障礙因素。

- Translated English Output -

[Cognitive Function Summary Report]

1. Memory Function
Status: Normal
Evidence: The subject obtained a z-score of -0.18 in HKLLT-4 (10-minute delayed recall), recalling 7 words with 2 semantic clusterings; in HKLLT-5 (30-minute delayed recall), obtained a z-score of 0.08, recalling 7 words with 4 semantic clusterings.
Interpretation: The subject"s memory function is within normal range, with z-scores above -1.0. Although semantic clustering frequency in HKLLT-4 was relatively low, overall recall performance remained stable, showing no significant memory consolidation or long-term retention deficits, inconsistent with the typical rapid forgetting pattern of early Alzheimer"s disease.

2. Executive Function
Status: Normal
Evidence: In the animal naming test, the subject named 24 animals within 1 minute (criterion: 11 for normal); in the abstraction test, Q1 was correct, Q2 was incorrect.
Interpretation: The subject demonstrated excellent performance in semantic fluency, far exceeding the normal threshold, suggesting semantic memory and executive function are not significantly affected. Although Q2 in the abstraction test was incorrect, overall executive function remains within normal range, with no indication of frontal-temporal lobe dysfunction.

3. Attention & Working Memory
Status: Normal
Evidence: Scored 3/3 in Serial 7s (5/5 correct); in digit span test, passed forward span, passed backward span, total score 2/2.
Interpretation: The subject performed excellently in both Serial 7s and digit span tests, indicating good attention concentration and sufficient working memory capacity, with no evidence of attention or working memory impairment.

4. Language Function
Status: Normal
Evidence: Scored 3/3 in naming test; scored 2/2 in sentence repetition test.
Interpretation: The subject performed normally in both visual naming and sentence repetition tests, indicating intact language function with no significant language impairment or decline in repetition ability.

Comprehensive Summary:
Integrating performance across cognitive domains, the subject"s overall cognitive function presents a normal pattern, with no significant impairment found in any major cognitive domain. Combined with MODERATE risk level, this suggests possible mild cognitive changes, but has not yet met the diagnostic criteria for Alzheimer"s disease. Further observation and follow-up are recommended to rule out other potential cognitive impairment factors.

Figure 12: An example of the model-generated cognitive function summary report for a normal case. The left column shows the original Chinese output, and the right column shows the English translation.

Prompt Example: LLM-CoT Baseline (Bilingual Contrast)

- Original Chinese Prompt -

[Instruction]:

臨床協議背景

...

輸出要求

請嚴格按照以下JSON格式輸出診斷結果：

```
{
"cognitive_assessment": {
"memory_function": {"status": "正常/輕度損害/中度損害/重度損害",
"evidence": ["HKLLT-4表 現: ...", "HKLLT-5表 現: ..."],
"interpretation": "簡要解釋"},
"executive_function": {"status": "正 常/損 害", ..."},
"attention_working_memory": {"status": "正 常/損害", ..."},
"language_function": {"status": "正 常/損 害", ..."}
},
"risk_analysis": {"main_findings": [...],
"risk_factors": [...], "protective_factors": [...]},
"step_by_step_reasoning": {
"step_1": {"focus": "記 憶 功 能 分 析", "observation": "...", "conclusion": "..."},
"step_2": {"focus": "多認知域綜合分析", ..."},
"step_3": {"focus": "人口統計學因素與風險整合", ..."}
},
"diagnosis": {"reasoning": "...", "confidence": 0.0-1.0, "is_alzheimer_likely": true/false}
}
```

注意事項：

1. step_by_step_reasoning必須按照cognitive_assessment結果逐步推理要明確引用前麵3個step
3. 保持客觀中立，隻有當多個認知域損害時才判斷AD可能

[User Input]:

當前案例

受試者信息

ID: TBS0134, 年齡: 70歲, 性別: 女, 教育年限: 0.0年

認知測試轉錄

HKLLT-4 (10分鐘延遲**):** 完全唔記得囉，繫啊，完全唔記得，一個都記唔到，
HKLLT-5 (30分鐘延遲**):** 啊，唔記得啦，繫啊，繫啊，
MoCA動物命名**:** 貓狗老虎豹，獅子，大笨象，牛，馬，狗，老虎，...
MoCA連續減法**:** 一百減七啊，九啊七，九啊七啊，八八十三，八十三，再又...
MoCA數字廣度**:** 二一八五四，四，四七二，
MoCA命名**:** 獅子，啊，牛牛，駱駝，
MoCA句子重複**:** 姨丈買魚腸，西施四十四歲，
MoCA抽象思維**:** 大家可以交通用佢囉，即繫可以用可以，代代步啦，繫，...

請基於以上信息進行綜合評估。

[Model Output]:

```
{
"cognitive_assessment": {
"memory_function": {"status": "重度損害",
"evidence": ["HKLLT-4: 完全唔記得...", "HKLLT-5: 唔記得啦..."],
"interpretation": "嚴重記憶障礙，情景記憶功能嚴重受損"},
"executive_function": {"status": "損 害",
"interpretation": "抽象思維混亂"},
"attention_working_memory": {"status": "損 害",
"interpretation": "計算錯誤頻繁"},
"language_function": {"status": "損 害",
"interpretation": "命名和重複錯誤"}
},
"step_by_step_reasoning": {
"step_1": {"focus": "記憶功能分析", "conclusion": "記憶功能呈現重度損害"},
"step_2": {"focus": "多認知域分析", "conclusion": "多認知域損害模式"},
"step_3": {"focus": "風 險 整 合", "conclusion": "存在較高認知下降風險"}
},
"diagnosis": {"reasoning": "(1)記 憶 重 度 損 害，(2)多認知域異常，(3)高風險因素",
"confidence": 0.85, "is_alzheimer_likely": true}
}
```

- Translated English Prompt -

[Instruction]:

Clinical Protocol Context

...

Output Requirements

Please strictly output diagnostic results in the following JSON format:

```
{
"cognitive_assessment": {
"memory_function": {"status": "normal/mild/moderate/severe impairment",
"evidence": ["HKLLT-4 performance: ...", "HKLLT-5 performance: ..."],
"interpretation": "brief explanation"},
"executive_function": {"status": "normal/impaired", ..."},
"attention_working_memory": {"status": "normal/impaired", ..."},
"language_function": {"status": "normal/impaired", ..."}
},
"risk_analysis": {"main_findings": [...],
"risk_factors": [...], "protective_factors": [...]},
"step_by_step_reasoning": {
"step_1": {"focus": "memory function analysis", "observation": "...", "conclusion": "..."},
"step_2": {"focus": "multi-domain cognitive analysis", ..."},
"step_3": {"focus": "demographic factors & risk integration", ..."}
},
"diagnosis": {"reasoning": "...", "confidence": 0.0-1.0, "is_alzheimer_likely": true/false}
}
```

Notes:

1. step_by_step_reasoning must follow cognitive_assessment results step-by-step
2. diagnosis reasoning must cite conclusions from all 3 steps
3. Remain objective; only judge as AD likely when multiple domains are impaired

---

[User Input]:

Current Case

Subject Information

ID: TBS0134, Age: 70, Gender: Female, Education: 0.0 years

Cognitive Test Transcripts

HKLLT-4 (10-min delay): Completely can"t remember, yes, completely forget, can"t recall any...
HKLLT-5 (30-min delay): Ah, don"t remember, yes, yes...
MoCA Animal Naming: Cat dog tiger leopard, lion, elephant, cow, horse, dog, tiger...
MoCA Serial 7s: Hundred minus seven, ninety-seven, ninety-seven, eighty-three...
MoCA Digit Span: Two one eight five four, four, four seven two...
MoCA Naming: Lion, ah, cow cow, camel...
MoCA Sentence Repeat: Uncle bought fish sausage, Xishi forty-four years old...
MoCA Abstraction: Everyone can use for transport, can use, commute...

Please conduct comprehensive assessment based on above.

[Model Output]:

```
{
"cognitive_assessment": {
"memory_function": {"status": "severe impairment",
"evidence": ["HKLLT-4: Completely can''t remember...", "HKLLT-5: Don''t remember..."],
"interpretation": "Severe memory deficit, episodic memory severely impaired"},
"executive_function": {"status": "impaired",
"interpretation": "Abstract thinking confused"},
"attention_working_memory": {"status": "impaired", "interpretation": "Frequent calculation errors"},
"language_function": {"status": "impaired",
"interpretation": "Naming and repetition errors"}
},
"step_by_step_reasoning": {
"step_1": {"focus": "memory analysis",
"conclusion": "Severe memory impairment"},
"step_2": {"focus": "multi-domain analysis",
"conclusion": "Multi-domain damage pattern"},
"step_3": {"focus": "risk integration",
"conclusion": "High cognitive decline risk"}
},
"diagnosis": {"reasoning": "(1)Severe memory damage, (2)Multi-domain abnormal, (3)High risk",
"confidence": 0.85, "is_alzheimer_likely": true}
}
```

Figure 13: An example of the LLM-CoT baselineprompt with step-by-step reasoning for Alzheimer's diagnosis assessment. The left column shows the original Chinese prompt, and the right column shows the English translation.

Error Case 1: Hallucination (Bilingual Contrast)

- Original Prompt -

[Instruction]:

任務介紹
你是一位醫學助理，評估MoCA連續減法任務。任務是從轉錄中提取受試者出的所有有效數字回答。

注意事項
- 任務要求從100開始連續減7
- 隻提取轉錄中明確出現的數字
- 忽略重複、猶豫或修正的數字
- 轉錄粵語口語
- 不要捏造或推測未出現的數字

格式要求
有效回答: [數字1, 數字2, ...]
`<tool_call>`
...
`</tool_call>`

範例
轉錄: "九十三， 八十六， 七十九， 七十二"
有效回答: [93, 86, 79, 72]

[User Input]:
轉錄
請按格式要求輸出有效回答並調用函數。

轉錄: "由一百減七，百一百開始囉喎，一百減七仲有九啊三九啊三減七就仲有，九啊七減，九啊七減，減個七咋喎繫咪啊，一路減減個即繫一路減七咋繫咪啊，一百減七仲有九啊三九啊三減七就仲有，啊八啊八啊四，八啊四減七仲有八啊一，八啊一減七就繫仲有六十，六十四六十四減七就仲有，五十，嗯，五十，五十，五十七，五十七減誒"

[Model Output (1st attempt):]
有效回答: [93, 84, 76, 69, 62]
`<tool_call>`
...
`</tool_call>`

[Verifier:]
數字76 未在轉錄中找到，可能是錯誤識別或捏造。
數字69 未在轉錄中找到，可能是錯誤識別或捏造。
數字62 未在轉錄中找到，可能是錯誤識別或捏造。

這次忽略之前的回答，重新思考。

[Model Output (2nd attempt):]
有效回答: [93, 84, 79, 72, 65]
`<tool_call>`
...
`</tool_call>`

[Verifier:]
數字79 未在轉錄中找到，可能是錯誤識別或捏造。
數字72 未在轉錄中找到，可能是錯誤識別或捏造。
數字65 未在轉錄中找到，可能是錯誤識別或捏造。

這次忽略之前的回答，重新思考。

[Final Model Output:]
有效回答: [93, 84]
`<tool_call>`
...
`</tool_call>`

[Verifier:]
pass, 所有數字都在轉錄中找到。

- Translated English Prompt -

[Instruction]:
Task Description
You are a medical assistant evaluating the MoCA Serial Subtraction task. Extract all valid numerical answers spoken by the subject from the transcript.

Notes
- Task requires subtracting 7 continuously from 100
- Only extract numbers explicitly present in transcript
- Ignore repeated, hesitant, or corrected numbers
- Transcript is in Cantonese colloquial
- Do not fabricate or infer numbers not present

Format Requirements
Valid responses: [number1, number2, ...]
`<tool_call>`
...
`</tool_call>`

Example
Transcript: `"Ninety-three, eighty-six, seventy-nine, seventy-two"`
Valid responses: [93, 86, 79, 72]

[User Input:]
Transcript
Please output valid responses in required format and call function.

Transcript: `"From one hundred minus seven, hundred start from hundred lo, hundred minus seven still have ninety-three, ninety-three minus seven still have, ninety-seven minus, ninety-seven minus, minus seven right, continuously minus means continuously minus seven right, hundred minus seven still have ninety-three, ninety-three minus seven still have, ah eighty ah eighty-four, eighty-four minus seven still have eighty-one, eighty-one minus seven is still have sixty, sixty-four, sixty-four minus seven still have, fifty, um, fifty, fifty, fifty-seven, fifty-seven minus eh"`

[Model Output (1st attempt):]
Valid responses: [93, 84, 76, 69, 62]
`<tool_call>`
...
`</tool_call>`

[Verifier:]
Number 76 not found in transcript, may be misidentified or fabricated.
Number 69 not found in transcript, may be misidentified or fabricated.
Number 62 not found in transcript, may be misidentified or fabricated.

Ignore previous answer and rethink.

[Model Output (2nd attempt):]
Valid responses: [93, 84, 79, 72, 65]
`<tool_call>`
...
`</tool_call>`

[Verifier:]
Number 79 not found in transcript, may be misidentified or fabricated.
Number 72 not found in transcript, may be misidentified or fabricated.
Number 65 not found in transcript, may be misidentified or fabricated.

Ignore previous answer and rethink.

[Final Model Output:]
Valid responses: [93, 84]
`<tool_call>`
...
`</tool_call>`

[Verifier:]
Pass, all numbers found in transcript.

Figure 14: An error case in the MoCA Serial Subtraction task. The examiner generates hallucinated numbers. With the verifier's feedback, the examiner simply removes hallucinated numbers and gives up on extracting other new numbers. The left column shows the original Cantonese prompt, and the right column shows the English translation.

Error Case 2: Boundary Case (Bilingual Contrast)

- Original Prompt -

[Instruction]:
任務介紹
你是一位醫學助理，評估MoCA抽象概唸任務。任務是從轉錄中提取受試者對兩個問題的回答（Q1:火車-單車相似處, Q2:手表-直尺相似處），並判斷是否正確。

注意事項
- Q1和Q2之間有<|question-change|> 分隔
- Q1正確答案：交通工具、運輸工具、代步工具等（錯誤：有輪子、會動、沒區別、唔相似）
- Q2正確答案：度量工具、測量工具、計量器具等（錯誤：有數字、可以看時間、不知道、諗唔到）
- 轉錄粵語口語

格式要求
Q1回答: <提取概唸> Q1判斷: <true/false>
Q2回答: <提取概唸> Q2判斷: <true/false>
{ "Q1": {"response": ["..."], "is_correct": <true/false>}, "Q2": {...} }

範例
轉錄: "單車同火車，都繫交通 工具，<|question-change|>，手表，間尺，大家都繫，計數目..."
Q1回答: 交通工具 Q1判斷: true
Q2回答: 計數目 Q2判斷: false
{ "Q1": {..., "is_correct": true}, "Q2": {..., "is_correct": false} }

[User Input]:
轉錄
請按格式要求輸出Q1、Q2回答和判斷，並輸出JSON結果。

轉錄: "嗯，都繫用轆去行囉，誒，單車同火車，用轆去行就可以繫，啲我唔知啊，<|question-change|>，手表同間尺相似地方邊度，哦都繫長方形囉，手表啊睇時間間尺繫量度，哦都唔繫，量度囉我覺得就繫，其實我真繫唔都唔知有咩分別有咩相同，大家都繫長方形囉"

[Model Output (Intermediate)]:
Q1回答: 用轆去行 Q1判斷: true
Q2回答: 量度 Q2判斷: true
{
"Q1": {"response": ["用 轆 去 行"], "is_correct": true},
"Q2": {"response": ["量 度 "], "is_correct": true}
}

[Verifier]:
Q1判斷錯誤（假陽性）：回答"用轆去行" 不是抽象概唸，應該判錯誤
- 模型判斷: is_correct = true
- 正確判斷: is_correct = false
- 原因: 回答未包含交通工具/代步工具等抽象概唸，可能是具體特徵（如"有輪子"）
修正建議：將Q1.is_correct 改false

[Final Model Output]:
{
"Q1": {"response": ["用 轆 去 行"], "is_correct": false},
"Q2": {"response": ["量 度 "], "is_correct": true}
}

- Translated English Prompt -

[Instruction]:
Task Description
You are a medical assistant evaluating the MoCA Abstraction task. Extract the subject"s answers to two questions (Q1: Train-Bicycle similarity, Q2: Watch-Ruler similarity) from the transcript and determine correctness.

Notes
- Q1 and Q2 are separated by <|question-change|>.
- Q1 Correct: Transportation, Vehicle, Commute tools, etc. (Wrong: have wheels, move, no difference, not similar).
- Q2 Correct: Measuring tools, Measurement instruments, Gauges, etc. (Wrong: have numbers, tells time, don"t know).
- Transcript is in Cantonese colloquial.

Format Requirements
Q1 Resp: <Concept> Q1 Judge: <true/false>
Q2 Resp: <Concept> Q2 Judge: <true/false>
{ "Q1": {"response": ["..."], "is_correct": <true/false>}, "Q2": {...} }

Example
Transcript: "Bicycle and train, both are transport tools, <|question-change|>, watch, ruler, everyone is, counting numbers..."
Q1 Resp: Transport tools Q1 Judge: true
Q2 Resp: Counting numbers Q2 Judge: false
{ "Q1": {..., "is_correct": true}, "Q2": {..., "is_correct": false} }

[User Input]:
Transcript
Please output Q1/Q2 response and judgment in JSON format.

Transcript: "Um, both use wheels to move, eh, bicycle and train, use wheels to move can be, then I don''t know, <|question-change|>, where''s the similarity between watch and ruler, oh both are rectangular, watch ah tells time ruler is for measuring, oh no, things for measuring lo I think is, actually I really don''t know what''s the difference what''s similar, both are rectangular"

[Model Output (Intermediate)]:
Q1 Resp: use wheels to move Q1 Judge: true
Q2 Resp: things for measuring Q2 Judge: true
{
"Q1": {"response": ["use wheels to move"], "is_correct": true},
"Q2": {"response": ["things for measuring"], "is_correct": true}
}

[Verifier]:
Q1 Judgment Error (False Positive): Response "use wheels to move" is not an abstract concept, should be incorrect
- Model: is_correct = true
- Ground Truth: is_correct = false
- Reason: Response doesn"t include abstract concepts like transportation/vehicle, likely a concrete feature (e.g., "have wheels")
Correction: Change Q1.is_correct to false

[Final Model Output]:
{
"Q1": {"response": ["use wheels to move"], "is_correct": false},
"Q2": {"response": ["things for measuring"], "is_correct": true}
}

Figure 15: An error case in the MoCA Abstraction task. In this boundary case, it"s hard to judge if "use wheels to move" is an abstract concept, while the assessor judges it as correct. The left column shows the original Cantonese prompt, and the right column shows the English translation.

Table 10: Full metric breakdown for MoCA-SL and HKLLT tasks. This table reports all available performance indicators, including exact score matching (SMR Exact), tolerance-based matching (SMR ± 1), and granular error metrics (MAE, RMSE) for both final scores and intermediate counts (e.g., recall number, item accuracy).

Task	Metric	Value
MoCA-SL		
Picture Naming	SMR (Exact)	97.1%
	SMR (± 1)	100.0%
	MAE (Score)	0.029
	RMSE (Score)	0.171
Digit Span	SMR (Exact)	98.5%
	SMR (± 1)	100.0%
	MAE (Score)	0.015
	RMSE (Score)	0.123
	Forward Accuracy	100.0%
	Backward Accuracy	98.5%
Serial 7 Subtraction	SMR (Exact)	82.4%
	SMR (± 1)	98.5%
	MAE (Score)	0.191
	RMSE (Score)	0.470
	MAE (Count Correct)	0.368
	RMSE (Count Correct)	0.813
Sentence Repetition	SMR (Exact)	89.7%
	SMR (± 1)	100.0%
	MAE (Score)	0.103
	RMSE (Score)	0.321
	Q-1 Accuracy	89.7%
	Q-2 Accuracy	100.0%
Animal Fluency	SMR (Exact)	98.5%
	SMR (± 1)	100.0%
	MAE (Score)	0.015
	RMSE (Score)	0.121
	MAE (Count)	0.544
Abstraction	SMR (Exact)	82.4%
	SMR (± 1)	100.0%
	MAE (Score)	0.176
	RMSE (Score)	0.420
	Q-1 Accuracy	86.8%
	Q-2 Accuracy	86.8%
HKLLT		
Trial-4	MAE (Recall N)	0.059
	RMSE (Recall N)	0.243
	SMR (Recall N)	94.1%
	SMR (Recall N, ± 1)	100.0%
	MAE (Z-Score)	0.084
	RMSE (Z-Score)	0.380
	SMR (Z-Score)	88.2%
Trial-5	MAE (Recall N)	0.074
	RMSE (Recall N)	0.271
	SMR (Recall N)	92.6%
	SMR (Recall N, ± 1)	100.0%
	MAE (Z-Score)	0.109
	RMSE (Z-Score)	0.453
	SMR (Z-Score)	86.8%