

Automated scoring of a free speech task using a large language model

Natasha Clarke^{1,2}, Andres Salcedo³, James E Galvin³, and Michael J Kleiman³

¹Department of Psychology, Université de Montréal, Montréal, QC, Canada,
²Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, QC, Canada
³University of Miami Miller School of Medicine, Department of Neurology, Comprehensive Center for Brain Health, Boca Raton, FL USA



Main findings: Targeted prompts scored the content of a free response question: “*Describe your typical morning routine.*” A five-question composite detects MCI.

Background

- Free response questions capture rich cognitive information, but scoring open-ended questions objectively is difficult. Manual scoring is subjective and labor-intensive
- Automated linguistic and acoustic speech markers tell only part of the story: they miss content and are vulnerable to confounds (age, sex, education)
- Open-ended questions are easy for patients to answer

Objective

- Test whether targeted LLM prompts can score open-ended responses and detect MCI
- Determine whether a focused, few-item content score is robust to demographic matching and response length, where some linguistic features are not

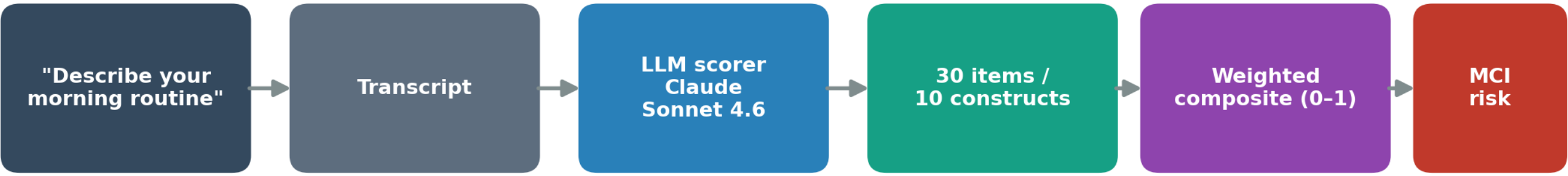
Methods

Participants: Cognitive-aging research cohort (65+) with clinical diagnosis: 275 cognitively normal (CN), 91 MCI, 36 impaired-but-not-MCI (127 impaired cases)

Procedure: Each participant answered the question “Describe your typical morning routine”. Targeted LLM prompts scored specific content dimensions using *Claude Sonnet 4.6*. 30 questions were generated targeted 9 categories:

executive, genericity, lexical, temporal, density, scaffolding, monitoring, discourse, and affective

Analysis: Discrimination measured by AUC, sensitivity, specificity in a 1:1 matched cohort on age, sex, and education. Robustness tested by adjusting for word count and comparing content vs linguistic/acoustic features



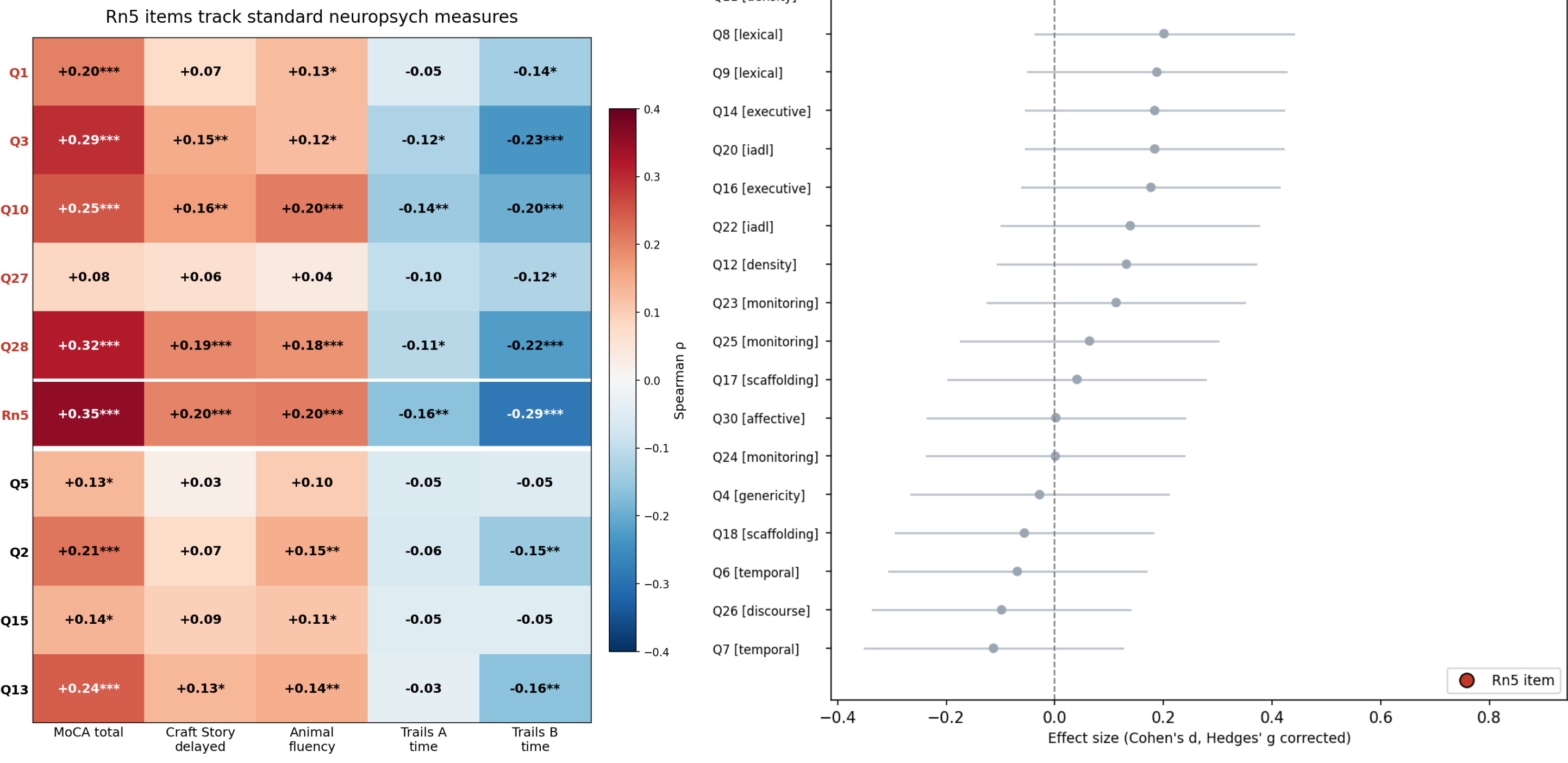
DISCLOSURES:
MJK is the inventor of IP that is utilized in this study. The University of Miami also stands to gain royalties from commercialization of this IP

Results

- 5 question composite (“Rn5”) reached AUC = 0.676 for CN-vs-CI and survived word-count adjustment (p=.02); the 30-item score did not (p=.07)
- Both content scores added information over demographics (LR p <.001) but not over MoCA (LR p=.19)
- When excluding Impaired-not-MCI, linguistic/acoustic features collapsed toward chance (AUC .652 → .573) while the content score held (AUC .647 → .634)

Discussion

- Targeted LLM prompts make objective, scalable content scoring of a single open-ended answer feasible
- Carries signal that linguistic/acoustic speech features miss and survives correction
- Limitations:* single exploratory cohort, item selection partly in-sample, a small heterogenous impaired-not-MCI subgroup, and moderate AUCs with wide confidence intervals
- Next steps:* expand sample to more impaired participants; free response questions likely have a greater potential to capture more advanced forms of cognitive impairment



Q	Category	Summary
1	genericity (3)	Has idiosyncratic personal specifics
2	genericity (3) ↓	Purely generic/script-level
3	genericity (2)	Marks day-to-day variability
4	genericity (2)	Cites a specific recent instance
5	temporal (2)	Anchors to clock times/durations
6	temporal (2)	Stated times internally consistent
7	temporal (3) ↓	Afternoon/evening activities in morning
8	lexical (2)	All references have clear antecedents
9	lexical (2) ↓	Circumlocution (names function, not object)
10	lexical (1) ↓	Relies on generic verbs (get/do/have)
11	density (2)	Multi-proposition sentences
12	density (1)	Explains why (causal language)
13	density (1)	Uses subordinate/conditional clauses
14	executive (2)	Interleaves concurrent tasks
15	executive (2)	Prepares for later events / plans ahead
16	executive (2)	Handles routine deviations / contingencies
17	scaffolding (3) ↓	Needs another person's help/supervision
18	scaffolding (2) ↓	Uses external memory aids (lists/alarms/pill box)
19	scaffolding (2) ↓	Fixed order needed to avoid forgetting
20	iadl (2)	Performs complex IADL unaided
21	iadl (2)	Manages own medications
22	iadl (1)	Engages effortful content (reading/puzzles)
23	monitoring (1)	Self-repairs a narration error
24	monitoring (3) ↓	Unacknowledged contradiction / impossible order
25	monitoring (3) ↓	Unwitting, unacknowledged repetition
26	discourse (2)	Digressions bridged back correctly
27	discourse (1)	Reaches a natural narrative endpoint
28	discourse (1)	Calibrated detail (not too sparse/off-topic)
29	affective (2) ↓	Apathy / behavioral shutdown
30	affective (1)	Includes social interaction