

Automatic Analysis of Informant-Based Collateral Information for Dementia Detection: A Large Language Model Study

Labhpreet Kaur¹, Judith Harrison², Xingyi Song³, Siqi Sun³, Alys Griffiths¹, Daniel Blackburn¹

¹Sheffield Institute for Translational Neuroscience, University of Sheffield, UK; ²Translational and Clinical Neurosciences, Newcastle University, UK;

³School of Computer Science, University of Sheffield, UK.

NIHR National Institute for Health Research

Introduction

Caregivers are often the first to notice dementia-related changes and provide valuable context through informant measures such as the IQCODE and CBI-R. However, these reports are inconsistently collected and underused due to clinical time and resource constraints. Large Language Models offer an automated, standardised and scalable approach for analysing caregiver narratives and extracting clinically meaningful information from collateral histories.

Aims:

- (1) Assess whether an LLM can analyse caregiver collateral histories to identify cognitive impairment, domain-specific deficits and dementia.
- (2) Evaluate agreement between LLM-derived interpretations and expert clinician ratings.

Methods

39 structured caregivers interviews were conducted to obtain collateral histories.



(target n = 50)

Anonymised transcripts were independently reviewed by two expert clinicians: consultant neurologist and old age psychiatrist.

All raters identified: (1) evidence of cognitive impairment, (2) domain-specific deficits and severity, (3) likely differential diagnoses and (4) signs of caregiver stress.

Transcripts were analysed using LLaMA 3.2-3B Instruct with standardised prompts.



Model outputs were compared against clinician ratings to assess agreement, diagnostic classification and domain-level detection accuracy.

Results: Early Findings

Sample Characteristics

- **Mean Age (SD):** 61 ± 13.2
- **Female:** 79.4%; **Male:** 20.5%
- **Mean Time Since Diagnosis:** 5.9 years
- **Relationship:** Child 51.3%; Spouse/ Partner 30.8%; Other 17.9%
- **Ethnicity:** White British 87.1%; Other 12.8%

Global Cognitive Impairment

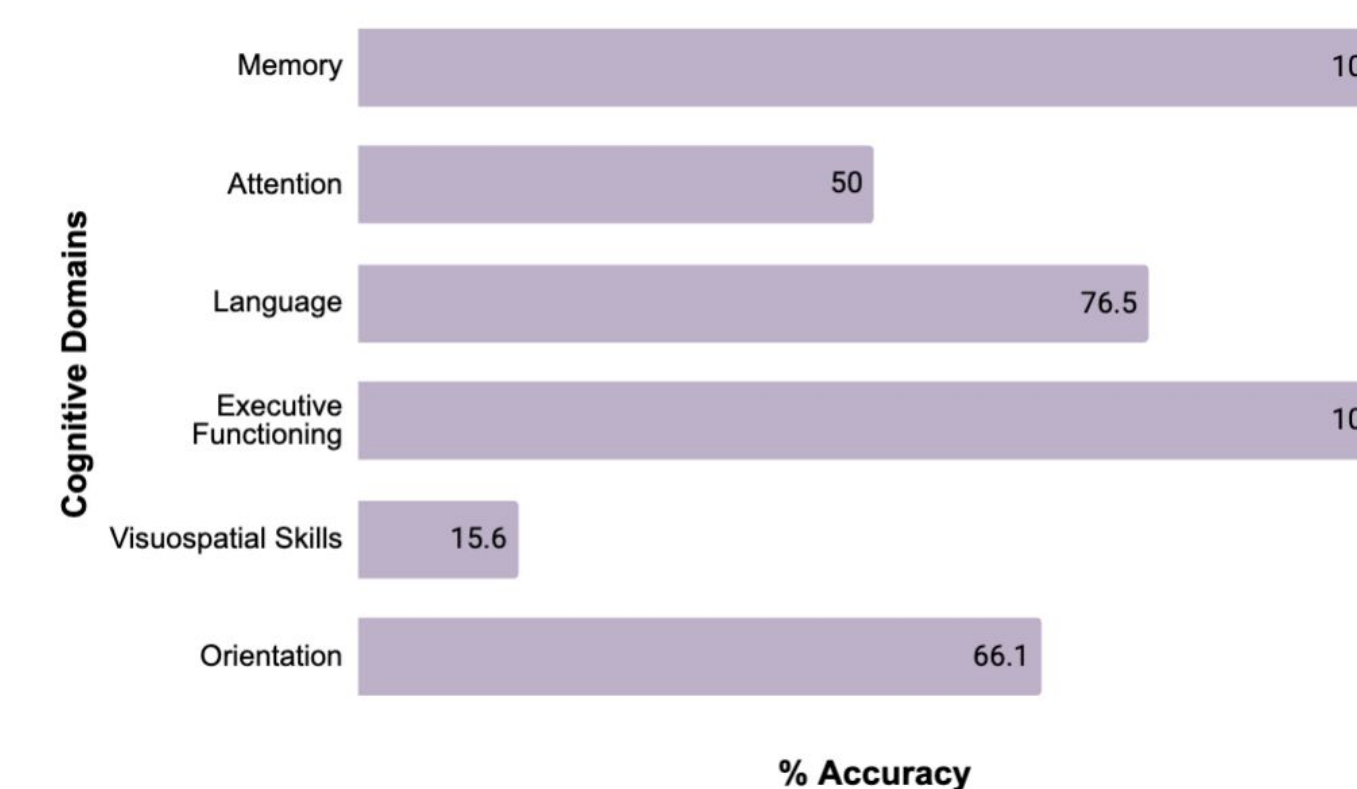
100% raw categorical agreement (all cases identified as impaired) Confidence ratings differed across raters:

- **Clinician 1:** 0.97 ± 0.04
- **Clinician 2:** 0.73 ± 0.08
- **LLM:** 0.64 ± 0.12

*All ratings differed significantly (Friedman test: $\chi^2(2) = 63.09$, $p < 0.001$).

Identifying Cognitive Deficits

Clinician Inter-Rater Reliability: 76.9% agreement and fair agreement ($\kappa = 0.38$) for identifying presence. 45.0% exact agreement and fair agreement ($\kappa = 0.23$) for severity rating.

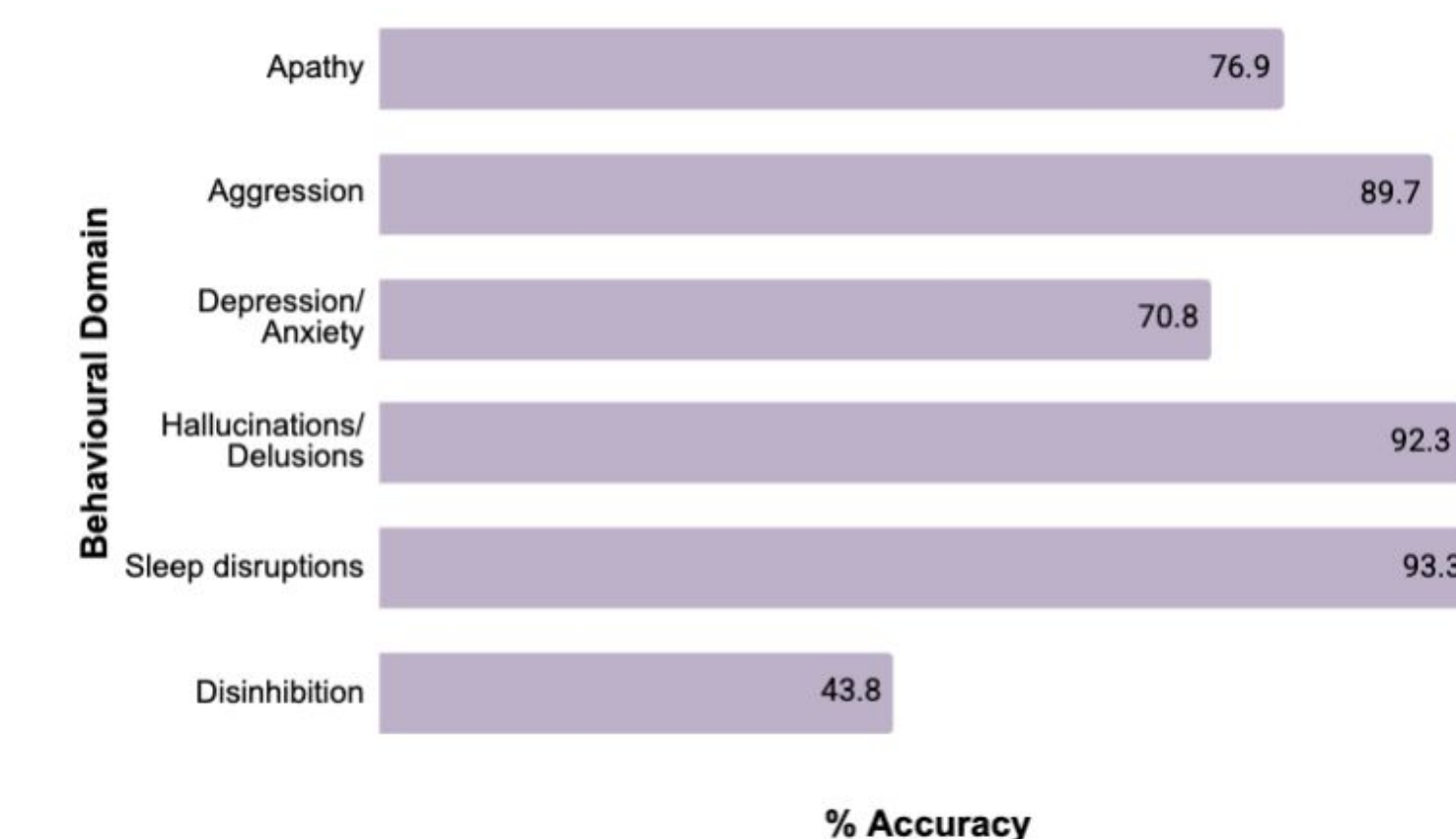


LLM Accuracy: Presence Detection

- Fair agreement (66.1% exact agreement; macro-F1 = 0.45; $\kappa = 0.30$).
- Errors mainly Yes → Uncertain rather than Yes → No, which is clinically safer.

Identifying Behavioural Deficits

Clinician Inter-rater Reliability: Moderate agreement for presence ($\kappa = 0.45$); moderate for severity ($\kappa = 0.55$). Highest agreement for observable behaviours (aggression, disinhibition) and lowest for internal states (depression/ anxiety).

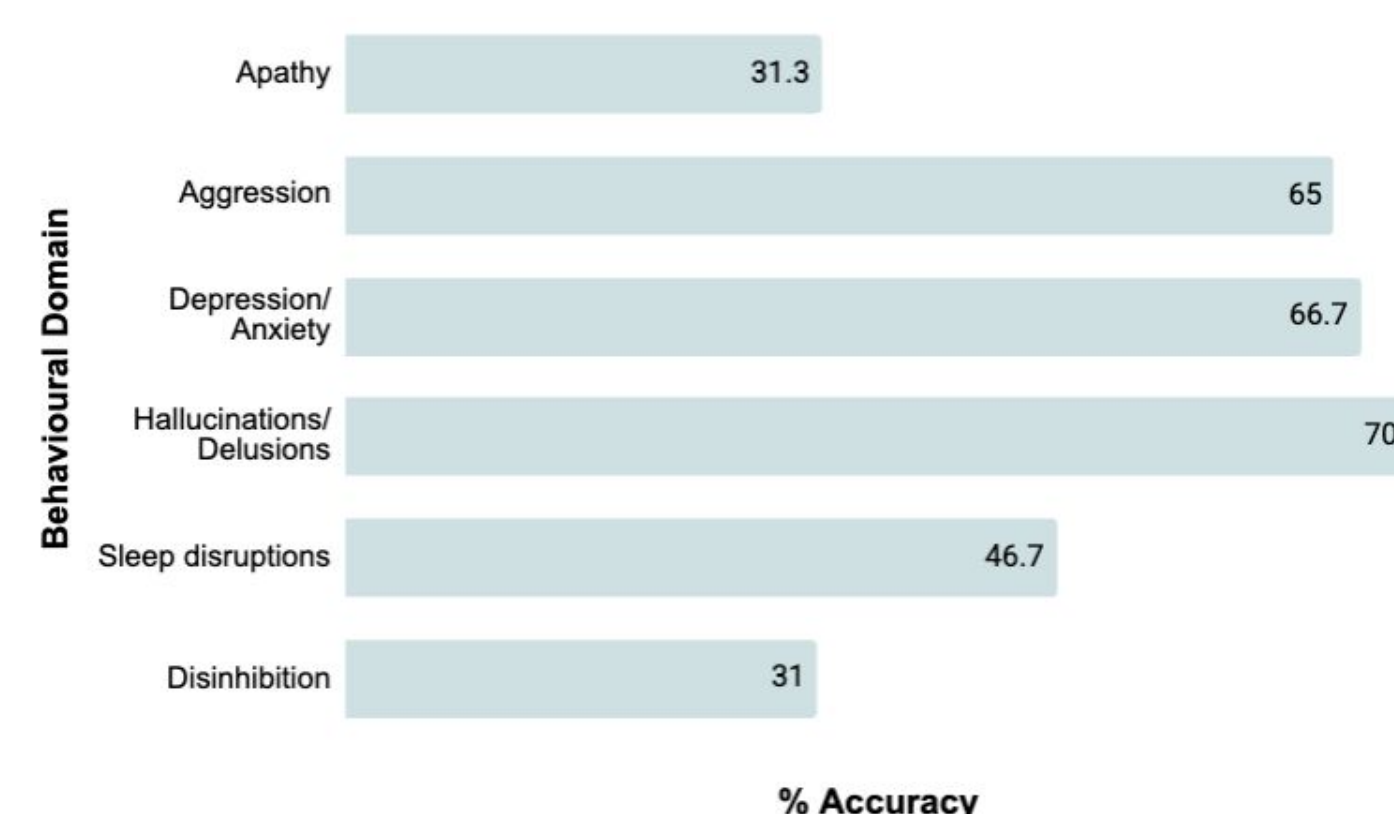


LLM Accuracy: Presence Detection

- Moderate agreement (77.2% exact agreement; macro-F1 = 0.76; $\kappa = 0.54$). Strongest: hallucinations and delusions, sleep disruptions. Weakest: disinhibition, where the LLM over-identified.

Severity Classification

- **Moderate agreement** (51.2% exact agreement; quadratic-weighted $\kappa = 0.46$)
- LLM deviating by an MAE of 0.56 severity levels, indicating most errors were adjacent-level misclassifications (95.9% within ±1 severity level).



LLM's Agreement with Clinicians in Detecting Dementia-Related Impairments

Identifying Functional Impairments

Clinician Inter-rater Reliability: Identifying BADL deficits with moderate agreement ($\kappa = 0.46$; 73.3% agreement) and IADL deficits with fair ($\kappa = 0.26$; 62.6% agreement).

LLM Accuracy:

- **BADL:** Accuracy of 0.49, with high sensitivity (1.00) but low specificity (0.18), indicating the LLM identified all clinician-confirmed deficits but had a tendency to over-identify BADL functional impairments
- **IADL:** Accuracy of 0.77, with high sensitivity (0.89) but moderate specificity (0.57)

Conclusion

- LLM demonstrated fair-moderate agreement with clinicians in identifying and rating dementia-related impairments, on par with clinician inter-rater reliability
- LLMs may have value as a supportive tool for collecting and summarising collateral histories to support assessment such as triage or highlighting areas requiring clinical follow-up.
- **Limitations:** No objective patient measures. Retrospective caregiver report.
- **Future:** Recruit caregivers of newly diagnosed cases to assess early detection performance. Clinicians assess LLM-generated summaries to test fidelity, completeness and safety?

Email: lkaur2@sheffield.ac.uk

LinkedIn:

