Brigham and Women's Hospital
Founding Member, Mass General Brigham

Massachusetts General Hospital
Founding Member, Mass General Brigham

HARVARD MEDICAL SCHOOL

Using LLMs for Thematic Analysis of Cognitive Concerns in Subjective Cognitive Decline

Liqin Wang, PhD; Taiyang Chen, MS; Justin Wang; Sheril Varghese, BS; Jiazi Tian, MS; Rebecca E. Amariglio, PhD; Li Zhou, MD, PhD; Gad A. Marshall, MD

Brigham and Women's Hospital, Boston, USA; Massachusetts General Hospital, Boston, USA; Harvard Medical School, Boston, USA

BACKGROUND & OBJECTIVES

Subjective cognitive decline (SCD) refers to self- or informant-reported memory or cognitive concerns despite normal test performance.
Recognizing SCD provides a critical window for early intervention, risk reduction, and recruitment into prevention or clinical trials.
Because SCD lacks standardized diagnostic codes, and cognitive concerns are often documented only in free-text EHR notes, it is frequently overlooked by traditional data queries.
Research examining documented cognitive concerns in electronic health records (EHRs) at this early, preclinical stage remains limited.
Developing a taxonomy of cognitive concern themes:

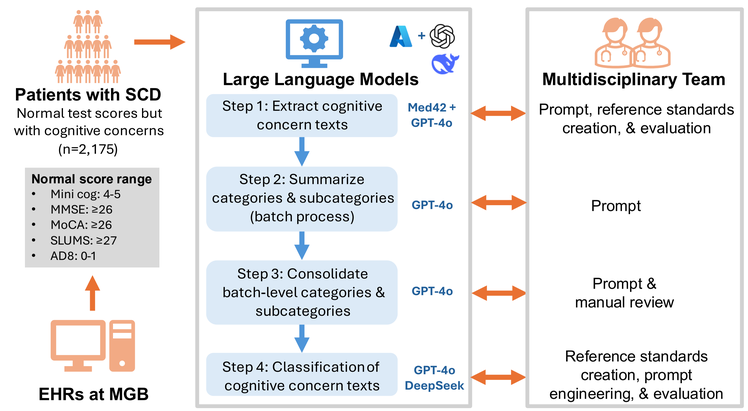
- Provides *a conceptual framework* for characterizing early cognitive symptoms.
- Informs the *design of structured intake forms* and patient-reported measures.
- Supports *systematic and standardized documentation* in clinical workflows.
- Enables *automated detection and monitoring* of SCD in EHRs.
- Facilitates *research on early detection, risk stratification, and intervention timing*.

Objectives:

✅ Develop large language model (LLM)-based pipeline to identify themes and sub-themes of cognitive concerns among patients with SCD.
✅ Develop and evaluate LLM-based classifier of cognitive concerns.
✅ Analyze the distribution of cognitive concern themes and subthemes in SCD.

METHODS

Clinical setting, data source, & study population: EHR data from patients ≥65 years undergoing first cognitive test at Mass General Brigham (Jan 1, 2019–Apr 10, 2024).
Inclusion criteria: normal first test scores, no prior cognitive decline, cognitive concern in notes within 1 year. Final cohort: 2,175 patients.
Cognitive concern extraction: Two-stage LLM approach: Med42-v2-8B for high-sensitivity screening; GPT-4o (Azure, HIPAA-compliant) for precision filtering.
Category extraction & summarization: Removed duplicates; split concern texts into batches of 100; GPT-4o summarized and generated themes/sub-themes; consolidated and reviewed by two experts.
Reference standards: Random 300 texts manually classified by two experts; conflicts resolved by third reviewer; 100 for prompt refinement, 200 for evaluation.
Multi-class classification: Mapped concern texts to finalized themes/sub-themes using GPT-4o and DeepSeek.
Evaluation: Compared GPT-4o vs. DeepSeek on 200 labeled texts using precision, recall, and F1-score.
Statistical analysis: Calculated patient-level theme prevalence, sub-theme proportions within each theme, and theme prevalence by socio-demographic characteristics.

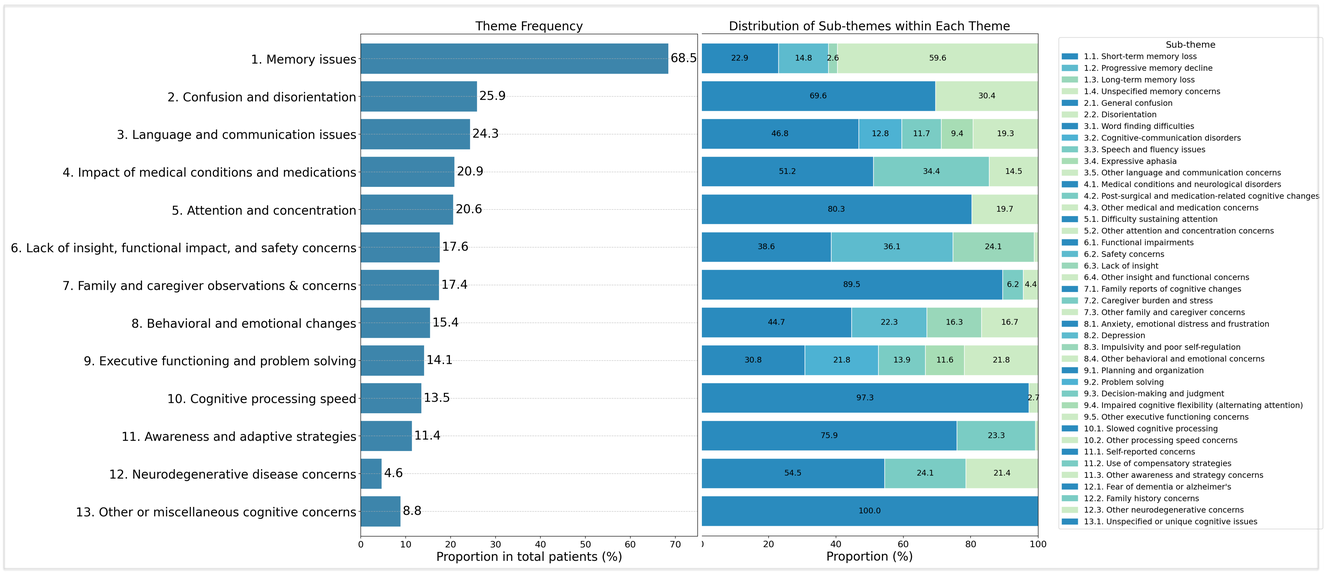


RESULTS

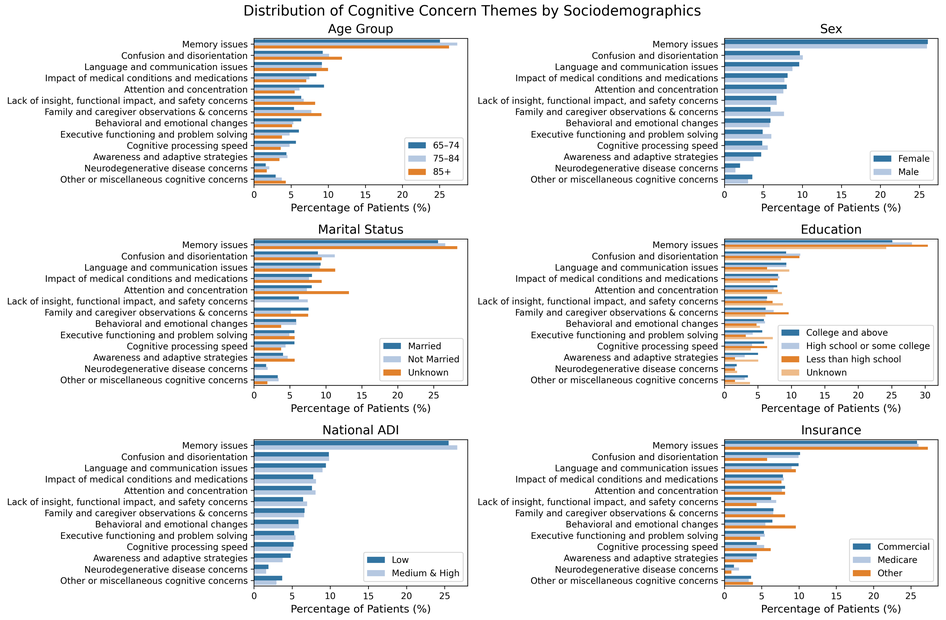
- LLM pipeline using Med42-v2-8B + GPT-4o achieved high precision (92.4%), recall (95.5%), and F1-score (0.939) for identifying cognitive concerns.
- 7,033 unique cognitive concern texts were identified from 2,175 patients with SCD
- Analyzing a sample of 3000 unique concern text identified 12 themes and 40 sub-themes.
- GPT-4o demonstrated high annotation performance, achieving micro F1: 0.848 and macro F1: 0.726 for theme classification, comparable to human annotators.
- Most frequent themes, including *memory issues* (68.5%), *confusion/disorientation* (25.9%), *language/communication* (24.3%), *impact of medical conditions/medications* (20.9%), and *attention/concentration* (20.6%).
- 60% of patients had cognitive concerns spanning multiple themes; 8.8% exhibited issues outside existing categories.

Table 1. Theme- and subtheme-level performance of large language models (LLMs) and human experts in cognitive concern classification (n = 200)

Level	Model/Expert	Micro Score			Macro Score		
		Precision (%)	Recall (%)	F1 Score	Precision (%)	Recall (%)	F1 Score
Category	GPT-4o	90.7	79.7	0.848	80.5	71.5	0.726
	DeepSeek	88.4	73.1	0.801	78.1	62.7	0.669
	Expert1	89.6	82.1	0.857	84.7	80.3	0.802
	Expert2	94.0	83.4	0.884	85.6	72.6	0.775
Subcategory	GPT-4o	85.0	70.8	0.773	77.1	63.7	0.658
	DeepSeek	77.8	62.2	0.691	69.0	48.2	0.510
	Expert1	84.2	74.3	0.789	77.5	70.7	0.710
	Expert2	87.5	75.7	0.812	73.4	66.3	0.671



Distribution of Cognitive Concern Themes by Sociodemographics



CONCLUSION

✅ LLMs enable scalable detection of SCD from EHRs.
✅ GPT-4o achieved higher classification performance than DeepSeek; with accuracy comparable to human annotators, supporting its use for large-scale cognitive concern analysis.
✅ Thematic analysis revealed *memory issues* as the most frequent concern, followed by *confusion and disorientation*, *language/communication issues*, and other cognitive functional themes.
✅ Future work will link concern themes to dementia progression in SCD and evaluate their utility for risk stratification and early intervention.

ACKNOWLEDGEMENTS

This work was supported by the Alzheimer's Association (AARF-22-924992) & National Institute of Health (R00AG075190).