



LLM-Derived Metrics of Surprisal in Alzheimer's Spectrum Disorders

Annette Glotfelty¹, Rene L. Utianski¹, Leland R. Barnard¹, Joseph R. Duffy¹, John L. Stricker¹, Jonathan Graff-Radford¹, Stuart J. McCarter¹, Gregory S. Day², Neill R. Graff-Radford², Bradley F. Boeve¹, David T. Jones¹, Prashanthi Vemuri¹, Val Lowe¹, Clifford R. Jack¹, David S. Knopman¹, Ronald C. Petersen¹, Vijay K. Ramanan¹, Hugo Botha¹

¹Mayo Clinic, Rochester, MN ²Mayo Clinic, Jacksonville, FL

BACKGROUND

LANGUAGE CHANGES IN AD

Language deficits – such as anomia, circumlocution, perseveration, and tangential language – are well documented in patients with AD

DISCOURSE SAMPLES

Open-ended connected speech production involves ongoing interactions among diverse cognitive processes – including working memory, executive function, and semantic retrieval

Picture description tasks enable quick collection of rich data that represent these overlapping cognitive processes, with minimal burden on patients (1)

AUTOMATED ANALYSES

LLMs provide an opportunity to automate the formerly onerous process of transcription and linguistic feature extraction from open-ended speech

SURPRISAL

Surprisal tracks the overall unexpectedness of a word choice in the context of the full utterance. Surprising words can arise from unexpected choices at multiple linguistic levels – semantic, syntactic, or pragmatic

SURPRISAL IN APHASIA

One study of patients with aphasia – neurogenic language disorder – found higher surprisal values when speech is structurally deficient or “agrammatic” (2)

Another study on agrammatic aphasia found that surprisal values were attenuated by the process of lexical condensation – using richer words to compensate for impoverished sentence structure (3)

RESEARCH QUESTION

Can LLM-derived surprisal values in an open-ended picture description task distinguish different severities of cognitive impairment due to AD?

METHODS

Participants: 112 patients from clinical practice and Mayo Clinic Study of Aging

- 60 amyloid-negative controls
- 52 cognitively impaired with biomarker confirmed AD
 - 32 Mild Cognitive Impairment (MCI)
 - 20 Dementia (ADD)

Task: Open-ended picture description

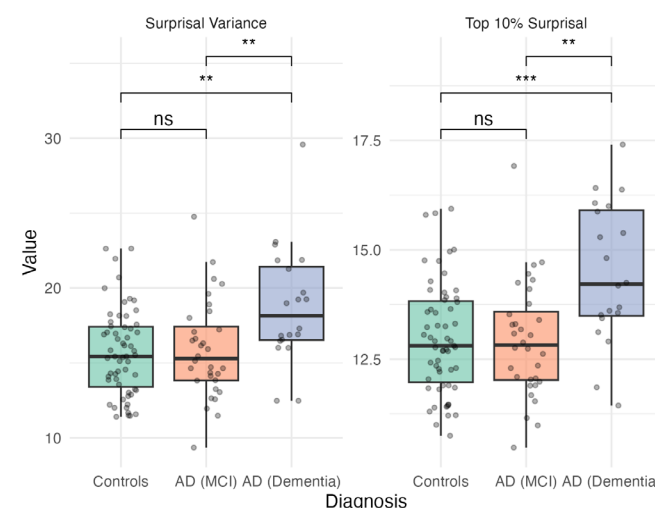
Transcriptions created with CrisperWhisper

Surprisal value = negative log-probability of a given token given its preceding context, using Gemma-7B

GROUP DIFFERENCES

- **Mean Surprisal**
 - No differences between groups
- **Surprisal Variance**
 - Kruskal-Wallis H = 9.9, p = 0.0071
- **Top 10% Surprisal**
 - Kruskal-Wallis H = 11.3492, p = 0.0034

Surprisal Metrics: AD Clinical Continuum
Pairwise comparisons via Wilcoxon test



% OF VARIANCE EXPLAINED

While holding age, gender, and education level constant, diagnosis (Controls, MCI, ADD) explained:

- 12.4% of variance in **Surprisal Variance** values (p = 0.003)
- 11.4% of variance in **Top 10% Surprisal** values (p = 0.006)

DISCUSSION

- Average surprisal across entire picture description does not differ among groups
- Control and MCI groups do not differ on any metric of surprisal
- ADD patients have higher surprisal variance, or more jumps between predictable and unexpected words
- The top 10% most surprising words produced were significantly more unexpected in the ADD group

CONCLUSIONS

- Elevated surprisal metrics in discourse are not specific to aphasia
- Language produced at advanced levels of AD cognitive impairment differs from typical English in quantifiable ways
- Discourse produced by patients with ADD is not more surprising overall; it has a broader range of surprising words and more extreme peaks
- Lack of elevated surprisal in MCI may indicate that unexpected word choices are a hallmark of later disease progression, or that a compensatory mechanism present in MCI is no longer effective in dementia

REFERENCES

1. Mueller, K. D., Hermann, B., Mecollari, J., & Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of clinical and experimental neuropsychology*, 40(9), 917-939.
2. Cong, Y., LaCroix, A. N., & Lee J. (2024). Clinical efficacy of pre-trained large language models through the lens of aphasia. *Scientific reports*, 14(1), 15573.
3. Rezaei, N., Michaelov, J., Josephy-Hernandez, S., Ren, B., Hochberg, D., Quimby, M., & Dickerson, B. C. (2023). Measuring Sentence Information via Surprisal: Theoretical and Clinical Implications in Nonfluent Aphasia. *Annals of neurology*, 94(4), 647-657.

CONTACT

GLOTFELTY.ANNETTE@MAYO.EDU