

LEVERAGING LLMs FOR THE PREDICTION OF APHASIA SEVERITY DIRECTLY FROM SPEECH

Emily Tumacder (CoE) and Dr. Yan Cong (Linguistics)

ABSTRACT

Aphasia is an acquired language impairment, affecting the production and comprehension of speech.

Current aphasia detection relies on diagnoses from speech pathologists examination. However, we can lower costs of clinical diagnoses with automatic aphasia diagnosis.

We propose a user-friendly, end-to-end method of aphasia severity prediction through direct input of patient speech into pre-trained LLMs, without a need for intermediate score calculations or feature extractions.

We designed various prompts to produce a multi-class classification (severe, moderate, mild, latent) [Kertesz, 2007] of the patient's conversation. We tested these prompts on the publicly available, pre-trained LLM Llama 3.2-1B. We demonstrate the effects of prompting styles on the model accuracy and precision, noticing that with our prompts the pre-trained LLM has little variance in its predictions.

PRE-CLEANING

Semi-cleaned conversation:

- Includes some disfluencies in speech (repetition, filler words, misspeak) to retain identifiers of aphasia (paraphasia)
- Over 50 characters to exclude filler conversation

Severities [Kertesz, 2007]:

- Latent: No reported WAB-AQ score

- Mild: 76+ WAB-AQ

- Moderate: 51-75 WAB-AQ

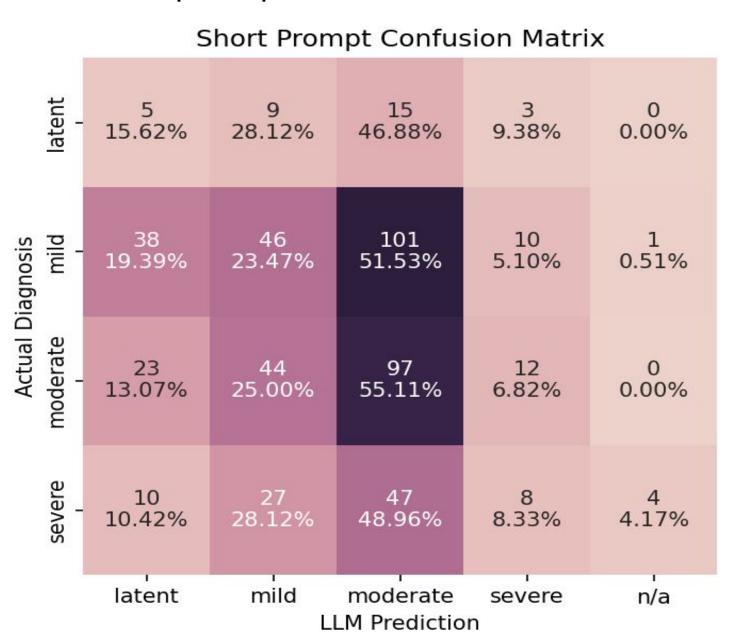
- Severe: 0-50 WAB-AQ

	Aphasic (n = 482)
Male/ Female	283/199
Age* (range)	61.3 (25.0 - 90.7)
Duration of Aphasia** (range)	5.7 (0.2 - 44.0)
Education*** (range)	15.6 (7.0 - 25.0)
number of participants with age reporte	d = 479
** number of participants with duration of	f aphasia reported = 360
*** number of participants with education	reported = 401
Demographics of partic	ipants (AphasiaBank)

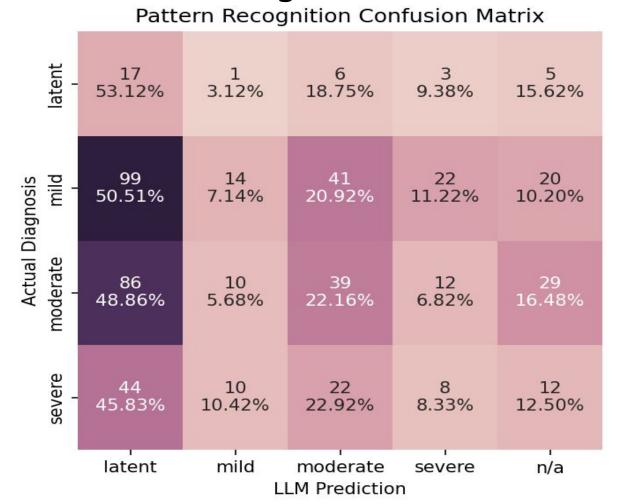
EXPERIMENT: PROMPTING

Each prompt was applied onto 500 conversations of varying lengths and topics.

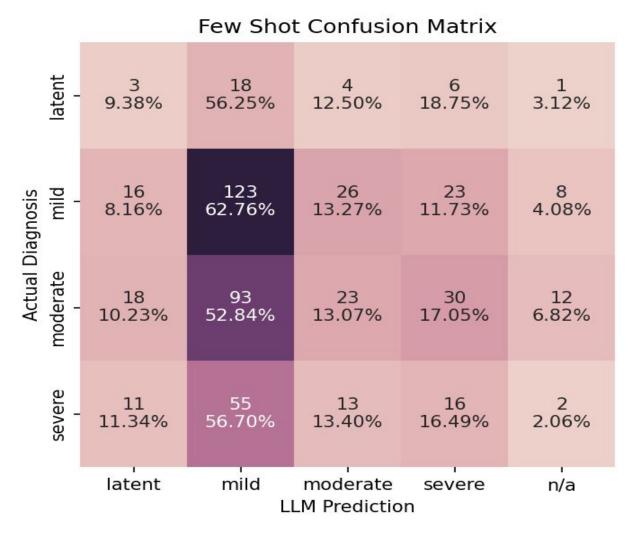
1. Short prompt



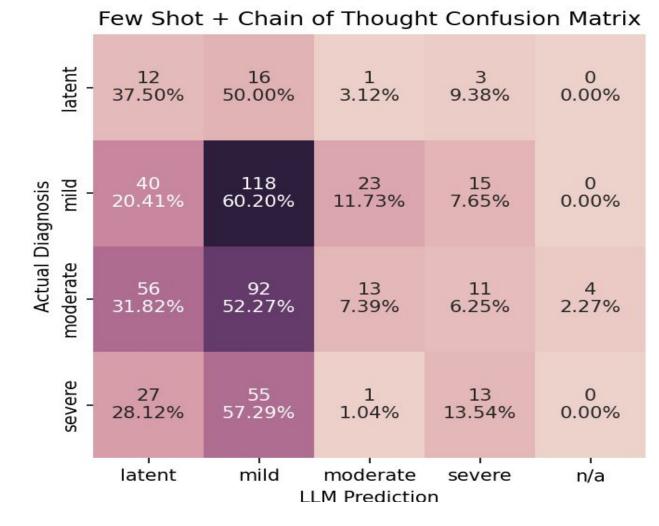
2. Pattern Recognition



3. Few Shot



4. Few Shot + Chain of Thought



- Single categorization and generalization to all conversations
 - structure of the prompt
 - general nature of LLMs.
- LLM may not be sensitive to nuance in aphasia severity
- May pick up "random" language patterns

Example of successful LLM response:

A conversation with a person with severe aphasia about floods.

Aphasia Severity: Severe
The provided conversation is a dialogue
between two people, Flood and Nan. In this
conversation, Flood is repeating the same
words and phrases and speaking about
things not related to the topic of
conversation. These characteristics are
indicative of a severe aphasia.

The repetition of words and phrases, such as "oh" and "yeah", suggests a significant loss of fluent speech production. The use of similar sounds or words to replace correct phrases, such as "one" and "nan", further supports the diagnosis of aphasia.

Example of unsuccessful LLM response:

A conversation with a person with mild aphasia about a boy kicking a soccer ball.

- **Category:** Aphasia (Severe)
- **Explanation:** The conversation appears to be a simple exchange between two people, with the speaker using a **repetitive** and nonsensical phrase...

The speaker uses similar sounds and words to replace correct phrases ("like" and "it" are replaced by "um"), and uses words that do not sound normal ("soccer" instead of "football"). The conversation is also characterized by a lack of coherence and a lack of context...

FINE TUNING

It's expected that, if trained on the data, Llama 3.2-1B should perform better with all prompts at recognizing aphasic text. For fine tuning, we would train the model on 80% of the data, and test it on 20%.

DATASETS AND RESOURCES

AphasiaBank (MacWhinney et al., 2011) Western Aphasia Battery WAB-R-AQ (Kertesz, 2007)

ACKNOWLEDGEMENTS:

We acknowledge Jiyeon Lee and Arianna N. LaCroix for their inspiration. This project is funded by the Faculty Research Development Award, the Center on Aging and the Life Course at Purdue University.