Anomia is a prominent feature of aphasia (Goodglass & Wingfield, 1997) that manifests in all communicative contexts, from single word productions to complex conversations. Work has been done to automate scoring of single word responses on confrontation picture-naming tests (Salem et al., 2022, Fergadiotis et al., 2016; McKinney-Bock & Bedrick, 2019), but to date there are no automated tools for the classification of paraphasias within discourse. Analyzing discourse directly may provide clinical insights not gained via decontextualized tasks such as confrontation naming (Hickin et al., 2001, Mayer & Murray, 2003, Pashek & Tompkins, 2002). However, discourse is harder to analyze because unlike in picture-naming tests, we do not know the ground-truth targets, and first need to identify them in order to classify the paraphasias. Recent advancements in computer hardware (specifically development of GPUs) have led to the development of large language models (LLMs): machine learning based models pre-trained on very large general-purpose datasets, which can be fine-tuned for specific language automation tasks. In this work, we make a first attempt at automating a process for predicting the intended targets of paraphasias in discourse using a LLM called Big Bird (Zaheer et al., 2021).

Aims

The purpose of the current study was to create a baseline model for automated target word prediction of paraphasias within transcribed spoken discourse using only the narrative context around the paraphasia itself (excluding gestures and phonological information). To this end, we fine-tuned Big Bird to automatically predict paraphasia targets in story retellings. We had two research objectives: 1) assess the feasibility of applying a modern LLM to this task and establish a performance baseline; 2) explore the impact of clinical factors (specifically fluency and aphasia severity) and intended target ambiguity (according to human transcribers) on model performance.

Method

Data consisted of 332 Cinderella story retelling transcripts from 240 people with aphasia from the English AphasiaBank database (MacWhinney et al., 2011). These sessions contained 2,489 paraphasias. First, three research assistants identified the intended targets of each paraphasia and provided a confidence rating (1-4). Then, a certified SLP resolved all disagreements to determine the intended target. For each paraphasia, we used the research assistants' data to calculate the percent agreement between annotators and average confidence. Participants' demographic and clinical data can be seen in Table 1.

To prepare the transcripts, we replaced paraphasias with a "blank" token. We accessed the pre-trained LLM (Wolf et al., 2019) and fine-tuned it to automatically fill in the blank with a predicted target, based on the bidirectional context of the rest of the Cinderella story retelling. We compared this performance with the pre-trained LLM without fine-tuning. We tested the models' predictions against our human-identified paraphasia targets by calculating accuracy for exact match (the top model prediction was the same as the human-identified target) or within X accuracy (the human-identified target was within the top X model predictions, where X ranged from 1-20). We determined whether disagreements between the pre-trained and fine-tuned models were significant using McNemar's test (McNemar, 1947). Finally, we stratified our results by Western Aphasia Battery-Revised (WAB-R) severity (Kertesz, 2012), fluent vs non-fluent aphasia, whether humans had perfect agreement in target identification, and human confidence in target identification. We tested whether differences in performance between these stratifications were significant using two-sided *z*-tests for independent proportions.

Results

Accuracy results are shown in Table 2. The pre-trained LLM achieved 25.5% accuracy at exactly matching the human-identified target, and the fine-tuned LLM achieved 46.8% accuracy. The difference in performance was significant, with McNemar's *p*-value < 0.001. Performance of the two models within top 20 predictions is shown in Figure 1. The fine-tuned model performed better on targets with perfect human agreement (59.5% vs 34.2% accuracy) and higher human confidence (60.5% vs 36.2%). It also performed better on paraphasias from participants with less severe aphasia (52.7% vs 41.6%) or fluent aphasia (48.7% vs 41.2%). All differences in performance were significant ($p < 0.01$, Table 3).

Discussion

We were able to automatically identify the intended target of paraphasias in discourse using just semantic information about half of the time. Model performance was higher on targets for which human annotators had less difficulty, and on participants with fluent or less severe aphasia. These findings take us a step closer to automatic aphasic discourse analysis, and open up possibilities for novel applications that extend beyond assessment (e.g., AAC). In future work, we will incorporate phonological information to further improve predictive utility. The findings will be discussed with an emphasis on their implications for research and clinical practice.

Tables

Table 1

Clinical and demographic information for the 240 participants at their first session.

Characteristic	**Value**
Age (years)	
M (*SD*)	61.478 (12.494)
Min - Max	25.600 - 91.718
Missing (*N*)	23
Gender	
M (*N*)	124
F (*N*)	96
Missing (*N*)	20
Race	
White (*N*)	189
African American (*N*)	23
Asian (*N*)	2
Hispanic/Latino (*N*)	4
Native Hawaiian/Pacific Islander (*N*)	1
Mixed (*N*)	1
Unavailable (*N*)	20
Education (years)	
M (*SD*)	15.439 (2.811)
Min - Max	8.000 - 25.000
Missing (*N*)	28
Years post onset	
M (*SD*)	5.389 (4.731)
Min - Max	0.080 - 30.000
Missing (*N*)	23
WAB-R AQ	
M (*SD*)	72.771 (17.659)
Min - Max	10.800 - 99.600
Missing (*N*)	11
BNT	
M (*SD*)	7.517 (4.475)
Min - Max	0.000 - 15.000
Missing (*N*)	31
VNT	
M (*SD*)	15.200 (6.084)
Min - Max	0.000 - 22.000
Missing (*N*)	31

Note. WAB-R AQ is the Western Aphasia Battery-Revised Aphasia Quotient. BNT is the raw score from the Boston Naming Test-Short Form (Kaplan et al., 2001). VNT is the raw score from the Verb Naming Test (Cho-Reyes et al., 2012).

Table 2

Accuracy of pre-trained and fine-tuned LLMs matching the human-identified target, across test sets.

		Pre-trained		**Fine-tuned**	
Test set	***N* paraphasias**	**Accuracy exact match**	**Accuracy within 5**	**Accuracy exact match**	**Accuracy within 5**
All paraphasias	2489	0.255	0.379	0.468	0.657
Human agreement = 100%	1244	0.309	0.405	0.595	0.767
Human agreement < 100%	1245	0.201	0.353	0.342	0.548
Human confidence > median (3.3)	1089	0.319	0.419	0.605	0.768
Human confidence <= median (3.3)	1400	0.206	0.348	0.362	0.571
WAB-R AQ > median (74.6)	1039	0.294	0.410	0.527	0.703
WAB-R AQ <= median (74.6)	1076	0.204	0.325	0.416	0.621
Fluent participants	1666	0.261	0.385	0.487	0.670
Non-fluent participants	449	0.198	0.301	0.412	0.626

Note. WAB-R AQ is the Western Aphasia Battery-Revised Aphasia Quotient. Fluent participants are those with Wernicke, Anomic, Conduction, or Transcortical Sensory aphasia, or those considered “non aphasic” by the WAB-R. Non-fluent participants are those with the Broca, Global, or Transcortical Motor aphasia. 46 out of 332 total sessions had unavailable WAB-R results and were excluded just from analyses involving WAB-R scores. Accuracy exact match refers to the top model prediction of target word matching the human-identified target word. Accuracy within 5 refers to the human-identified target word being one of the top five model predictions.

Table 3

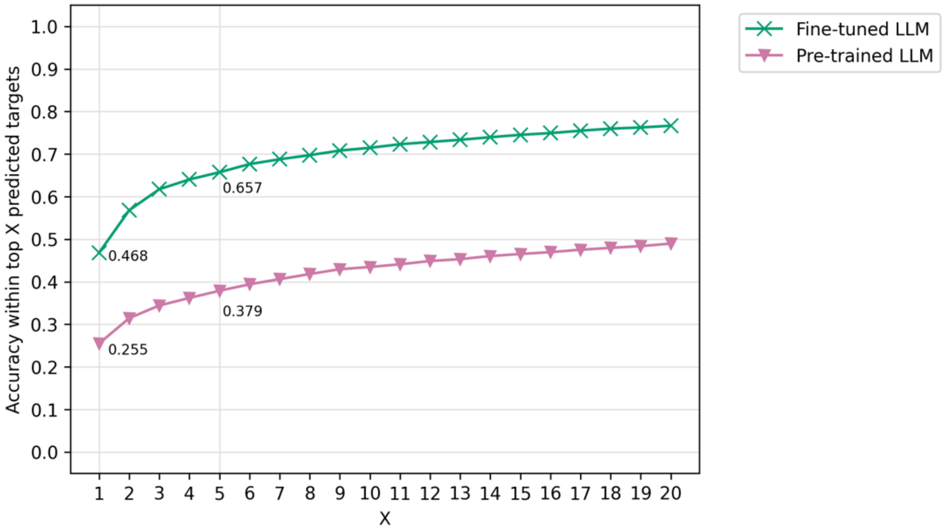
Two-sided *z*-tests for independent proportions for test set stratifications of exact match accuracy for the fine-tuned LLM.

Comparison	***z***	***p***
Human agreement = 100% vs Human agreement < 100%	11.353	<0.001
Human confidence > median (3.3) vs Human confidence <= median (3.3)	11.121	<0.001
WAB-R AQ > median (74.6) vs WAB-R AQ <= median (74.6)	4.793	<0.001
Fluent participants vs Non-fluent participants	2.581	0.010

Figures

Figure 1

Accuracy of pre-trained and fine-tuned LLMs matching the human-identified target within top 1-20 model predictions.



References

1. Cho-Reyes, S., & Thompson, C. K. (2012). Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). *Aphasiology*, *26*(10), 1250–1277. https://doi.org/10.1080/02687038.2012.693584
2. Fergadiotis, G., Gorman, K., & Bedrick, S. (2016). Algorithmic Classification of Five Characteristic Types of Paraphasias. American Journal of Speech-Language Pathology, 25(4S). https://doi.org/10.1044/2016_AJSLP-15-0147
3. Goodglass, H., & Wingfield, A. (Eds.). (1997). Anomia: Neuroanatomical and cognitive correlates. Academic Press.
4. Hickin, J., Best, W., Herbert, R., Howard, D., & Osborne, F. (2001). Treatment of Word Retrieval in Aphasia: Generalisation to Conversational Speech. International Journal of Language & Communication Disorders, 36(s1), 13–18. https://doi.org/10.3109/13682820109177851
5. Kaplan, E., Goodglass, H., & Weintraub, S. (Eds.). (2001). *Boston naming test* (2. ed). Lippincott, Williams & Wilkins.
6. Kertesz, A. (2012). Western Aphasia Battery—Revised [Data set]. American Psychological Association. https://doi.org/10.1037/t15168-000
7. MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk (3rd ed.). Lawrence Erlbaum Associates.
8. Mayer, J., & Murray, L. (2003). Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. Aphasiology, 17(5), 481–497. https://doi.org/10.1080/02687030344000148
9. McKinney-Bock, K., & Bedrick, S. (2019). Classification of Semantic Paraphasias: Optimization of a Word Embedding Model. Proceedings of the 3rd Workshop on Evaluating Vector Space Representations For, 52–62. https://doi.org/10.18653/v1/W19-2007
10. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996
11. Pashek, G. V., & Tompkins, C. A. (2002). Context and word class influences on lexical retrieval in aphasia. Aphasiology, 16(3), 261–286. https://doi.org/10.1080/02687040143000573
12. Salem, A. C., Gale, R., Casilio, M., Fleegle, M., Fergadiotis, G., & Bedrick, S. (2022). Refining Semantic Similarity of Paraphasias Using a Contextual Language Model. Journal of Speech, Language, and Hearing Research, 1–15. https://doi.org/10.1044/2022_JSLHR-22-00277
13. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6
14. Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for longer sequences. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.