

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

Abstract

Purpose: To date there are no automated tools for the identification and fine-grained classification of paraphasias within discourse, the production of which is the hallmark characteristic of most people with aphasia (PWA). In this work we fine-tune a large language model (LLM) to automatically predict paraphasia targets in Cinderella story retellings.

Method: Data consisted of 353 Cinderella story retellings containing 2,489 paraphasias from PWA, for which research assistants identified their intended targets. We supplemented this training data with 256 sessions from control participants, to which we added 2,427 synthetic paraphasias. We conducted four experiments using different training data configurations to fine-tune the LLM to automatically “fill in the blank” of the paraphasia with a predicted target, given the context of the rest of the story retelling. We tested the experiments’ predictions against our human-identified targets and stratified our results by ambiguity of the targets and clinical factors.

Results: The model trained on controls and PWA achieved 46.8% accuracy at exactly matching the human-identified target. Fine-tuning on PWA data, with or without controls, led to comparable performance. The model performed better on targets with less human ambiguity, and on paraphasias from participants with less severe or fluent aphasia.

Conclusion: We were able to automatically identify the intended target of paraphasias in discourse using just the surrounding language about half of the time. These findings take us a step closer to automatic aphasic discourse analysis. In future work, we will incorporate phonological information from the paraphasia to further improve predictive utility.

37 Anomia or word-finding difficulty is a prominent and persistent feature of aphasia
38 (Goodglass and Wingfield, 1997) and manifests in all communicative contexts, from single word
39 responses to complex conversations. Given the ubiquitous nature of anomia, anomia assessments
40 are given in most clinical settings and are of high practical value for quantifying performance
41 and monitoring outcomes. Typically, anomia assessments include confrontation picture naming
42 tests (Rabin et al., 2005; Simmons-Mackie, Threats, & Kagan, 2005), in which a person with
43 aphasia is asked to name a series of pictured objects and/or actions. The popularity of
44 confrontation picture naming tests can be attributed to their well-documented validity and
45 reliability (e.g., Roach et al., 1996; Strauss, Sherman, & Spreen, 2006; Walker & Schwartz,
46 2012), and also to their relatively low testing burden, particularly in the context of short forms
47 and simple accuracy scoring schemes. Other sources of diagnostic information such as discourse-
48 level analyses may provide additional clinically useful information for completing a patient's
49 clinical profile (Fergadiotis et al., 2019; Richardson et al., 2018) but such analyses are not
50 performed routinely in clinical settings. Viewed through an implementation science lens
51 (Damschroder et al., 2009; Breimaier et al., 2015), several barriers hinder the utilization of
52 discourse-based analyses including their complexity, reliability, and time burden. The latter
53 factor especially can be an insurmountable barrier for implementation in most real-world clinical
54 settings. Therefore, there is a need to develop new approaches that will enable professionals to
55 assess people with aphasia (PWA) in a more objective, precise, efficient, and ecologically valid
56 manner.

57 Computational methods, especially those from the field of Natural Language Processing
58 (NLP), have the potential to be essential tools in designing such approaches. Recent work has
59 demonstrated these methods' efficacy in automating certain aspects of confrontation naming test

60 scoring (Casilio et al., 2023; Salem et al., 2022; Fergadiotis et al., 2016; McKinney-Bock &
61 Bedrick, 2019; described later in more detail). In this work, we report on a crucial first step in
62 applying such methods to discourse samples. Specifically, we describe the results of a
63 computational model that analyzes the context in which a paraphasia occurs in a discourse
64 sample and predicts the speaker's intended word (or a set of possible intended words). Below, we
65 describe the key role that this specific task of target word prediction plays in the clinical
66 assessment of discourse samples from PWA, motivate our overall computational approach, and
67 describe our model and its behavior. In addition, we evaluate the impact of clinical features of
68 the speaker on our model's ability to correctly predict target words. This part of the work
69 highlights specific areas where current technology falls short and points to missing pieces that
70 the field must address.

71 **Assessing Anomia at Discourse Level**

72 It is well documented in the literature that the ability to produce discourse is what matters
73 most to PWA and their families (Cruice et al., 2003; Mayer & Murray, 2003). Yet, despite their
74 popularity, there is evidence that confrontation naming tests cannot fully account for the severity
75 and patterns of anomia exhibited during connected speech. First, connectionist accounts of word
76 retrieval at the discourse level highlight how lexical characteristics of target words interact with
77 activated representations within and across different linguistic levels (e.g., phonological,
78 semantic) (Bock, 1995; Dell, 1986; Dell et al., 1999; Schwartz et al., 2006; Levelt, 1999; Levelt
79 et al., 1999). In addition, several models (e.g. MacDonald, 1994; Tabor et al., 1997) emphasize
80 the influence and relative strength of naturally occurring probabilistic constraints in language use
81 on the activation of linguistic representations. In fact, there seems to be a general consensus in
82 recent empirical investigations that while performance in confrontation naming tests is related to

83 discourse-level performance, analyzing discourse directly may provide unique and useful clinical
84 insights not gained via confrontation naming tests (Fergadiotis et al., 2019; Hickin et al., 2001;
85 Mayer & Murray, 2003; Pashek & Tompkins, 2002). Therefore, relevant assessment tools for
86 aphasia should a) operate at the discourse level, b) be able to capture changes in language skills
87 over time, and c) be routinely included as therapy outcome measures.

88 At the level of single words, anomia severity is commonly assessed using picture naming
89 tests and reported in terms of overall accuracy scores or ability estimates. Further, a more in-
90 depth analysis of the types and frequencies of word production errors can reveal which linguistic
91 processes that support word access and retrieval are more or less disrupted (Dell et al., 1997).
92 Theoretical accounts of word production allow professionals and/or algorithms to classify an
93 individual's collection of paraphasias in order to create a detailed profile of that individual's
94 anomia. This paraphasia classification process requires a series of binary judgments with regards
95 to the paraphasia and its relationship to the intended target word. Specifically, those judgments
96 are: 1) lexicality, i.e., whether or not the paraphasia is a real word; 2) semantic similarity, i.e.,
97 whether or not the paraphasia is semantically related to the target; and 3) phonological similarity,
98 i.e., whether or not the paraphasia is phonologically related to the target. To highlight a couple of
99 classification examples, a Semantic paraphasia is a real word that is semantically related to its
100 intended target but phonologically unrelated (e.g., "beard" for "mustache"); whereas a neologism
101 is a nonword, not semantically related by definition, that is phonologically related to the target
102 (e.g., "mustaff" for "mustache"). Lexical or real word paraphasias are understood to represent
103 mostly impairments in lexical-semantic access while nonword paraphasias are thought to reflect
104 deficits in phonological encoding. To help make this time- and labor- intensive assessment
105 process more efficient and therefore more feasible for clinical settings, our research team has

106 developed a paraphasia classification algorithm called ParAlg (Paraphasia Algorithms) that
107 automatically classifies word production errors in the context of object picture naming tests
108 (Casilio et al., 2023; Salem et al., 2022; Fergadiotis et al., 2016; McKinney-Bock & Bedrick,
109 2019). ParAlg's paraphasia classifiers algorithmically mirror the main paraphasia classification
110 criteria of the Philadelphia Naming Test (Roach et al., 1996), which includes one of the most
111 well-established and thorough frameworks for error classification during object picture naming.

112 The accuracy of this multistep paraphasia classification process, however, is entirely
113 predicated on successfully identifying a given paraphasia's intended target. Target identification
114 is relatively straightforward in the context of confrontation picture naming tests, where the target
115 is presumed to be the word depicted in the picture, but in the context of discourse, determining
116 the target is not as straightforward. Researchers and clinicians undertake this task by applying
117 background knowledge of word production disorders and common anomic patterns (Martin,
118 2017), as well as general knowledge of the discourse task itself, such as the expected lexicon and
119 the expected temporal arrangement of that lexicon given the overall narrative structure.
120 Furthermore, target prediction can incorporate a multitude of localized contextual factors such as
121 timely gestures, re-tracings from the paraphasia to or toward the intended target, phonological
122 fragments or false starts leading up to the paraphasia, syntactic/semantic information
123 immediately surrounding the paraphasia, and/or semantic and phonological similarities between
124 the paraphasia and its working hypothesis target.

125 In light of this highly variable and complex process, the preliminary focus of this
126 automation work and of the current paper is to leverage and model the semantic information
127 surrounding word production breakdowns. Elegantly enough, this approach mirrors widely
128 accepted models of spoken word production, such as Dell's model described earlier where step

129 one involves identification and activation of semantic representations surrounding the target
130 word. One additional and imminent aim of this work, though outside of the scope of this paper, is
131 the exploration of a more fully-automated and naturalistic application of ParAlg - classification
132 of paraphasias in discourse using machine-generated targets. While the present paper explores
133 automatic target prediction for a full range of content words (nouns, verbs, adverbs, adjectives),
134 we do not anticipate being able to classify paraphasias with non-noun targets until equally robust
135 psycholinguistic models are developed for additional parts of speech.

136 **Novel Approaches for Assessing Paraphasias at Discourse Level**

137 Given the resource-intensive nature of discourse analysis, several computational
138 approaches have been developed to assist researchers and clinicians in analyzing discourse such
139 as automated speech and language measures (e.g., Fergadiotis & Wright, 2011; Bryant et al.,
140 2013; Miller & Iglesias, 2012; Forbes et al., 2014; Day et al., 2021; Chatzoudis et al., 2022). An
141 active area of research is establishing automatic speech recognition (ASR) systems that are
142 effective on aphasic speech (e.g., Le & Provost, 2016; Perez et al., 2020; Gale et al., 2022), some
143 of which are developed and used for diagnosing aphasia or aphasia subtypes (e.g., Fraser et al.,
144 2013; Le et al., 2018). Some preliminary attempts have been made at automated classification of
145 paraphasias in connected speech, but these studies have focused solely on the task of *detecting*
146 paraphasias and determining if they are real words or neologisms (Le et al., 2017; Pai et al.,
147 2020), as opposed to complete classification. Despite the recent advances in automated
148 approaches, to this date there are no computer assisted discourse analyses for the identification
149 and fine-grained classification of paraphasias, the production of which is the hallmark
150 characteristic of most PWA.

151 Our first attempts at predicting targets of paraphasias in discourse were made using more
152 traditional n-gram and early neural net based language models (Adams et al., 2017), but since
153 then, there have been significant developments in the field of language modeling. In this work, to
154 automatically predict the intended targets of paraphasias in discourse using the surrounding
155 language, we use a machine learning-based transformer language model. Transformer models
156 were first introduced in 2017 (Vaswani et al., 2017) and have since become ubiquitous in NLP
157 research due to their high performance; their structure allows them to be trained on large scale
158 datasets with graphical processing units (GPUs). The introduction of transformer models led to
159 the development of BERT (Bidirectional Encoder Representations from Transformers; Devlin et
160 al., 2019), a large language model (LLM) which has been successful on a variety of NLP tasks
161 such as Google search, text summarization, and question answering (Devlin et al., 2019; Liu &
162 Lapata, 2019; B. Schwartz, 2020). BERT is designed to be pre-trained on a very large scale
163 general purpose dataset and can then be used in its out-of-the-box pre-trained format, or one can
164 use transfer learning to adapt them for a specific domain and task with a process called fine-
165 tuning. During fine-tuning, the model is trained further on a downstream task with domain-
166 specific data. This process allows the models to work well even on tasks with fewer data
167 resources (Zaheer et al, 2021).

168 LLMs have been successfully applied to a variety of biomedical language tasks. For
169 example, by fine-tuning BERT with PubMed abstracts and clinical notes, Peng et al. (2019)
170 outperformed previous state-of-the-art on five biomedical tasks (e.g., similarity of two sentences
171 from Mayo Clinic clinical data). Researchers have also found success applying these models to
172 clinical language research. For instance, Balagopalan et al. (2020) fine-tuned BERT to detect
173 Alzheimer's disease from transcribed spontaneous speech. They found that BERT performed

174 better than a standard model based on hand-crafted features. Gale et al. (2021) fine-tuned a
175 variation of BERT called DistilBERT (Sanh et al., 2019) to automatically score commonly used
176 expressive language tasks on a diverse group of children (Autism Spectrum Disorder, Attention-
177 Deficit Hyperactivity Disorder, Developmental Language Disorder, and typical development;
178 age 5-9 years) with high accuracy (83-99%). In previous work developing ParAlg, our group
179 fine-tuned DistilBERT to automatically determine the semantic similarity of lexical paraphasias
180 to the target word with 95.3% accuracy (Salem et al., 2022).

181 While models like BERT have been very successful, one drawback is that they are
182 designed for relatively short sequences of words; in fact, BERT has a hard limit of taking
183 sequences of text of maximum length 512 tokens. Our data, which consists of retellings of the
184 Cinderella story, includes many sessions longer than that limit. In this work, we instead use a
185 recent LLM called BigBird (Zaheer et al., 2021) which was specifically designed to address this
186 limitation of BERT. Importantly, BigBird, like its predecessor BERT, was trained using “masked
187 language modeling”, a type of sentence cloze task. In this task, randomly selected words from
188 the corpus are masked (i.e., removed and replaced with a special blank token [MASK]), and the
189 model learns to fill in the blank and predict those masked words using the surrounding context,
190 allowing it to learn what words occur in what contexts. This task is in fact similar to our task at
191 hand: we want to predict what target word a person with aphasia was intending to say, given the
192 context of their discourse. Thus, considering the wide success of LLMs, the adaptation of this
193 model to long sequences, and the similarity of its training process to our task, we hypothesized
194 that BigBird would be a good fit for automatically predicting paraphasia targets in discourse.

195 Given that the current study represents a novel application of a LLM to data from a
196 clinical population, it is worthwhile to explore factors that might influence the accuracy of that

197 approach. It is generally accepted that PWA represent a heterogeneous group in terms of the
198 nature and severity of deficits exhibited during discourse production. For example, some
199 individuals on the mild end of the ability continuum may present with well-constructed
200 utterances during connected speech with only occasional hesitations and single word
201 paraphasias. On the other hand, people on the more severe end of the distribution may exhibit
202 morphosyntactic disturbances as well as significant manifestations of word retrieval deficits
203 including abandoned phrases, revisions, retracings, reformulations, as well as multiple
204 paraphasias. Therefore, given that the LLM relies on the surrounding context of a masked word
205 for prediction, it is conceivable that the success of the model may depend on overall aphasia
206 severity of the speaker. In addition to overall aphasia severity, the predictive utility of the LLM
207 may also depend on the nature of the syntactic deficits exhibited by people with aphasia.
208 Specifically, connected speech from PWA can be characterized as agrammatic or paragrammatic
209 (Butterworth & Howard, 1987; Goodglass, 1993; Saffran et al., 1989; Thompson et al., 1997).
210 Agrammatic speech is typically characterized by an overall reduction of grammatical
211 morphology, simplification of syntactic structure, and overreliance on content words, primarily
212 nouns. On the other hand, paragrammatism is associated with misuse of grammatical aspects
213 including inflectional morphology, significant word substitutions that cross word class, as well as
214 pronounced errors in word ordering. Finally, during discourse production, there are instances
215 where a speaker's intended target is clear, but that is not always the case, and different raters can
216 disagree. In this study, in addition to clinical factors, we investigated the performance of our
217 LLM as a function of the certainty with which raters can perform the same task.

218 **Purpose of Study**

238 256 transcripts from control participants without aphasia in AphasiaBank. Our data preparation
239 pipeline is illustrated in Figure 1. More details are provided in the sections below.

240 ***Paraphasia Identification***

241 Archival audiovisual recordings and CHAT transcript files (Codes for the Human
242 Analysis of Transcripts; MacWhinney, 2000) of the Cinderella story retell task were retrieved
243 from the English AphasiaBank database on May 4, 2022 for any and all PWA whose sample
244 contained at least one word-level error as annotated by AphasiaBank.¹ We defined paraphasias as
245 word-level errors made to the lemma of content words (i.e., nouns, verbs, adjectives, adverbs)
246 and excluded from target prediction all other kinds of word-level errors, including those related
247 to disfluency, morphological markings (e.g., plurality, tense), and non-content words (e.g.,
248 articles, pronouns). Referencing the CHAT manual (MacWhinney, 2000) accessed on April 13,
249 2022, we developed a list of word-level error codes for preliminary inclusion and exclusion.

250 ***Target Identification***

251 Target words were identified and annotated in ELAN transcription software (version 6.2),
252 using custom generated templates that also allowed for review of the retellings' transcripts as
253 well as playback of audiovisual recordings. To maximize transcript readability and efficacy for
254 this task, AphasiaBank transcripts were preprocessed to remove from view additional
255 annotations irrelevant to the task (e.g., utterance-level error coding) as well as the original
256 annotator's target prediction, if provided.

257 Target word identifications were completed by five trained student research assistants in
258 pseudorandom order under the supervision of a research SLP, resulting in a total of three

¹ Although the content of the transcripts is based on the AphasiaBank database on May 4, 2022, we applied updates to the clinical scores that were unavailable on AphasiaBank until December, 2022.

259 independent target identifications for each paraphasia. Research assistants were instructed to
260 watch the audiovisual recordings of the Cinderella story retell task and make their paraphasia
261 target predictions based on a number of contextual factors, including background knowledge
262 related to word production disorders and the Cinderella story. For each identified target, a
263 confidence rating ranging from 1 to 4 was assigned with 1 signifying very unconfident, 2
264 unconfident, 3 confident, and 4 very confident. In the process, research assistants flagged for
265 potential exclusion any word errors believed to be outside the scope of this project (e.g., the
266 predicted target is not a noun, verb, adjective, or adverb) or produced in the context of personal
267 commentary (e.g., a comment about the difficulty of the task, performance on the task, etc.).

268 Identified targets from our research assistants as well as AphasiaBank annotators were
269 automatically extracted and compiled for side-by-side comparison and resolution in a
270 spreadsheet. Discrepancies in target words and word errors flagged for exclusion were resolved
271 by a research SLP to arrive at a single, best target identification and in some cases multiple
272 viable target words were provided (e.g., shoe vs. slipper, coach vs. carriage). If there was
273 universal agreement among all three raters and AphasiaBank, then that target was not subject to
274 resolution. If there was disagreement among raters, rater confidence was low, and the resolver
275 could not arrive at a suitable prediction upon review, then the target was listed as “unknown”.
276 All paraphasia-target pairs were reviewed by the research SLP for phonological similarity and
277 whether or not an intermediary target was readily apparent (e.g., the paraphasia “bot”, where
278 “bot” could be interpreted as phonemic paraphasia of “boot”, the intermediary target, and “boot”
279 could be interpreted as a semantic paraphasia of “slipper”, the ultimate target). We calculated
280 average confidence scores (between the three research assistants) and percent agreement
281 (between the three research assistants and the original AphasiaBank target, where available) for

282 each identified target. After filtering to content word paraphasias and excluding paraphasias with
283 unknown targets, we were left with 353 Cinderella story sessions from 254 participants, with a
284 total of 2489 paraphasias.

285 *Session Text Cleaning*

286 We compiled our target identifications as well as human rater confidence and percent
287 agreement in the CHAT file format. We added our annotations within the “comment on main
288 line” markers specified in the CHAT manual, formatted in a structured notation (YAML) which
289 can be parsed in common programming languages such as Python. The following example shows
290 one such transcript, with our additional annotations highlighted in boldface type:

```
291 *PAR: and she rode off with the pmts@u [: prince] [% {target: a, agreement:  
292 1.0, confidence: 3.33}] [* p:n] . •680333_684666•
```

293 To prepare the transcripts for use with our LLM, we automated a process to convert the
294 transcripts to a more natural-looking written English. Motivated by the long-term goal of a fully
295 automated anomia system, we generally aimed to prepare the transcripts to look like those an
296 automatic speech recognition system would produce. Markings indicating prosodic (e.g. pauses)
297 and paralinguistic details (e.g. gestures) were removed. The CHAT format also uses special
298 markers to indicate phenomena peculiar to the spoken modality, such as retracing and repeats.
299 For situations like these, we omitted the special markers, but retained most of the spoken content,
300 though we discarded extraneous words that could be identified by simple rules (e.g. a list of filler
301 words like “um”).

302 In the AphasiaBank files, the transcripts are segmented into units called “utterances” or
303 “conversational units.” These units look similar to sentences—they are delimited by periods—
304 but tend to be shorter and more fragmentary, owing to the inherent differences between spoken

305 and written language. Especially as compared to the written text used to pre-train LLMs, the
306 utterance segmentation guidelines laid out by the CHAT manual would not reliably contain a
307 substantial amount of semantic context for our masked word prediction task. So, while popular
308 LLMs (e.g. BERT) typically process a sentence or two at a time, our transcripts do not divide
309 cleanly into sentences. Rather than attempt to redraw the AphasiaBank-provided utterance
310 boundaries to suit our task, we chose to prepare our data with a full context. In other words, for
311 each paraphasia shown to the LLM, the model was working with a participant’s complete
312 retelling of the Cinderella story.

313 Each paraphasia was prepared for training or testing by replacing it with a “blank” token
314 (also known as a “mask”) and filling in the other paraphasias in the session with the human
315 identified target word. The following example from above illustrates the cleaned sentence in
316 context, where the paraphasia has been replaced with a mask token:

317 ... and then and and she put her foot in the. and she rode off with the [MASK].
318 Cinderella was pretty girl. ...

319 During fine-tuning and testing, the model learned to fill in the blank of the mask token with the
320 most likely word given the context of the rest of the Cinderella story retelling.

321 ***Data Splitting***

322 We used ten-fold cross validation of the PWA data in order to reduce model overfitting.
323 That is, we divided the 2,489 instances into ten groups and trained ten separate models for each
324 experiment, in each of which one group was held out as testing data. This was done in such a
325 way that for each of the ten iterations, a participant’s responses were only in either the training
326 data or the testing data to prevent the models from learning participant-specific information, and
327 the distribution of Western Aphasia Battery-Revised (WAB-R; Kertesz, 2007) Aphasia Quotient

328 (AQ) scores in training and testing was as close as possible. When evaluating overall
329 performance, the results from the ten test set splits were concatenated, and performance on the
330 entire set of 2489 paraphasias was examined. The same ten-fold splits were used for all
331 experiments.

332 ***Control Data Augmentation***

333 To add additional training data for our experiments and reduce overfitting, we conducted
334 data augmentation (a method of adding synthetic data; see Feng et al., 2021 for more
335 background) on sessions of the Cinderella retelling task from control participants without
336 aphasia. We retrieved all files in AphasiaBank from control participants with a Cinderella story
337 task on April 12, 2022 and added synthetic paraphasias to these sessions. For each session, for
338 each utterance spoken by the participant, with a 20% chance we randomly assigned a content
339 word (one of: noun, verb, adjective, adverb) to be a “paraphasia” to be predicted. This left a
340 control dataset with 256 sessions from 248 participants, with a total of 2427 synthetic
341 paraphasias, which was very close to the number of paraphasias from the PWA data (2489). We
342 cleaned and prepared these sessions using the same process as for PWA data, described in the
343 subsection Session Text Cleaning.

344 **Model Training and Experiments**

345 In all experiments we used a pre-trained version of the LLM BigBird (Zaheer et al.,
346 2021). This model is a machine learning-based transformer model. Specifically, it is a sparse-
347 attention version of BERT designed for longer sequences of text. As previously mentioned, it
348 was pre-trained on masked language modeling. During masked language model training, the
349 model is given sentences from the corpus where 15% of the tokens are masked (i.e., removed
350 and replaced with a special non-word token, “[MASK]”), and the model attempts to predict what

351 those masked words were given the context of the surrounding sentence. By doing this on the
352 whole corpus of sentences, the model learns what words occur in what contexts. We accessed
353 this pre-trained BigBird from the HuggingFace transformer library (Wolf et al., 2020).

354 For each experiment (excluding the baseline experiment), we fine-tuned the LLM using
355 another masked language modeling task. Specifically, given the context of the whole Cinderella
356 story transcript, the model tried to fill in the blank of the mask token with the intended target.²
357 The model then compared that prediction with the human-determined ground truth intended
358 target (or the original word for control participants), and learned from its correct and incorrect
359 predictions. The fine-tuning process was repeated on the whole training data set until early-
360 stopping occurred, meaning performance stopped improving on a small portion of the testing
361 data that was held out. Once the model was fine-tuned, we tested it on the PWA paraphasias,
362 which were prepared in the same way as the training data, with each paraphasia sequentially
363 replaced with a mask, and all others filled in with their target. At test time, we pulled out the
364 model's top prediction, as well as its nineteen next most likely predictions, giving us its top
365 twenty predictions for the target, sorted from most likely to least likely. We considered more
366 than just the top prediction because there is inherent ambiguity in target identification, and in
367 future work we may consider multiple possible targets when classifying paraphasias in discourse.

368 We conducted four experiments using different preparations of training data, which are
369 summarized in Table 2. In Experiment 1, we used the pre-trained BigBird model without any
370 fine-tuning using Cinderella story data. We considered this our “baseline” model to beat. In
371 Experiment 2, we fine-tuned the LLM using just the Cinderella story sessions from control

² There exist certain subtleties to how this is done at a technical level, which we describe in detail in Appendix A. The precise manner in which we performed our masking, and ensuing prediction experiments, would be slightly different had we chosen a different neural model, but the overall methodology would be the same.

372 participants with synthetic paraphasias. In Experiment 3, the pre-trained model was fine-tuned
373 using Cinderella story sessions from PWA. Finally, in Experiment 4, the model was fine-tuned
374 using a combined data set of control participant data *and* PWA data.

375 **Evaluation**

376 We evaluated performance of the four experiments using accuracy. We calculated the
377 accuracy of “exact match” between the model’s top predicted intended word and the human
378 determined target word by counting up the number of matches and dividing by the total number
379 of test instances. Additionally, we calculated the accuracy within the top one-20 model
380 predictions. That is, we counted up how many times out of all test instances the human
381 determined target word was: the top model prediction (i.e., top one or exact match); the first or
382 second model prediction (top two); the first, second or third model prediction (top three); and so
383 on for up to 20 chances to predict the right target. We primarily compared accuracy within one
384 chance (exact match) and accuracy within five chances for the four experiments. We determined
385 whether disagreements between exact match accuracy of the models were significant using
386 McNemar’s test with continuity correction (McNemar, 1947).

387 First, we calculated accuracy on all 2489 paraphasias. To determine what factors
388 influenced model performance, we also calculated exact match and within five accuracy on
389 several different test set stratifications for each model. We calculated performance separately on
390 sessions from participants with WAB-R AQ above or below the median, participants with fluent
391 aphasia (Wernicke, Anomic, Conduction, or Transcortical Sensory aphasia, or those considered
392 “non aphasic” by the WAB-R) and non-fluent aphasia (Broca, Global, or Transcortical Motor
393 aphasia), test instances where the human raters had high confidence (above median) or low
394 confidence (below median) in intended target determination, and test instances where human

395 raters had perfect agreement in determining the intended target, or imperfect agreement. We
396 tested whether differences in performance between these stratifications were significant using
397 two-sided z-tests for independent proportions. Throughout, a p -value of <0.05 was retained as a
398 level of statistical significance.

399 **Results**

400 Accuracy results from Experiments 1-4 are shown in Tables 3, 4, 5, and 6, respectively.
401 Experiment 1, our baseline model, achieved 25.5% for exact match accuracy on all paraphasias.
402 Experiment 2, the model fine-tuned on control data, achieved 34.6% exact match accuracy.
403 Experiments 3 and 4 (fine-tuned on PWA data and controls plus PWA data respectively) both
404 achieved exact match accuracy of 46.8%, 21.3 points above the baseline model. According to
405 McNemar's test, Experiment 3 and Experiment 4's exact match accuracy levels were
406 significantly different than both Experiment 1 (the baseline model) and Experiment 2, all with p
407 < 0.001 . Experiment 3's exact match accuracy was not significantly different from Experiment
408 4's exact match accuracy ($p = 0.963$).

409 Figure 3 shows accuracy within the top 20 model predictions for all four experiments.
410 Accuracy of all experiments saw the sharpest increase within the top one (exact match) and top
411 five model predictions, and then slower increase when allowing the remaining 15 chances to find
412 the correct target. As stated previously, Experiments 3 and 4 achieved the highest performance of
413 46.8% exact match accuracy on all paraphasias. Considering within five accuracy, experiment 4
414 obtained 66.8% accuracy within its top five predictions, which was just one point higher than
415 Experiment 3, which obtained 65.7% accuracy within top five predictions. Regardless of the
416 number of top predicted targets we considered, the baseline performed the lowest, followed by
417 Experiment 2 (trained on controls), and then the two experiments fine-tuned with PWA data

418 were our highest performing models. When looking across accuracy within top one through 20
419 predictions, the difference in performance between Experiment 4 (fine-tuned on PWA and
420 controls data) and Experiment 3 (fine-tuned on PWA data) was an increase of just one point or
421 less. These findings indicate that performance between these two models was not significantly
422 different. So, without loss of generality, we discuss Experiment 4 in more detail below.

423 We explored the impact of clinical factors and intended target ambiguity on model
424 performance by sequentially calculating accuracy of the test set stratified by these factors.
425 Considering exact match accuracy, performance in Experiment 4 was higher (59.5%) on the
426 paraphasias with targets humans all agreed upon and lower (34.2%) on the paraphasias with less
427 than perfect agreement. A similar pattern emerged for human confidence, with higher accuracy
428 (60.5%) on paraphasias with targets humans were more confident at identifying and lower
429 accuracy (36.2%) on targets with lower human confidence. We also saw higher performance on
430 sessions where the participant had a WAB-R AQ higher than the median (52.7% accuracy)
431 versus those where the participant had a WAB-R AQ below the median (41.6% accuracy).
432 Similarly, we saw higher performance on the participants with fluent aphasia (48.7% accuracy)
433 than the participants with non-fluent aphasia (41.2% accuracy). Overall, the highest accuracy out
434 of all test sets was on the paraphasias with high human confidence in target determination. For
435 each of these four comparisons, the two test set stratifications (e.g., perfect human agreement vs
436 imperfect human agreement) obtained significantly different performance levels according to the
437 two-sided *z*-test for independent proportions (see Supplemental Table 1 in the Supplemental
438 Material). *P*-values were all ≤ 0.001 except for the fluent versus non-fluent stratification, which
439 had $p = 0.016$. The same directions of performance difference were seen for the accuracy within
440 the top five predictions of these comparisons. The highest within-five accuracy out of all test set

441 stratifications was also seen for the above median human confidence paraphasias, which
442 Experiment 4 got correct 76.8% of the time within the top five model predictions.

443 **Discussion**

444 In this study, we trained a LLM to automatically predict the intended targets for
445 paraphasias in discourse during the Cinderella story retelling task. We tried various training data
446 configurations and our two best performing experiments were fine-tuned using PWA data, with
447 or without controls data, and achieved exact match accuracy 47%, and accuracy within top five
448 predictions between 66-67%. Considering just one of these (Experiment 4, fine-tuned on PWA
449 and controls data), the model performed better on paraphasias which had targets that were easier
450 for humans to identify. It also performed better on paraphasias from participants with less severe
451 aphasia and fluent aphasia. Overall, this work produced a relatively high performing model for
452 automatically determining paraphasia targets in connected speech, while just using the
453 surrounding context.

454 Our baseline model achieved an overall exact match accuracy of 25.5%. This model,
455 which was not fine-tuned to our data at all, was able to use its general-purpose recognition of
456 language patterns to make some correct predictions, without having been exposed to the specific
457 vocabulary and structure of the Cinderella story retellings. It is likely that the original corpus of
458 text used in pre-training the LLM would have included examples of various forms of the
459 Cinderella story, but to a much lesser degree had it been fine-tuned to it. The model used in
460 Experiment 2, fine-tuned using data from control-group participants with the addition of
461 synthesized paraphasias, improved by almost ten points beyond the baseline model with exact
462 match accuracy 34.6%. In this experiment, the pre-trained LLM was specifically exposed to the
463 vocabulary and structure of the Cinderella story, as well as the general task of filling in words in

464 it, but it was not exposed to any real-world examples of paraphasias. In contrast, Experiment 3,
465 fine-tuned on just PWA data, saw a 21 point increase in exact match accuracy over the baseline
466 model. Thus, training the model for this task required not just exposing the pre-trained model to
467 the vocabulary of the Cinderella story, but also specifically examples of real-world paraphasias
468 that occur in that task. Somewhat surprisingly, the model using both PWA data and controls data
469 (Experiment 4) did not improve beyond the model fine-tuned with just PWA data (Experiment
470 3). This likely indicates that the PWA data gave enough of that vocabulary knowledge to the
471 LLM, and the controls data did not provide any further information. However, more work could
472 be done to synthesize paraphasias in the controls data to make them more similar to real-world
473 paraphasias. As described in the Control Data Augmentation subsection, we attempted to make
474 them more “realistic” by only making content words paraphasias, but there are other possibilities
475 that could be explored in future work: adding synthetic re-tracings, for example, as well as
476 utilizing psycholinguistic variables (e.g. length in phonemes, frequency of occurrence,
477 imageability, etc.) to produce more realistic synthetic training data.

478 We found that human certainty about paraphasia targets was associated with model
479 performance. Specifically, our best performing model (Experiment 4) performed significantly
480 better on paraphasias with targets that humans were more confident on or had perfect agreement
481 on. This association is reassuring and acts as a simple validity check, since it indicates that our
482 trained models had an easier time with the more obvious targets. There is inherent ambiguity in
483 determining targets for paraphasias in discourse. Half of the paraphasias had percent agreement
484 below 100%, and in fact, average percent agreement on target identification was 76.8%.
485 Moreover, this percentage agreement is only on the paraphasias for which we were able to
486 resolve a target and excludes targets where ground truth could not be determined. Considering

487 76.8% agreement as a stand-in for the obtainable human accuracy on this task, obtaining 46.8%
488 accuracy on paraphasias with known targets appears high. Relatedly, while the LLM was
489 designed to rely exclusively on the surrounding language for its predictions, human raters had
490 access to audiovisual recordings and transcripts and thus were able predict targets utilizing
491 additional sources of information such as phonological similarity and gestures.

492 We also found that, as expected, Experiment 4 saw significantly different performance
493 between participants with above median severity and below median severity, according to the
494 WAB-R AQ, with exact match accuracy 8.4% higher on participants with less severe aphasia.
495 The exact reason for this difference in performance, whether it be factors such as increased
496 occurrence of abandoned phrasings or multiple paraphasias from more severe participants, could
497 be examined further. Relatedly, Experiment 4 performed significantly better on fluent
498 participants than non-fluent participants. Our fluent (Wernicke, Anomic, Conduction,
499 Transcortical Sensory, or non-aphasic by WAB-R) and non-fluent (Broca, Global, or
500 Transcortical Motor) stratifications acted as a proxy for capturing paragrammatic and
501 agrammatic aphasia types respectively. The non-fluent (and perhaps agrammatic) participants
502 may have harder to identify targets because of a lack of content words and context for the LLM
503 to rely on. However, we recognize limitations with this approach. We had substantially fewer
504 training examples from non-fluent participants (449 paraphasias) than fluent participants (1666
505 paraphasias), which may have impacted that performance difference. Additionally, classification
506 based on the WAB-R is not perfect as there is both classification error and considerable
507 heterogeneity within groups. Finally, the mapping between fluency types and type of
508 grammatical deficits is not perfect. Nonetheless, these stratifications of the test set provided
509 some clues on what features impact performance and where the models can improve. It is also

510 possible that, particularly with more training data, separate models trained for use on specific
511 types of aphasia could see higher performance and better clinical utility.

512 After our quantitative analyses, we conducted an informal review of Experiment 4’s
513 output, observing some of the more apparent patterns. Some errors were rather unsurprising, like
514 swapping similar verbs (e.g. “sweeping” for “cleaning”). Others were random and garbled (e.g.
515 “Cinderellaipper” for “slipper”) and obviously a consequence of the text encoding constraints
516 (see Appendix A). Where larger patterns stood out, though, they tended to point to a few
517 peculiarities of the dataset.

518 For example, about 26% of the samples in our dataset involved paraphasias which
519 AphasiaBank had annotated as part of a “retracing” event. Retracing is when a speaker abandons
520 a segment of speech and then retries that segment again (e.g. “Cinderella <put on> [//] tried on
521 the slipper”). When a target word was involved in a retracing event, our LLM’s top-five
522 accuracy for target prediction increased to 80% (vs. 62% when it was not). Since we fill in all
523 the paraphasia targets except the current target (see Model Training and Experiments) any other
524 paraphasias in the immediate context would have been filled in with the correct target word,
525 which provides an advantage for the task at hand. However, this can also work against the model
526 when a target was not actually a part of a retracing event. Informally, we observed that the model
527 sometimes incorrectly chose a word from the immediate context, predicting a retracing where
528 there was none.

529 Another peculiarity of our dataset was the storytelling task itself, marked by a Cinderella-
530 centric distribution of target words. Out of the 523 unique target words, about 30% of targets
531 were one of five salient words from the fairy tale (“Cinderella,” “prince,” “slipper,” “ball,” or
532 “godmother”). For the most common word, “Cinderella” (265 examples, 11% of total), the LLM

533 was correct 170 times (64%) within the first guess and 227 times (86%) within five guesses.
534 However, this advantage was largely canceled out when the correct target was not the
535 protagonist's name: the model incorrectly predicted "Cinderella" 157 times as a first guess, and
536 443 times as a top-five guess. Looking at a subset of the data unaffected by the above factors, we
537 find 233 samples which had a unique target word (occurring only once) and also were not part of
538 a retracing event. The first-guess accuracy for these samples dropped from 39% to 15% between
539 the baseline and fine-tuned models, respectively.

540 These three patterns—predicting targets that were repeats from the surrounding context,
541 frequently predicting common words from the task, and having difficulty with more rare
542 words—are all consequences of fine-tuning a model. There is a tradeoff between the desirable
543 outcome of improving performance by following common patterns in the training data and the
544 loss in performance when new data points break that pattern; this is known as the bias-variance
545 tradeoff and is well documented in machine learning literature (Geman et al., 1992; Belkin et al.,
546 2019). We employed techniques to reduce overfitting to the training data (data augmentation,
547 cross validation, early stopping), but more strategies could be explored.

548 Given the architecture of our LLM, we suspect various utterance-related measures would
549 also influence target prediction accuracy for a given speaker and/or utterance. For example, we
550 would predict that speakers with longer utterances, i.e., mean length of utterance in words, would
551 be supplying the model with more linguistic information and therefore increase the likelihood of
552 target prediction success. Another set of hypotheses relates to the quality of the speaker's
553 utterances in terms of completeness, percentage of utterances that are complete sentences;
554 correctness, percentage of syntactically and/or semantically correct sentences; complexity,
555 number of embedded clauses per sentence, sentence complexity ratio (Thompson et al., 1995),

556 and verbs per utterance; as well as lexical diversity measures like type-token ratio and vocd
557 (Malvern, Richards, Chipere, & Purán, 2004). As mentioned previously, these factors may
558 further explain why performance was affected by fluency and aphasia severity. All of the
559 aforementioned speaker outcome measures can be automatically calculated using CLAN
560 software (MacWhinney, 2000), and we posit all of them would be positive predictors of target
561 prediction accuracy. To deepen our understanding and interpretation of our results, therefore, a
562 future direction of this work is to employ a generalized linear mixed effects model to test these
563 hypothesizes and quantify the magnitude of any significant predictors.

564 There are many other future directions for this work. Currently, we achieve 46.8%
565 accuracy at predicting paraphasia targets by just using the text of the story, excluding the
566 paraphasia. However, in many cases the details of the paraphasia itself would provide useful
567 information for determining the target. In future work, we plan to develop a model that uses both
568 the semantic context surrounding the paraphasia as well as the phonemes of the paraphasia itself
569 to further improve predictive utility. Considering the difficulty of the task at hand, our
570 performance using just the surrounding language is surprisingly high. However, as mentioned,
571 the Cinderella retelling task is a highly constrained activity, with a much smaller expected target
572 vocabulary than in standard speech. In the context of test and scale development for clinical
573 assessment, when batteries typically include one or two specific stories, gains due to the
574 constrained nature of the stimuli are advantageous. However, in the future, it could be beneficial
575 to train models for less constrained tasks or more naturalistic speech. Additionally, these findings
576 open up possibilities for novel applications that extend beyond assessment, such as augmentative
577 and alternative communication systems. Finally, as previously mentioned, we intend to
578 eventually extend ParAlg, our automated system for classifying paraphasias, to use it on

579 discourse. This work generates a preliminary model for the first step in that process:
580 automatically identifying the most likely targets for paraphasias in discourse.

581 **Acknowledgments**

582 This work was supported by National Institute on Deafness and Other Communication
583 Disorders Grant R01DC015999 (Principal Investigators: Steven Bedrick and Gerasimos
584 Fergadiotis). We would also like to thank Mia Cywinski, Samuel Hedine, Lidiya Khoroshenikh,
585 Jonathan Madrigal, and Anya Russell for their crucial work identifying paraphasia targets.

586 **Data Availability Statement**

587 Data from PWA and controls is available from AphasiaBank to all members of the AphasiaBank
588 consortium group (<https://aphasia.talkbank.org/>).

589 **References**

590 Adams, J., Bedrick, S., Fergadiotis, G., Gorman, K., & van Santen, J. (2017). Target word
591 prediction and paraphasia classification in spoken discourse. *BioNLP 2017*, 1–8.
592 <https://doi.org/10.18653/v1/W17-2301>

593 Balagopalan, A., Eyre, B., Rudzicz, F., & Novikova, J. (2020). To BERT or not to BERT:
594 Comparing speech and language-based approaches for Alzheimer’s disease detection.
595 *Interspeech 2020*, 2167–2171. <https://doi.org/10.21437/Interspeech.2020-2557>

596 Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning
597 practice and the classical bias-variance trade-off. *Proceedings of the National Academy*
598 *of Sciences*, 116(32), 15849–15854. <https://doi.org/10.1073/pnas.1903070116>

599

- 600 Bock, K. (1995). Sentence production: From mind to mouth. In *Speech, Language, and*
601 *Communication* (pp. 181–216). Elsevier. [https://doi.org/10.1016/B978-012497770-](https://doi.org/10.1016/B978-012497770-9/50008-X)
602 [9/50008-X](https://doi.org/10.1016/B978-012497770-9/50008-X)
- 603 Breimaier, H. E., Heckemann, B., Halfens, R. J. G., & Lohrmann, C. (2015). The Consolidated
604 Framework for Implementation Research (CFIR): A useful theoretical framework for
605 guiding and evaluating a guideline implementation process in a hospital-based nursing
606 practice. *BMC Nursing*, *14*(1), 43. <https://doi.org/10.1186/s12912-015-0088-4>
- 607 Bryant, L., Spencer, E., Ferguson, A., Craig, H., Colyvas, K., & Worrall, L. (2013).
608 Propositional Idea Density in aphasic discourse. *Aphasiology*, *27*(8), 992–1009.
609 <https://doi.org/10.1080/02687038.2013.803514>
- 610 Butterworth, B., & Howard, D. (1987). Paragrammatisms. *Cognition*, *26*(1), 1–37.
611 [https://doi.org/10.1016/0010-0277\(87\)90012-6](https://doi.org/10.1016/0010-0277(87)90012-6)
- 612 Casilio, M., Fergadiotis, G., Salem, A. C., Gale, R., McKinney-Bock, K., & Bedrick, S. (2023).
613 ParAlg: A paraphasia algorithm for multinomial classification of picture naming errors.
614 *Journal of Speech, Language, and Hearing Research*.
615 https://doi.org/10.1044/2022_JSLHR-22-00255
- 616 Chatzoudis, G., Plitsis, M., Stamouli, S., Dimou, A., Katsamanis, N., & Katsouros, V. (2022).
617 Zero-shot cross-lingual aphasia detection using automatic speech recognition.
618 *Interspeech 2022*, 2178–2182. <https://doi.org/10.21437/Interspeech.2022-10681>
- 619 Cruice, M., Worrall, L., Hickson, L., & Murison, R. (2003). Finding a focus for quality of life
620 with aphasia: Social and emotional health, and psychological well-being. *Aphasiology*,
621 *17*(4), 333–353. <https://doi.org/10.1080/02687030244000707>

- 622 Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C.
623 (2009). Fostering implementation of health services research findings into practice: A
624 consolidated framework for advancing implementation science. *Implementation Science*,
625 4(1), 50. <https://doi.org/10.1186/1748-5908-4-50>
- 626 Day, M., Dey, R. K., Baucum, M., Paek, E. J., Park, H., & Khojandi, A. (2021). Predicting
627 severity in people with aphasia: A natural language processing and machine learning
628 approach. *2021 43rd Annual International Conference of the IEEE Engineering in*
629 *Medicine & Biology Society (EMBC)*, 2299–2302.
630 <https://doi.org/10.1109/EMBC46164.2021.9630694>
- 631 Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production.
632 *Psychological Review*, 93(3), 283–321. <https://doi.org/10.1037/0033-295X.93.3.283>
- 633 Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production:
634 Lexical access and grammatical encoding. *Cognitive Science*, 23(4), 517–542.
635 https://doi.org/10.1207/s15516709cog2304_6
- 636 Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep
637 bidirectional transformers for language understanding. *Proceedings of the 2019*
638 *Conference of the North American Chapter of the Association for Computational*
639 *Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–
640 4186. <https://doi.org/10.18653/v1/N19-1423>
- 641 Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A
642 survey of data augmentation approaches for NLP. Findings of the Association for
643 Computational Linguistics: ACL-IJCNLP 2021, 968–988.
644 <https://doi.org/10.18653/v1/2021.findings-acl.84>

- 645 Fergadiotis, G., Gorman, K., & Bedrick, S. (2016). Algorithmic classification of five
646 characteristic types of paraphasias. *American Journal of Speech-Language Pathology*,
647 25(4S). https://doi.org/10.1044/2016_AJSLP-15-0147
- 648 Fergadiotis, G., Kapantzoglou, M., Kintz, S., & Wright, H. H. (2019). Modeling confrontation
649 naming and discourse informativeness using structural equation modeling. *Aphasiology*,
650 33(5), 544–560. <https://doi.org/10.1080/02687038.2018.1482404>
- 651 Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia
652 across discourse elicitation tasks. *Aphasiology*, 25(11), 1414–1430.
653 <https://doi.org/10.1080/02687038.2011.603898>
- 654 Fergadiotis, G., Wright, H. H., & Capilouto, G. J. (2011). Productive vocabulary across
655 discourse types. *Aphasiology*, 25(10), 1261–1278.
656 <https://doi.org/10.1080/02687038.2011.606974>
- 657 Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative
658 discourse of people with aphasia. *American Journal of Speech-Language Pathology*,
659 22(2). [https://doi.org/10.1044/1058-0360\(2013/12-0083](https://doi.org/10.1044/1058-0360(2013/12-0083)
- 660 Forbes, M., Fromm, D., Holland, A., & MacWhinney, B. (2014). EVAL: A tool for clinicians
661 from AphasiaBank. *Clinical Aphasiology Conference, St. Simons Island, GA*.
- 662 Fraser, K., Rudzicz, F., Graham, N., & Rochon, E. (2013). Automatic speech recognition in the
663 diagnosis of primary progressive aphasia. *Proceedings of the Fourth Workshop on*
664 *Speech and Language Processing for Assistive Technologies*, 47–54.
665 <https://www.aclweb.org/anthology/W13-3909>

- 666 Gale, R., Bird, J., Wang, Y., van Santen, J., Prud'hommeaux, E., Dolata, J., & Asgari, M. (2021).
667 Automated scoring of tablet-administered expressive language tests. *Frontiers in*
668 *Psychology*, 12, 668401. <https://doi.org/10.3389/fpsyg.2021.668401>
- 669 Gale, R. C., Fleegle, M., Fergadiotis, G., & Bedrick, S. (2022). The Post-Stroke Speech
670 Transcription (PSST) Challenge. *Proceedings of the RaPID Workshop - Resources and*
671 *ProcessIng of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with*
672 *Various Forms of Cognitive/Psychiatric/Developmental Impairments - within the 13th*
673 *Language Resources and Evaluation Conference*, 41–55.
674 <https://aclanthology.org/2022.rapid-1.6>
- 675 Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the Bias/Variance
676 dilemma. *Neural Computation*, 4(1), 1–58. <https://doi.org/10.1162/neco.1992.4.1.1>
- 677 Goodglass, H. (1993). *Understanding aphasia*. Academic Press.
- 678 Goodglass, H., & Wingfield, A. (Eds.). (1997). *Anomia: Neuroanatomical and cognitive*
679 *correlates*. Academic Press.
- 680 Hickin, J., Best, W., Herbert, R., Howard, D., & Osborne, F. (2001). Treatment of word retrieval
681 in aphasia: Generalisation to conversational speech. *International Journal of Language &*
682 *Communication Disorders*, 36(s1), 13–18. <https://doi.org/10.3109/13682820109177851>
- 683 Kertesz, A. (2012). Western Aphasia Battery—Revised [Data set]. American Psychological
684 Association. <https://doi.org/10.1037/t15168-000>
- 685 Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword
686 tokenizer and detokenizer for Neural Text Processing. *Proceedings of the 2018*
687 *Conference on Empirical Methods in Natural Language Processing: System*
688 *Demonstrations*, 66–71. <https://doi.org/10.18653/v1/D18-2012>

- 689 Le, D., Licata, K., & Mower Provost, E. (2018). Automatic quantitative analysis of spontaneous
690 aphasic speech. *Speech Communication, 100*, 1–12.
691 <https://doi.org/10.1016/j.specom.2018.04.001>
- 692 Le, D., Licata, K., & Provost, E. M. (2017). Automatic paraphasia detection from aphasic
693 speech: A preliminary study. *Proc. Interspeech 2017*, 294–298.
694 <https://doi.org/10.21437/Interspeech.2017-626>
- 695 Le, D., & Provost, E. M. (2016). Improving automatic recognition of aphasic speech with
696 AphasiaBank. *Interspeech 2016*, 2681–2685. [https://doi.org/10.21437/Interspeech.2016-](https://doi.org/10.21437/Interspeech.2016-213)
697 [213](https://doi.org/10.21437/Interspeech.2016-213)
- 698 Levelt, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences, 3*(6), 223–
699 232. [https://doi.org/10.1016/S1364-6613\(99\)01319-4](https://doi.org/10.1016/S1364-6613(99)01319-4)
- 700 Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech
701 production. *Behavioral and Brain Sciences, 22*(01).
702 <https://doi.org/10.1017/S0140525X99001776>
- 703 Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. *Proceedings of the*
704 *2019 Conference on Empirical Methods in Natural Language Processing and the 9th*
705 *International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*,
706 3728–3738. <https://doi.org/10.18653/v1/D19-1387>
- 707 Lowerre, T. B. (1976). The Harpy speech recognition system [Ph.D. Thesis]. Carnegie Mellon
708 University.
- 709 MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of
710 syntactic ambiguity resolution. *Psychological Review, 101*(4), 676–703.
711 <https://doi.org/10.1037/0033-295X.101.4.676>

- 712 MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence
713 Erlbaum Associates.
- 714 MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for
715 studying discourse. *Aphasiology*, 25(11), 1286–1307.
716 <https://doi.org/10.1080/02687038.2011.589893>
- 717 Malvern, D. (Ed.). (2008). *Lexical diversity and language development: Quantification and*
718 *assessment*. Palgrave Macmillan.
- 719 Mayer, J., & Murray, L. (2003). Functional measures of naming in aphasia: Word retrieval in
720 confrontation naming versus connected speech. *Aphasiology*, 17(5), 481–497.
721 <https://doi.org/10.1080/02687030344000148>
- 722 McNemar, Q. (1947). Note on the sampling error of the difference between correlated
723 proportions or percentages. *Psychometrika*, 12(2), 153–157.
724 <https://doi.org/10.1007/BF02295996>
- 725 Miller, J., & Iglesias, A. (2012). Systematic Analysis of Language Transcripts (SALT), research
726 version 2012 [computer software]. SALT Software, LLC.
- 727 Papathanasiou, I., & Coppens, P. (2017). Disorders of word production. In *Aphasia And Related*
728 *Neurogenic Communication Disorders* (1st ed., pp. 169–195). Jones & Bartlett Learning.
- 729 Pashek, G. V., & Tompkins, C. A. (2002). Context and word class influences on lexical retrieval
730 in aphasia. *Aphasiology*, 16(3), 261–286. <https://doi.org/10.1080/02687040143000573>
- 731 Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing:
732 An evaluation of BERT and ELMo on ten benchmarking datasets. *Proceedings of the*
733 *18th BioNLP Workshop and Shared Task*, 58–65. <https://doi.org/10.18653/v1/W19-5006>

- 734 Perez, M., Aldeneh, Z., & Provost, E. M. (2020). Aphasic speech recognition using a mixture of
735 speech intelligibility experts. *Interspeech 2020*, 4986–4990.
736 <https://doi.org/10.21437/Interspeech.2020-2049>
- 737 Rabin, L., Barr, W., & Burton, L. (2005). Assessment practices of clinical neuropsychologists in
738 the United States and Canada: A survey of INS, NAN, and APA Division 40 members.
739 *Archives of Clinical Neuropsychology*, 20(1), 33–65.
740 <https://doi.org/10.1016/j.acn.2004.02.005>
- 741 Richardson, J. D., Hudspeth Dalton, S. G., Fromm, D., Forbes, M., Holland, A., & MacWhinney,
742 B. (2018). The relationship between confrontation naming and story gist production in
743 aphasia. *American Journal of Speech-Language Pathology*, 27(1S), 406–422.
744 https://doi.org/10.1044/2017_AJSLP-16-0211
- 745 Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia
746 Naming Test: Scoring and rationale. *Clinical Aphasiology*, 24, 121–133.
- 747 Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic
748 production: Procedure and data. *Brain and Language*, 37(3), 440–479.
749 [https://doi.org/10.1016/0093-934X\(89\)90030-8](https://doi.org/10.1016/0093-934X(89)90030-8)
- 750 Salem, A. C., Gale, R., Casilio, M., Fleegle, M., Fergadiotis, G., & Bedrick, S. (2022). Refining
751 semantic similarity of paraphasias using a contextual language model. *Journal of Speech,*
752 *Language, and Hearing Research*, 1–15. https://doi.org/10.1044/2022_JSLHR-22-00277
- 753 Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT:
754 Smaller, faster, cheaper and lighter. <https://doi.org/10.48550/ARXIV.1910.01108>

- 755 Schwartz, B. (2020, October 15). Google: BERT now used on almost every English query.
756 *Search Engine Land*. [https://searchengineland.com/google-bert-used-on-almost-every-](https://searchengineland.com/google-bert-used-on-almost-every-english-query-342193)
757 [english-query-342193](https://searchengineland.com/google-bert-used-on-almost-every-english-query-342193)
- 758 Schwartz, M., Dell, G., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the
759 interactive two-step model of lexical access: Evidence from picture naming. *Journal of*
760 *Memory and Language*, 54(2), 228–264. <https://doi.org/10.1016/j.jml.2005.10.001>
- 761 Simmons-Mackie, N., Threats, T. T., & Kagan, A. (2005). Outcome assessment in aphasia: A
762 survey. *Journal of Communication Disorders*, 38(1), 1–27.
763 <https://doi.org/10.1016/j.jcomdis.2004.03.007>
- 764 Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests:*
765 *Administration, norms, and commentary* (E. M. S. Sherman, E. Strauss, & O. Spreen,
766 Eds.; 3rd ed). Oxford University Press.
- 767 Tabor, W., Juliano, C., & Tanenhaus, M. K. (1997). Parsing in a dynamical system: An
768 attractor-based account of the interaction of lexical and structural constraints in sentence
769 processing. *Language and Cognitive Processes*, 12(2–3), 211–271.
770 <https://doi.org/10.1080/016909697386853>
- 771 Thompson, C. K., Shapiro, L. P., Tait, M. E., Jacobs, B., Schneider, S. L., & Ballard, K. (1995).
772 A system for the linguistic analysis of agrammatic language production. *Brain and*
773 *Language*, 51(1), 124–129.
- 774 Thompson, C. K., Lange, K. L., Schneider, S. L., & Shapiro, L. P. (1997). Agrammatic and non-
775 brain-damaged subjects' verb and verb argument structure production. *Aphasiology*,
776 11(4–5), 473–490. <https://doi.org/10.1080/02687039708248485>

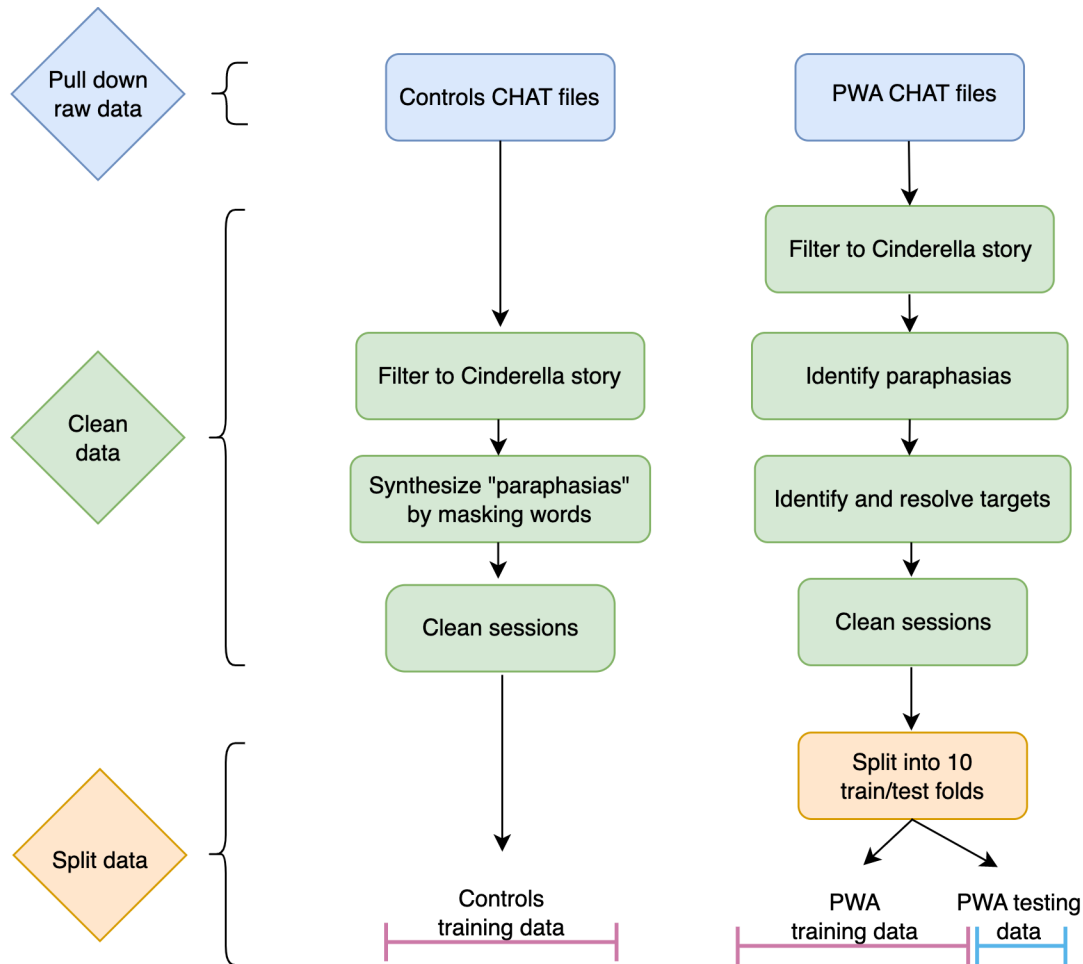
- 777 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., &
778 Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International*
779 *Conference on Neural Information Processing Systems*, 6000–6010.
- 780 Walker, G. M., & Schwartz, M. F. (2012). Short-form Philadelphia Naming Test: Rationale and
781 empirical evaluation. *American Journal of Speech-Language Pathology*, 21(2).
782 [https://doi.org/10.1044/1058-0360\(2012/11-0089\)](https://doi.org/10.1044/1058-0360(2012/11-0089))
- 783 Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf,
784 R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J.,
785 Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-art
786 natural language processing. *Proceedings of the 2020 Conference on Empirical Methods*
787 *in Natural Language Processing: System Demonstrations*, 38–45.
788 <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- 789 Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula,
790 A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for longer
791 sequences. *Proceedings of the 34th International Conference on Neural Information*
792 *Processing Systems*.

793

Figures

794 **Figure 1**

795 *Data preparation pipeline*



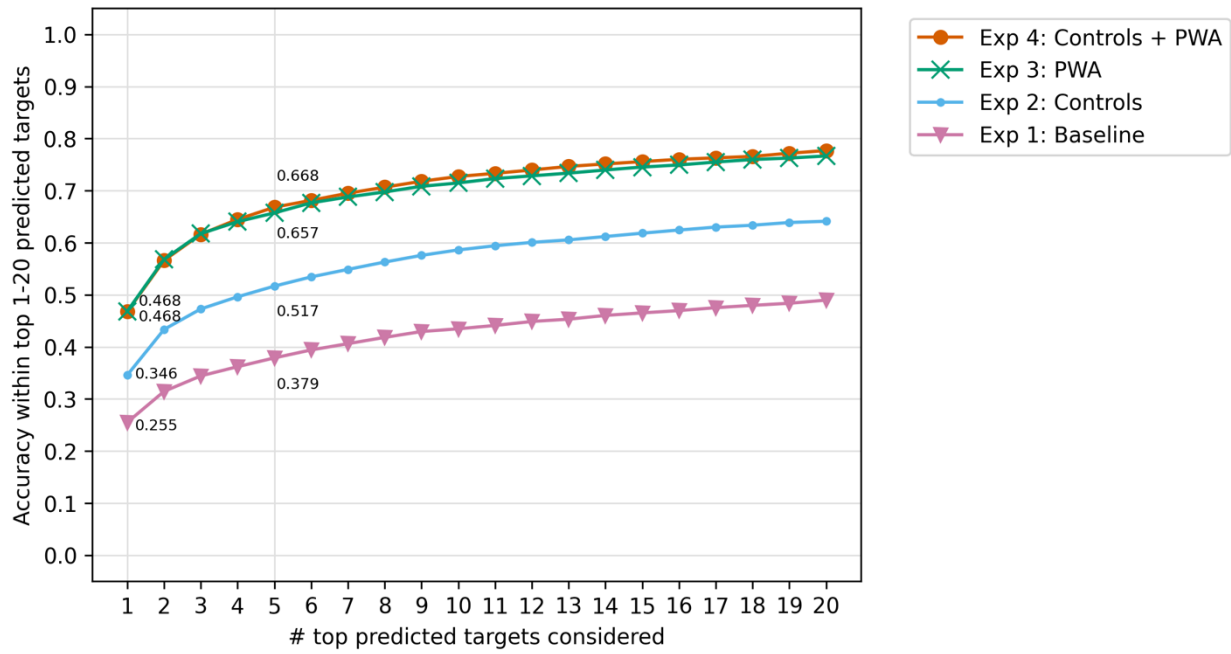
796

797 *Note.* CHAT stands for Codes for the Human Analysis of Transcripts, and is a format for

798 transcription. PWA stands for people with aphasia.

799 **Figure 2**

800 *Accuracy within top 1-20 predicted targets for experiments 1-4*



801

802 *Note.* PWA stands for people with aphasia.

803

Tables

804 **Table 1**

805 *Clinical and demographic information for the 254 participants at their first session.*

Characteristic	Value
Age (years)	
<i>M (SD)</i>	61.916 (12.408)
Min - Max	25.600 - 91.718
Missing (<i>N</i>)	24
Gender	
M (<i>N</i>)	133
F (<i>N</i>)	100
Missing (<i>N</i>)	21
Race	
White (<i>N</i>)	201
African American (<i>N</i>)	23
Asian (<i>N</i>)	2
Hispanic/Latino (<i>N</i>)	5
Native Hawaiian/ Pacific Islander (<i>N</i>)	1
Mixed (<i>N</i>)	1
Unavailable (<i>N</i>)	21
Education (years)	
<i>M (SD)</i>	15.498 (2.828)
Min - Max	8.000 - 25.000
Missing (<i>N</i>)	31
Aphasia duration	
<i>M (SD)</i>	5.429 (4.829)
Min - Max	0.080 - 30.000
Missing (<i>N</i>)	24
WAB-R AQ	
<i>M (SD)</i>	72.271 (17.992)
Min - Max	10.800 - 99.600
Missing (<i>N</i>)	11
BNT-SF	
<i>M (SD)</i>	7.369 (4.512)
Min - Max	0.000 - 15.000
Missing (<i>N</i>)	32
VNT	
<i>M (SD)</i>	15.000 (6.275)
Min - Max	0.000 - 22.000
Missing (<i>N</i>)	32

806 *Note.* WAB-R AQ is the Western Aphasia Battery-Revised Aphasia Quotient (Kertesz, 2012).

807 BNT-SF is the raw score from the Boston Naming Test-Short Form (Kaplan et al., 2001). VNT

808 is the raw score from the Verb Naming Test (Cho-Reyes et al., 2012).

809 **Table 2**

810 *Descriptions of experiments 1-4*

Experiment Number	Experiment Name	Description	Training data	Testing data
1	Baseline	Pre-trained LLM, without any fine-tuning to our data	N/A	PWA testing data
2	Controls	Pre-trained LLM, fine-tuned using all data from the control participants of the Cinderella story task	Controls training data	PWA testing data
3	PWA	Pre-trained LLM, fine-tuned using all PWA data from the Cinderella story task	PWA training data	PWA testing data
4	Controls + PWA	Pre-trained LLM, fine-tuned using all data from the control participants and PWA, from the Cinderella story task	Controls training data + PWA training data	PWA testing data

811 *Note.* PWA stands for people with aphasia. LLM stands for large language model. Note that all

812 models are tested on PWA testing data.

813 **Table 3**

814 *Experiment 1: Baseline*

Test set	Number of paraphasias	Accuracy exact match	Accuracy within 5
All paraphasias	2489	0.255	0.379
Human agreement = 100%	1244	0.309	0.405
Human agreement < 100%	1245	0.201	0.353
Human confidence > median (3.3)	1089	0.319	0.419
Humans confidence <= median (3.3)	1400	0.206	0.348
WAB-R AQ > median (74.6)	1039	0.294	0.410
WAB-R AQ <= median (74.6)	1076	0.204	0.325
Fluent participants	1666	0.261	0.385
Non-fluent participants	449	0.198	0.301

815 *Note.* WAB-R AQ is the Western Aphasia Battery-Revised Aphasia Quotient (Kertesz, 2012).

816 Fluent participants are those with Wernicke, Anomic, Conduction, or Transcortical Sensory

817 aphasia, or those considered “non aphasic” by the WAB-R. Non-fluent participants are those

818 with the Broca, Global, or Transcortical Motor aphasia. 48 out of 353 total sessions had

819 unavailable WAB-R results and were excluded just from analyses involving WAB-R scores.

820 Accuracy exact match refers to the top model prediction of target word matching the human-

821 identified target word. Accuracy within 5 refers to the human-identified target word being one of

822 the top five model predictions.

823 **Table 4**

824 *Experiment 2: Fine-tuned on controls data*

Test set	Number of paraphasias	Accuracy exact match	Accuracy within 5
All paraphasias	2489	0.346	0.517
Human agreement = 100%	1244	0.436	0.600
Human agreement < 100%	1245	0.255	0.434
Human confidence > median (3.3)	1089	0.453	0.614
Humans confidence <= median (3.3)	1400	0.263	0.441
WAB-R AQ > median (74.6)	1039	0.398	0.580
WAB-R AQ <= median (74.6)	1076	0.290	0.453
Fluent participants	1666	0.362	0.543
Non-fluent participants	449	0.274	0.414

825
826 *Note.* WAB-R AQ is the Western Aphasia Battery-Revised Aphasia Quotient (Kertesz, 2012).
827 Fluent participants are those with Wernicke, Anomic, Conduction, or Transcortical Sensory
828 aphasia, or those considered “non aphasic” by the WAB-R. Non-fluent participants are those
829 with the Broca, Global, or Transcortical Motor aphasia. 48 out of 353 total sessions had
830 unavailable WAB-R results and were excluded just from analyses involving WAB-R scores.
831 Accuracy exact match refers to the top model prediction of target word matching the human-
832 identified target word. Accuracy within 5 refers to the human-identified target word being one of
833 the top five model predictions.

834 **Table 5**

835 *Experiment 3: Fine-tuned on PWA data*

Test set	Number of paraphasias	Accuracy exact match	Accuracy within 5
All paraphasias	2489	0.468	0.657
Human agreement = 100%	1244	0.595	0.767
Human agreement < 100%	1245	0.342	0.548
Human confidence > median (3.3)	1089	0.605	0.768
Humans confidence <= median (3.3)	1400	0.362	0.571
WAB-R AQ > median (74.6)	1039	0.527	0.703
WAB-R AQ <= median (74.6)	1076	0.416	0.621
Fluent participants	1666	0.487	0.670
Non-fluent participants	449	0.412	0.626

836 *Note.* PWA stands for people with aphasia. WAB-R AQ is the Western Aphasia Battery-Revised
837 Aphasia Quotient (Kertesz, 2012). Fluent participants are those with Wernicke, Anomic,
838 Conduction, or Transcortical Sensory aphasia, or those considered “non aphasic” by the WAB-R.
839 Non-fluent participants are those with the Broca, Global, or Transcortical Motor aphasia. 48 out
840 of 353 total sessions had unavailable WAB-R results and were excluded just from analyses
841 involving WAB-R scores. Accuracy exact match refers to the top model prediction of target
842 word matching the human-identified target word. Accuracy within 5 refers to the human-
843 identified target word being one of the top five model predictions.

844 **Table 6**

845 *Experiment 4: Fine-tuned on controls and PWA data*

Test set	Number of paraphasias	Accuracy exact match	Accuracy within 5
All paraphasias	2489	0.468	0.668
Human agreement = 100%	1244	0.572	0.767
Human agreement < 100%	1245	0.363	0.569
Human confidence > median (3.3)	1089	0.600	0.792
Humans confidence <= median (3.3)	1400	0.365	0.572
WAB-R AQ > median (74.6)	1039	0.510	0.700
WAB-R AQ <= median (74.6)	1076	0.426	0.638
Fluent participants	1666	0.478	0.681
Non-fluent participants	449	0.425	0.624

846 *Note.* PWA stands for people with aphasia. WAB-R AQ is the Western Aphasia Battery-Revised
847 Aphasia Quotient (Kertesz, 2012). Fluent participants are those with Wernicke, Anomic,
848 Conduction, or Transcortical Sensory aphasia, or those considered “non aphasic” by the WAB-R.
849 Non-fluent participants are those with the Broca, Global, or Transcortical Motor aphasia. 48 out
850 of 353 total sessions had unavailable WAB-R results and were excluded just from analyses
851 involving WAB-R scores. Accuracy exact match refers to the top model prediction of target
852 word matching the human-identified target word. Accuracy within 5 refers to the human-
853 identified target word being one of the top five model predictions.

854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876

Appendix

Appendix A: Details of Masking and Decoding

To encode our inputs and outputs into a discrete numerical form recognizable to our specific choice of LLM, the text is encoded as sub-word units called SentencePieces (Kudo & Richardson, 2018). For example, the word “slipper” is represented by two tokens: “sl” and “ipper”. The SentencePieces algorithm identifies token boundaries using an unsupervised statistical algorithm, and its outputs reflect patterns of corpus frequency rather than morphology or any other linguistic principle (though, in practice, on English text there is often some incidental overlap with morphology). For most purposes, these SentencePieces and their contents are an implementation detail, encoded and decoded automatically by tools included with the language modeling software. However, the detail is relevant to two of our methodological choices. First, due to input and output constraints imposed by the architecture of the baseline model, each target word was masked with as many [MASK] tokens as corresponded to its SentencePiece-encoded length. Relatedly, upon decoding our model’s target word predictions, the model produced as many SentencePieces as there were [MASK] tokens in the input sequence. In other words, for our present experimental setup, the model could not produce a prediction with too many or too few SentencePieces. Second, for outputs requiring more than one SentencePiece, we decoded the output using a standard technique known as “beam search” (Lowerre, 1976). Given that the number of possible SentencePiece permutations grows exponentially with each additional [MASK] token, a beam search allows us to efficiently identify possible combinations of SentencePieces by estimating conditional probabilities for only the n most likely tokens at each step in the sequence. We used a limit (“beam width”) of n=20 while decoding our model’s output.

877

Supplemental Material

878 Supplemental Table 1

879 *Two-sided z-tests for independent proportions for test set stratifications of exact match accuracy*

880 *for all experiments*

Exp	Comparison	z	p
	Human agreement = 100% vs Human agreement < 100%	4.891	<0.001
	Human confidence > median vs Human confidence <= median	5.692	<0.001
1. Baseline	WAB-R AQ > median vs WAB-R AQ <= median	4.170	<0.001
	Fluent participants vs Non-fluent participants	2.879	0.004
	Human agreement = 100% vs Human agreement < 100%	8.471	<0.001
	Human confidence > median vs Human confidence <= median	9.532	<0.001
2. Controls	WAB-R AQ > median vs WAB-R AQ <= median	5.795	<0.001
	Fluent participants vs Non-fluent participants	4.746	<0.001
	Human agreement = 100% vs Human agreement < 100%	11.353	<0.001
	Human confidence > median vs Human confidence <= median	11.121	<0.001
3. PWA	WAB-R AQ > median vs WAB-R AQ <= median	4.793	<0.001
	Fluent participants vs Non-fluent participants	2.581	0.010
4. Controls + PWA	Human agreement = 100% vs Human agreement < 100%	10.336	<0.001

Human confidence > median vs Human confidence <= median	11.783	<0.001
WAB-R AQ > median vs WAB-R AQ <= median	3.335	0.001
Fluent participants vs Non-fluent participants	2.419	0.016

881 *Note.* Exp stands for experiment. PWA stands for people with aphasia. WAB-R AQ is the
882 Western Aphasia Battery-Revised Aphasia Quotient. Fluent participants are those with
883 Wernicke, Anomic, Conduction, or Transcortical Sensory aphasia, or those considered “non
884 aphasic” by the WAB-R. Non-fluent participants are those with the Broca, Global, or
885 Transcortical Motor aphasia. 48 out of 353 total sessions had unavailable WAB-R results and
886 were excluded just from analyses involving WAB-R scores.