Leveraging pre-trained large language models for aphasia detection in English and Chinese speakers

Yan Cong[1], **Jiyeon Lee**[2], **Arianna LaCroix**[2]
[1]School of languages and cultures; Linguistics, Purdue University
[2]Department of Speech, Language, and Hearing Sciences, Purdue University
{cong4, lee1704, anlacroi}@purdue.edu

Abstract

We explore the utility of pre-trained Large Language Models (LLMs) in detecting the presence, subtypes, and severity of aphasia across English and Mandarin Chinese speakers. Our investigation suggests that even without fine-tuning or domain-specific training, pre-trained LLMs can offer *some* insights on language disorders, regardless of speakers' first language. Our analysis also reveals noticeable differences between English and Chinese LLMs. While the English LLMs exhibit near-chance level accuracy in subtyping aphasia, the Chinese counterparts demonstrate less than satisfactory performance in distinguishing between individuals with and without aphasia. This research advocates for the importance of linguistically tailored and specified approaches in leveraging LLMs for clinical applications, especially in the context of multilingual populations.

1 Introduction

Large language models (LLMs) are transformative in various tasks (Tran, 2020; Chang et al., 2023; Hadi et al., 2023; Rezaii et al., 2023b, 2021). It remains understudied how to leverage non-English LLMs in a clinical context such as aphasia detection. Aphasia is an acquired neurogenic language disorder, most often caused by stroke, with devastating impact on one's communication abilities. Most aphasia studies with NLP perspectives focus on monolingual English speakers (Salem et al., 2023; Purohit et al., 2023; Sanguedolce et al., 2023; Ortiz-Perez et al., 2023). Fewer studies with NLP methods focus on the non-English population (Smaïli et al., 2022; Chatzoudis et al., 2022; Balagopalan et al., 2020). To bridge the gap, we leverage pre-trained LLMs to detect aphasia in English and Mandarin Chinese speakers. Given LLMs' widely claimed adaptability and linguistic competence (Zhao et al., 2023a; Bommasani et al., 2021), we hypothesize that integrating LLMs would enhance clinical diagnosis of language disorders in aphasia.

Aphasia in Chinese speakers has recently been studied from NLP perspectives. Balagopalan et al. (2020) utilized optimal transport domain adaptation to detect aphasia in Chinese and French. Shivkumar et al. (2020) developed an open-source python library called BlaBla to automatically extract linguistic features in English, Chinese and French aphasia data. Mahmoud et al. (2020) focused on deep learning's application to speech assessment of Chinese speakers with aphasia. Qin et al. (2022) used LLMs to derive embeddings, and fine-tuned LLMs for detection tasks. Their findings suggest that fine-tuned models outperform acoustic features and static embeddings.

As far as our knowledge goes, there is no study utilizing pre-trained LLMs derived surprisals to detect aphasia in Chinese speakers. Surprisal can be calculated by the negative likelihood of a token given previous context. Conceptually, it measures the unexpectedness of a sequence in a context. Surprisals' cognitive plausibility has been discussed in both psycholinguistic and clinical literature (Futrell et al., 2018; Rezaii et al., 2023a, 2022; Van Schijndel and Linzen, 2018; Wilcox et al., 2018; Michaelov and Bergen, 2020, 2022a,b; Michaelov et al., 2023; Ryu and Lewis, 2021; Cong et al., 2023; De Varda and Marelli, 2022). This motivates us to implement LLMs derived surprisals for aphasia detection in Chinese speakers. We additionally compare LLMs surprisals in Chinese datasets with those in English, given that English is a dominant language in NLP, English speakers are the most studied population in clinical contexts, and we hope to establish an interpretation baseline on how LLMs surprisals behave in English aphasia speakers.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 238–245
June 21, 2024 ©2024 Association for Computational Linguistics

2 Experiments

2.1 Datasets

All the datasets were drawn from the AphasiaBank[1] (MacWhinney et al., 2011), and all the observations are from participants who are monolingual speakers whose first language is English or Mandarin Chinese, with a Western Aphasia Battery-Aphasia Quotient (WAB-AQ (Kertesz, 2007)) of 92 or lower in the aphasia group.

For the Chinese dataset, we matched the aphasia with the control group on age, education, and sex using the R *matchit* package to perform optimal pair matching. The matched sample contains an equal amount of observations (N=1756) for each group, with similar tasks such as picture description and story retelling. The same aphasia sample was used in detecting aphasia severity. As for aphasia subtypes detection, we focused on Broca's and anomic aphasia, which are two of the most representative subtypes in the dataset. Since Broca's contains 86 observations in total, we randomly sampled 86 observations from the anomic aphasia group to get a balanced dataset.

For the English dataset, we conducted the same matching procedures with similar sample size. We compiled 1586 observations for each group, since that is the maximum of the control group. The selected aphasia sample was used in detecting aphasia severity. We randomly sampled 86 observations for each of the Broca's and anomic aphasia types.

2.2 Aphasia detection

We leveraged pre-trained LLMs in three tasks for both English and Chinese datasets: (1) detecting the presence of aphasia; (2) detecting aphasia subtypes (diagnosis labels provided by the AphasiaBank); (3) detecting aphasia severity (WAB-AQ, provided in the AphasiaBank). We constructed and optimized machine learning models. Logistic regression classifiers were used to classify aphasia and control (task 1) and Broca's and anomic aphasia (task 2). Elastic net was used to predict WAB-AQ scores (task 3). All the machine learning models were developed and evaluated in scikit-learn (Buitinck et al., 2013). Considering the limited sample size, for all the machine learning models, we focused on linear models and used default parameter settings without fine-grained hyperparameter tuning.

2.3 LLMs details

Each LLM read in utterance and output a surprisal score for that utterance. Specifically, we first computed token-wise surprisals, summed them for each utterance, then divided it by the utterance length (the number of tokens) to get mean surprisals. We hypothesize that higher surprisals, as an indicator of larger amount of grammatical unacceptability, are associated with higher severity of aphasia. Three pre-trained LLMs were used to generate token-wise surprisals in both the Chinese and English datasets: GPT2[2] (Radford et al., 2019; Zhao et al., 2019, 2023b), Llama2-7B (Touvron et al., 2023), and BERT (*bert-base-chinese* for Chinese and *bert-base-uncased* for English) (Devlin et al., 2019, 2018). We chose these Chinese LLMs because they are among the most widely used open-source LLMs according to the HuggingFace leaderboard[3]. We used the corresponding comparable pre-trained LLMs in English. To keep consistency, we used minicons (Misra, 2022), a utility for analyzing transformer-based representations of language. We make all code and meta-data available for additional testing[4].

2.4 Feature selection

We chose the following features as the predictor variable: utterance length and utterance level mean surprisal computed by pre-trained LLMs. This is because surprisial can measure language abilities at the utterance level and has been shown to be correlated with the features of agrammatism in aphasia (Rezaii et al., 2023a). Besides GPT2 surprisals, which have been investigated in previous studies, we attempt to examine the clinical capability of multiple pre-trained LLMs with difference scales in a non-English setting, and to investigate how these LLMs' surprisals relate to the clinical manifestation of aphasia. We chose utterance length as another independent variable. This is because, as a clinical indicator of linguistic productivity (MacWhinney et al., 2011; Fromm and MacWhinney, 2023; Fromm et al., 2022, 2020), utterance length can be informative of aphasia detection. Ut-

[1] https://talkbank.org/DB/

[2] We acknowledge that technically speaking, GPT2 may not be considered as a "large" language model, compared to other LLMs used in this study. Here, in order to keep the naming convention consistent and easy to follow, by "LLMs", we meant language models that have a transformer architecture as opposed to the classic n-gram paradigm.

[3] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[4] https://github.com/yancong222/ClinicalNLP2024

terance length will also greatly influence LLMs surprisals calculation, since utterance suprisal score is normalized by sequence length. We did not include other language measures in this study, due to the scope of this preliminary experiment. In this exploratory analysis, we intend to focus on one utility (i.e., LLMs) in a cross-linguistic clinical setting. The existing language measures such as verb ratio, noun percentage, sentence complexity, and so on, will need an additional utility to derive (e.g., the CLAN software for Computerized Language Analysis by MacWhinney et al. (2011)).

3 Results and discussion

3.1 LLMs' performance in aphasia presence and subtypes detection

Table 1 illustrated the logistic regression classifiers' performance in detecting the presence and subtypes of aphasia in Chinese speakers. Notation: Acc: accuracy; Prec: precision; Rec: recall; AUC: area under the curve. Results suggest that pre-trained LLMs are more effective in subtyping (F1-score 0.86) than detecting the presence of aphasia in Chinese speakers (F1-score 0.61). On the other hand, pre-trained LLMs showed the inverse pattern for detecting aphasia in English speakers (Table 2). Findings reveal that LLMs are less effective in detecting subtypes (F1-score 0.54) than the presence of aphasia in English speakers (F1-score 0.79). The two classification report tables contain weighted average values (averaging the sample-weighted mean per label, e.g., aphasia versus healthy; Broca's and anomic aphasia).

Task	Acc	Prec	Rec	F1-score	AUC
Presence	0.61	0.61	0.61	0.61	0.63
Subtype	0.86	0.86	0.86	0.86	0.93

Table 1: Evaluation of logistic regression classifiers using LLMs surprisals in Chinese aphasia detection.

Task	Acc	Prec	Rec	F1-score	AUC
Presence	0.79	0.79	0.79	0.79	0.86
Subtype	0.54	0.54	0.54	0.54	0.51

Table 2: Evaluation of logistic regression classifiers using LLMs surprisals in English aphasia detection.

Our interpretation is that using matched datasets and LLMs surprisals, LLMs pre-trained in Chinese are sensitive in separating non-fluent Broca's aphasia from anomic aphasia in Chinese speakers, whereas English LLMs showed efficacy in classifying aphasia versus control in English speakers. We infer that this result has something to do with crosslinguistic differences. The basic unit of grammar in Chinese is *zì* "character", but it is a *word* in English (Duanmu, 2017; Tsai and McConkie, 2003). Most Chinese words are made of two characters. Studies in psycholinguistic and NLP (Bai et al., 2008; Li et al., 2019) suggest that characters, rather than words, are considered the fundamental units of Chinese language processing. As far as our knowledge goes, most of the pre-trained LLMs for Chinese are based on character-level tokenization (Si et al., 2023). This character-based processing in LLMs could influence aphasia subtyping. Since LLMs' vocabularies for Chinese are consisted of characters, their representation of *word* meanings is not intrinsic. LLMs have to combine multiple characters to represent a word's meaning (Tsai and McConkie, 2003; Bai et al., 2008). It is likely that such character-based representation enables Chinese LLMs to get better tuned to pinpoint word retrieval difficulties, hence Chinese LLMs may be capable to identify more fine-grained differences such as specific aphasia subtypes.

Why do Chinese LLMs performed less effectively in detecting the presence of aphasia? The availability and size of training datasets for crosslinguistic LLMs (such as Chinese) can vary, but we maintain that typically English LLMs may have access to larger training datasets. Accordingly, we stipulate that non-English pre-trained LLMs are hypothetically less flexible and harder to generalize to domain-specific data (e.g., aphasia). Therefore, compared to English LLMs in English aphasia detection, Chinese LLMs are likely to be less sensitive to the broad linguistic disturbances associated with aphasia in Chinese speakers, leading to lower efficacy in detecting aphasia overall. Further, we infer that the low efficacy may be due to Chinese not having verb conjugations. Studies show that a hallmark in aphasia is the main verb problem, which is associated with morphological impairment (Bates et al., 1991; Pak-Hin Kong, 2011). In English, larger morphological load carried by verbs (compared with nouns) likely cause such impairment. The lack of verb conjugations and rich morphological markings in Chinese may lead to difficulties for

LLMs, since these commonly seen signs of aphasia in English are absent in Chinese.

The distinct patterns suggest that subtler linguistic features captured by LLMs are more discriminative in identifying specific subtypes of aphasia in Chinese. Conversely, a contrasting scenario was found in English speakers, where the LLMs exhibit superior performance in detecting the presence of aphasia compared to subtype classification. This discrepancy makes us wonder if language-specific nuances influence the performance of LLMs in aphasia detection. The findings emphasize the importance of tailored approaches for leveraging LLMs in clinical applications across diverse linguistic populations. The inverse patterns observed between English and Chinese speakers indicate the necessity of language-specific model adaptations and fine-tuning strategies, which will likely optimize the utility of LLMs in clinical practice. To sum up, we found *some* clinical efficacy in Chinese pre-trained LLMs for aphasia subtyping. Crosslinguistic LLMs are promising utilities for clinical diagnosis. However, we are cautiously optimistic since these LLMs showed less than satisfactory accuracy (0.61) when detecting the presence of aphasia, a task we think is fundamental to benchmark LLMs' clinical reliability.

3.2 LLMs' performance in aphasia severity detection

Given that we have a relatively small sample size and only a handful of features which are related, to handle multicollinearity, we used elastic net regression to model LLMs' efficacy in predicting aphasia severity (WAB-AQ scores). Elastic net model was evaluated using repeated 10-fold cross-validation. We report the average mean absolute error (MAE) and predictor variables' coefficients in Table 3.

Dataset	MAE	utterance length	GPT2	Llama2	BERT
English	14.97	0.00	-0.55	-3.05	1.56
Chinese	7.61	0.55	-0.03	-0.37	-0.06

Table 3: Elastic net regression models in predicting English and Chinese aphasia severity.

Model coefficients in Table 3 suggest that for the English dataset tasks, the role of utterance length as a predictor of aphasia severity is trivial. The two decoder LLMs (GPT2 and Llama2) showed negative effects, namely higher surprisals are associated with lower WAB-AQ (higher severity). BERT showed the inverse, which is unexpected and hard to interpret. For all three LLMs, Llama2 showed the strongest coefficients. For the Chinese dataset, utterance length played a role in predicting aphasia severity. All the LLMs' surprisals showed negative coefficients for the Chinese dataset. Llama2, the largest LLM, gave the largest coefficient again. This implies that larger LLMs tend to outperform smaller ones, and scaling improves LLMs' performance in both English and Chinese tasks. We do not find sufficient evidence showing that bidirectional LLMs' surprisals such as BERT are less effective than unidirectional LLMs' like GPT2 in clinical tasks, although GPT type LLMs' pretraining task (next token prediction given previous context) appears to be more suitable for surprisals computation (Shain et al., 2024).

Additionally, MAEs, an average measure of how far the model's predictions are from the actual target values in the test set, suggest that elastic net regression model is a better fit for the Chinese than the English tasks. This indicates that to operationalize pre-trained LLMs and help healthcare practitioners make clinical decisions for the non-English aphasia population, we need LLMs pre-trained in corresponding languages. Open-source crosslinguistic pre-trained LLMs have the potential to improve LLMs' ecological validity in a clinical setting.

Note that the analysis of LLMs' performance in aphasia severity detection is based on the raw data irrespective of whether the initial classification of aphasia presence and subtype was correct. There are two primary motivations. First, the sample size is already small. Selecting only cases that are correctly identified as having aphasia may further shrink the dataset. Second, we intend to independently examine how much LLMs surprisals can measure aphasia severity, based on raw data. This approach will also enable reproducibility and model applicability, since no intermediate pipelines are needed to filter data based on previous tasks' efficacy. However, we acknowledge that it is open to discussion how much noise from misclassified cases potentially may skew the severity models' performance metrics. For future research, we hope to expand the datasets, and construct and compare multiple models with and without initial classification.

3.3 Qualitative error analysis

In order to increase interpretability, we conducted qualitative error analyses. Concrete examples highlighting certain unexpected outputs from LLMs are given in Table (4, 5), for which a higher surprisal is unexpectedly found in the control group.

Results suggest that extremely short utterances turn out to give rise to large surprisal scores for both Chinese and English datasets, especially for Llama2 and GPT2 (example 4). Interestingly for BERT, the utterance length effect is not strong. It is also likely that English interjection or filler words like “gee”, low frequency verb "startle", and Chinese sentence final particles such as “呢” “呀” lead to higher surprisals (examples (2,4)). The level of cleaning and pre-processing of the inpu text may play a role. We hope to independently test this hypothesis for future research.

4 Conclusion

This study leveraged pre-trained LLMs to detect the presence, subtypes, and severity of aphasia in English and Mandarin Chinese speakers. Our findings suggest that without fine-tuning, taking pre-trained LLMs off-the-shelf can already inform us how surprisals distribute in aphasic individuals whose first language is or is not English. That said, we also found that Chinese LLMs showed less decent performance in classifying healthy control versus aphasia, and that English LLMs show almost chance level accuracy in subtyping aphasia. We plan to fine-tune crosslinguistic LLMs using aphasia datasets to improve the models’ competence in clinical tasks.

Our study highlights the clinical application of pre-trained LLMs in English and non-English aphasia individuals. There is a critical need for automatic aphasia diagnosis, since manually assessing language disturbances is labor and cost intensive, especially in low-resource non-English settings. The advent of LLMs has the potential to advance the field of aphasia detection. As a case study of utilizing pre-trained LLMs in Chinese and English datasets, our investigation advocates for refining clinical NLP pipelines via incorporating LLMs pre-trained in non-English languages.

5 Limitation

Given the relatively small sample size, the current study is meant to be a proof of concept, rather than providing any end-to-end or predictive models or analytical frameworks. We hope to showcase how much we can gain from pre-trained LLMs in non-English speakers with aphasia, advocating for clinical crosslinguistic LLMs in low-resource settings, for example languages other than English.

Our findings suggest that larger LLMs gave higher clinical efficacy. This implies that scaling could matter. We are aware that scaling up is not necessarily a feasible option for most researchers, given its demanding computation requirement (Schick and Schütze, 2020). Exactly how much scaling and sample size matter is open to discussion and out of the scope of the current study. We maintain that dataset size may play a role in how well LLMs perform in classifying and subtyping aphasia. We hope to examine this with a more comprehensive set of pre-trained LLMs and larger sample size.

Moreover, we acknowledge that our study only showed that there is difference when using LLMs pre-trained in different languages, but we did not show its magnitude and specifically what linguistic properties (e.g., argument structure, word order) differ in LLMs’ detection of Chinese and English speakers with aphasia. Also, in aphasia studies, overlapping patterns were found in Chinese and English speakers: although there are crosslinguistic differences, a previous study has reproduced the impairment caused by the syntactic complexity of utterances produced by Chinese speakers with aphasia (Wang and Thompson, 2016). We plan to expand our datasets and examine to what extent the crosslinguistic impairment similarities can be detected when using crosslinguistic LLMs.

6 Acknowledgments

We acknowledge Cameron Pilla’s and Emily Tumacder’s help with compiling the datasets and optimizing the machine learning pipeline. We thank Emmanuele Chersoni, Sunny Tang, Phillip Wolff, and Sunghye Cho for their inspirations. We appreciate anonymous reviewers’ constructive and helpful comments. This project is supported by the College of Liberal Arts, Purdue University.

References

Xuejun Bai, Guoli Yan, Simon P Liversedge, Chuanli Zang, and Keith Rayner. 2008. Reading Spaced and Unspaced Chinese Text: Evidence from Eye Movements. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1277.

Participant group	Utterance	LLMs surprisals
(1) Aphasia	mhm. okay. mhm. okay. it's fun. the kid throws the ball. and it. oops. on the window. and the guy's dad wasn't just boom. and here comes the ball. and then he looks up. that's pretty funny.	*Llama2* 3.11; *GPT2* 3.81
(2) Healthy control	a young boy is kicking a ball and crashed through a window. startled the man. and he looked up at the cracked window.	*Llama2* 3.44; *GPT2* 4.51
(3) Aphasia	yeah. yeah. alright. book. mow oh boy. boy. woe. shakes. ball. hey books. yeah. jay balls. balls six. oh boy balls. bugs.	*Llama2* 4.2; *BERT* 19.16
(4) Healthy control	oh gee.	*Llama2* 5.43; *GPT2* 5.26

Table 4: Unexpected output given by the English LLMs in picture description tasks.

Participant group	Utterance	Literal translation	LLMs surprisals
(1) Aphasia	两个人两只动物比谁走跑得快	two people two animals compare who walk run faster	*Llama2* 3.25; *GPT2* 4.4
(2) Healthy control	后来呢兔子和小乌龟比赛跑	then hare and small tortoise compete to run	*Llama2* 4.06; *GPT2* 5.1
(3) Aphasia	就是到医院医院医院然后就是做了这个就是看了这个头	so to the hospital hospital hospital then just do this just look at this head	*Llama2* 3.09; *GPT2* 3.61
(4) Healthy control	不识字呀	do not recognize characters	*Llama2* 5.15; *GPT2* 5.2

Table 5: Unexpected output given by the Chinese LLMs in picture description tasks.

Aparna Balagopalan, Jekaterina Novikova, Matthew BA Mcdermott, Bret Nestor, Tristan Naumann, and Marzyeh Ghassemi. 2020. Cross-language aphasia detection using optimal transport domain adaptation. In *Machine Learning for Health Workshop*, pages 202–219. PMLR.

Elizabeth Bates, Sylvia Chen, Ovid Tzeng, Ping Li, and Meiti Opie. 1991. The noun-verb problem in chinese aphasia. *Brain and language*, 41(2):203–233.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, et al. 2013. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Gerasimos Chatzoudis, Manos Plitsis, Spyridoula Stamouli, Athanasia-Lida Dimou, Athanasios Katsamanis, and Vassilis Katsouros. 2022. Zero-shot cross-lingual aphasia detection using automatic speech recognition. *arXiv preprint arXiv:2204.00448*.

Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu, and Alessandro Lenci. 2023. Are Language Models Sensitive to Semantic Attraction? A Study on Surprisal. In *Proceedings of *SEM*.

Andrea De Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

San Duanmu. 2017. Word and Wordhood, Modern. *Encyclopedia of Chinese Language and Linguistics*, 4:543–49.

Davida Fromm, Joel Greenhouse, Mitchell Pudil, Yichun Shi, and Brian MacWhinney. 2022. Enhancing the classification of aphasia: a statistical analysis using connected speech. *Aphasiology*, 36(12):1492–1519.

Davida Fromm and Brian MacWhinney. 2023. Discourse databases for use with clinical populations.

Davida Fromm, Brian MacWhinney, and Cynthia K Thompson. 2020. Automation of the northwestern narrative language analysis system. *Journal of Speech, Language, and Hearing Research*, 63(6):1835–1844.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as Psycholinguistic Subjects: Syntactic State and Grammatical Dependency. *arXiv preprint arXiv:1809.01329*.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Andrew Kertesz. 2007. Western aphasia battery–revised.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? In *Proceedings of ACL*.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Seedahmed S Mahmoud, Akshay Kumar, Yiting Tang, Youcun Li, Xudong Gu, Jianming Fu, and Qiang Fang. 2020. An efficient deep learning based method for speech assessment of mandarin-speaking aphasic patients. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3191–3202.

James A Michaelov and Benjamin K Bergen. 2020. How Well Does Surprisal Explain N400 Amplitude under Different Experimental Conditions? In *Proceedings of CONLL*.

James A Michaelov and Benjamin K Bergen. 2022a. Collateral Facilitation in Humans and Language Models. In *Proceedings of CONLL*.

James A Michaelov and Benjamin K Bergen. 2022b. 'Rarely'a Problem? Language Models Exhibit Inverse Scaling in their Predictions Following 'Few'-type Quantifiers. *arXiv preprint arXiv:2212.08700*.

James A Michaelov, Seana Coulson, and Benjamin K Bergen. 2023. Can Peanuts Fall in Love with Distributional Semantics? *arXiv preprint arXiv:2301.08731*.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

David Ortiz-Perez, Pablo Ruiz-Ponce, Javier Rodríguez-Juan, David Tomás, Jose Garcia-Rodriguez, and Grzegorz J Nalepa. 2023. Deep learning-based emotion detection in aphasia patients. In *International Conference on Soft Computing Models in Industrial and Environmental Applications*, pages 195–204. Springer.

Anthony Pak-Hin Kong. 2011. Aphasia assessment in chinese speakers. *The ASHA Leader*, 16(13):36–38.

Aditya kumar Purohit, Aditya Upadhyaya, and Adrian Holzer. 2023. Chatgpt in healthcare: Exploring ai chatbot for spontaneous word retrieval in aphasia. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 1–5.

Ying Qin, Tan Lee, Anthony Pak Hin Kong, and Feng Lin. 2022. Aphasia detection for cantonese-speaking and mandarin-speaking patients using pre-trained language models. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 359–363. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Neguine Rezaii, Nicole Carvalho, Michael Brickhouse, Emmaleigh Loyer, Phillip Wolff, Alexandra Touroutoglou, Bonnie Wong, Megan Quimby, and Brad C Dickerson. 2021. Neuroanatomical mapping of artificial intelligence-based classification of language in ppa. *Alzheimer's & Dementia*, 17:e055340.

Neguine Rezaii, Kyle Mahowald, Rachel Ryskin, Bradford Dickerson, and Edward Gibson. 2022. A syntax–lexicon trade-off in language production. *Proceedings of the National Academy of Sciences*, 119(25):e2120203119.

Neguine Rezaii, James Michaelov, Sylvia Josephy-Hernandez, Boyu Ren, Daisy Hochberg, Megan Quimby, and Bradford C Dickerson. 2023a. Measuring sentence information via surprisal: theoretical and clinical implications in nonfluent aphasia. *Annals of Neurology*, 94(4):647–657.

Neguine Rezaii, Megan Quimby, Bonnie Wong, Daisy Hochberg, Michael Brickhouse, Alexandra Touroutoglou, Bradford C Dickerson, and Phillip Wolff. 2023b. Using generative artificial intelligence to classify primary progressive aphasia from connected speech. *medRxiv*, pages 2023–12.

Soo Hyun Ryu and Richard L Lewis. 2021. Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Alexandra C Salem, Robert C Gale, Mikala Fleegle, Gerasimos Fergadiotis, and Steven Bedrick. 2023. Automating intended target identification for paraphasias in discourse using a large language model. *Journal of Speech, Language, and Hearing Research*, 66(12):4949–4966.

Giulia Sanguedolce, Patrick A Naylor, and Fatemeh Geranmayeh. 2023. Uncovering the potential for a weakly supervised end-to-end model in recognising speech from patient with post-stroke aphasia. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 182–190.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Abhishek Shivkumar, Jack Weston, Raphael Lenain, and Emil Fristed. 2020. Blabla: Linguistic feature extraction for clinical analysis in multiple languages. *arXiv preprint arXiv:2005.10219*.

Chenglei Si, Zhengyan Zhang, Yingfa Chen, Fanchao Qi, Xiaozhi Wang, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2023. Sub-character Tokenization for Chinese Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 11:469–487.

Kamel Smaïli, David Langlois, and Peter Pribil. 2022. Language rehabilitation of people with broca aphasia using deep neural machine translation. In *Fifth International Conference on Computational Linguistics in Bulgaria*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ke Tran. 2020. From english to foreign languages: Transferring pre-trained language models. *arXiv preprint arXiv:2002.07306*.

Jie-Li Tsai and George W McConkie. 2003. Where Do Chinese Readers Send Their Eyes? In *The Mind's Eye*, pages 159–176. Elsevier.

Marten Van Schijndel and Tal Linzen. 2018. Modeling Garden Path Effects without Explicit Hierarchical Syntax. In *Proceedings of CogSci*.

Honglei Wang and Cynthia K Thompson. 2016. Assessing syntactic deficits in chinese broca's aphasia using the northwestern assessment of verbs and sentences-chinese (navs-c). *Aphasiology*, 30(7):815–840.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What Do RNN Language Models Learn about Filler-gap Dependencies? *arXiv preprint arXiv:1809.00042*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023a. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

Zhe Zhao, Yudong Li, Cheng Hou, Jing Zhao, et al. 2023b. Tencentpretrain: A scalable and flexible toolkit for pre-training models of different modalities. *ACL 2023*, page 217.