Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Analysis: Serving Individuals with Language Impairments using Automatic Speech Recognition Models and Large Language Models: Challenges and Opportunities

Yiyu Shi

yshi4@nd.edu

University of Notre Dame https://orcid.org/0000-0002-6788-9823

Ruiyang Qin

Villanova University

Haoxinran Yu

University of Notre Dame

Lixuan Wei

University of Notre Dame

Yuxuan Liu

University of Notre Dame

Dancheng Liu

State University of New York at Buffalo https://orcid.org/0009-0003-6107-5612

Chenhui Xu

State University of New York at Buffalo

Jiajie Li

State University of New York at Buffalo

Gelei Xu

University of Notre Dame

Ahmed Abbasi

University of Notre Dame

Jinjun Xiong

State University of New York at Buffalo

Xiufan Yu

University of Notre Dame

Zhi Zheng

University of Notre Dame

Article

Keywords:

Posted Date: July 24th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7069967/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
There is **NO** Competing Interest.

This is determined to be not a human subject study by our IRB as it is based on public domain dataset with no identifiable information.

Version of Record: A version of this preprint was published at Nature Communications on September 10th, 2026. See the published version at https://doi.org/10.1038/s41467-026-76677-z.

Serving Individuals with Language Impairments using Automatic Speech Recognition Models and Large Language Models: Challenges and Opportunities

Anonymous Author(s)[1]

[1]Anonymous Affiliation(s).

Abstract

Large language models (LLMs) have attracted much attention for healthcare applications, demonstrating strong potential in automating conversational interactions. However, cloud-hosted LLMs pose major data privacy concerns when processing Protected Health Information. Moreover, current LLM-based systems rely on text input/output, creating substantial barriers for users, such as children and older adults, who may have difficulty typing. To mitigate these challenges, there has been growing interest in developing edge device-based, voice-enabled LLM systems. Running LLMs on edge devices minimizes the risks of PHI leaking to the cloud, while automatic speech recognition (ASR) eliminates the need for text-based inputs. Despite these advantages, existing ASRs convert speech into word-by-word text, which often contains disfluencies and fillers (e.g., "um", "hum") and grammatical errors, especially for individuals with language impairments. This noisy input can significantly degrade the performance of LLMs, yet this chained issue remains underexplored in healthcare applications. To address this critical gap, we conducted a systematic analysis through comparison studies and ablation experiments to identify key factors affecting the performance of edge-based ASR-LLM systems when used by individuals with language impairments. Furthermore, we proposed an evaluation framework for speech-enabled AI healthcare to emphasize both interpretability and robustness, paving the way for more inclusive and secure conversational healthcare solutions.

1 Introduction

The global healthcare sector is facing the crisis of staff shortage [1]. To address this challenge, innovators are leveraging cutting-edge Artificial Intelligence (AI) to bridge the gap between healthcare demand and accessibility [2]. Among these efforts, Large Language Models (LLMs) are being explored as a solution to address the bottleneck of conversational communication, a cornerstone of healthcare delivery. However, like most AIs, the performance of LLMs highly depends on the data used for model training. Given that very limited training data has been collected from people with language impairments, the efficacy of current LLMs for this population remains uncertain. This technical uncertainty poses the risk of significant disruptions and exclusion, particularly in vulnerable individuals, considering that more than 5% of the general population experiences communication impairments [3, 4].

Applying a LLM to healthcare communication for individuals with language impairments presents additional challenges beyond the model itself. First, many users, such as older adults living with dementia and children with language impairments, often face difficulties with text-based input, which remains the default interface for most LLMs. This limitation can hinder communication efficiency and reduce the effectiveness of care compared to traditional human-provided services[5]. Second, the interaction between users and LLMs often involves sensitive personal data, raising privacy concerns and potentially discouraging users from adopting LLM-driven healthcare solutions[6]. A promising approach to addressing the first challenge is the integration of Automatic Speech Recognition (ASR), which converts speech audio into text that can be processed by LLMs. This ASR-LLM strategy has been adopted by major technology companies such as OpenAI [7] and Google [8]. In our study, the term ASR-LLM refers to a pipeline where the LLM takes the transcript from the ASR as input, and produces a text response as output. To mitigate privacy concerns, deploying LLMs directly on edge devices, such as tablets or smartphones, offers a viable solution. By ensuring that all data processing occurs locally without cloud-based data exchange, edge-based LLMs can significantly reduce privacy risks and improve user trust [9, 10].

While deploying LLMs on edge devices and integrating ASR can improve accessibility and privacy, this combination raise a critical question: whether an edge-based ASR-LLM can effectively communicate to those with language impairments. Current ASRs convert speech into word-by-word text, often containing disfluencies, fillers (e.g., "um", "hum"), and grammatical errors, especially in individuals with language impairments. Additionally, compared to full-scale LLMs hosted on cloud servers, LLMs on an edge device are much smaller with reduced model parameters, resulting in weaker reasoning capabilities[11, 12]. Similarly, on-device ASRs undergo performance degradation due to size compression, which may further compound the performance limitations of the edge LLMs.

Our study utilized data collected from various clinical trials in three representative diseases: Aphasia, language impairments in children (LIC), and Dementia in older adults. Aphasia is a language disorder resulting from damage to one or more language-processing areas of the brain, often resulting from stroke. Approximately one-third of stroke cases lead to aphasia [13]. The selected LIC datasets include communication

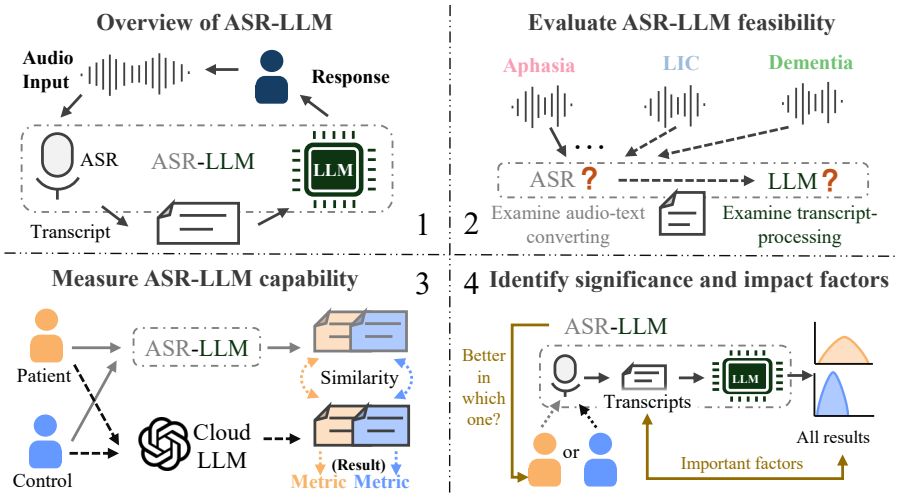


Figure 1: Overview of our study. (1) We showed the ASR-LLM where the ASR took the audio input and generated the transcript as the input to LLM. (2) We evaluated the feasibility of the ASR-LLM on three diseases: Aphasia, language impairments in children (LIC), and Dementia. (3) We measured the capability of the ASR-LLM on the three disease, using the response generated by Cloud LLM as the reference. (4) We identified the ASR-LLM significant difference (significance) between patients and controls and the factors leading to the significance.

disorder that interferes with the development of language skills in children without hearing loss [14], affecting about 1 in 14 children [15]. It also includes data recorded from children who suffer language impairments due to epilepsy. Dementia encompasses a group of conditions characterized by a progressive decline in language and cognitive functions, such as memory, reasoning, and problem solving. It is estimated that 4% of older adults aged 65 years and older have been diagnosed with dementia, with prevalence increasing with age [16]. Based on the context, we built up ASR-LLM simulators for edge device usage, as shown in Figure 1. In the ASR-LLM, the ASR converted the audio input to the transcript. The LLM generated the response corresponding to the audio based on the transcript. These data were analyzed quantitatively to evaluate the ASR-LLM's performance. By analyzing the conversational data, we aim to demonstrate how atypical language patterns across different groups impact the performance of edge ASR-LLM.

The study of ASR-LLM performance consisted of three main steps, as illustrated in Figure 1. First, we evaluated the feasibility of applying ASR-LLM to each dataset, focusing on whether the ASR and LLM components could process the input appropriately. This was a crucial step because ASR-LLMs deployed on edge devices have limited model weights that inherently constrain the models' capabilities. Second, we measured the capability of edge ASR-LLM, compared to a well-performing cloud-based LLM, which possesses broader knowledge and stronger reasoning ability. We measured this capability using text similarity between the outputs of ASR-LLM and the cloud LLM. This allowed us to quantify both the absolute performance of ASR-LLM on each dataset and its relative performance on patients vs. on controls. Finally, we tested

whether the performance differences between patients and controls were statistically significant. For the datasets where significances were detected, we further analyzed the ASR-generated transcripts to identify the linguistic factors that influenced ASR-LLM capability.

In summary, the major contributions of this work are three-fold: 1) We developed a novel framework (parts 2 - 4 in Figure 1) to evaluate the performance of ASR-LLM for individuals with language impairments; 2) Through this framework, we conducted in-depth comparison (patients vs. controls) analyses to identify a set of key factors that determine the performance of ASR-LLM; 3) Building on these findings, we further explored strategies to enhance ASR-LLM performance. We believe that our framework and the insights gained from data analyses offer a solid foundation for future research, enabling researchers to assess the viability of using ASR-LLM in conditions involving language impairments and to develop targeted improvements based on critical factors.

2 Results

We conducted a systematic analysis comparing how ASR-LLMs perform for patients vs. controls. Twenty TalkBank datasets across Aphasia (patient N = 227; control N = 210), LIC (patient N = 348; control N = 508), and Dementia (patient N = 448; control N = 356) were used. These datasets were collected from various clinical trials and included audio/videos of participants engaged in healthcare conversations, such as those during symptom assessments and therapy sessions. We selected 23 state-of-the-art (SOTA) LLM models of various sizes to be combined with the popular SOTA ASR model, Whisper. Five different sizes of Whisper were used, thus formalizing 115 possible edge-based combinations of ASR-LLM for validation.

Three widely used semantical similarity metrics, ROUGE, BERTScore, and BLEURT, were employed to measure performance in terms of one-to-one text match, embedding using pretrained models, and under fine-tuning. Computing each metric score requires a reference (the best case possible). Since the TalkBank datasets did not include such references, for each ASR-LLM, the metrics were computed based on the semantic similarities between the ASR-LLM's output and the output of a top-notch reference: combining human annotations (text inputs, counterpart of ASR) and a well-performing large cloud-based LLM (GPT-4o, which outperforms all edge LLMs). Each ASR-LLM took the audio as input and produced a summarization. Conversely, the GPT-4o used the human-made transcript directly to produce summarizations. Summarization is considered an important and suitable task because it can reflect an LLM's understanding of the input text [17, 18]. For consistency, the same reference was used across all metric computations. Hence, higher semantic similarity scores between the ASR-LLM's output and that of the reference were considered desirable.

Due to varied data distributions, the non-parametric Wilcoxon tests with Bonferroni correction were applied to compare each ASR-LLM's performance on patients and controls. Given the presence of multiple comparisons using various ASR-LLM combinations, we applied the Bonferroni correction to adjust the significance threshold and

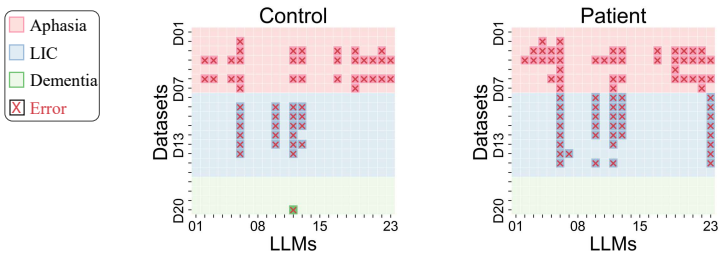


Figure 2: Evaluation on 20 datasets (Table 9) across 23 LLMs, each comprising one LLM (Table 8) paired with five ASR models. A grid is crossed if any of its ASR-LLM combination fails.

control the family-wise error rate. The Bonferroni correction is known for its conservative nature, therefore any significance that does emerge under the Bonferroni-adjusted threshold is considered highly reliable.

2.1 Evaluating ASR-LLM Feasibility for Clinical Audio Applications

First, we analyzed the capacity of ASR-LLM sets in processing the datasets of three health conditions. The bottom line is that audio inputs from a dataset should be processed by the computer programs without causing execution errors (e.g., memory overflow). We went through datasets across ASR-LLM sets and found that all datasets could be processed by all tested ASRs. However, some ASR-generated transcripts could not be processed by certain LLMs. These **"errors"** are highlighted in Figure 2, where each grid represents the aggregated results of a particular LLM combined with the five tested ASR on a particular dataset.

Program errors happened in processing both controls' and patients' data, while the control's data caused fewer errors. The errors of the Aphasia datasets concentrated on three datasets (NEURAL, NEURAL_2, and UNH) across multiple LLMs, while the errors for the Childes datasets concentrated on four (for control group) or five (for patient group) LLMs across multiple datasets.

To further investigate the errors and their underlying causes, these representative LLMs and datasets were selected to: 1) measure the error ratio of ASR-LLM from the perspective of LLMs (Equation 1) and the perspective of Datasets (Equation 2) and 2) identify the specific error types.

$$\text{For each dataset, Error Ratio (L)} = \text{Avg}_{\text{identified LLMs}} \left(\frac{N_{\text{error}}}{\text{Sample Size}} \right) / N_{\text{ASR}} \quad (1)$$

$$\text{For each LLM, Error Ratio (D)} = \text{Avg}_{\text{identified datasets}} \left(\frac{N_{\text{error}}}{\text{Sample Size}} \right) / N_{\text{ASR}} \quad (2)$$

N_{error} is the number of samples (participants) in a dataset that encountered errors. $N_{samples}$ is the sample size of that dataset. N_{ASR} is 5, the number of ASR tested.

The most frequently encountered type of errors was **"CUDA out of memory" (ERROR1)**. As shown in Figure 3, the ERROR1 was at its highest on Aphasia datasets in both patients and controls. Meanwhile, ERROR1 rarely happened on

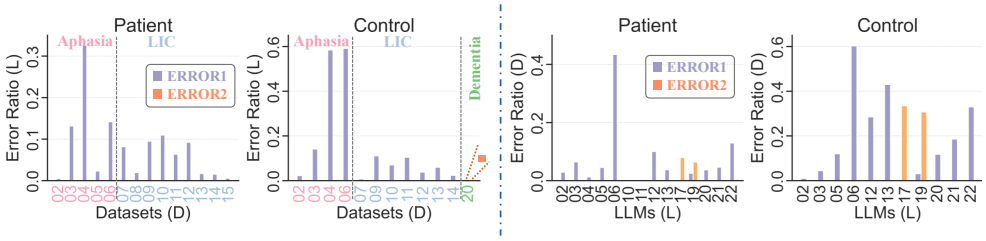


Figure 3: Error ratio of datasets based on the selected LLMs (Left). Error ratio of LLMs based on the selected datasets (Right).

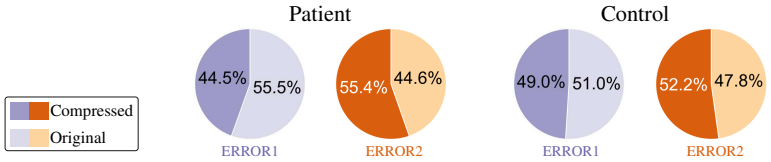


Figure 4: Error distributions between compressed and original LLMs

Dementia datasets. The other noticeable type of errors are **"Probability tensor contains either inf, 'nan' or element < 0" (ERROR2)**. ERROR2 was mostly found in Dementia controls.

There are two common types of LLMs for on-device usage: 1) original moderate-to-small-sized LLMs pretrained from scratch and 2) compressed LLMs made via compressing large pretrained LLMs into smaller ones for computational efficiency. Types of the tested LLMs are listed in Table 8 in the Appendix. Analysis showed that the types of LLMs are related to the errors. As shown in Figure 4, in both patients and controls, original LLMs dominated ERROR1, and the compressed LLMs dominated ERROR2.

Both ERROR1 and ERROR2 were closely linked to the length of ASR-generated transcripts. For instance, in dataset NEURAL_2, all ERROR1s involved transcripts exceeding 10,000 tokens as shown in Figure 5, suggesting that long inputs overwhelmed the LLM's rotary positional encoding [19]. Rotary positional encoding is a highly memory-consuming step in the LLM pipeline that adds positional information into the LLM's computation, enabling the model to understand relationships between words. In contrast, ERROR2 occurred consistently around the 6000-token mark, indicating inference failed beyond a model-specific length threshold. Inference is when an LLM performs computations to predict the most likely output based on a given input.

In addition, larger ASR models reduced errors in both patients and controls. Considering testing one ASR-LLM combination on one dataset is a trial, then for each ASR, the general error rate measured the proportion of failed trials within each disease domain. In the LIC datasets, error ratios declined markedly with increasing ASR size for both patients and controls. In contrast, with increasing ASR sizes, Aphasia datasets showed a more pronounced reduction in error ratio for patients than for controls.

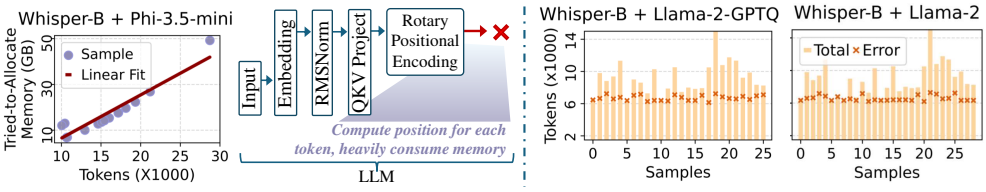


Figure 5: Left: Illustration of memory consumption vs. token size and where ERROR1 happens in the LLM pipeline (memory overflow); Right: Visualization of ERROR2 index in the token stack.

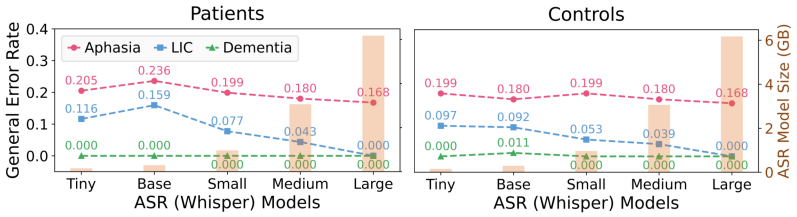


Figure 6: Examine the relation between error rate and ASR size.

2.2 Assess the ASR-LLM performances and the significant differences between patients and controls

Using three commonly used LLM performance metrics (ROUGE [20], BERTScore [21], and BLEURT [22]), the ASR-LLM median performances in Table 1 show that they generally performed the best in Dementia, and performed better in LIC than in Aphasia. In Dementia, ASR-LLM performances were generally better in controls. In Aphasia and LIC, the ASR-LLM performances between patients and controls were more mixed; although not consistently better in controls, the differences were typically small, often within a margin of 0.01.

To assess the differences in performance between controls (C) and patients (P), we considered a two-sided hypothesis test focusing on the null hypothesis

$$H_0 : \text{there was no difference in ASR-LLM performance between controls and patients} \tag{3}$$

and the two-sided alternative hypothesis

$$H_{a1} : \text{ the ASR-LLM performed differently between controls and patients.}$$

For brevity, we denote the above two-sided test as

$$H_0 : P = C \quad \text{versus} \quad H_{a1} : P \neq C \tag{4}$$

where "=" and "≠" represent equivalence and difference, respectively, in ASR-LLM performances.

The two-sided Wilcoxon test with Bonferroni-adjusted p-values (Table 2) showed statistically significant differences between patients and controls in all Aphasia and Dementia datasets, as well as two LIC datasets (ENNI and Poler). Especially, significances were found in a large number of ASR-LLMs on Richardson (Aphasia), ENNI (LIC), and Pitt/cookie (Dementia), as illustrated by the detailed raw p-values in Figures 10–12 in the Appendix. The results provided significant evidence against the null hypothesis (H_0), supporting the alternative hypothesis (H_1).

2.3 Assess the ASR-LLM preference: patients or controls

To further assess the directional differences in ASR-LLM performance, we considered two one-sided hypothesis tests to separately evaluate whether the performance is significantly better in patients compared to groups, and vice versa. Specifically, one-sided Wilcoxon tests were employed to test the null hypothesis H_0 (defined in (3)) against two one-sided alternative hypotheses

$$H_{a2} : \text{the ASR-LLM performed better in controls than in patients.}$$
$$H_{a3} : \text{the ASR-LLM performed better in patients than in controls.}$$

Similarly, we denote

$$H_0 : P = C \quad \text{versus} \quad H_{a2} : C > P \tag{5}$$

and

$$H_0 : P = C \quad \text{versus} \quad H_{a3} : P > C \tag{6}$$

where "$>$" indicates superior ASR-LLM performance in one group compared to the other.

Bonferroni adjusted p-values from one-sided Wilcoxon tests on (5) and (6) were reported in Table 3. ASR-LLMs performed significantly better on controls in most

	Dataset	ROUGE		BERTSCORE		BLEURT	
		Patients	Controls	Patients	Controls	Patients	Controls
Aphasia	Baycrest	0.36	**0.41**	0.87	**0.88**	-0.54	**-0.52**
	MSU	0.37	**0.38**	0.87	**0.88**	**-0.56**	-0.59
	NEURAL	0.38	**0.39**	0.88	0.88	**-0.51**	-0.57
	NEURAL_2	0.33	**0.35**	0.87	0.87	**-0.70**	-0.72
	UMD	0.36	**0.38**	0.87	**0.88**	**-0.56**	-0.64
	UNH	**0.37**	0.35	**0.87**	0.86	**-0.64**	-0.74
	Richardson	0.36	**0.40**	0.87	0.87	**-0.63**	-0.66
LIC	ENNI	0.37	**0.41**	0.86	**0.87**	-0.73	**-0.64**
	Ellis/30ec	**0.34**	0.33	0.86	0.86	-0.70	**-0.68**
	Ellis/30pc	**0.31**	0.30	0.85	0.85	-0.72	**-0.71**
	Ellis/42ec	0.33	0.33	0.86	0.86	-0.67	-0.67
	Ellis/42pc	**0.32**	0.31	0.85	0.85	**-0.69**	-0.71
	Ellis/54ec	0.34	0.34	0.86	0.86	-0.60	-0.60
	Ellis/54int	**0.39**	0.38	0.88	0.88	**-0.58**	-0.61
	Ellis/66conv	0.37	0.37	0.88	0.88	**-0.63**	-0.64
	Poler	0.39	**0.43**	0.87	**0.88**	-0.65	**-0.57**
Dementia	ADReSS-IS2020	0.47	**0.51**	0.89	**0.89**	-0.43	**-0.40**
	Delaware	0.39	**0.39**	0.88	**0.88**	-0.65	**-0.65**
	Lu	0.48	**0.54**	0.88	**0.90**	-0.44	**-0.27**
	Pitt/cookie	0.50	**0.54**	0.89	**0.90**	-0.41	**-0.36**

Table 1: The median performance of all feasible ASR-LLMs on patients and controls on metrics: ROUGE, BERTScore, Bleurt. The higher the values of the three metrics, the better the performances.

Aphasia datasets (Baycrest, MSU, NEURAL, UMD, UNH, Richardson), several LIC

	Dataset	p-value (ROUGE)	p-value (BERTSCORE)	p-value (BLEURT)
Aphasia	Baycrest	1.4e-05*	1.2e-04*	1.7e-04*
	MSU	1.3e-05*	7.8e-05*	1.9e-04*
	NEURAL	2.5e-05*	3.6e-03*	5.9e-03*
	NEURAL_2	8.6e-02	7.7e-03*	1.1e-02*
	UMD	3.1e-05*	8.0e-05*	3.9e-04*
	UNH	7.1e-06*	2.1e-05*	2.4e-05*
	Richardson	8.2e-14*	3.0e-08*	7.8e-07*
LIC	ENNI	3.3e-04*	5.5e-06*	6.2e-04*
	Ellis/30ec	1.6e-01	1.0e+00	5.2e-01
	Ellis/30pc	6.2e-01	6.3e-01	3.7e-01
	Ellis/42ec	1.0e+00	6.0e-01	1.0e+00
	Ellis/42pc	6.4e-01	2.2e-01	5.5e-01
	Ellis/54ec	6.3e-01	3.4e-01	1.0e+00
	Ellis/54int	6.4e-01	1.0e+00	7.0e-01
	Ellis/66conv	1.0e+00	1.0e+00	1.0e+00
	Poler	6.8e-08*	1.1e-07*	2.8e-07*
Dementia	ADReSS-IS2020	3.1e-18*	3.9e-17*	3.9e-17*
	Delaware	5.4e-18*	1.2e-13*	1.3e-10*
	Lu	8.2e-09*	3.6e-08*	3.6e-08*
	Pitt/cookie	1.5e-97*	6.1e-88*	1.1e-87*

Table 2: Bonferroni adjusted p-values from two-sided Wilcoxon tests on hypotheses (4) to assess the performance difference between patients and controls. * indicates statistical significance at the significance level 0.05 after Bonferroni correction. Raw p-values for each dataset are presented in Figures 10, 11, 12 of the Appendix.

	Dataset	p-value (ROUGE)		p-value (BERTSCORE)		p-value (BLEURT)	
		H_{a2}	H_{a3}	H_{a2}	H_{a3}	H_{a2}	H_{a3}
Aphasia	Baycrest	1.5e-02*	7.1e-06*	9.7e-03*	6.1e-05*	1.8e-01	8.6e-05*
	MSU	1.3e-01	6.4e-06*	1.3e-01	3.9e-05*	3.2e-01	9.4e-05*
	NEURAL	1.2e-05*	1.7e-01	7.2e-03*	1.8e-03*	8.3e-01	3.0e-03*
	NEURAL_2	4.3e-02*	1.0e+00	3.8e-03*	1.1e-01	5.6e-03*	3.8e-02*
	UMD	1.4e-01	1.6e-05*	3.3e-02*	4.0e-05*	1.0e+00	1.9e-04*
	UNH	3.5e-06*	4.3e-04*	1.0e-05*	3.5e-04*	1.2e-05*	2.0e-04*
	Richardson	2.3e-07*	4.1e-14*	3.0e-05*	1.5e-08*	1.7e-03*	3.9e-07*
LIC	ENNI	6.4e-05*	1.0e+00	2.8e-06*	1.0e+00	2.0e-04*	1.0e+00
	Ellis/30ec	2.2e-03*	8.0e-02	5.8e-04*	1.0e+00	3.0e-06*	7.9e-01
	Ellis/30pc	4.6e-05*	3.7e-01	6.2e-06*	3.1e-01	1.1e-04*	2.2e-01
	Ellis/42ec	9.9e-01	1.0e+00	7.3e-01	3.0e-01	6.3e-01	1.0e+00
	Ellis/42pc	9.1e-01	3.2e-01	1.0e+00	1.1e-01	4.7e-01	2.8e-01
	Ellis/54ec	3.9e-01	3.2e-01	1.0e+00	1.7e-01	1.0e+00	1.0e+00
	Ellis/54int	3.2e-01	1.0e+00	1.0e+00	1.0e+00	1.0e+00	3.5e-01
	Ellis/66conv	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00
	Poler	4.9e-02*	3.4e-08*	1.9e-01	5.6e-08*	7.6e-02	1.4e-07*
Dementia	ADReSS-IS2020	1.5e-18*	1.7e-01	1.9e-17*	1.0e+00	1.9e-17*	2.5e-01
	Delaware	2.7e-18*	9.5e-01	6.0e-14*	7.7e-01	6.7e-11*	1.0e+00
	Lu	4.1e-09*	1.0e+00	1.8e-08*	1.0e+00	1.8e-08*	1.0e+00
	Pitt/cookie	7.5e-98*	5.4e-03*	3.1e-88*	1.0e+00	5.3e-88*	4.0e-01

Table 3: Bonferroni adjusted p-values from one-sided Wilcoxon tests on (5) and (6), with the alternative hypotheses being $H_{a2} : C > P$ and $H_{a3} : P > C$, respectively. * indicates statistical significance at the significance level 0.05 after Bonferroni correction. For tests against H_{a2}, raw p-values for each dataset are presented in Figures 13, 14, 15 of the Appendix. For tests against H_{a3}, raw p-values are presented in Figures 16, 17, 18 of the Appendix.

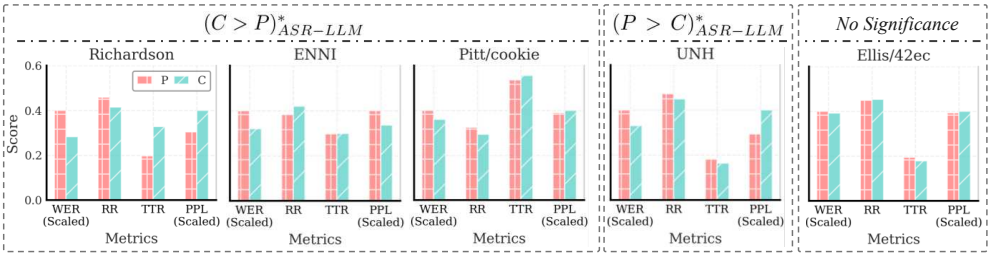


Figure 7: Visualize the average scores of WER, RR, TTR, PPL on all ASR-generated transcripts on five datasets from three cases: ASR-LLM performance was found significantly better on controls ($(C > P)^*_{ASR-LLM}$), ASR-LLM performance was found significantly better on patients,($(P > C)^*_{ASR-LLM}$) and no significant difference was found between patients and controls. WER and PPL were scaled to the range of 0 to 0.4 so they could share the y-axis with the RR and TTR.

datasets (ENNI, Ellis/30ec, Ellis/30pc, Poler), and all Dementia datasets (ADReSS-IS2020, Delaware, Lu, Pitt/cookie). In contrast, significant patient advantages were also observed in the same set of Aphasia datasets, particularly in UNH, where more ASR-LLMs favored patients, highlighting its distinct characteristics. This bidirectional result in Aphasia reflects that different ASR-LLMs support opposite one-sided hypotheses, rather than a contradiction within a single model.

In cases where both one-sided tests ($C > P$ and $P > C$) were significant, the results were supported by different ASR-LLMs, as visualized in Figures 14–18. These findings provide significant evidence against the null hypothesis (H_0), supporting the alternative hypotheses H_{a2} and H_{a3}.

2.4 Assess the quality of ASR-generated transcripts: How they were related to ASR-LLM performance

The quality of ASR-generated transcripts impacted the ASR-LLM performance. To quantify transcript quality and assess its relation to ASR-LLM performance, four quality metrics were applied. Word Error Rate [23, 24] (WER) directly measured the proportion of transcription errors; higher scores indicated lower transcript accuracy, meaning the ASR made more mistakes during transcription. Repetition Ratio [25–27] (RR) quantified vocabulary repetition; higher scores indicated greater levels of stuttering or disfluency presented in the audio. Type-Token Ratio [28, 29] (TTR) measured vocabulary diversity; higher scores indicated a richer vocabulary within the text. In this study, due to the fact that ASR may incorrectly recognize language fillers (e.g., hum, eh, etc.) as words in another language, TTR might also reflect ASR detection noise. Perplexity [30, 31] (PPL) reflected the semantic complexity of the text; higher scores indicate that the text was more difficult to understand.

Representative datasets: Richardson (Aphasia), ENNI (LIC), and Pitt/cookie (Dementia) for the case of ASR-LLM performance significantly better on controls; UNH (Aphasia) for the case of LLM perfomance significantly better on patients; and Ellis-42ec (LIC) for the case of no significance between Controls and Patients, were selected

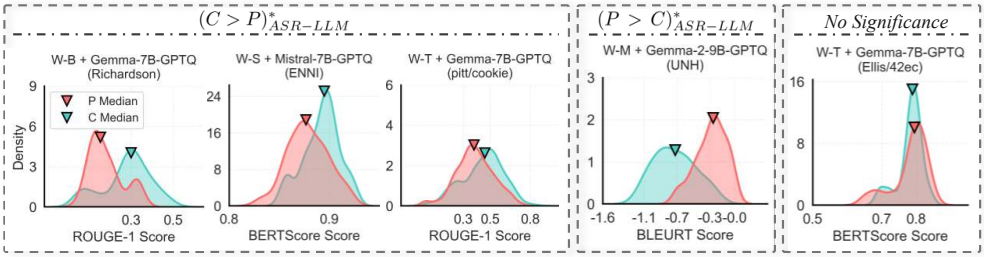


Figure 8: Visualize the performance distributions of exemplary ASR-LLMs on the five datasets in three cases: significantly better performance on controls than patients ($(C > P)^*_{ASR-LLM}$), significantly better performance on patients than controls ($(P > C)^*_{ASR-LLM}$), and no significant differences.

to demonstrate how the ASR-generated transcripts were reflected on the matrics. The WER, RR, TTR, and PPL scores computed for the ASR-generated transcripts from these five datasets are presented in Figure 7.

Across Richardson, ENNI, Pitt/Cookie, and UNH, the average WER was higher in patients than in controls. However, this pattern did not align with the ASR-LLM performance significance (i.e., $(C > P)^*_{ASR-LLM}$ and $(P > C)^*_{ASR-LLM}$). This may suggest that ASR errors were not positively associated with ASR-LLM performance. In Ellis/42EC, there was no difference in WER between patients and controls, which was consistent with the patterns observed in RR, TTR, and PPL.

Average RR scores were higher in Richardson (Aphasia), UNH (Aphasia), and Pitt/cookie (Dementia), due to the fact that increased vocabulary repetition was evident in these two datasets. In Ellis/42ec (LIC), Average RR scores were roughly equivalent between patients and controls. This may explain the lack of significant ASR-LLM performance differences. Patterns of PPL appeared to align with those observed for RR.

For TTR, average scores were higher in controls for Richardson (Aphasia) and Pitt/cookie (Dementia). This pattern was reversed in UNH (Aphasia). In Ellis/42ec (LIC) and ENNI (LIC), patients and controls exhibited similar TTR average scores.

2.5 Verify the identified transcripts patterns to improve the ASR-LLM performance

To demonstrate representative transcript patterns from the five datasets analyzed in the previous section, one patient sample and one control sample with median ASR-LLM performance were selected, as shown in Figure 8. These selected samples are presented in Table 4 ($(C > P)^*_{ASR-LLM}$), Table 5 ($(P > C)^*_{ASR-LLM}$), and Table 6 (no significant differences). In these sample transcripts, clear patterns of repetition and noise were observed.

In the example of $(C > P)^*_{ASR-LLM}$, removing repetitions and noise from patient transcripts improved ASR-LLM performance (Table 4). Notably, removing noise led to greater performance improvements than removing repetitions alone, suggesting that

noise had a stronger negative impact on the LLM. After removing the language issues, ASR-LLM performances were much closer between patients and controls.

In the example of $(P > C)^*_{ASR-LLM}$, repetitions and noise were present in both patient and control transcripts. While removing these factors from the control transcript improved ASR-LLM performance, the resulting performance (BLEURT: -0.3854) remained significantly lower than that on the patient sample (BLEURT: -0.2599). The control transcript contained substantially more tokens than the patient transcript (5109 vs. 1209). As shown in Table 5, truncating the control transcript to match the length of the patient transcript (1209 tokens) further improved performance.

In the example of no significant ASR-LLM performance differences between patients and controls, as shown in Table 6, the lengths of the control and patient transcripts were comparable. After removing language pattern issues, both the patient's example and the control's example improved, and were still comparable. These findings were consistent with our previous analyses, suggesting that (1) repetition and noise degrade LLM performance and (2) comparable transcript lengths between patient and control groups contribute to similar performance levels.

Table 4: On the $(C > P)^*_{ASR-LLM}$ examples, illustrating how the ASR-LLM performance changed after removing identified issues in ASR-generated transcripts.

Richardson via Whisper-Base + Gemma-2B-GPTQ		
Patient: 2053 tokens		**ROUGE**
Original	Like, do you know you're in a coma? [...] She's three days late. **MBC 뉴스 김지훈입니다.** Yep. [...] gets over the ball. **It's , It's , [... It's , ...] It's , It's ,** So what you'd be missing? you	0.1294
Remove Repetition	Like, do you know you're in a coma? [...] She's three days late. **MBC 뉴스 김지훈입니다.** Yep. [...] gets over the ball. [...] So what you'd be missing? you	0.1764
Remove Noise	Like, do you know you're in a coma? [...] She's three days late. Yep. [...] gets over the ball. **It's , It's , [... It's , ...] It's , It's ,** So what you'd be missing? you	0.2470
Remove Both	Like, do you know you're in a coma? [...] She's three days late. Yep. [...] gets over the ball. So what you'd be missing? you	0.2709
Control: 1120 tokens		**ROUGE**
Original	Okay. So there was this boy. and he liked to play soccer. [...] on your second slice of bread. And then you put the two slices of bread together. and that's it.	**0.3118**
ENNI via Whisper-Small + Mistral-7B-GPTQ		
Patient: 629 tokens		**BERTScore**
Original	The elephant is playing with apple. [...] **ac ydych chi'n gwybod i'r ffordd i'r ffordd i'r ffordd. Rwyf yn ddweud y [... ddweud y ...] ddweud y** He wants to go [...] **그래서, 이거 아삭한 것 같아요 www[... www ...]www** They both got the envelope.	0.8685
Remove Repetition	The elephant is playing with apple. [...] **ac ydych chi'n gwybod i'r ffordd i'r ffordd i'r ffordd. Rwyf yn** He wants [...] **그래서, 이거 아삭한 것 같아요** They both got the envelope.	0.8739
Remove Noise	The elephant is playing with apple. [...] **ddweud y [... ddweud y ...] ddweud y** He wants to go [...] **www[... www ...]www** They both got the envelope.	0.8820
Remove Both	The elephant is playing with apple. [...] He wants to go see the circus bunny again to get us a pony [...] They both got the envelope.	0.8917
Control: 492 tokens		**BERTScore**
Original	at swimming pool. They saw a ball in the water. It's in a horse or a jolly. [...] The giraffe is trying to swim through the ball.	**0.8950**
Pitt/cookie via Whisper-Tiny + Gemma-7B-GPTQ		
Patient: 106 tokens		**ROUGE**
Original	You see [...] **Ja, jag vet inte.** Yeah, [...] and a little boys trying to get **cookies jar up and [...] and the cookie jar up.** and there's an official girl with him. **And this is the bancher of what he had called it.** He looks like it's going to fall. **I'm going to... [...] I'm going to...** I'm going to borrow, [...] Okay, that's fine.	0.41509
Remove Repetition	You see [...] **Ja, jag vet inte.** Yeah, [...] and a little boys trying to get cookies up. and there's an official girl with him. **And this is the bancher of what he had called it.** He looks like it's going to fall. I'm going to borrow, [...] Okay, that's fine.	0.4528
Remove Noise	You see [...] Yeah, [...] and a little boys trying to get cookies up. and there's an official girl with him. He looks like it's going to fall. **I'm going to... [...] I'm going to...** I'm going to borrow, [...] Okay, that's fine.	0.4716
Remove Both	You see [...] Yeah, [...] and a little boys trying to get cookies up. and there's an official girl with him. He looks like it's going to fall. I'm going to borrow, [...] Okay, that's fine.	0.5094
Control: 83 tokens		**ROUGE**
Original	There's a girl waiting to get a cookie from I guess her brother chairs tip and over where he's in. He's in the cookie jar and they cover it. [...] Windows open.	**0.4999**

Table 5: On the $(P > C)^*_{ASR-LLM}$ examples , illustrating how the ASR-LLM performance changed after removing identified issues in ASR-generated transcripts

UNH via Whisper-Medium + Gemma-2-9B-GPTQ		
Patient: 1209 tokens		**BLEURT**
Original	So for this part, I want to ask you [...] **MBC 뉴스 김수근입니다**. going to the... um... **the... [...] the...** No. Yeah. **빨리** eating. Awesome.	**-0.2599**
Control: 5109 tokens		**BLEURT**
Original	Something, ... **KATHRYN** The last time I had a really bad case of the flu was last year. [...] I knew him through **ORP** you [...] **amily** was only ever have like tiny dogs **아, [...] 아,** or just other [**... 1013 Tokens ...**]	-0.7138
Remove Repetition	Something, ... **KATHRYN** The last time I had a really bad case of the flu was last year. [...] I knew him through **ORP** you [...] **amily** was only ever have like tiny dogs. or just other [**... 1013 Tokens ...**]	-0.4139
Remove Noise	Something, ... The last time I had a really bad case of the flu was last year. [...] I knew him through you [...] was only ever have like tiny dogs **아, [...] 아,** or just other [**... 1013 Tokens ...**]	-0.5601
Remove Both	Something, ... The last time I had a really bad case of the flu was last year. [...] I knew him through you [...] was only ever have like tiny dogs. or just other [**... 1013 Tokens ...**]	-0.3854
Truncate	Something, ... The last time I had a really bad case of the flu was last year. [...] I knew him through you [...] was only ever have like tiny dogs. or just other.	-0.3245

Table 6: On the *No Significance* examples, illustrating how the ASR-LLM performance changed after removing identified issues in ASR-generated transcripts.

Ellis/42ec via Whisper-Tiny + Gemma-7B-GPTQ		
Patient: 1501 tokens		**BERTScore**
Original	It's too dangerous. **Приходите, что это променное** Oh, maybe she can just sit here on these stairs. So that'd be our Yes. Oh, this. Okay. **안녕** Oh there. أرَا **[...]** أرَا [...] Thank you. you're just gonna have to put	0.7949
Remove Repetition	It's too dangerous. **Приходите, что это променное** Oh, maybe she can just sit here on these stairs. So that'd be our Yes. Oh, this. Okay. **안녕** Oh there. [...] Thank you. you're just gonna have to put	0.8020
Remove Noise	It's too dangerous. Oh, maybe she can just sit here on these stairs. So that'd be our Yes. Oh, this. Okay. Oh there. أرَا **[...]** أرَا [...] Thank you. you're just gonna have to put	0.8204
Remove Both	It's too dangerous. Oh, maybe she can just sit here on these stairs. So that'd be our Yes. Oh, this. Okay. Oh there. [...] Thank you. you're just gonna have to put	0.8214
Control: 1478 tokens		**BERTScore**
Original	Thank you all the best. 我來的氣球 I have a lot of stuff in there, right? and sit in bed in each この後 I don't like dying hair. الُ كَيفَ فِ رَسِهِ That's a good place for you. **Oh, [...] oh,** [...] We're going to look at this.	0.7889
Remove Repetition	Thank you all the best. 我來的氣球 I have a lot of stuff in there, right? and sit in bed in each この後 I don't like dying hair. الُ كَيفَ فِ رَسِهِ That's a good place for you. [...] We're going to look at this.	0.8043
Remove Noise	Thank you all the best. I have a lot of stuff in there, right? and sit in bed in each. I don't like dying hair. That's a good place for you. **Oh, [...] oh,** [...] We're going to look at this.	0.8002
Remove Both	Thank you all the best. I have a lot of stuff in there, right? and sit in bed in each. I don't like dying hair. That's a good place for you. [...] We're going to look at this.	0.8017

3 Discussion

Evaluate ASR-LLM Feasibility: We evaluated 115 ASR-LLMs across three speech-related conditions (Aphasia, LIC, and Dementia) over 20 datasets. Therefore, 2300 experimental trials, each using one ASR-LLM on one dataset, were conducted. Trials were flagged as errors if any sample failed to process. Errors were widespread in Aphasia datasets, suggesting limitations of ASR-LLMs on procedding language inputs related to this condition. In LIC, errors were restricted to specific LLMs, indicating model-related rather than disease-related issues. Errors occurred in both patients and controls, though controls triggered fewer errors, suggesting errors were not solely caused by impaired speech.

To understand error distribution, we examined trials with widespread errors vs. those with few errors. The errors concentrated on certain datasets and certain LLMs. On the datasets that had more errors, ASR-LLM showed consistently higher error rates on controls' speeches than on the patients', suggesting that certain characteristics of controls' speech were more prone to errors. Among the LLMs, error rates were substantially higher in Aphasia datasets, especially among controls. Analysis on the results demonstrated that factors, such as sentence length and noise, tended to cause errors. However, these factors might not directly link to language impairments, as all selected Aphasia datasets exhibited higher error rates on controls than on patients.

Two common error types were identified. The most dominant error, ERROR1, appeared in samples exceeding 10,000 tokens, linked to rotary positional encoding and memory demand. The instability was caused by excessive sequence length, which led to out-of-memory issues during inference. ERROR2, seen in two trials, arose during gradient calculation [32, 33], where the denominator became zero, triggered around 6,000 tokens. These findings suggested that transcripts should be limited to a certain number of tokens (6,000 in this study) to avoid inference errors. This is expected for on-device usage, which has limited memory and processing capability.

We further evaluated the impact of LLM compression on ASR-LLM errors using all error trials. Compressed LLMs showed fewer ERROR1 errors, potentially due to reduced computational demands [34, 35], which leave more memory headroom during inference. However, compressed LLMs exhibited a slight increase in ERROR2, though ERROR1 remained the dominant error type.

Additionally, we analyzed the impact of ASR model size on error rates. Larger ASRs produced more accurate transcripts and reduced errors. In LIC, the largest ASR eliminated all errors. In Aphasia, error reduction was less pronounced using larger ASR, likely due to intrinsic trial characteristics. Even optimal ASR models could not fully resolve errors in long/complex Aphasia transcripts as the main constrain was on LLM, underscoring the need for on-device LLM that can handle longer sentences directly or other mechanisms that concise long sentences into short ones without losing key information.

Measure ASR-LLM Performance: After evaluating the feasibility of ASR-LLMs, we developed a framework to measure their performance. In each dataset, every sample included the audio and the corresponding human-annotated transcript. For each audio, the ASR-LLM generated a textual output. Calculating LLM performances required a reference (the best case possible). Since the TalkBank datasets did not

include such references, we constructed references by inputting the human-annotated transcript into a well-performing cloud-based LLM (GPT-4o), which offers a top-notch reasoning capability of cloud-based LLMs. Then the similarity between the on-edge-device ASR-LLM output and the cloud-based LLM output was used to generate edge ASR-LLM performance measures.

Identify Differences and Key Factors: For ROUGE, scores above 0.4 indicate good surface-level matching between the ASR-LLM output and the cloud-based LLM reference, while scores above 0.3 are still considered acceptable [36]. BERTScore, which captures sentence-level semantic similarity (as opposed to ROUGE's token-level matching), typically indicates good performance when ranging between 0.85 and 0.9 [21]. BLEURT, a learned evaluation metric, is calibrated to assess overall text quality, with lower (often negative) values being common [22]. Based on the three matrices, results showed that ASR-LLMs performed the best on Dementia datasets and the worst on Aphasia, across both patients and controls.

Within each disease, ASR-LLM performance differences between patients and controls were also examined. Two-sided Wilcoxon tests showed significant differences in all the Aphasia and Dementia datasets. In LIC, significant differences were found in ENNI and Polar only. This showed that ASR-LLM is less likely to perform differently in children living with LIC, compared to children with out language impairments. One-sided Wilcoxon tests further revealed that, in cases where ASR-LLM performed differently on patients vs. controls, the performance were mostly better on controls across all three diseases.

Repetition, noise, and transcript length affect ASR-LLM performance: As ASR models produced transcripts as LLMs' inputs, text matrices of the generated transcripts were analyzed to assess potential correlations between transcript quality and ASR-LLM performance. Two factors, repetition and noise, emerged as relevant. From examples extracted from both patients and controls. we could see that when ASR-LLM performed lower on patients, removing repetition and noise improved ASR-LLM performance, bring it closer to that of the controls. In the cases when ASR-LLM performed better on patients, similar removal on controls' ASR transcripts improved the performance to some extent. However, a more important phenomenon in such cases was that the controls' transcripts were usually much longer than the patients'. Due to the fact that on-device ASR-LLMs cannot handle long transcripts, truncating controls' transcripts to match with patients' transcript length yielded even greater ASR-LLM improvements. In detail, this aligned with our earlier error analysis: transcripts $>$ 6,000 tokens increase LLM errors. To ensure proper ASR-LLM use, designs should prioritize brevity and conciseness.

References

[1] Jones, C.H., Dolsten, M.: Healthcare on the brink: navigating the challenges of an aging society in the united states. npj Aging **10**(1), 22 (2024)

[2] Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature medicine **29**(8), 1930–1940 (2023)

[3] National Institute on Deafness and Other Communication Disorders: Quick Statistics About Hearing, Balance, & Dizziness. https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing. Accessed: March 12, 2025 (2020)

[4] Notara, V., Vagka, E., Gnardellis, C., Lagiou, A.: The emerging phenomenon of nomophobia in young adults: a systematic review study. Addiction & health **13**(2), 120 (2021)

[5] Moyle, W.: Grand challenge of maintaining meaningful communication in dementia care. Frontiers in Dementia **2**, 1137897 (2023) https://doi.org/10.3389/frdem.2023.1137897 . PMID: 39081994; PMCID: PMC11285537

[6] Murdoch, B.: Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC medical ethics **22**, 1–5 (2021)

[7] Whisper — OpenAI Research. https://openai.com/research/whisper. Accessed: March 12, 2025

[8] Google Cloud Speech-to-Text. https://cloud.google.com/speech-to-text. Accessed: March 12, 2025

[9] Asaithambi, S., Nallusamy, S., Yang, J., Prajapat, S., Kumar, G., Rathore, P.S.: A secure and trustworthy blockchain-assisted edge computing architecture for industrial internet of things. Scientific Reports **15**(1), 15410 (2025)

[10] Qin, R., Xia, J., Jia, Z., Jiang, M., Abbasi, A., Zhou, P., Hu, J., Shi, Y.: Enabling on-device large language model personalization with self-supervised data selection and synthesis. In: Proceedings of the 61st ACM/IEEE Design Automation Conference, pp. 1–6 (2024)

[11] Shridhar, K., Stolfo, A., Sachan, M.: Distilling reasoning capabilities into smaller language models. Findings of the Association for Computational Linguistics: ACL 2023, 7059–7073 (2023)

[12] Qin, R., Liu, D., Xu, C., Yan, Z., Tan, Z., Jia, Z., Nassereldine, A., Li, J., Jiang, M., Abbasi, A., et al.: Empirical guidelines for deploying llms onto resource-constrained edge devices. ACM Transactions on Design Automation of Electronic Systems (2024)

[13] Rehabilitation, People With Aphasia After Stroke (RELEASE) Collaborators, R.: Predictors of poststroke aphasia recovery: A systematic review-informed individual participant data meta-analysis. Stroke **52**(5), 1778–1787 (2021)

[14] Leonard, L.B.: Specific language impairment across languages. Speech and language impairments in children, 115–129 (2014)

[15] Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E., O'Brien, M.: Prevalence of specific language impairment in kindergarten children. Journal of speech, language, and hearing research **40**(6), 1245–1260 (1997)

[16] Kramarow, E.A.: Diagnosed dementia in adults age 65 and older: United states, 2022. National Health Statistics Report 203, National Center for Health Statistics, CDC (June 2024). https://doi.org/10.15620/cdc/154766

[17] Liu, Y., Shi, K., He, K.S., Ye, L., Fabbri, A.R., Liu, P., Radev, D., Cohan, A.: On learning to summarize with large language models as references. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2024). https://arxiv.org/abs/2305.14239

[18] Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., Raffel, C.: Evaluating the factual consistency of large language models through news summarization. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, pp. 5220–5255. Association for Computational Linguistics, Toronto, Canada (2023). https://aclanthology.org/2023.findings-acl.322/

[19] Kazemnejad, A., Padhi, I., Natesan Ramamurthy, K., Das, P., Reddy, S.: The impact of positional encoding on length generalization in transformers. Advances in Neural Information Processing Systems **36**, 24892–24928 (2023)

[20] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[21] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)

[22] Sellam, T., Das, D., Parikh, A.P.: Bleurt: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696 (2020)

[23] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., Goel, S.: Racial disparities in automated speech recognition. Proceedings of the national academy of sciences **117**(14), 7684–7689 (2020)

[24] Holmlund, T.B., Chandler, C., Foltz, P.W., Cohen, A.S., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Elvevåg, B.: Applying speech technologies to assess verbal

memory in patients with serious mental illness. NPJ digital medicine **3**(1), 33 (2020)

[25] Cettolo, M., Bertoldi, N., Federico, M.: The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In: Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track, pp. 166–179 (2014)

[26] Fraser, K.C., Meltzer, J.A., Graham, N.L., Leonard, C., Hirst, G., Black, S.E., Rochon, E.: Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex **55**, 43–60 (2014)

[27] Huang, D., Nguyen, T.-S., Liausvia, F., Wang, Z.: Rap: A metric for balancing repetition and performance in open-source large language models. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 1479–1496 (2025)

[28] Tweedie, F.J., Baayen, R.H.: How variable may a constant be? measures of lexical richness in perspective. Computers and the Humanities **32**, 323–352 (1998)

[29] Covington, M.A., McFall, J.D.: Cutting the gordian knot: The moving-average type–token ratio (mattr). Journal of quantitative linguistics **17**(2), 94–100 (2010)

[30] Jelinek, F., Mercer, R.L., Bahl, L.R., Baker, J.K.: Perplexity—a measure of the difficulty of speech recognition tasks. The Journal of the Acoustical Society of America **62**(S1), 63–63 (1977)

[31] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

[32] Li, S., Xue, F., Baranwal, C., Li, Y., You, Y.: Sequence parallelism: Long sequence training from system perspective. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2391–2404. Association for Computational Linguistics, Toronto, Canada (2023). ACL Anthology ID: 2023.acl-long.134

[33] Variš, D., Bojar, O.: Sequence length is a domain: Length-based overfitting in transformer models. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021). Association for Computational Linguistics

[34] Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: International Conference on Learning Representations (ICLR) (2016)

[35] Dettmers, T., Zettlemoyer, L., Lewis, M., Shleifer, S.: 8-bit optimizers via block-wise quantization. In: International Conference on Learning Representations (ICLR) (2022)

[36] Ganesan, K.: Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. arXiv preprint arXiv:1803.01937 (2018)

[37] MacWhinney, B.: Understanding spoken language through talkbank. Behavior research methods **51**, 1919–1927 (2019)

[38] NVIDIA: Canary-1B: A multilingual ASR/AST model from NVIDIA NeMo. Hugging Face model card. https://huggingface.co/nvidia/canary-1b (2024)

[39] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518 (2023). PMLR

[40] NVIDIA, Suno.ai: Parakeet-RNNT-1.1B: FastConformer-Transducer ASR model from NVIDIA NeMo & Suno.ai. Hugging Face model card. https://huggingface.co/nvidia/parakeet-rnnt-1.1b (2024)

[41] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[42] Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D.: Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323 (2022)

[43] Renze, M.: The effect of sampling temperature on problem solving in large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 7346–7356 (2024)

[44] Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[45] Brown, P.F., Pietra, V.J.D., Mercer, R.L., Pietra, S.A.D., Lai, J.C.: An estimate of an upper bound for the entropy of english. Computational Linguistics **18**(1), 31–40 (1992)

[46] Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines **30**, 681–694 (2020)

A Method

A.1 Preparing datasets for the ASR-LLM study

We included 2,097 participants from the Talkbank database [37]. As shown in Table 9, Aphasia had seven datasets, including 227 patients and 210 controls in total; LIC had nine datasets, including 363 patients and 607 controls; Dementia had four datasets, including 448 patients and 356 controls. The dataset was selected only if it contained the patient and control samples. And each sample had the corresponding transcript made by professional human annotators. All datasets included the participant and the investigator. The participant could be either the patient or the control. The investigator was a doctor or a person related to clinical investigation.

The Aphasia datasets involved participants describing a story or a picture, with the investigator providing an initial prompt and limited further involvement. The Dementia datasets used similar methods, with the investigator guiding the conversation at the beginning and allowing participants to speak freely. The LIC datasets recorded active interactions between the investigator and the child participant throughout the session. These included play, toy descriptions, and storytelling tasks designed to induce natural language from the child.

In our study, each audio in each dataset was converted into a transcript via the ASR models. During the conversion, while the audio could include the participant and the investigator, the participant's part was not split from the investigator's to avoid context loss. Additionally, since the original recordings were professionally designed for clinical studies within TalkBank, and featured utterance-aligned transcripts directly linked to audio media, splitting the audio could compromise the integrity of the data [37].

A.2 Building up the proper ASR-LLM

To enable audio input for the LLM, an ASR model was used to convert the audio into a transcript. In the ASR-LLM pipeline, the transcript was automatically forwarded to the LLM, which then generated a response based on the transcript. Three aspects required explicit definition: (1) ASR model selection, (2) LLM selection, and (3) default settings.

For ASR model selection, four commonly used ASR models were considered: Canary-1B [38] (NVIDIA), Whisper-base [39] (OpenAI), Parakeet-RNNT-1.1B [40] (NVIDIA), and Wav2Vec2-base [41] (Meta). These four ASR models had comparable model sizes. Each model family included multiple versions with varying sizes. To identify which models would be suitable for constructing the ASR-LLM pipeline, each ASR model was paired with an LLM (Phi-3.5-mini-instruct) to generate text from audio in three representative datasets: Baycrest (Aphasia), Poler (LIC), and Delaware (Dementia). As shown in Figure 9, the combination of Whisper-base with the LLM yielded the best performance. In addition, the Whisper family offers five versions of varying sizes, is widely used, and is easy to deploy. Therefore, Whisper models were selected for use in this study.

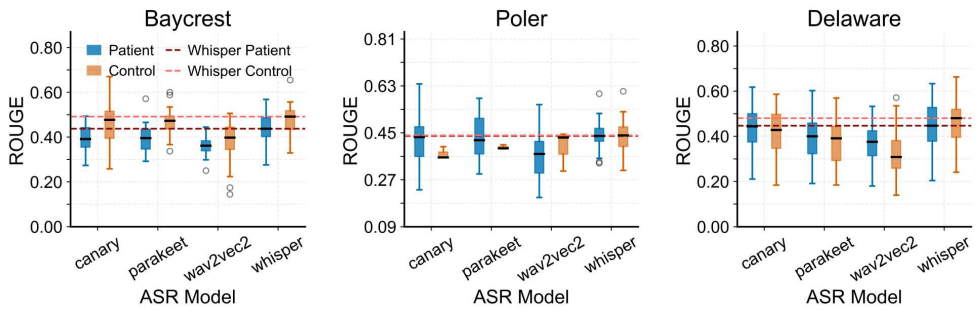


Figure 9: Examine the performance of LLM (Phi-3.5-mini-instruct) on the transcript generated by different types of ASR models: Canary-1B (canary), Parakeet-RNNT-0.6B (parakeet), Wav2vec2-Large-960h (wav2vec2), and Whisper-Medium (whisper)

For LLM selection, five LLM families were included: Gemma (Google), Mistral (Mistral AI, France), Phi (Microsoft), Llama (Meta), and Qwen (Alibaba, China). These LLMs are widely used, with each having over 100,000 downloads on Hugging Face. Their model sizes were below 20 GB, allowing them to run on commonly available hardware such as NVIDIA Jetson Orin or NVIDIA Jetson AGX. For each LLM family, compressed versions based on quantization were also included.

There were two types of LLMs being tested: original LLMs and compressed LLMs, as shown in Table 8. For original LLMs, they were trained from scratch. Their parameter size was initialized to be small. For compressed LLMs, they were compressed from the original LLMs. The parameter size of compressed LLMs could be initialized to be large. They were compressed by lowering the precision of parameters, so to lower to resource usage. The compression technique used GPTQ [42], a widely used LLM compression method. For all tested models, GPTQ could be well compatible with them. When we used the compressed LLMs, their parameter precision was set to 4-bit integer, with the activations as 16-bit floating-point (FP16).

For the default settings used to construct the ASR-LLM pipeline, the following configurations were applied across all experiments. The ASR sampling rate was set to 16,000 Hz. For the LLM, no input length limit was imposed to preserve the integrity of the transcripts. The LLM output was constrained to 100 tokens. The generation temperature and top_p were set to 0.1 and 0.9, respectively, to ensure stable and consistent generation [43].

A.3 Measuring the ASR-LLM performance

The Aphasia, LIC, and Dementia datasets only contained the transcript made by the human annotator and the audio for each sample. The human-made transcript was sent to the cloud-based LLM for summarization. In the ASR-LLM, the audio was converted into the ASR-made transcript, which was summarized by the LLM. For both cloud-based LLM (GPT-4o) and the LLM in ASR-LLM, the summarization was based on the text format:

"*Please summarize the following transcript made by an experiment participant.*

Transcript: <ASR-made/human annotator-made transcript>

Summary:"

The performance of ASR-LLM was calculated by measuring the similarity between the two summarizations made by the cloud-based LLM and the LLM in ASR-LLM. Higher similarity indicates better performance, as it suggests the ASR-LLM can generate outputs more aligned with the cloud-based reference. The good performance of the ASR-LLM indicated that it may serve individuals with language impairments in the same way as the human annotators and the cloud-based LLM.

There were three metrics to measure the similarity. (1) ROUGE [20] measure the token-to-token similarity. It computes the overlap between the LLM-produced output and the reference based on n-grams, sequences, and word pairs, providing a straightforward and interpretable estimate of the textual similarity. (2) BERTScore [21] extends beyond token-level matching by leveraging contextual embeddings from BERT[44]. It calculates the cosine similarity between the contextual representations of the output and the reference, thus reflecting deeper semantic alignment. (3) BLEURT [22] uses pretrained and fine-tuned BERT models to produce a more human-like judgment of text quality. By adapting BERT to the domain-specific data, BLEURT often better aligns with human evaluations compared to traditional methods. Collectively, these three metrics offer a robust and reliable framework for assessing the similarity between the ASR-LLM output and the reference summary.

A.4 Measuring the quality of the ASR-generated transcripts

ASR-generated transcripts, as the input to the LLM, could be the origin of the ASR-LLM performance difference. State-of-the-art ASRs directly convert a speech audio word-by-word into text, which often includes noise, grammar issues (e.g., repeated words), and speech fillers (e.g., "uh", "hmm"). These "noises" can be complicated structures and massive parameters for LLM to extract semantic meanings.

To analyze the quality of the transcripts, four well-established metrics were used. Perplexity (PPL) measures how well a language model predicts a given text, with lower values indicating more natural language flow [45]. In this analysis, WER was used to measure the errors in ASR-generated transcripts. It reflected whether the ASR could recognize the words from the input audio. PPL was used to detect speech fluency disruptions and unusual word sequences, which are common in aphasic speech. Type-Token Ratio (TTR), defined as the ratio of unique words to total words [29], reflects the diversity and richness of vocabulary. Lower TTR values can highlight limitations in lexical variety, indicating potential word-finding difficulties or constrained language use. Finally, Repetition Ratio (RR) quantifies the proportion of repeated words within the transcript, regardless of overall vocabulary diversity [26]. This metric specifically captures repetitive word usage (perseveration), a marker of impaired word retrieval,

even when overall vocabulary diversity may remain moderate. Example computations for these metrics are provided in Table 7.

Table 7: Comparison of Linguistic Metrics for ASR Transcripts

Metric	Example Transcript	Computation and Values
PPL	**Good:** *I went to the store and bought apples. Then I met my friend at the cafe.* **Impaired:** *Uh... I, uh, went... uh, to, um, the store... and, uh, uh, got some, um, uh, apples.*	$PPL = e^{\text{Loss}}$, computed using a language model where loss is the negative log-likelihood of the input text. **Good:** 18.2 (lower = more structured language) **Impaired:** 78.5 (higher = less predictable, fragmented speech)
TTR	**Good:** *I went to the store and bought apples. Then I met my friend at the cafe.* **Impaired:** *Uh... I, uh, went... uh, to, um, the store... and, uh, uh, got some, um, uh, apples.*	$TTR = \frac{\text{Unique Words}}{\text{Total Words}}$ **Good:** $14/16 = 0.85$ (high = diverse vocabulary) **Impaired:** $6/14 = 0.43$ (low = frequent word repetition)
RR	**Good:** *I went to the store and bought apples. Then I met my friend at the the cafe.* **Impaired:** *Uh... I, uh, went... uh, to, um, the store... and, uh, uh, got some, um, uh, apples.*	$RR = \frac{\text{Repeated Words Count}}{\text{Total Unique Words}}$ **Good:** Repeated word (the counted once): $1/14 = 0.03$ (low = minimal word repetition) **Impaired:** Repeated word (uh counted for each extra occurrence): $5/6 = 0.21$ (high = excessive repetition)
WER	**Reference:** *I went to the store and bought apples.* **ASR Output (Good):** *I went to the store and bought apples.* **ASR Output (Impaired):** *I go store bought apple.*	$WER = \frac{S + D + I}{N}$ where S = substitutions, D = deletions, I = insertions, and N = reference word count. **Good:** $0/7 = 0.00$ (perfect ASR match) **Impaired:** (3 subs + 1 deletion)$/7 = 0.57$ (high = more ASR errors)

B Tables and Figures

Table 8: LLM Information

[L]**LLM**	**Weights (GB)**
[L01] Gemma-2-2B-GPTQ[1]	2.080
[L02] Gemma-2-2B	5.231
[L03] Gemma-2-9B-GPTQ	6.170
[L04] Gemma-7B-GPTQ	7.180
[L05] Gemma-7B	17.070
[L06] Gemma-2-9B	18.480
[L07] Ministral-8B-GPTQ	5.760
[L08] Mistral-7B-GPTQ	4.170
[L09] Ministral-3B	6.618
[L10] Mistral-7B	14.500
[L11] Ministral-8B	16.030
[L12] Phi-3.5-mini-GPTQ	2.280
[L13] Phi-3.5-mini	7.640
[L14] Llama-3.2-1B-GPTQ	1.520
[L15] Llama-3.2-3B-GPTQ	2.260
[L16] Llama-3.2-1B	2.470
[L17] Llama-2-7B-GPTQ	3.900
[L18] Llama-3.2-3B	6.430
[L19] Llama-2-7B	13.480
[L20] Qwen2-2B-GPTQ	2.480
[L21] Qwen2-2B	4.419
[L22] Qwen2-7B-GPTQ	6.920
[L23] Qwen2-7B	12.674

[1]GPTQ [42] is a widely used compression method, originally designed to compress Generative Pre-trained Transformer models (GPT) [46].

Table 9: Dataset information, including the study purpose of the dataset and the sample size of the patient (P) and control (C)

[D] **Dataset**	**Study Purpose**	**P**	**C**
[D01] Baycrest	Given the Cinderella story book, the participants were asked to retell the story in their own words.	14	20
[D02] MSU	Ask participants about their experience of stroke.	16	18
[D03] NEURAL	Primarily interested in spoken language, especially discourse (connected speech).	48	48
[D04] NEURAL_2	Continuous investigations, same format as NEURAL.	74	42
[D05] UMD	This project compared standardized test and narrative language performance (AphasiaBank protocol).	18	16
[D06] UNH	Same as above.	35	12
[D07] Richardson	Same as above.	22	54
[D08] ENNI	The Edmonton Narrative Norms Instrument (ENNI) database was collected to develop measures and norms for storytelling using stimuli designed for the purpose.	66	287
[D09] Ellis/30ec	30-month child exposed to novel word learning tasks under varying processing demands	40	65
[D10] Ellis/30pc	30-month child interacting with parent during free play with toys	51	64
[D11] Ellis/42ec	42-month child interacting with examiner during semi-structured play with toys	41	43
[D12] Ellis/42pc	42-month child interacting with parent during free play with toys	39	29
[D13] Ellis/54ec	54-month child interacting with examiner during semi-structured play with toys	33	20
[D14] Ellis/54int	54-month child participating in a structured interview with examiner	31	36
[D15] Ellis/66conv	66-month child engaging in free conversation with examiner without toys	36	37
[D16] Poler	25 children with epilepsy (CWE) matched by age and gender to unaffected peers.	26	26
[D17] ADReSS -IS2020	"Cookie Theft" picture description task. Designed for classifying Alzheimer from normal.	54	54
[D18] Delaware	Picture descriptions and procedural narrative tasks (e.g., "Cookie Theft").	60	34
[D19] Lu	Cookie Theft picture descriptions.	28	26
[D20] Pitt/cookie	Cookie Theft picture descriptions.	306	242

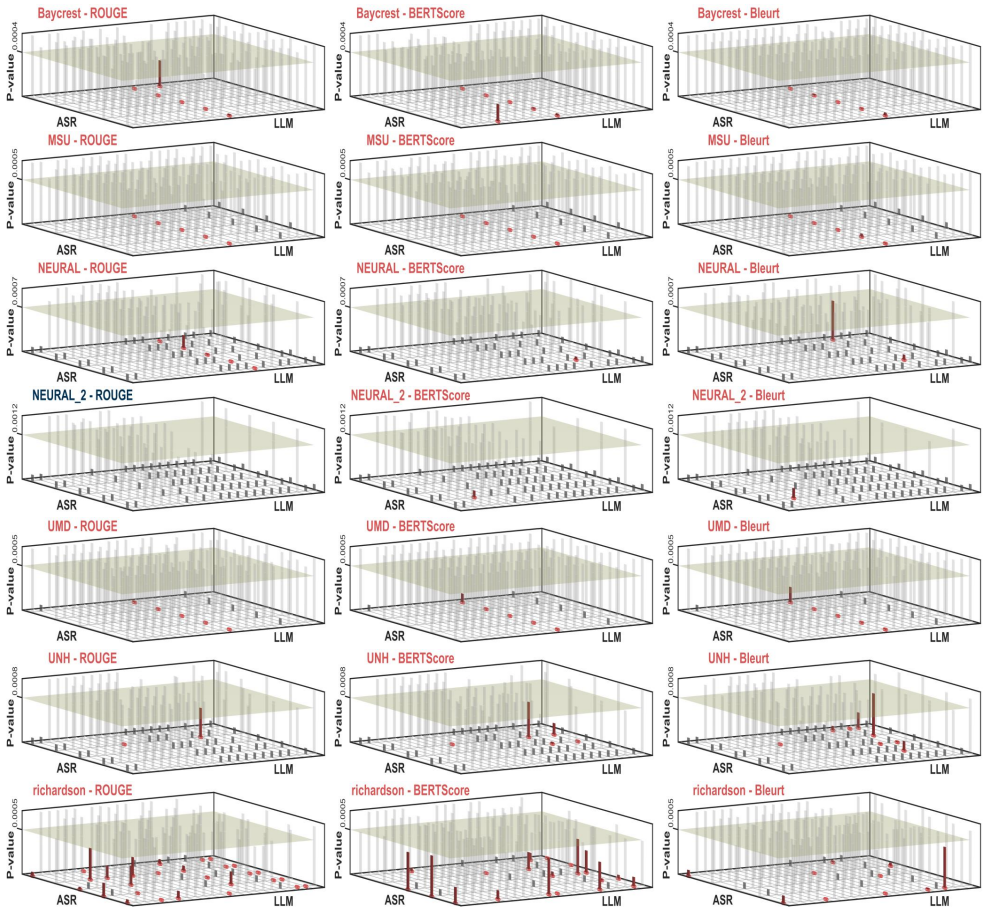


Figure 10: p-values from the two-sided Wilcoxon tests on hypotheses (4) to assess the performance difference between patients and controls in Aphasia datasets.

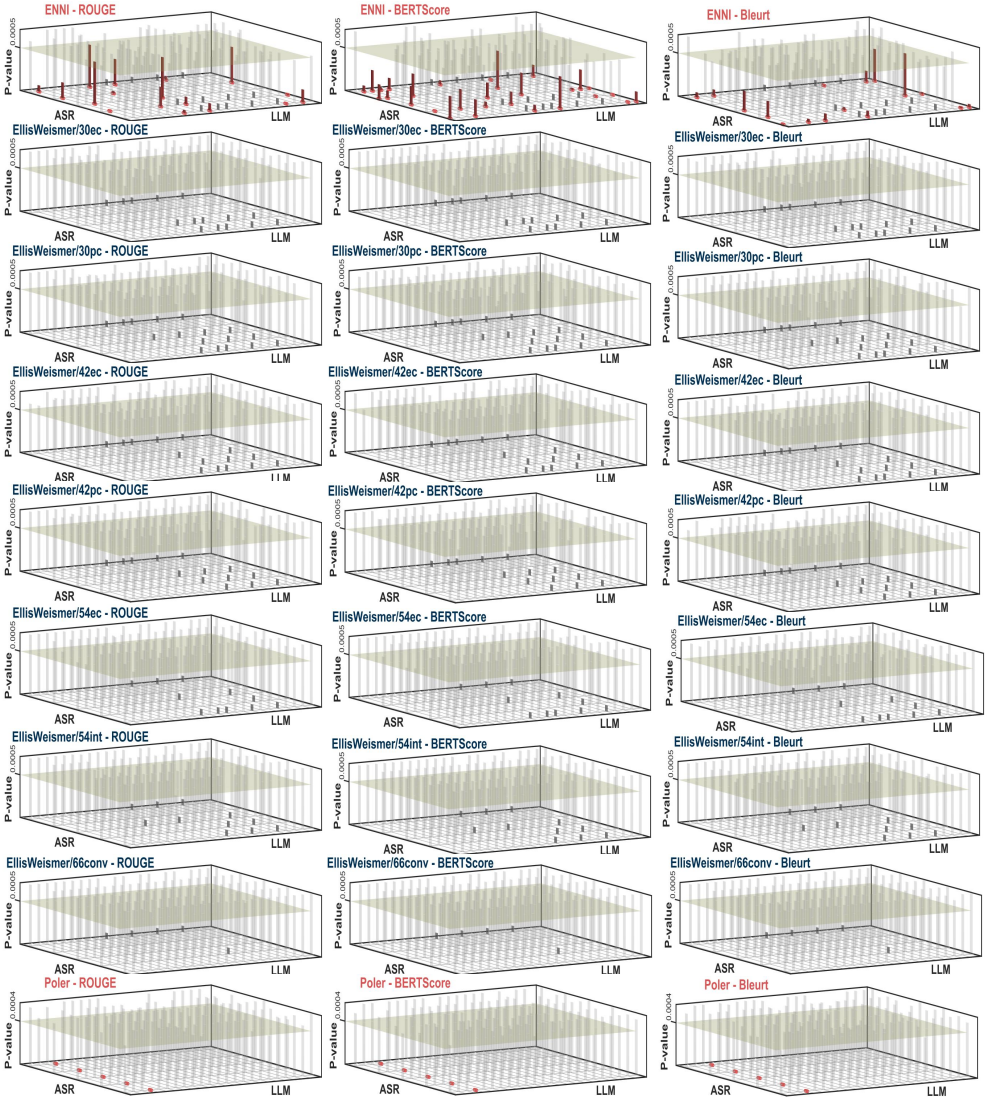


Figure 11: p-values from the two-sided Wilcoxon tests on hypotheses (4) to assess the performance difference between patients and controls in LIC datasets.

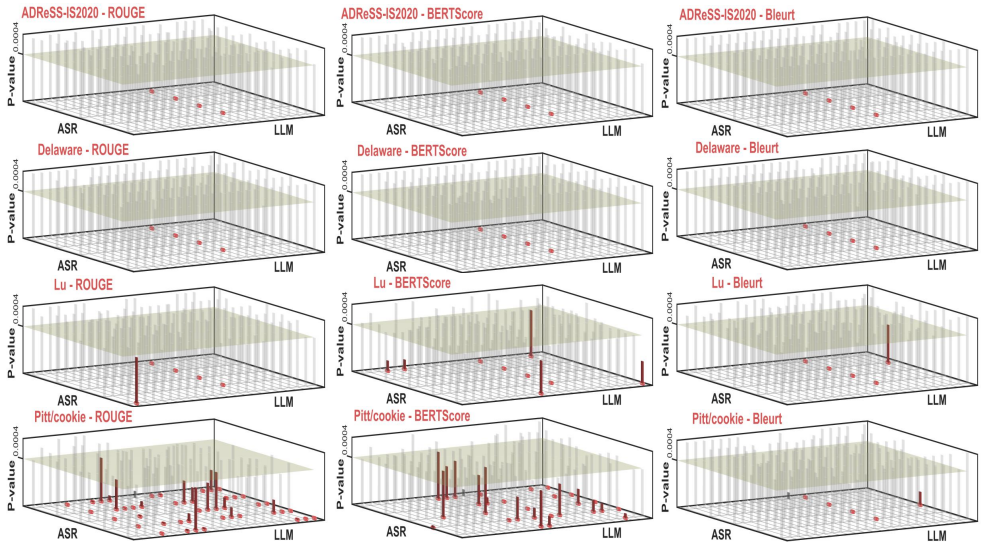


Figure 12: p-values from the two-sided Wilcoxon tests on hypotheses (4) to assess the performance difference between patients and controls in Dementia datasets.

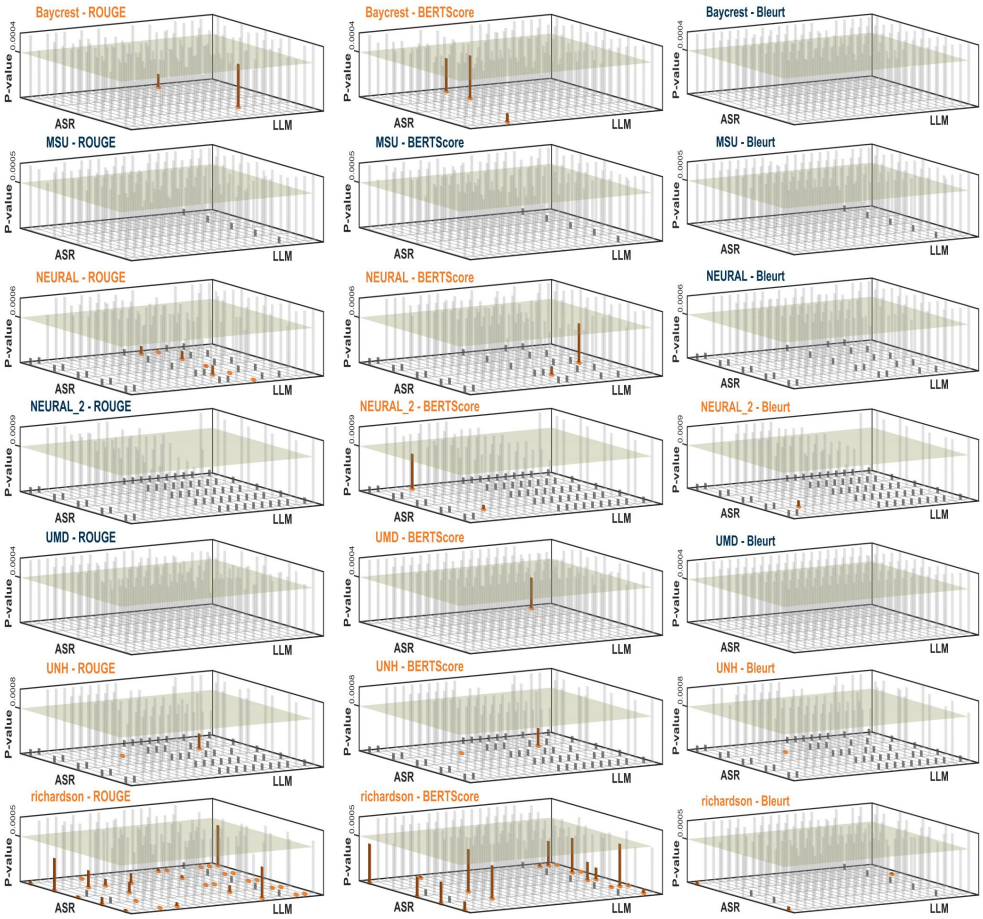


Figure 13: p-values from the one-sided Wilcoxon tests on hypotheses (5) to assess whether ASR-LLM performed better in controls than in patients in Aphasia datasets.

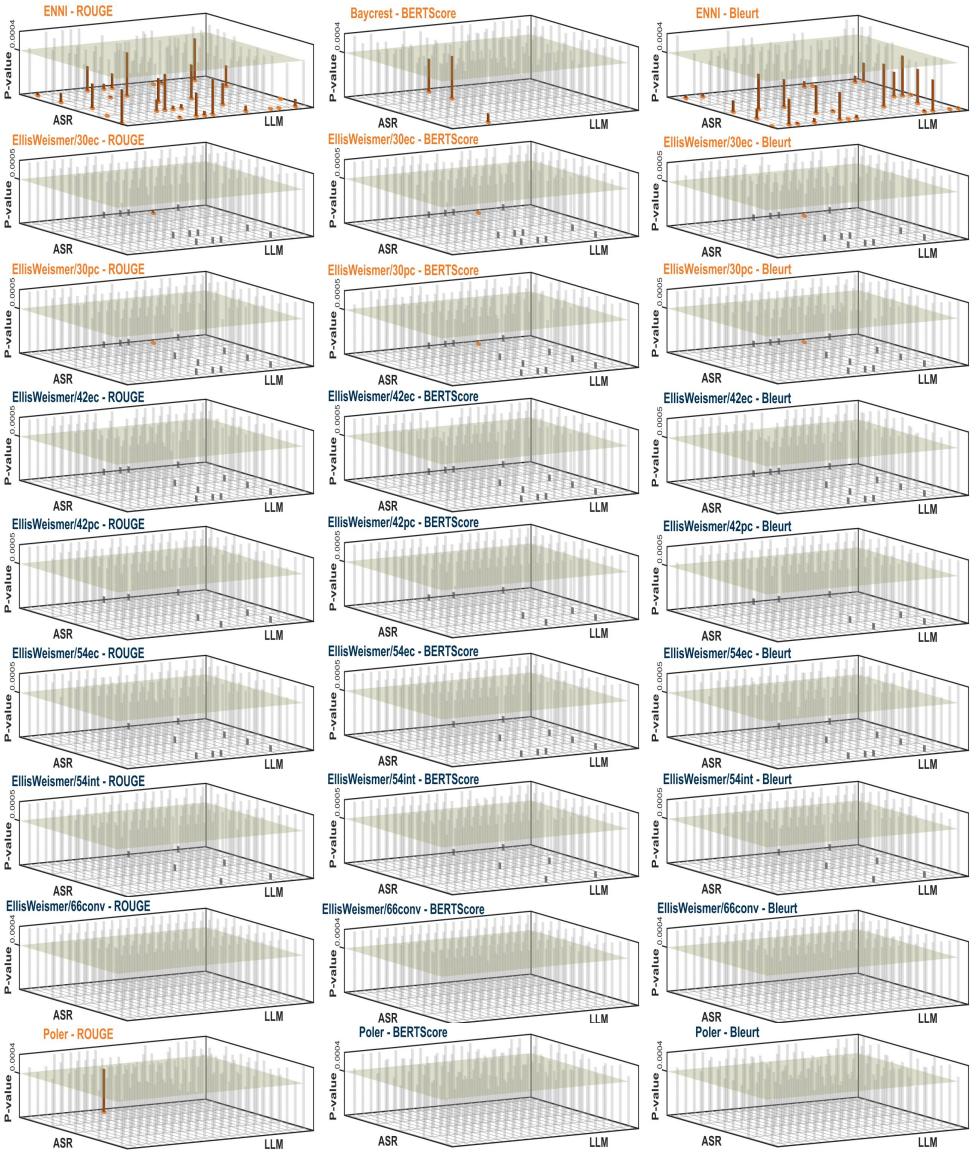


Figure 14: p-values from the one-sided Wilcoxon tests on hypotheses (5) to assess whether ASR-LLM performed better in controls than in patients in Childes datasets.

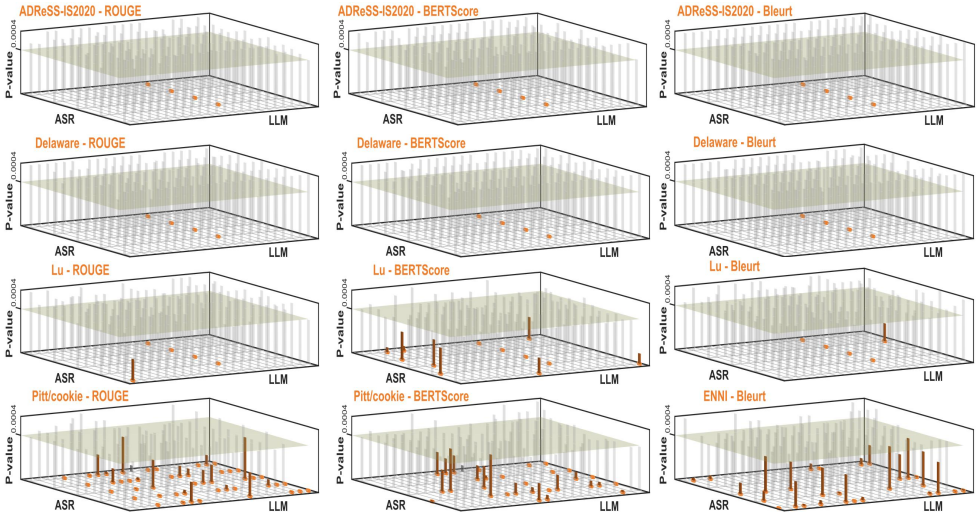


Figure 15: p-values from the one-sided Wilcoxon tests on hypotheses (5) to assess whether ASR-LLM performed better in controls than in patients in Dementia datasets.

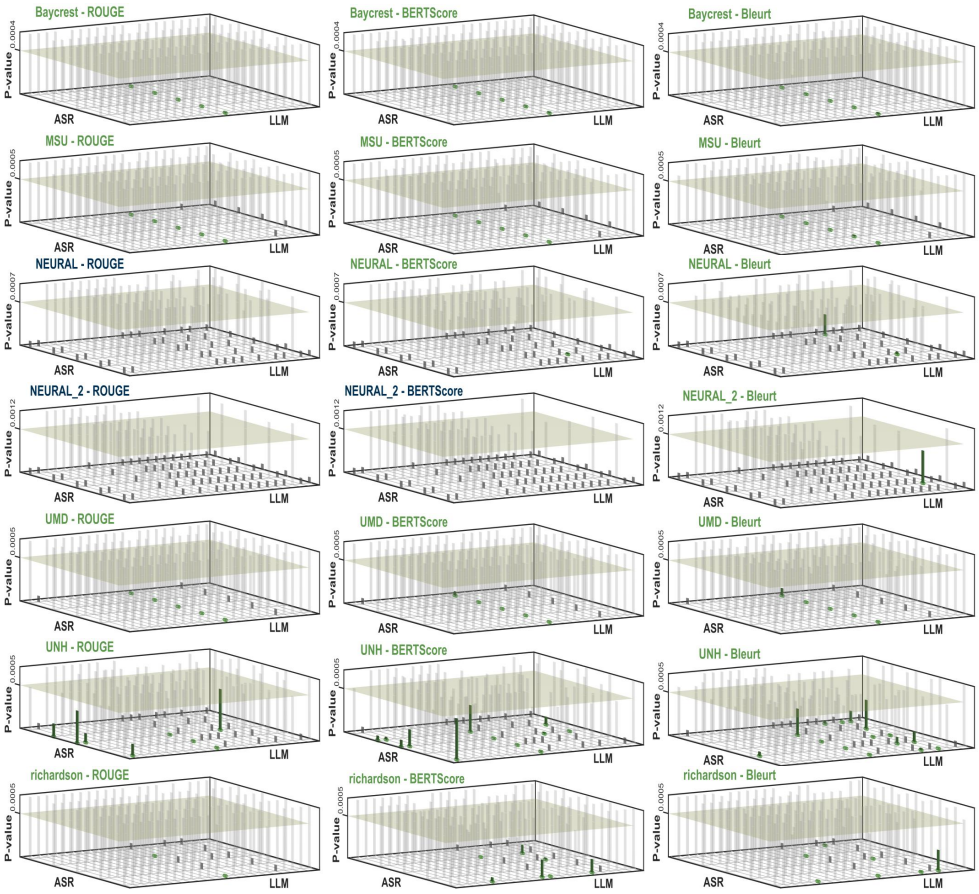


Figure 16: p-values from the one-sided Wilcoxon tests on hypotheses (6) to assess whether ASR-LLM performed better in patients than in controls in Aphasia datasets.

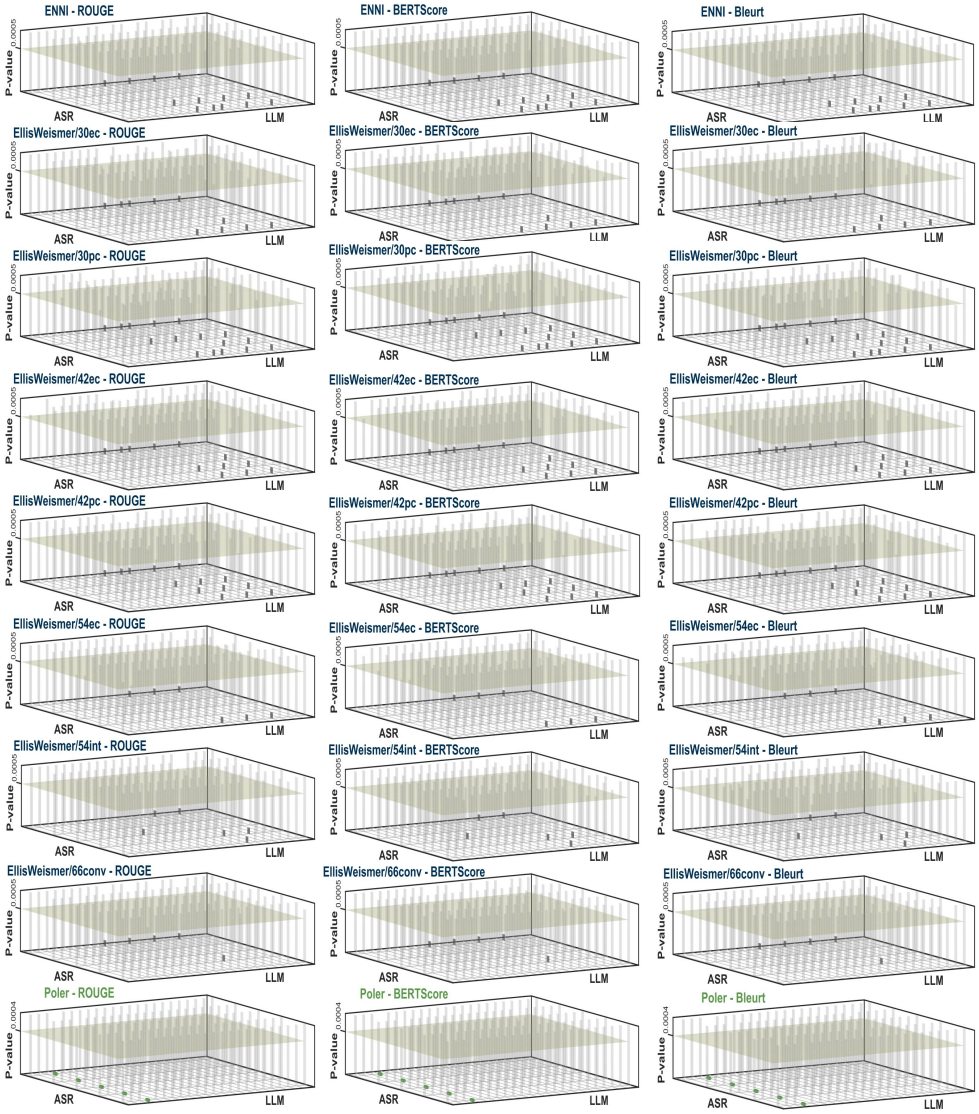


Figure 17: p-values from the one-sided Wilcoxon tests on hypotheses (6) to assess whether ASR-LLM performed better in patients than in controls in LIC datasets.

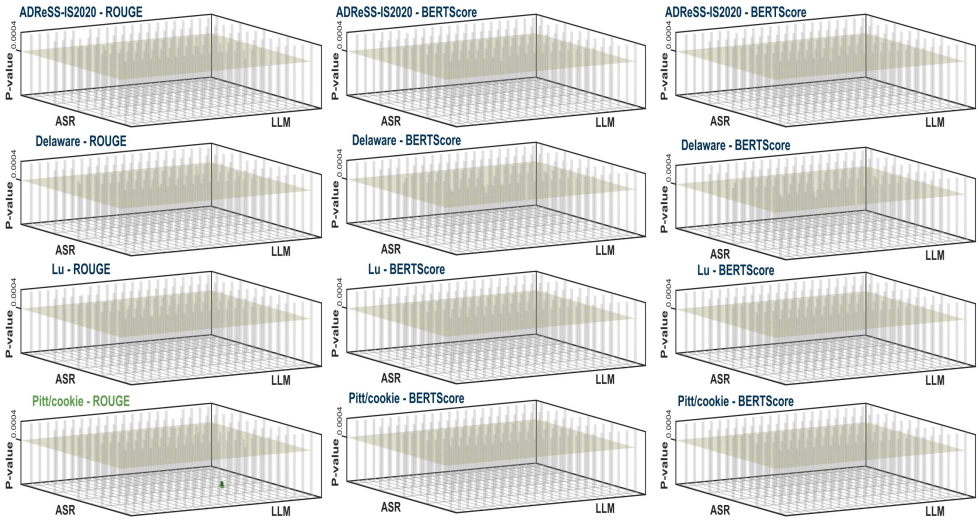


Figure 18: p-values from the one-sided Wilcoxon tests on hypotheses (6) to assess whether ASR-LLM performed better in patients than in controls in Dementia datasets.