

Enhancing Aphasia Speech Interpretation using Small Language Models (SLMs)*

Abdel-Karim Al-Tamimi^{1,4,*}, Kate Radford^{2,5}, Jacqueline Benfield^{2,6}, Jacob A Andrews², and Catherine Sweby³

¹ Sheffield Hallam University, Sheffield S11WB, UK

² University of Nottingham, Nottingham NG7 2RD, UK

³ Northern Care Alliance (NCA), Salford M6 8HD, UK

⁴ Yarmouk University, Irbid 21136, Jordan

⁵ Nottingham Biomedical Research Centre, Nottingham, UK

⁶ Derbyshire Community Health Services NHS Foundation Trust, Derby, UK

Abstract

Aphasia is a severe communication disorder that significantly impairs an individual's ability to convey and process language, often resulting from stroke-related damage to brain regions critical for speech and language functions. With the emergence of Large Language Models (LLMs), their potential has been explored in various text-based tasks due to their exceptional language understanding capabilities, which are particularly valuable in medical applications where access to specialised data is crucial yet frequently restricted. In this paper, we present our research on leveraging Tiny and Small Language Models (SLMs) to improve speech interpretation for people living with aphasia (PwA). Through benchmarking several LLMs, we established performance benchmarks to guide the development of our SLM-based solution. Our findings indicate that chain-of-thought prompting significantly enhances interpretation accuracy (median similarity score: 0.68 vs. 0.64 for zero-shot), with larger SLMs (e.g., Phi4-mini:3.8b) outperforming smaller counterparts while maintaining clinical utility. Notably, compact models like Qwen2.5:1.5b achieved competitive results, demonstrating feasibility for re-source-constrained settings. This work advances accessible, privacy-preserving assistive technology for aphasia, balancing computational efficiency with clinical relevance.

Keywords

Aphasia, Large Language Models, Small Language Models, SLLMs, AAC

1. Introduction


Aphasia is a profoundly under-recognised yet widespread condition that significantly impacts millions worldwide. In the USA alone, over two million individuals live with aphasia, a prevalence surpassing that of multiple sclerosis, Parkinson's disease, and muscular dystrophy [1] [2]. Similarly, conservative estimates suggest that at least 350,000 individuals are affected in the UK, with approximately 66 new cases per 100,000 people each year [3]. Stroke, a leading cause of aphasia, remains the third most common cause of death in both the USA and Great Britain, with one-third of stroke survivors experiencing aphasia and 12% remaining aphasic six months post-stroke [3]. Despite these alarming figures, awareness remains critically low, with 84.5% of people having never encountered the term aphasia, and only 8.8% can correctly identify it as a language disorder [2]. The consequences extend beyond communication difficulties; aphasia has been shown to have a greater

*AIME'25: 23rd International Conference on Artificial Intelligence in Medicine, June 23–26, 2025, Pavia, Italy
SLM4Health Workshop: Improving Healthcare with Small Language Models

^{1*} Corresponding author.

✉ a.al-tamimi@shu.ac.uk (A. Al-Tamimi); kate.radford@nottingham.ac.uk (K. Radford);
jacqueline.benfield1@nottingham.ac.uk (J. Benfield); jacob.andrews@nottingham.ac.uk (J. A. Andrews);
cath.sweby@nca.nhs.uk (C. Sweby)

ORCID 0000-0003-2459-0298 (A. Al-Tamimi); 0000-0001-6246-3180 (K. Radford); 0000-0002-8807-3049 (J. Benfield); 0000-0001-8408-5782 (J. A. Andrews); 0009-0008-3510-2193 (C. Sweby)

 © 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

negative impact on quality of life than both cancer and Alzheimer's disease [2]. Aphasia severely impacts communication, leading to challenges in employment, social participation, and relationships, ultimately reducing quality of life [4]. Post-stroke, individuals with aphasia face worse recovery outcomes, including longer hospital stays and higher mortality rates [5].

Large Language Models (LLMs) are advanced AI systems trained on vast text data, enabling human-like language understanding and generation. In digital health, they enhance diagnostics, medical documentation, education, and project management [6]. Integrated into healthcare applications, LLMs improve patient engagement, symptom analysis, and real-time health insights, supporting precision medicine and personalised care [7].

The authors in [8] proposed a novel framework for detecting and analysing speech dysfluencies (e.g., stuttering, repetitions, blocks) using articulatory gestures and a connectionist subsequence aligner (CSA), achieving state-of-the-art accuracy in dysfluency detection and alignment. The research concludes that while LLMs enhanced usability (e.g., generating diagnostic reports), the study highlighted that scalability and accuracy depended more on the gestural and alignment modules, suggesting LLMs serve best as interactive interfaces rather than core dysfluency analysers.

In [9], the study found that LLMs when integrated into AAC systems like SocializeChat, can enhance Augmentative and Alternative Communication (AAC) supported social communication through personalised, context-aware responses, but face limitations in handling open-ended dialogue, accurately modelling user preferences, and adapting to cultural and contextual nuances.

Small Language Models (SLMs) and Tiny LMs offer a more efficient alternative to LLMs, operating with reduced computational resources. While not as powerful as full-scale LLMs, they provide key advantages in security and privacy by processing data locally rather than relying on cloud-based servers. Additionally, SLMs and Tiny LMs deliver faster response times and function in low-connectivity environments, making them ideal for real-time health applications [10].

This paper introduces our innovative approach to use Natural Language Processing (NLP) and SLMs to explore how we can improve communication for individuals with mild-to-moderate expressive aphasia. Our proposed AI-based assistive tool helps users express their thoughts more clearly by interpreting and extracting meanings from their speech. Designed through a collaborative co-design process with experts and healthcare professionals, it aims to enhance stroke survivors' independence and quality of life.

2. Methodology

Our AI-based solution leverages the recent breakthroughs NLP, specifically the advent of LLMs, and their small and tiny variants, to augment the speech comprehension of people with aphasia (PwA). By harnessing the language-understanding capabilities of LLMs, our proposed system generates coherent speech by building upon the utterances of PwA. The process, illustrated in Figure 1, begins with recording and transcribing the speech of people with aphasia (PwA). These transcripts, along with responses from their conversation partners, are used to craft precise prompts and provide contextual data for LLMs. The LLMs then generate coherent interpretations by inferring unclear terms, filling in missing words, and removing filler content. The user selects the most appropriate output, which is subsequently rendered as speech, optionally using a synthesised voice that mimics the user's own voice. This study focuses on evaluating the performance of several open-source SLMs within this pipeline, particularly their potential to replace LLMs in low-resource or embedded systems.

The use of tiny and SLMs is critical to enable deployment on mobile devices, ensuring widespread accessibility and real-time assistance in daily communication. Furthermore, SLMs address pressing privacy and security concerns by processing sensitive speech data locally, minimising reliance on cloud-based servers and reducing risks associated with transmitting identifiable health information. This innovative approach ensures the production of intelligible sentences readily understood by the conversational partner, offering advantages such as personalisation opportunities, continuous improvement, potential integration with existing assistive technologies and services.

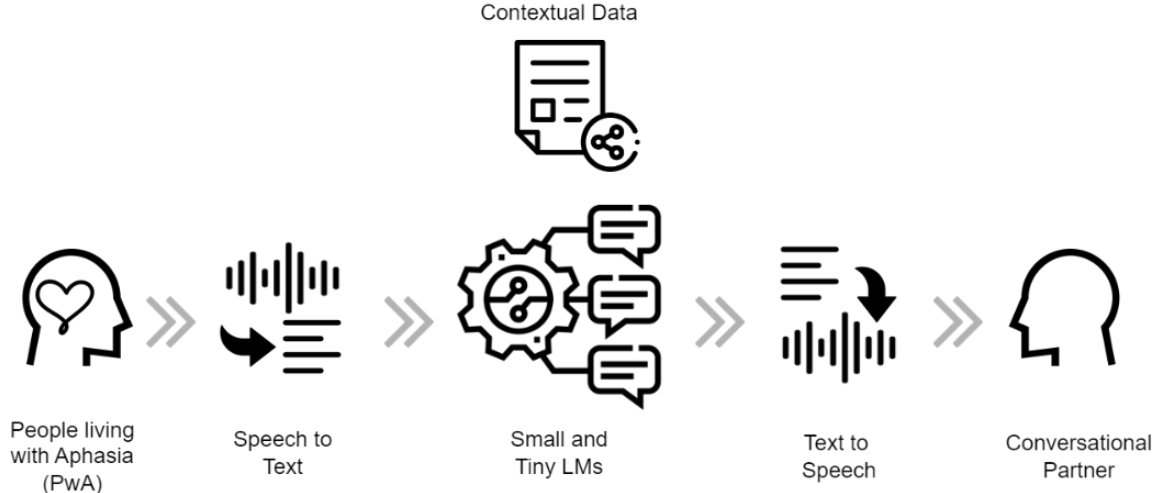


Figure 1: Pipeline of the Proposed AI-Assisted Communication System for Aphasia.

To establish a ground truth or a consensus-derived reference for benchmarking the performance of the selected tiny and SLMs in transforming spoken utterances from PwA into intelligible sentences, we collected 30 question-answer pairs transcribed from AphasiaBank [11]. These pairs were processed using six LLMs: Mix-tral:22x7b, Gemma2:9b, Qwen2:7b, Llama3:8b, Phi3:3.8b, and WizardLM2:7b. These LLMs were chosen for their open-source nature, allowing local deployment without the need for high-end hardware. This enabled multiple experimental runs to assess result consistency. They are also known for their strong reasoning capabilities, which is essential for the task at hand. Moreover, running entirely offline ensured that potentially sensitive medical data remained secure, avoiding transmission over cloud-based services and supporting ethical compliance.

Five experts, specialising in speech-language therapy, clinical rehabilitation, and digital health transformation, independently evaluated the LLM interpretations of aphasic speech for each pair, selecting the output they deemed most accurate. Consensus analysis revealed that Mixtral outperformed all other models, as shown in Figure 2, followed by Gemma and Qwen. The highest-rated interpretations were then designated as the ground truth for subsequent evaluations. This expert-derived reference standard enabled systematic evaluation of ten widely used tiny and SLMs against clinically validated interpretations, as detailed in the following section.

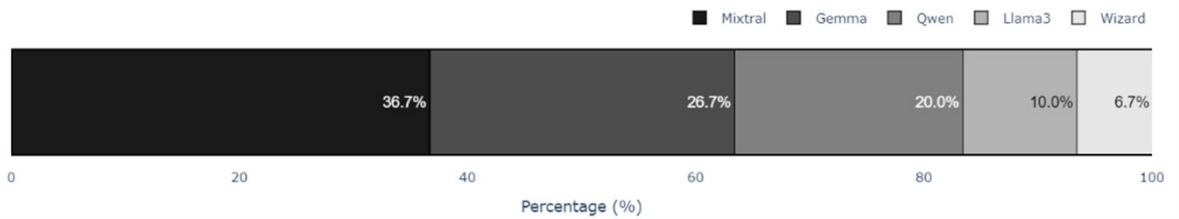


Figure 2: Consensus Performance of LLMs in Interpreting Aphasic Speech (% Agreement with Expert Judgements).

3. Implementation and Results

Building on the expert-validated benchmark established in Section 2, we evaluated the performance of 10 tiny and small language models (SLMs) in interpreting aphasic speech by comparing their outputs to the consensus-derived interpretations. We implemented a structured evaluation framework measuring cosine similarity between model-generated outputs and the consensus-derived interpretations. The chosen models (*gemma3:1b*, *llama3.2:1b*, *llama3.2:3b*, *qwen2.5:0.5b*, *qwen2.5:1.5b*, *qwen2.5:3b*, *smollm2:1.7b*, *phi3:3.8b*, *phi4-mini:3.8b*, and *hermes3:3b*), ranging from 0.5B

to 3.8B parameters, were tested across three prompting techniques: zero-shot, zero-shot chain-of-thought (CoT), and explicit CoT prompting designed to mimic clinical reasoning [12].

This chosen parameter range, spanning two orders of magnitude, was deliberately selected to assess the balance between computational efficiency and clinical efficacy, with smaller models (e.g., 0.5B) optimised for deployment in re-source-limited environments (e.g., mobile devices), and larger architectures (e.g., 3.8B) targeting enhanced linguistic reasoning for complex aphasic speech patterns, deployable on higher-capacity portable clinical systems. The spectrum directly informs real-world applicability, balancing latency-sensitive environments against scenarios requiring nuanced semantic reconstruction. Table 1 defines these prompting techniques within the study context and shares the CoT prompt used in this study.

Table 1
Prompt Techniques Definitions.

| Prompt Technique | Definition/Description |
|------------------|--|
| Zero-Shot | The model is directly instructed to provide interpretations without examples or in-context guidance or hints |
| Zero-Shot CoT | A variation of Zero-Shot where the model is instructed to generate its own CoT steps before arriving to its final answer |
| CoT | <p>The model is provided with explicit reasoning framework to guide the analysis steps to help the model break down the problem and reach more accurate results.</p> <p><i>Prompt: You are a speech-language pathologist interpreting responses from a person with aphasia. Aphasia is a communication disorder that can affect a person ability to speak, understand, read, or write. People with aphasia often use incomplete phrases, incorrect words, or fragmented speech, yet their intended meaning can often be inferred from context and key words. Your task is to reconstruct the person intended message based on their response.</i></p> <p><i>Question: '{question}'</i></p> <p><i>Answer from patient: '{answer}'</i></p> <p><i>Internally analyse the response step by step:</i></p> <ol style="list-style-type: none"><i>1. Identify meaningful or relevant words in the patient answer</i><i>2. Infer what the patient is trying to communicate</i><i>3. Consider how the response relates to the context of the question</i><i>4. Construct a clear and natural interpretation of the patient intended message</i> <p><i>Only output the final interpretation, written in the first person as if the patient is saying it themselves. Do not include any internal reasoning, explanations, or additional commentary.</i></p> <p><i>Interpretation:</i></p> |

Table 2 provides an example of the variations in aphasic speech interpretation across the ten language models, contrasting a representative question-answer pair with model-generated outputs. The *ground-truth* interpretation in this example (Wizardlm2:7b, zero-shot) reflects comprehensive inference, while the best-performing SLM (Phi4-mini:3.8b, chain-of-thought) achieves high semantic similarity (0.82) through concise synthesis of key medical details. In contrast, the weakest interpretation (Llama3.2:1b, zero-shot) scores substantially lower similarity score (0.45), demonstrating typical failure patterns where models overlook temporal progression and diagnostic markers. The comparison highlights how model architecture and prompting strategies influence the model ability in reconstructing communicative intent from fragmented aphasic speech.

Table 2

Example of LLM Interpretations of Aphasic Speech.

| Question | Aphasic Speech Response |
|---|--|
| <i>Do you remember much about your strokes?</i> | <i>"well the first one . no . I could hear the doctors that morning . I could hear them talking . I could hear what they were saying . but I could not do things . I I could listen to them . and then later it was just hard to say things . I couldn't take words . I didn't know words ."</i> |
| <i>Ground Truth</i> Wizardlm2:7b {zero_shot} | When I had my stroke, I recall hearing the doctors that morning as they discussed my condition, but at the time, I was unable to respond or perform tasks on my own. Later, it became particularly challenging to articulate words and express what I needed to say. |
| <i>Best SLM Interpretation</i> Phi4-mini:3.8b {cot} [acc=0.82] | I remember some details about my strokes; I could hear what was happening when they occurred but couldn't respond physically at that time and later found speaking difficult because of word-finding issues. |
| <i>Worst SLM Interpretation</i> Llama3.2:1b {zero_shot}[acc=0.45] | I remember hearing doctors talking about me in the hospital room. |

For each model-prompt combination, we compared the semantic similarity scores using cosine similarity between sentence embeddings generated by the *all-MiniLM-L6-v2* sentence-transformer model [13]. We also compared the performance two widely used evaluation metrics, *BertScore* [14] and *ROUGE* scores [15]. Our analysis reveals that these existing metrics yield inconsistent evaluations and fail to adequately capture the semantic similarity that this research seeks to measure. To illustrate this limitation, we examined the results obtained for the sample texts presented in Table 2. When assessed using *BertScore*, the best and worst SLMs responses received scores of 0.89 and 0.85, respectively, suggesting only a marginal difference in quality. In contrast, ROUGE-1 scores for the same responses were substantially lower (0.5 and 0.3), while ROUGE-2 scores exhibited an even greater discrepancy (0.06 and 0.00). Additionally, ROUGE-L scores displayed inconsistency, assigning values of 0.32 and 0.36 to the respective responses. These divergent results highlight the lack of agreement between metrics and underscore their inadequacy in reliably assessing semantic similarity.

Performance was analysed through aggregate statistics (mean, median similarity scores) and visualised to compare models and prompting strategies. Figure 3 demonstrates how larger language models (e.g., *Phi3:3.8b* and *Phi4-mini:3.8b*) achieve significantly higher semantic similarity scores compared to other models, while the smallest model (*Qwen2.5:0.5b*) performs weakest. Notably, *Qwen2.5:1.5b* delivers competitive performance relative to its compact architecture, suggesting a favourable trade-off between model size and accuracy.

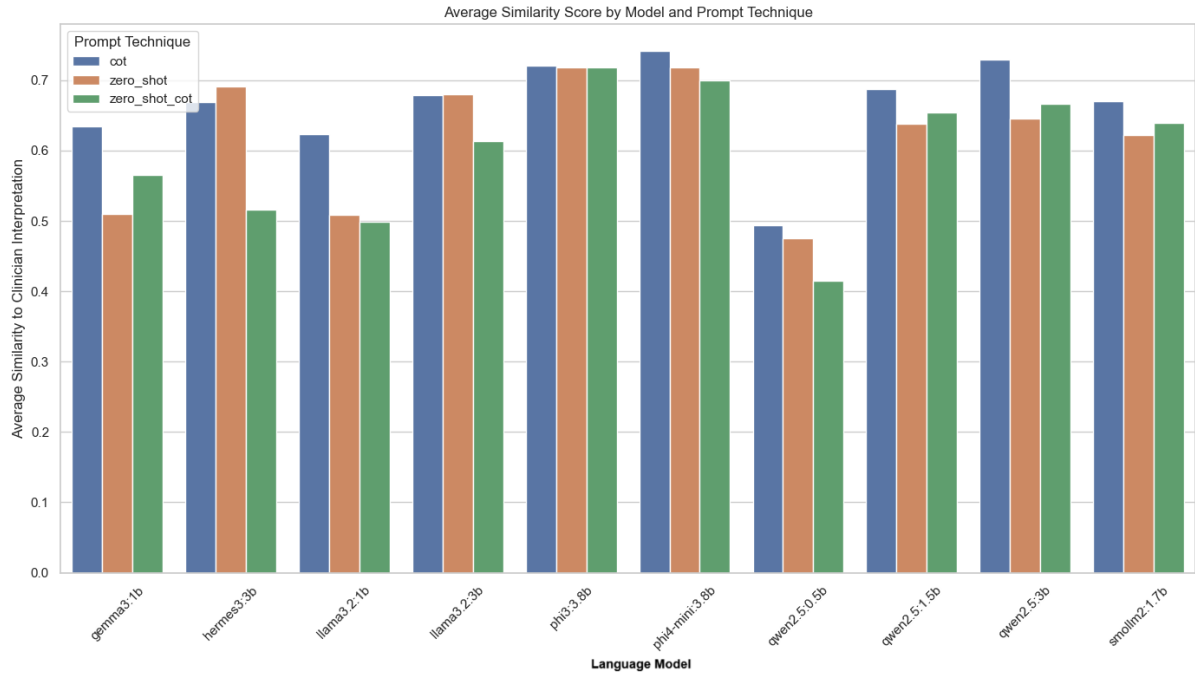


Figure 3: Semantic Similarity of Model-Prompt Combinations for Aphasic Speech Interpretation.

The boxplot analysis, shown in Figure 4, reveals distinct performance patterns among the tested prompting techniques. CoT prompting achieves the highest median semantic similarity score (~0.68), with a tighter interquartile range (IQR) suggesting more consistent performance compared to zero-shot methods. Zero-shot techniques show broader score distributions (median ~0.64), indicating higher variability in interpretation quality. Notably, zero-shot CoT (median ~0.63) bridges this gap, demonstrating that implicit step-by-step reasoning improves reliability over basic zero-shot approaches while remaining less constrained than full CoT. These results underscore that explicit reasoning frameworks, represented in CoT, enhance both accuracy and consistency in aphasic speech interpretation.

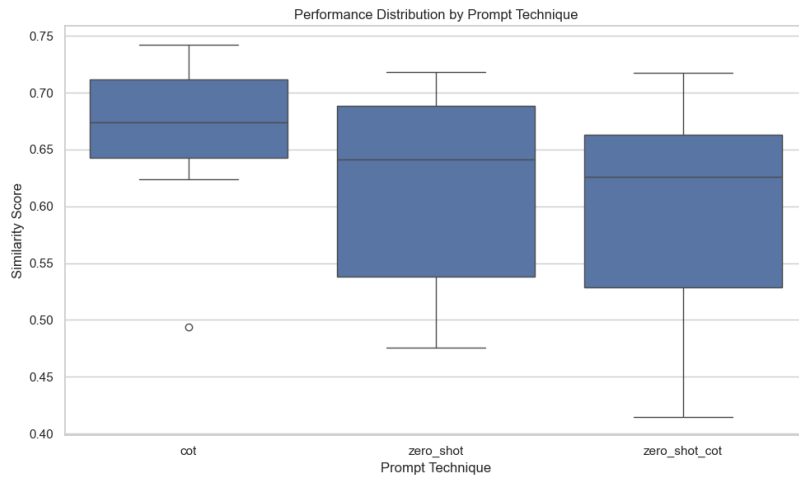


Figure 4: Distribution of Semantic Similarity Scores Across Prompting Techniques.

4. Conclusions

This study demonstrates the potential of tiny and small language models (SLMs) to enhance communication for individuals with aphasia, offering a balance between computational efficiency and clinical utility. Our findings indicate that model size and prompting strategy significantly influence interpretation accuracy, with chain-of-thought (CoT) techniques yielding the most reliable results. While larger SLMs (e.g., *Phi4-mini:3.8b*) achieved higher semantic similarity to clinician

benchmarks, smaller models like *Qwen2.5:1.5b* showed promising performance relative to their reduced size, suggesting feasibility for real-world deployment on mobile devices. However, it is important to note this research limitations that include the modest sample size of expert-validated utterances and the focus on expressive aphasia, which may not fully capture the diversity of aphasic speech patterns. Additionally, the study's reliance on cosine similarity, though widely adopted, may overlook nuanced semantic differences critical in clinical contexts.

Future work would expand the dataset to include broader aphasia subtypes and multilingual contexts, while incorporating real-time user feedback to refine model outputs dynamically. Investigating hybrid approaches, by combining SLMs with rule-based systems or personalised LLM fine-tuning, could further improve the solution accuracy, particularly for complex conversational scenarios. Furthermore, participatory design methodologies should integrate stroke survivors as co-evaluators in assessing system utility, ensuring solutions align with lived experiences of aphasic communication challenges. By addressing these challenges, SLM-based assistive tools could evolve into scalable solutions, bridging gaps in accessible communication support for underserved populations.

Acknowledgements

This study was partially funded by EPSRC/Next Generation Rehabilitation Technologies (EP/W000679/1. R-SPEAK: 27598876 R02794). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript versions of this paper arising from this submission

Declaration on Generative AI

During the preparation of this work, the author(s) used Qwen2.5 in order to: Grammar and spelling check.

References

- [1] National Aphasia Association, Aphasia statistics, 2023. URL: <https://aphasia.org/aphasia-resources/aphasia-statistics/>.
- [2] National Aphasia Association, Aphasia fact sheet, 2023. URL: <https://aphasia.org/aphasia-resources/aphasia-factsheet/>.
- [3] Royal College of Speech and Language Therapists, Resource manual for commissioning and planning services for SLCN, 2023. URL: <https://rcslt.org/wp-content/uploads/media/Project/RCSLT/slcN-resource-manual.pdf>.
- [4] M. Konishi, et al., Exploring the impact of aphasia severity on employment, social participation, and quality of life, medRxiv (2025). doi:10.1101/2025.01.08.25320231.
- [5] R. M. Lazar, A. K. Boehme, Aphasia as a predictor of stroke outcome, Curr. Neurol. Neurosci. Rep. 17 (2017) 83. doi:10.1007/s11910-017-0797-z.
- [6] X. Meng, et al., The application of large language models in medicine: A scoping review, iScience 27(5) (2024).
- [7] K. Holley, M. Mathur, LLMs and generative AI for healthcare: The next frontier, O'Reilly Media, Inc., 2024.
- [8] J. Lian et al, SSDM: Scalable speech dysfluency modelling, Advances in neural information processing systems 37 (2024): 101818-101855.
- [9] Y. Fang et al., SocializeChat: A GPT-based AAC tool for social communication through eye gazing, Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing, 2023.

- [10] Healthcare Digital, Personal health LLMs to emerge in HealthTech 2024, 2025. URL: <https://www.healthcare.digital/single-post/personal-health-llm-s-to-emerge-in-healthtech-2024>.
- [11] AphasiaBank, 2025. URL: <https://aphasia.talkbank.org/>.
- [12] J. Ziqi, W. Lu, Tab-CoT: Zero-shot tabular chain of thought, Findings of the Association for Computational Linguistics: ACL 2023 (2023) 10259–10277. doi:10.18653/v1/2023.findings-acl.651.
- [13] all-MiniLM-L6-v2, 2025. URL: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>.
- [14] T. Zhang et al., Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675, 2019.
- [15] M. Grusky, Rogue scores, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Volume 1: Long Papers. 2023.