Towards Reconstructing Impaired Language for People with Aphasia in Spanish

Stribor Pracic[1], Horacio Saggion[1]

[1]Natural Language Processing Group (TALN), Universitat Pompeu Fabra, Carrer de Tànger 122, 08018 Barcelona, Spain

Abstract

Assisting people with language impairments by means of Artificial Intelligence (AI) has become a tantalising prospect in the past few years. Large Language Models (LLMs) have shown considerable performance in certain language-oriented tasks. To date, however, the research of LLMs in reconstructing impaired speech has been concentrated primarily on English. In this paper, we propose a method for reconstructing fragmented speech produced by Spanish aphasic patients. Employing data extracted from the AphasiaBank repository, we tested the smallest version of the *Mistral* model using zero-shot and few-shot in-context prompting to identify conversational errors and produce corrected outputs of aphasic sentences. We utilised the Langchain architecture which enabled the model to maintain a curated window of prior sentences in memory, helping to preserve context. We also examined the relationship between different levels of aphasia severity and its impact on the reconstruction ability of our language model.

Keywords

Aphasia, Natural Language Processing, LLMs for Aphasia, Assistive Communication

1. Introduction

Aphasia is a disturbance in the physiological processes indispensable for language production, usually caused by a stroke or brain injury. It manifests itself in various degrees of language impairment severity, and its symptoms can affect some or all language modalities, such as production of speech, writing, as well as speech and reading comprehension [1, 2, 3, 4].

It is estimated that around one third of all stroke patients are diagnosed with some form of aphasia [5, 6]. The intensity and subtype of aphasia is primarily linked to the size and location of the stroke, the rapidness of medical attention immediately following a stroke, as well as the patient's medical history (e.g. prior strokes or brain injuries) [6]. The most prominent types of aphasia, according to the Boston classification developed by Goodglass and Kaplan [7], are Broca's (expressive), Wernicke's (receptive), global, anomic, and conduction aphasia [6].

Aphasia is characterised by numerous language impairments, such as a difficulty for the patients in finding the right words to express thoughts and ideas, in addition to challenges pertaining to the overall comprehension of language [4]. Language output produced by people with aphasia (PWA) is severely compromised, and is often filled with syntactic, grammatical, and phonological errors, word repetitions, and neologisms, all of which are dependent on the type and severity of the aphasia [2, 4].

As a result, aphasia not only has a detrimental effect on the patients' everyday communication, but also on their overall quality of life, which often leads to frustration, social isolation, and depression [4, 8, 9]. According to the Portuguese Institute of Aphasia, around 150,000 people in Spain live with some form of language impairment. And around the world, it is estimated that tens of millions of people suffer from aphasia [10, 11]. Moreover, as the average age of the global population continues to rise, and given that age is a significant risk factor for strokes and other neurological detriments, the number of PWAs is likely to increase as well [11].

Over the years, many different methods of assistive communication devices for PWAs have been developed [12]. For example, Sentence-Shaper [13] is an innovative computer program that enables people with language impairments to create speech in their own voices by recording words or short phrases and arranging them into sentences. In addition, several tablet or phone-based therapy options have also been made available. For instance, Sentantics [14], which acts as a virtual assistant that guides users through exercises that ask them to match sentences to images and produce descriptive, target utterances, thereby improving both sentence comprehension and production [6].

Given the rapid development of artificial intelligence (AI) technologies, there has never been a better time to explore the use of generative AI in aiding people with communicative disorders. In recent years, large language models (LLMs) have excelled in tasks such as text generation and masked token prediction [4]. In particular, their ability to predict masked (missing) words in a sentence based on the surrounding context makes them an invaluable potential tool for language correction and as an assistive communication device. However, given the nascent nature of the field, the research on the matter is still in its infancy.

The main goal of this paper is to assess and propose a method for reconstructing impaired sentences produced by Spanish PWAs following the methodology established in Adikari et al. [4]. We rely on the availability of the AphasiaBank[1] dataset which provides curated transcripts of conversations between therapists and aphasia patients in 13 different languages.

We attest to the reconstruction performance of the *Mistral-7B* on the Spanish aphasic sentences extracted from the AphasiaBank repository. With our model, we utilise both zero-shot and few-shot inference settings and the Langchain technology [15]. This makes it possible to transform impaired language produced by PWAs into corrected sentences while maintaining memory and context from prior utterances in the conversation. Our results suggest that the reconstruction accuracy of the few-shot model surpasses that of the zero-shot model across all aphasia types.

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

stribor.pracic@upf.edu (S. Pracic); horacio.saggion@upf.edu (H. Saggion)

0009-0009-5811-5808 (S. Pracic); 0000-0003-0016-7807 (H. Saggion)

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

2. Background

Recent studies have explored the use of various different models (BERT, GPT-4o) for enhancing assistive communication technologies via language reconstruction [4, 16], question-answering [17], as well as transfer learning using neurolinguistic-based synthetic datasets [18]. The primary objective of these researches is a development of advanced LLM systems that can aid the rehabilitation of PWAs, enabling faster recovery times, ease of communication, and an access for therapists and caregivers to more robust interpretation strategies.

From a medical perspective, the practical implementations of these technologies are manifold. Reconstructing impaired language carries with it not only a more accessible way for PWAs to express themselves, but is also intended to do away with certain side-effects of aphasia such as stress, mood disturbance, and reduced quality of life [4, 9].

Adikari et al. [4] proposed a conversation framework using the GPT-4o model, implemented together with the Langchain technology, which enables preservation of previous utterances within the conversation via a buffer memory. Their approach utilised an in-context few-shot approach (five examples of erroneous sentences per error type and their corrected references) to guide the model at inference time. Their test set was comprised of 2000 sentences from the AphasiaBank dataset across five most prominent types of aphasia: Broca's, Wernicke's, conduction, global, and anomic. Their model managed to achieve an 80% cosine similarity score and an average of 82% reconstruction accuracy across all five aphasia types. In addition, they performed a qualitative study of their results and reported that severe language impairments noticeably impact the models' ability to reconstruct the intended meaning. In other words, models exhibit a lower accuracy score when tested on more severe types of aphasia (e.g. global) compared to less severe ones (e.g. conduction or anomic).

van Vaals et al. [16] explored the use of four seq2seq models from the *T5* [19] and *Flan-T5* [20] families on completing impaired Broca's aphasia sentences, all fine-tuned on a synthetically generated dataset set to mimic sentences spoken by Broca's aphasia patients. Thereafter, they tested the performance of their models on both synthetic and authentic (sentences produced by actual patients) data. On the synthetic dataset, the two best performing models (*T5-Base* and *Flan-T5-XL*) achieved respectable ChrF [21] (56.56% and 56.19%) and ROUGE-L [22] (0.71 and 0.69) scores. However, due to the lack of the ground reference truth for the authentic dataset, they only presented a qualitative analysis of the sentences completed by their models.

Manir et al. [17] investigated fine-tuning the BERT model on the sentences extracted from the AphasiaBank dataset. The model was used in two tasks: first, for next-word prediction in sentences for people with aphasia; and second, for question-answering with the goal of increasing context comprehension. Human annotators were then tasked with assigning a score between 1 and 4, ranging from completely unsuitable to highly suitable. Their model performed promisingly, with the pairwise correlation heatmap indicating a substantial agreement between the annotators. The minimum observed correlation was 0.61, while the maximum was 0.74. In addition, they obtained a Fleiss' Kappa score of 0.32, suggesting a moderate level of agreement between the three evaluators.

However, while these researches offer a promising starting point, they have only focused on the English language, and thus highlight the need for deeper exploration, chief among them an expansion into other languages available in the AphasiaBank dataset.

3. Methodology

3.1. Dataset

The AphasiaBank repository represents a shared comprehensive database of multimedia interactions with the purpose of studying communication in aphasia [23]. The dataset is available to all researches on the TalkBank Browser site.[2] It consists of curated conversations between speech therapists and patients in 13 languages, where each interview consists of four discourse tasks: personal narratives (the patient's stroke, recovery, etc.), picture descriptions, storytelling, and procedural discourse (e.g. how to make a sandwich) [23].

In this research, we utilised transcripts from 8 out of the available 32 patients in Spanish from the AphasiaBank dataset, two for each aphasia type—Broca's, Wernicke's, anomic, and global. To ensure the usability of our data, we cleaned and organised the transcripts from the AphasiaBank dataset using our preprocessing pipeline. First, instead of concentrating only on the sentences spoken by the patients, we extracted both the therapist and patient utterances in order to cover the entire conversation span necessitated by the Langchain architecture.

Following this, we implemented a regex script to clean out all instances of special characters, indicators of pause (xxx), and punctuation marks displayed in brackets. Thereafter, we cleaned out all descriptions of motions and gestures displayed in square brackets, such as "looks at", "raises hand", "counts with fingers", etc. We then deleted all text succeeding the markers "&=", "&*", and "&-" as they were used to label certain gesture markers and filler words such as "uhhum". Lastly, we removed all instances of markers for prolonged syllable pronunciation (":") unless it was preceded by a capital letter, and converted whitespace to single space where necessary.[3] Presented in Table 1 is an example of the original and processed transcript taken from one of our evaluation datasets.

3.2. Sentence Reconstruction Method

Following Adikari et al. [4], we employed few-shot learning in order to help guide our model in reconstructing the impaired utterances. However, in order to assess the benefits of this approach, we also implemented a zero-shot learning approach as baseline to attest to the differences in reconstruction accuracy between the two setups.

Zero-shot learning refers to a machine learning pattern whereby the model is not provided with any examples of the task during training, meaning the outputs are generated based solely on the model's knowledge acquired during the pre-training phase. The main goal of zero-shot learning is for the model to predict new, unseen classes of data without any training examples provided at inference time [24].

Few-shot learning enhances the zero-shot method by providing the model with a few samples of the task at hand at inference time, which serve as a template for the model to

[2] https://talkbank.org

[3] Our code is available at https://github.com/spracic/aphasia/

Table 1
Excerpt of a Preprocessed and Cleaned AphasiaBank Transcript Entry

Conversation Turn	Preprocessed Script	Cleaned Script
Line 1: Patient	PAR: no lo sé [ˆ shrugs his shoulder] (.) no sé que pasa↓ pero:→	PAR: no lo sé que pasa pero.
Line 2: Patient	PAR: (..) seguramente↓ (.) <después no xxx> [>].	PAR: seguramente después no.
Line 3: Therapist	INF1: <pero> [<] [Î looks at JMB] si la cámara no funciona↑ (.) ahora↓ .	INF1: pero si la cámara no funciona ahora.
Line 4: Patient	PAR: espero que no↑ [ˆ looks at BGP] [ˆ looks at MUJ] .	PAR: espero que no.

Table 2
Descriptions and Examples of the Different Error Types in The Dataset

Error Type	Error Subtype (error code)	Error Description	Spanish Example	English Example
Phonol. Errors	Word Substitution (p_w)	Substitution of a word with a real word that resembles the target word phonologically	"vendo" for "viendo"	"boater" for "butter"
	Non-word Substitution (p_n)	Substitution of a word with a non-word that resembles the target word	"primega" for "primera"	"buther" for "butter"
Semantic Errors	Related Word, Target Known (s_r)	Use of a semantically related word	"preguntaba" for "contestaba"	"mother" for "father"
	Unrelated Word, Target Unknown (s_ur)	Use of an unrelated word with a known target	"perro" for "pelota"	"comb" for "umbrella"
	Word, Unknown Target (s_uk)	Use of a word without a known target	"que se me hace queso como yo se"	"I go wolf"
Morph. Errors	Gender Agreement (m_ga)	Use of an incorrect gender marker	"al" for "lo"	N/A
	Verb Agreement (m_va)	Use of an incorrect verb form	"supriras" for "supieras"	"I goes to the store"
Neologisms	Neologism, Known Target (n_k)	Creation of a novel word with a known target	"tie" for "piensa"	"nomemba" for "remember"
	Neologism, Unknown Target (n_uk)	Creation of a novel word without a known target	"porque si como yo aa se que lo mos"	"and dudu make a peanut butter sandwich"

follow. The model is therefore able to learn the underlying pattern in the data from these few examples [25].

We extracted five examples per error type: phonological, semantic, morphological, and neologisms. The annotation of error examples was done with the assistance of one of the co-authors of this paper, a native Spanish speaker. The selected examples all contained the impaired utterance, its corrected form, and the corresponding error subtype. The example sentences were excluded from the evaluation set to prevent test leakage. Presented in Table 2 are the descriptions of the error types, with one example for Spanish (extracted from our dataset) and one for English (extracted from Adikari et al. [4] and AphasiaBank).

3.3. Langchain Architecture

Langchain [15] is a rapidly evolving framework that provides tools to developers for developing LLM-based applications. It facilitates seamless integration with LLMs by providing APIs for connecting various language models such as GPT, Gemini, and LLaMA. The justification for its implementation in this research is two-fold. First, using an LLM which simply accepts a long prompt is too expensive and slow, especially if the prompt consist of over 100 000 tokens of conversation history. Second, Langchain allows us to maintain the balance between relevance and recency. By limiting the relevant contextual scope, we can ensure the model will not get distracted or influenced by older, less relevant data.

One of Langchain's primary components is *memory*, which maintains state and continuity across conversation turns by storing conversation history, which enables models to retain information from past interactions and consider both the immediate and wider contextual window [15, 26].

Although Langchain supports various types of memory (such as conversation buffer memory, knowledge graph memory, and conversation summary memory), we opted to employ the conversation buffer memory due to its ease of

implementation and dynamic adaptability to the length of the conversation. The size of the buffer (i.e. the number of previous utterances retained by the model) is represented by a memory parameter *k*. This parameter can, in turn, be adjusted to accommodate for the number of utterances necessary to be retained [4]. In our experimentation, we set the buffer parameter *k* to 6, which equates to three conversation turns between the therapist and patient.

3.4. Model Architecture

We instantiated both inference setups using the *Mistral-7B* model, alongside the Langchain architecture, and base transformer parameter configuration as described in Vaswani [27]. The system prompt fed to the model at inference time differed slightly between the zero and one-shot setups. Since the zero-shot setup was not provided with any examples, we wrote more detailed instructions for the model to follow.[4] In both setups, the model was also prompted to return a list of the error types detected, as well as the list of changes made during reconstruction.

The model's output was based solely on the system prompt and the Langchain architecture. The sentence reconstruction was conducted using *Mistral-7B*,[5] the smallest version of the *Mistral* model, running on a T4 GPU in *Google Colab*.

3.5. Evaluation Metrics

We implemented two metrics used for measuring the quality of text generated by our model: cosine similarity [28] and BLEURT [29]. Given that the original BLEURT model (BLEURT-base-128) was trained primarily for English, we opted to use BLEURT-20, a text generation evaluation metric built on top of a multilingual layer-pruned mBERT, which supports Spanish out of the box [29].

Cosine similarity is a metric used to measure the similarity between two vectors representing two sentences, paragraphs, or passages of text. Its implementing here is motivated by our aim to observe how closely the reconstructed sentences predicted by the model align with the intended meaning produced by the patients.

BLEURT is a text generation metric proposed by Sellam et al. [29] based on BERT [30]. BLEURT combines expressivity and robustness via a pre-training scheme on large amounts of synthetic data using perturbations such as masking and back-translation, before being fine-tuned on human ratings. By modelling human judgment, it is able to assess semantic similarity and fluency with a superior performance ceiling compared to other metrics such as BLEU [31].

4. Results

We present the total number of utterances analysed, along with the rate of successful reconstructions, in Table 5 and 6. In the few-shot model, Broca's and Wernicke's aphasia display near identical success reconstruction rates. Anomic aphasia displays the highest reconstruction rate, which aligns with our initial expectations. However, the success rate for global aphasia was also almost identical to anomic aphasia. In the zero-shot model, this discrepancy is even more prominent, as the model has failed to reconstruct only two utterances. To provide an explanation for this, we performed an analysis of the reconstructed utterances in the global aphasia dataset.

The utterances produced by global aphasia patients are usually severely fragmented, often-times lacking basic sentence structure and cohesion, or are comprised solely by single words or exclamations (e.g. "ah aah um", "uuuh"). The model, unable to generate any meaningful reconstructions, erroneously labels these entries as successful reconstructions. It appears that the severe deficits in language production, an abundance of neologisms, and fragmented phrases beset by global aphasia hinder the LLM's reconstruction ability.

We compare the reconstruction accuracy of our model to the severity of each individual aphasia, measured by the Western Aphasia Battery - Aphasia Quotient (WAB-AQ) score [32, 33]. The WAB-AQ score is computed by assessing a patient's competence in four subdomains: (1) spontaneous speech, (2) auditory comprehension, (3) repetition, and (4) naming. A total of 20 points are awarded for speech, 200 for auditory comprehension, 100 for repetition, and 100 for naming [34]. The formula for the aphasia quotient is "AQ=(Speech + Comprehension/20 + Repetition/10 + Naming/10)x2". The possible scores range from 0 to 100, with scores below 93.8 being labelled as aphasia. Moreover, a lower WAB-AQ score indicates a more severe form of aphasia, and vice versa, as shown in Table 3.

We observe the highest BLEURT accuracy scores for anomic aphasia, which coincides with its high mean WAB-AQ score (87.625). The lowest BLEURT accuracy score in the few-shot setup is observed for Wernicke's aphasia, despite a second lowest WAB-AQ score (54.267). Global aphasia displayed the overall lowest results in the zero-shot setup and the second lowest results in the few-shot setup, which was in accordance with its severity level (mean WAB-AQ of 24.478). The model performed considerably well on Broca's and Wernicke's aphasia, with Broca's aphasia showing slightly higher results, which aligns with its higher mean WAB-AQ score.

The scores for all aphasia types in both of our metrics (see Table 4) are higher in the few-shot method compared to the zero-shot method. However, while the BERTScore results are comparable, the BLEURT scores show a more stark contrast between the two inference setups. These results fall in line with our expectations, as the examples in the few-shot setup provided the model with a small bit of context from which it could extrapolate the underlying pattern, thereby increasing its performance.

We also explore the relationship between aphasia severity and LLM's reconstruction ability. To illustrate this, we constructed a correlation matrix between the BLEURT scores and the mean WAB-AQ scores for both inference settings (see Figure 1 and Figure 2).

In the few-shot setting, the Pearson correlation showed a moderate positive linear relationship between aphasia severity and reconstruction accuracy (r =.555); however with a p-value of .444, indicating that there is a 44% probability of such a result occurring by chance.

The Spearman value (p=.80) showed a strong monotonic relationship between the two variables, meaning as one variable increases, the other tends to increase as well. However, the p-value was still above 0.05 (p=.20), indicating a probability of 20% of random chance for such a result. We conclude that these results are owing to the frugality of the

[4]Both of our inference system prompts can also be found at https://github.com/spracic/aphasia

[5]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Table 3
Taxonomic Table of the Western Aphasia Battery Scores[6]

WAB-AQ	Severity
0-25	Very Severe
26-50	Severe
51-75	Moderate
76+	Mild

Table 4
The final scores for each of our tested aphasia types and inference setups. Values = BERT F1 semantic similarity score/BLEURT. The numbers in brackets indicate the WAB-AQ score for each aphasia subtype. The numbers in **bold** indicate the highest scoring aphasia type; the numbers in *italic* indicate the lowest scoring aphasia type.

Aphasia Type / Inference	Broca's (61.578)	Wernicke's (54.267)	Anomic (87.625)	Global (24.478)
Zero-Shot	0.862 / 0.737	0.863 / 0.694	**0.873** / **0.762**	*0.829* / *0.648*
Few-Shot	0.881 / 0.809	*0.870* / *0.761*	**0.887** / **0.822**	0.876 / 0.790

dataset used in this research.

In the zero-shot setting, the Pearson correlation showed a strong positive linear relationship between the two variables (r=.967; p=.03). The Spearman value (p=1.0) also showed a strong monotonic relationship, suggesting that the increase in reconstruction accuracy is correlated with the increase in the WAB-AQ score (i.e. lower aphasia severity).

This suggest that, in the zero-shot setting, more severe forms of aphasia present a considerable strain on the LLM to reconstruct meaningful utterances from the provided input. In the few-shot setting, although we observe the same underlying pattern, the size of our dataset prevents us from confidently concluding that the observed results are not an effect of random chance.

Lastly, we report on the number of errors for each aphasia subtype. Although Adikari et al. [4] concluded that morphological errors did not reach the significance threshold in English, we observe a stark contrast in Spanish. For Broca's, Wernicke's, and anomic aphasia types, morphological errors (chiefly "verb agreement" and "gender agreement") were the most represented out of all the error types (82 and 40 gender agreement errors). Given that Spanish is a morphologically rich language, the prevalence of these error types is to be expected, particularly in comparison with a morphologically poor language such as English.

For global aphasia, neologisms were the most represented error type (37 in total), meaning that the patients with this type of aphasia produce novel words which do not meet phonological error criteria. Moreover, there were scarcely any verb or gender agreement errors for this aphasia type, perhaps owing to the already severely impaired nature of the language produced by these patients. As postulated by Adikari et al. [4], neologisms create vague and unspecified expressions from which the LLMs struggle to extract meaning. As these new words are often produced without a known target, and therefore lack any contextual clarity, the model often cannot produce a valuable substitution, and can even sometimes be forced to hazard a guess, leading to a drop in performance.

Adikari et al. [4] observed that more severe language

Table 5
Success Rate of the Zero-Shot Model on Each Aphasia Type

Aphasia Type / Error Count	Broca's	Wernicke's	Anomic	Global
Total PAR Utterances	960	623	419	596
Succ. Reconstructions	871	592	403	594
Failed Reconstructions	89	31	16	2
Success Rate	90.72%	95.02%	96.18%	99.66%

Table 6
Success Rate of the Few-Shot Model on Each Aphasia Type

Aphasia Type / Error Count	Broca's	Wernicke's	Anomic	Global
Total PAR Utterances	960	623	419	596
Succ. Reconstructions	842	543	395	561
Failed Reconstructions	118	104	24	35
Success Rate	87.7%	87.15%	94.27%	94.12%

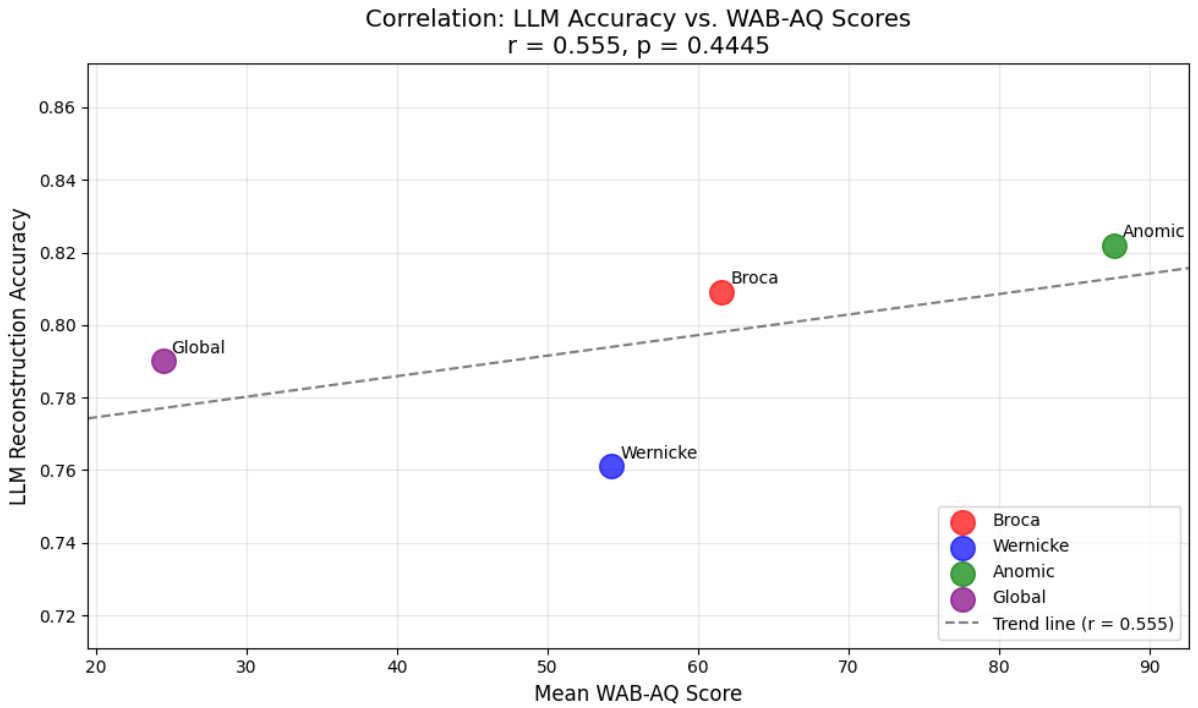

Figure 1: Correlation Matrix for the Four Tested Aphasia Types in the Few-Shot Setting. Y Axis = LLM Accuracy; X Axis = Mean WAB-AQ Score

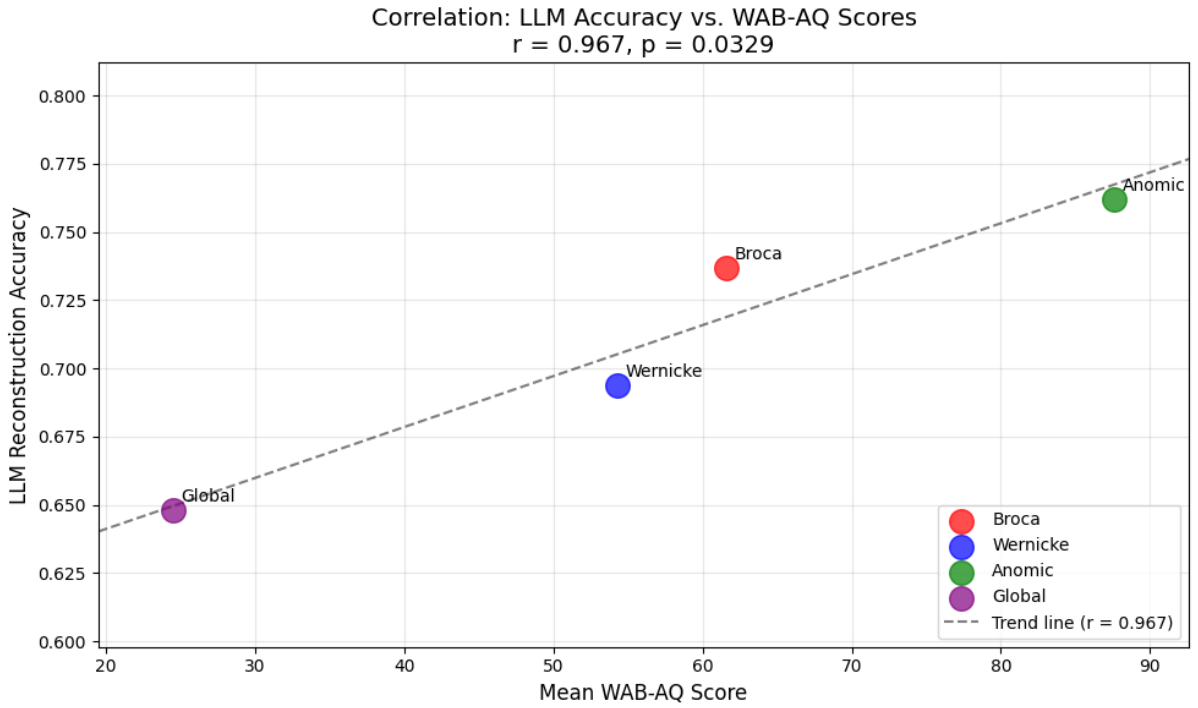

Figure 2: Correlation Matrix for the Four Tested Aphasia Types in the Zero-Shot Setting. Y Axis = LLM Accuracy; X Axis = Mean WAB-AQ Score

deficits significantly hinder the ability of LLMs to reconstruct impaired language. They also reported a high reconstruction accuracy for anomic aphasia, and the lowest reconstruction accuracy in global aphasia. What is more, the neologisms were marked as the most dominant error type, with no morphological errors for global aphasia, aligning with our observations in Spanish.

In conclusion, the patterns observed in both our inference setups show satisfactory results, and serve as a promising first step in examining the impact of aphasia severity on LLM reconstruction outside of English.

Table 7
Examples of Fragmented and LLM Reconstructed Speech

Fragmented Speech	Reconstructed Output	BLEURT Score
O conversaciones yo me ha dao cuenta de que es verdad.	O conversaciones, yo me he dado cuenta de que es verdad.	0.8629
Si es una tontería ya te vas a vamos a a no hagas nada ya.	Si es una tontería, ya te vas a ir, no hagas nada más.	0.7942
Entens y ya está y claro digo	Entiendo y está claro que digo	0.7832
Porque date cuenta tú que están transformando eh eh eh eh tiene que hacer insten instantáneamente.	Porque dices tú que están transformando esto y debe hacerlo inmediatamente.	0.5824

5. Discussion

In this paper, we aimed to expand the research of reconstructing impaired language for PWAs using modern NLP technologies. However, given the expanse of data availability in English, NLP research in aphasia reconstruction has stalled in other languages. Therefore, this is, to our knowledge, the first effort is implementing LLMs for reconstructing aphasic speech in Spanish.

In this paper, we employed the smallest version of the *Mistral* language model and the Langchain architecture in few-shot and zero-shot inference settings. Aligning with our initial expectations, we observed that the model shows better performance across all aphasia subtypes in the few-shot setting, although the performance in the zero-shot setting was still moderately high. We emphasise here that the BERT semantic similarity scores are, while optimistic, not to be taken at face value. Given that semantic similarity measures the semantic distance between two utterances, sentences that are paraphrased, but still convey the same meaning (e.g. “él pateó la pelota“; “él golpeó la pelota“), might get a lower score than sentences which convey opposite meanings, but differ by only a single word (e.g. “yo soy feliz; “yo no soy feliz“).

We also explored the relationship between WAB-AQ scores (aphasia severity) and LLM reconstruction accuracy. In the few-shot method, although we discovered a moderately positive correlation between the two variables, the frugality of the dataset was insufficient for us to claim that the results did not occur due to random chance alone ($r = .055$; p-value$>.05$). In the zero-shot method, however, we found a strong positive correlation between the two variables ($r=.967$; $p<0.03$), which correlates with the findings reported by Adikari et al. [4] that more severe aphasia types degrade the reconstruction performance of LLMs.

Lastly, we examined the error types present for each aphasia subtype and observed that morphological errors, chiefly gender and verb agreement errors, are, with the exception of global aphasia, represented across all aphasia types. For global aphasia, we observe that the highest number of errors are attributed to neologisms, which is explained by the predominance of this error type in English as well.

6. Ethical Considerations

Although the integration of AI in healthcare is a tantalising prospect, concerns pertaining to ethics and data bias may hinder the implementation of such technology. To ensure that the models do not develop latent biased patterns in their output, data must be gathered from a diverse spectrum of patients with different cultural and language backgrounds, types of aphasia, and varying degrees of aphasia severity [11]. Additional care must be taken to ensure patient privacy and security and upheld in accordance with the existing regulations, such as the General Data Protection Regulation (GDPR) in the European Union.

The AphasiaBank utilised in this research is open to all aphasia researchers. The access to the dataset is done via an online registration to the TalkBank website and an email request by the faculty advisors detailing information and affiliation of the researcher(s). Furthermore, all research conducted with the dataset must adhere to the code of ethics for data usage outlined on the TalkBank website.

The implementation of this technology should not be in the spirit of replacing human healthcare professionals, particularly speech therapists, but should be used in collaboration with them. This practice would ensure that this technology is integrated in concordance with the tenets of medical ethics, helping individuals in need and prioritising the upkeep of their integrity, dignity, and security.

7. Conclusion and Future Work

The goal of this paper was to assess and propose a method for reconstructing impaired speech produced by Spanish PWAs following the methodology described in Adikari et al. [4]. Using the data from the AphasiaBank repository, we tested the performance of a *Mistral-7B* model on four different aphasia types (Broca’s, Wernicke’s, anomic, and global) in a zero-shot and few-shot inference environment. Our results show that the model performs better across all aphasia types in the few-shot setting. The highest scoring aphasia type across both inference setups was anomic, which aligns with the findings of Adikari et al. [4] that aphasia severity is directly correlated to LLM performance.

In other words, more severe aphasia types hinder the reconstruction ability of LLMs to a higher extent compared to less severe aphasia types. Lastly, although inconclusive at this time, our initial results seem to indicate that this model behaviour translates into languages outside of English as well. However, further research will have to be conducted in order for us to fully corroborate this statement.

In the future, we plan to address the limitations faced in this paper by attesting to the performance of our *Mistral* model using all of the available data from AphasiaBank. Following this, we wish to explore the reconstruction ability of models on bilingual (English-Spanish) aphasic data made available by Chen et al. [35].

Lastly, the research of generating synthetic data using

LLMs is due further exploration across non-English aphasia datasets. If this strategy proves to be a viable approach to fine-tuning models for the reconstruction of impaired speech, the issue of data frugality could be bypassed in any language and aphasia subtype. This approach, however, demands further research before the goal of implementing this technology can be made manifest.

Declaration on Generative AI

The authors of this research hereby declare that no generative AI had been employed during its creation. All preparation and production had been conducted solely by its authors with no contribution by any third-party AI tools.

Acknowledgments

This document is part of a project that has received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101132431 (iDEM Project). The views and opinions expressed in this document are solely those of the author(s) and do not necessarily reflect the views of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

We also thank the Spanish State Research Agency under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M) and the Departament de Recerca i Universitats de la Generalitat de Catalunya.

References

[1] M. C. Brady, H. Kelly, J. Godwin, P. Enderby, P. Campbell, Speech and language therapy for aphasia following stroke, Cochrane database of systematic reviews (2016).

[2] C. Code, Speech automatism production in aphasia, Journal of Neurolinguistics 8 (1994) 135–148.

[3] D. F. Benson, A. Ardila, Aphasia: A clinical perspective, Oxford University Press, 1996.

[4] A. Adikari, D. Alahakoon, N. Pallewela, J. E. Pierce, N. J. Hernandez, M. L. Rose, Reconstructing impaired language using generative ai for people with aphasia, Scientific Reports 15 (2025) 40877.

[5] S. T. Engelter, M. Gostynski, S. Papa, M. Frei, C. Born, V. Ajdacic-Gross, F. Gutzwiller, P. A. Lyrer, Epidemiology of aphasia attributable to first ischemic stroke: incidence, severity, fluency, etiology, and thrombolysis, Stroke 37 (2006) 1379–1384.

[6] S. M. Sheppard, R. Sebastian, Diagnosing and managing post-stroke aphasia, Expert review of neurotherapeutics 21 (2021) 221–234.

[7] H. Goodglass, E. Kaplan, The Assessment of Aphasia and Related Disorders, Lea & Febiger, Philadelphia, 1972.

[8] C. Baker, L. Worrall, M. Rose, K. Hudson, B. Ryan, L. O'Byrne, A systematic review of rehabilitation interventions to prevent and treat depression in post-stroke aphasia, Disability and rehabilitation 40 (2018) 1870–1892.

[9] C. Code, M. Herrmann, The relevance of emotional and psychosocial factors in aphasia to rehabilitation, Neuropsychological rehabilitation 13 (2003) 109–132.

[10] A. Frederick, M. Jacobs, C. J. Adams-Mitchell, C. Ellis, The global rate of post-stroke aphasia, Perspectives of the ASHA Special Interest Groups 7 (2022) 1567–1572.

[11] A. J. Privitera, S. H. S. Ng, A. P.-H. Kong, B. S. Weekes, Ai and aphasia in the digital age: A critical review, Brain Sciences 14 (2024). URL: https://www.mdpi.com/2076-3425/14/4/383. doi:10.3390/brainsci14040383.

[12] H. L. Flowers, S. A. Skoretz, F. L. Silver, E. Rochon, J. Fang, C. Flamand-Roze, R. Martino, Poststroke aphasia frequency, recovery, and outcomes: a systematic review and meta-analysis, Archives of physical medicine and rehabilitation 97 (2016) 2188–2201.

[13] M. C. Linebarger, M. F. Schwartz, J. R. Romania, S. E. Kohn, D. L. Stephens, Grammatical encoding in aphasia: Evidence from a "processing prosthesis", Brain and Language 75 (2000) 416–427.

[14] C. K. Thompson, J. J. Choy, A. Holland, R. Cole, Sentactics®: Computer-automated treatment of underlying forms, Aphasiology 24 (2010) 1242–1266.

[15] O. Topsakal, T. C. Akinci, Creating large language model applications utilizing langchain: A primer on developing llm apps fast, in: International conference on applied engineering and natural sciences, volume 1, 2023, pp. 1050–1056.

[16] S. van Vaals, Y. Matusevych, F. Tsiwah, Generating completions for broca's aphasic sentences using large language models., IEEE Journal of Biomedical and Health Informatics (2025).

[17] S. B. Manir, K. S. Islam, P. Madiraju, P. Deshpande, Llm-based text prediction and question answer models for aphasia speech, IEEE Access (2024).

[18] M. M. Abbassy, W. M. Ead, A. I. El-Shora, A. M. Aboalndr, Uncovering advanced transfer learning strategies for deep neural networks in natural language processing, Scientific Reports 16 (2026) 14698.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[20] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, Journal of Machine Learning Research 25 (2024) 1–53. URL: http://jmlr.org/papers/v25/23-0870.html.

[21] M. Popović, chrf: character n-gram f-score for automatic mt evaluation, in: Proceedings of the tenth workshop on statistical machine translation, 2015, pp. 392–395.

[22] C.-Y. Lin, F. J. Och, Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in: Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04), 2004, pp. 605–612.

[23] B. MacWhinney, D. Fromm, M. Forbes, A. Holland, Aphasiabank: Methods for studying discourse, Aphasiology 25 (2011) 1286–1307.

[24] B. Romera-Paredes, P. Torr, An embarrassingly simple

approach to zero-shot learning, in: International conference on machine learning, PMLR, 2015, pp. 2152–2161.
[25] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM computing surveys (csur) 53 (2020) 1–34.
[26] A. Singh, A. Ehtesham, S. Mahmud, J.-H. Kim, Revolutionizing mental health care through langchain: A journey with a large language model, in: 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), 2024, pp. 0073–0078. doi:10.1109/CCWC60891.2024.10427865.
[27] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).
[28] D. Gunawan, C. Sembiring, M. A. Budiman, The implementation of cosine similarity to calculate text relevance between two documents, in: Journal of physics: conference series, volume 978, IOP Publishing, 2018, p. 012120.
[29] T. Sellam, D. Das, A. Parikh, Bleurt: Learning robust metrics for text generation, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 7881–7892.
[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[31] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[32] A. Kertesz, The western aphasia battery: A systematic review of research and clinical applications, Aphasiology 36 (2022) 21–50.
[33] H. M. Clark, R. L. Utianski, J. R. Duffy, E. A. Strand, H. Botha, K. A. Josephs, J. L. Whitwell, Western aphasia battery–revised profiles in primary progressive aphasia and primary progressive apraxia of speech, American journal of speech-language pathology 29 (2020) 498–510.
[34] Z. Yan, D. Wei, S. Xu, J. Zhang, C. Yang, X. He, C. Li, Y. Zhang, M. Chen, X. Li, et al., Determining levels of linguistic deficit by applying cluster analysis to the aphasia quotient of western aphasia battery in post-stroke aphasia, Scientific Reports 12 (2022) 15108.
[35] X. J. Chen, M. J. Marte, S. Kiran, E. Blanco-Elorrieta, Evidence for an integrated bilingual language system from discourse tasks in aphasia, Aphasiology (2026) 1–24.