Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/131/2024.20570

Capso—An LLM-Based Verbal Aphasia Assistance APP

Shangbin Dai[1,5,*], Yixiao Peng[2,6], Kaiwen Chen[3,7], Hengzhi Lyu[4,8]

[1]University of Bristol, Bristol, UK
[2]Fuzhou University, Xiamen, China
[3]Nanjing Normal University, Nanjing, China
[4]University of Minnesota Twin Cities, Minneapolis, USA

[5]shangbin0813@gmail.com
[6]eshaw_peng@qq.com
[7]chenkaiwen8112@gmail.com
[8]njluhz@hotmail.com
*corresponding author

Abstract. The increasing prevalence of verbal aphasia, exacerbated by the rise of digital communication and internet slang, has highlighted the need for innovative solutions to assist individuals with this condition. In response to this demand, we have developed Capso. This advanced application leverages cutting-edge AI technology, specifically the 'LLaMA 2 & BLIP-2' language model, to generate personalized and stylistically diverse captions from user-provided images. By addressing the challenges of verbal aphasia, Capso not only aids in communication but also enhances the user's ability to express themselves creatively on social media. This paper discusses the design principles, user workflow, and the technological core of Capso, ultimately demonstrating how it stands within the broader landscape of AI-driven aphasia interventions. The potential future enhancements of Capso underscore its promising role in the ongoing evolution of digital tools for aphasia therapy.

Keywords: Verbal Aphasia, AI-Generated Captions, Large language Model, Multimodal Interaction, User-Centric Design

1. Introduction

Communication abilities are affected to a large extent by verbal aphasia, which is a language disorder. Such a condition has become more serious with the advent of quick digital communication and abbreviations. According to studies, the constant usage of put-together sensory data (memes, emojis, and intangible Internet language) is connected with decreased verbal intelligence skills (MDPI). Apart from that, as communication gets more and more digitalized, the life issue of the linguistic and emotive requirements of people with aphasia has become more complicated [1].

The Capso application, made to solve the increasingly coming problems, scrutinizes the 'LLaMA 2 & BLIP-2' model to create text from the pictures that a user has uploaded. This groundbreaking method deals with the problem of verbal aphasia by converting image templates to appropriate and stylish text, thus boosting the communication process [2].

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/131/2024.20570

2. UI Design & User Workflow

Capso's UI design prioritizes simplicity and accessibility, ensuring ease of use for individuals of all abilities. The design process was guided by principles of user-centered design, emphasizing accessibility, personalization, and emotional engagement. The application consists of several key components:

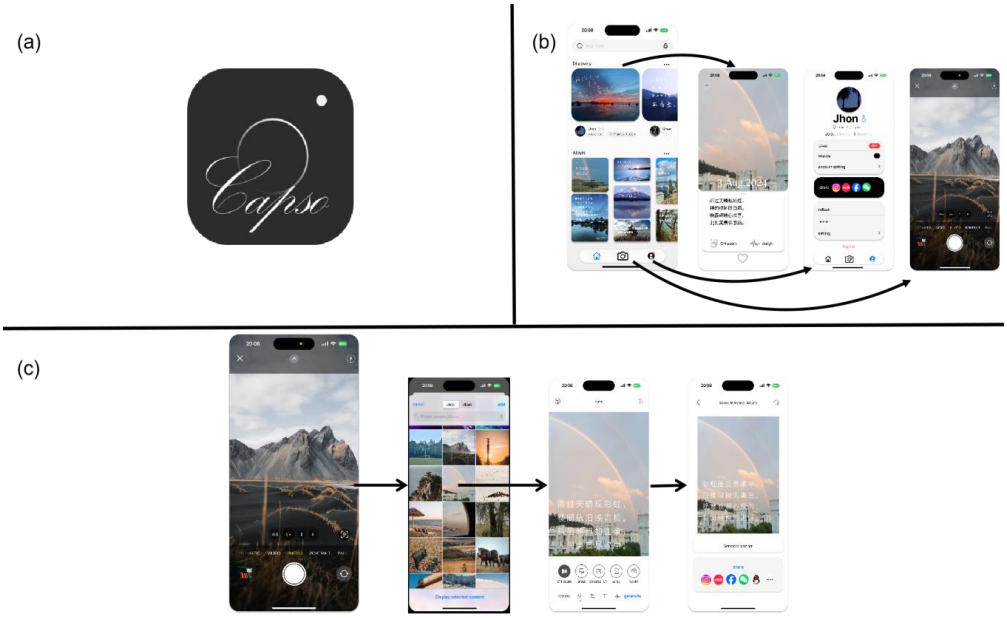

Figure 1. Image Import and Preview Interface. (a)The logo of Capso; (b)App homepage and 'Community', 'User', 'Camera' function pages; (c)The general process of taking photos or uploading pictures to generate text

2.1. Photo Import and Preview

The feature that lets users import and preview photos in Capso's workflow gives them the option to take or upload images from their device. This flexibility makes sure that they can pick the most suitable image for their needs. After they choose an image, the preview interface offers tools for cropping, rotating, and adjusting brightness. This is shown in Fig. 1. These tools allow users to focus on the key parts of the image, which improves the accuracy of the text generation process [3]. Users can also apply different filters. These filters can help make the image's style match the emotion they want in the caption. The zoom function helps users look closely at certain parts of the image. This feature allows them to highlight important details. At this stage, user control is important because it affects the quality and relevance of the captions generated by Capso's 'LLaMA 2 & BLIP-2' model. This makes the text fit better with the image and helps meet the user's needs.

2.2. Text Generation

The text creation stage is a key part of the Capso app. It is powered by the 'LLaMA 2 & BLIP-2' model. This step happens after you pick or edit an image. It makes captions, similar to those used by people with communication problems. Users can start creating captions by selecting classical Chinese poems, modern stories, famous sayings, or Instagram taglines. This gives them the chance to pick the writing style that fits them the best. Also, Capso lets users change the mood of the captions so they can shape the emotional tone of the text, like in Fig.2. This feature includes pictures, the user's chosen style, and emotional settings to make emotional and fitting captions.

DOI: 10.54254/2755-2721/131/2024.20570

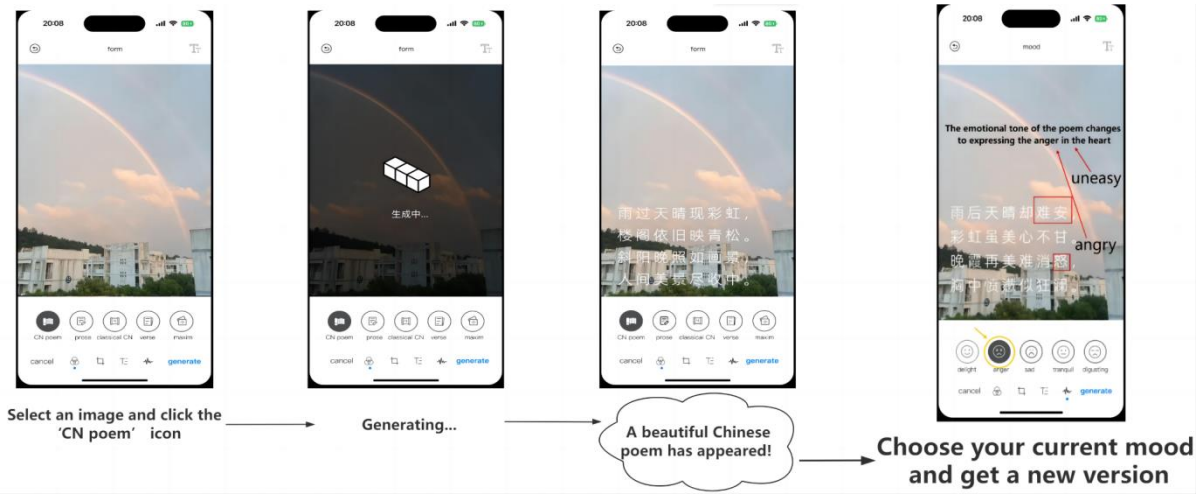

Figure 2. Text Generation and Customization Interface [4]

2.3. Text Editing and Customization

After creating the text, the users can make additional modifications and customize it. Capso has a variety of editing tools, which are important for users fine-tuning their text. These consist of text color adjustment, positioning through drag-and-drop, and word count limit setting. Sometimes the word count (such as in traditional Chinese poems) can be fixed in order to keep the poetic form of the poem. This shape-shifting ability allows the user to fine-tune the product, matching its results to the desired output, and so, turns it into a magnificent tool for personal expression [5].

2.4. Community Engagement

Community interaction is a core aspect of Capso. This feature helps you share posts that you have written, see the ones created by other users, like them, and comment on them. Thanks to this, users meet each other and inspire them to produce new content. Connecting with people who have similar interests also encourages individuals to try various expression forms, which could enhance their language skills.

2.5. Sharing

As soon as users are happy with their content, they can find themselves on social sites sharing it. With Capso being media-friendly, users can easily vent in an artistic manner that is compatible with text usage and content expansive visuals. The share option is intuitive in order to avoid complicating the process of interaction.

3. Technological Background

Capso's key architecture is composed of LLaMA 2 and BLIP-2 models, integrating the power of both:

LLaMA 2: A transformer-based language model whose training is specifically on simplified text, particularly targeted at the aphasia-prone. Its styles are inclusive of poetry, modern prose, and social media captions in more actionable ways [6].

BLIP-2: Here it could replace or complement photo picturing work: in generating captions that describe what is in the image accurately. Trained on a dataset of social media imagery, it provides captions that respect the emotional context, and are in line with the tone of the images in question, representing [7].

The hybrid approach we have developed utilizes state-of-the-art text generation as well as visionary capabilities to produce personalized and meaningful captions.

DOI: 10.54254/2755-2721/131/2024.20570

3.1. Fusion and Fine-Tuning Process

In order to construct a captioning system which produces contextually fine captions and works in a feedback-based manner, the two models went through an intensive fine-tuning stage. BLIP-2, an image-caption model adapted a dataset of 20,000 image-caption pairs, while LLaMA 2 was modified to focus on the kinds of sentences like the ones that are easier to understand by persons who have been diagnosed with aphasia [7]. Additionally, a conditional generation module was introduced to allow users to control the sentiment and style of the output.

3.2. Evaluation Results

The method was rated based on Perplexity (PPL) and BLEU measures, and these two bases should identify points to whether the caption generated is acceptable and fluent.

Table 1. 'LLaMA 2 & BLIP-2' Model Evaluation Data

Experiment	Metric	LLaMA 2 (Untrained)	LLaMA 2 (Fine-Tuned)	BLIP-2 (Untrained)	BLIP-2 (Fine-Tuned)
Text Perplexity (PPL)	Verbal Aphasia Texts	25.5	9.8	-	-
BLEU Score	Caption Generation	18.3	34.7	22.5	38.1
Sentiment Accuracy	Sentiment Control	65%	91%	-	-
Fluency Accuracy	Across Genres	60%	87%	-	-

Comparison of perplexity (the lower the better) and BLEU score (the higher the better) in Table 1. suggests that 'LLaMA 2 & BLIP 2'model have achieved its target and produced high quality captions.

4. Future Work

The existing platforms may pose difficulty for users who have specific needs. Nevertheless, there are some ways to improve Capso's tech core and cater the software as it assists a person with verbal aphasia:

4.1. Incorporation of Multilingual Capabilities

The fine-tuning process can additionally be expanded into the languages, which have the widely different sentence sentence structures. Then, the application can be useful for those users who communicate in other languages apart from Chinese or English [8].

4.2. Contextual and Adaptive Learning

With the incorporation of reinforcement learning into Capso, future releases could gain the ability to seamlessly adjust to user preferences. It will be based on the cues from user reactions and update how the captions will appear depending on how they interacted with it [9].

4.3. Emotionally Adaptive Models

A more complex emotional awareness may be elaborated by going through the dataset more attentively while including a richer set of emotional twists [9]. Capso can help users communicate complex feelings and messages by creating a model that will learn subtle emotional expressions.

4.4. Real-World Validation with Aphasia Patients

Random user testing with people who are aphasia diagnosed will be valuable as input to further refresh the models. This will familiarize us with what currently works and what does not and thus, possibly enhance communication using the tool for the intended users.

DOI: 10.54254/2755-2721/131/2024.20570

5. Conclusion

Capso is an application that uses the "LLaMA 2 & BLIP-2" model to help people with verbal aphasia. It allows users to create personalized captions and post content on social media, giving them the opportunity to express themselves and communicate more creatively. By combining the two models and fine-tuning them, we achieved significant improvements in text fluency and emotional accuracy. With further research in the future, it is hoped that Capso will actually help improve verbal aphasia.

Acknowledgement

Yixiao Peng, Kaiwen Chen, and Hengzhi Lyu contributed equally to this work and should be considered as co-second authors.

References

[1] Lin Cao, “Talking about verbal aphasia, don't let rubbish short videos waste The Ability to Think,” China Youth Daily. Accessed: Sep. 07, 2024. [Online]. Available: https://m.cyol.com/gb/articles/2023-07/04/content_LgLJpvtG7V.html

[2] Swathi Kiran, “Using Artificial Intelligence for Better Aphasia Therapies,” National Aphasia Association. Accessed: Sep. 07, 2024. [Online]. Available: https://aphasia.org/stories/using-artificial-intelligence-for-better-aphasia-therapies/

[3] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, and A. Jabbar, “A Review on Methods and Applications in Multimodal Deep Learning,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 19, no. 2s, pp. 1–41, Jun. 2023, doi: 10.1145/3545572/SUPPL_FILE/3545572-SUPP.PDF.

[4] A. Iannone and D. Giansanti, “Breaking Barriers—The Intersection of AI and Assistive Technology in Autism Care: A Narrative Review,” Jan. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/jpm14010041.

[5] J. A. G. García, L. V. G. Treviño, and J. C. F. López, “Quality of Life Perception on University Students: Comparative Study between Men and Women,” Journal of Education and Human Development, vol. 5, no. 3, 2016, doi: 10.15640/jehd.v5n3a16.

[6] Dave Bergmann, “What is Llama 2?,” IBM. Accessed: Sep. 07, 2024. [Online]. Available: https://www.ibm.com/topics/llama-2?mhsrc=ibmsearch_a&mhq=What%20is%20Llama%202

[7] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” Jul. 03, 2023, PMLR. Accessed: Sep. 07, 2024. [Online]. Available: https://proceedings.mlr.press/v202/li23q.html

[8] H. Rathnayake, J. Sumanapala, R. Rukshani, and S. Ranathunga, “Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification,” Knowl Inf Syst, vol. 64, no. 7, pp. 1937–1966, Jul. 2022, doi: 10.1007/S10115-022-01698-1/TABLES/19.

[9] S. Lamsiyah, A. El Mahdaouy, A. Nourbakhsh, and C. Schommer, “Fine-Tuning a Large Language Model with Reinforcement Learning for Educational Question Generation,” pp. 424–438, 2024, doi: 10.1007/978-3-031-64302-6_30/TABLES/4.