Text Style Transfer with Large Language Models: Enhancing Medical Anamnesis Transcriptions

Yanna Torres Gonçalves[1], **Ticiana L. Coelho da Silva**[1]

[1]Insight Data Science Lab – Universidade Federal do Ceará (UFC)

yannatorres@alu.ufc.br, ticianalc@insightlab.ufc.br

Abstract. *This study investigates the use of Large Language Models (LLMs) for enhancing medical anamnesis transcriptions via Text Style Transfer (TST). It involves benchmarking three models (Phi3, Llama, and Mistral) and fine-tuning Mistral based on evaluator feedback. Mistral achieved the best initial performance but showed limited improvement after fine-tuning. The work highlights the challenges of adapting LLMs to clinical tasks and discusses limitations such as data quality and evaluator bias. Future directions include extended training, dataset expansion, and exploring new machine learning techniques.*

1. Introduction

In medical practice, the anamnesis stands as a crucial element for patient assessment and diagnosis. Writing a detailed medical history is fundamental for clinical evaluation, guiding medical decision-making while providing a comprehensive patient overview that encompasses physical, social, and psychological aspects. It is also responsible for creating the patient-doctor relationship [Soares et al. 2016, Yehia et al. 2024]. However, significant challenges persist in current clinical settings. As reported by Hapvida NotreDame Intermédica, approximately half of the consultation time is consumed by documentation, often resulting in incomplete records that may lead to inaccurate diagnoses and inadequate treatments[1].

These challenges stem from bureaucratic processes in information gathering and the cognitive burden on healthcare professionals during documentation, which may lead to the omission of crucial details [Gür 2012]. Automatic Speech Recognition (ASR) technology emerges as a promising solution, enabling computers to interpret human speech and convert it to text [Reddy 1976]. While ASR systems have shown diverse applications in healthcare, their implementation for medical records faces limitations. [Gonçalves et al. 2024] highlights how Wav2Vec2 PT achieved superior performance compared to other ASR models in Brazilian Portuguese medical transcription, yet encountered persistent difficulties with phonetic ambiguities, silent letters, and medical abbreviations, even when employing decoding strategies like n-gram language models. These challenges, coupled with the unstructured nature of raw ASR outputs, make the direct clinical application of these technologies problematic without additional processing steps.

To address these limitations, this study proposes integrating ASR with Large Language Models (LLMs) to enhance outputs through Text Style Transfer (TST) tasks. While biomedical LLMs like BioGPT [Luo et al. 2022] and BioBERT [Lee et al. 2019]

[1]https://www.hapvida.com.br/

exist, they primarily train on scientific literature rather than clinical dialogues. Our work therefore pursues two primary objectives: (1) developing a fine-tuned LLM specifically adapted for medical history documentation to optimize both record quality and efficiency, and (2) evaluating its effectiveness in formatting and enhancing ASR outputs for clinical use. This approach addresses the critical gap in domain-specific adaptation, aiming to transform raw transcriptions into structured, clinically actionable medical records.

2. Related Work

Recent advances in LLM-based approaches have demonstrated significant potential for TST applications in medical documentation [Mukherjee and Dušek 2024, Liu et al. 2024]. While the effectiveness of LLMs remains debated, particularly in multilingual contexts requiring extensive labeled examples, their flexibility through prompting and fine-tuning offers new possibilities for clinical text processing. [Mukherjee et al. 2024] shows that while open LLMs underperform previous state-of-the-art methods in prompting scenarios, fine-tuning yields substantial improvements. However, their work focused exclusively on base models rather than instruction-tuned variants. Our approach extends this investigation by specifically examining instruction-based models through both zero-shot prompting and fine-tuning.

The technical implementation of LLM-based TST continues to evolve, with [Lai et al. 2024] proposing a novel neuron manipulation framework that identifies and deactivates style-specific neurons while preserving overlapping ones critical for content retention. This method demonstrates improved style transfer accuracy, though with noted trade-offs in text fluency. Such technical innovations complement traditional linguistic feature engineering and data-driven approaches [Jin et al. 2022], offering new pathways for medical applications where preserving clinical accuracy during style transformation remains paramount. However, evaluation challenges persist, as standard metrics like BLEU [Papineni et al. 2002] often fail to capture the precision requirements of medical documentation [Jin et al. 2022].

Our work builds upon these foundations while specifically addressing the unique requirements of Brazilian Portuguese medical histories through targeted adaptations of TST techniques.

3. Data and Methods

This study develops a medical-domain LLM for enhancing clinical documentation through four key phases: (1) Benchmarking LLM performance, (2) Training Data Preparation, (3) Model fine-tuning, and (4) Evaluation.

3.1. Benchmark Evaluation

To establish baseline performance, we evaluated three state-of-the-art open-source LLMs using a standardized test set of 70 medical history transcripts recorded by medical students from real clinical cases. The selected models were assessed using a prompt structure that combined single-turn instructions and negative prompting to prevent hallucinations. The prompt template explicitly defined the LLM's role as a "Portuguese medical record formatter" and included placeholders for both the anamnesis definition

({ANAMNESE_DEF}) and raw transcript ({transcript}). All models were then evaluated equally with the processed described in the Subsection 3.4. This benchmarking step enabled the identification of the most promising base model for subsequent fine-tuning.

3.2. Training Data Preparation

Recognizing the absence of suitable pre-existing datasets, we created a custom corpus of 100 doctor-patient interactions through a multi-stage process. Audio recordings were transcribed using Whisper and initially formatted with GPT-4. All outputs were then manually validated by the authors, ensuring three key quality criteria: strict avoidance of hallucinated content, organization into standard medical history sections, and standardized use of "No data" for missing information. A tailored prompt template guided fine-tuning, reinforcing the model's physician role and discouraging content fabrication.

3.3. Fine-tuning

Building on the benchmark results, we fine-tuned the Mistral model for a single epoch using a Tesla V100 GPU (32GB), with a batch size of 4, 2-step gradient accumulation, and a learning rate of 2e-4. The final training loss was 2.6, suggesting limited convergence due to the short training duration. The training was limited to a single epoch due to computational resource constraints. The fine-tuned model was evaluated using the same data and procedure described in Subsection 3.4.

3.4. Evaluation

Three medical evaluators (two final-year medical students and one nursing student) assessed the model outputs using a 5-point Likert scale (1 = poor, 5 = excellent) for formatting quality and clinical appropriateness. Evaluators provided optional comments regarding hallucinations or incorrect content placement. Descriptive statistics (mean, range, standard deviation) were calculated for each model's performance.

We calculated standard descriptive statistics (means, median, mode, and standard deviations) for each model's performance. To ensure rating consistency, we measured inter-rater reliability using both Kendall's Tau correlation coefficient [Kendall 1938] and Fleiss' Kappa [Fleiss 1971].

4. Experimental Results

To evaluate the performance of different language models in structuring medical histories, we conducted a benchmark comparison followed by a fine-tuning experiment. The benchmark results provide an overview of the models' baseline capabilities before any specialized training, while the fine-tuning section investigates how targeted training affects model performance, consistency, and reliability.

4.1. Benchmark Results

We evaluated three large language models on the task of formatting medical histories using 70 clinical transcripts assessed by three medical evaluators. The models tested were Llama-3.2-1B-Instruct[2], Mistral-Nemo-Instruct-2407[3], and Phi-3-mini-4k-instruct[4].

[2]https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct
[3]https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
[4]https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

Table 1 presents their performance on a 5-point rating scale based on clarity, completeness, and clinical relevance. Among the evaluated models, Mistral achieved the best overall results, with a mean score of 4.00, a median of 4, and a mode of 5. It also showed a moderate inter-rater agreement, reflected by a Fleiss' Kappa of 0.266 and Kendall's Tau values ranging from 0.27 to 0.59 across evaluators. These results suggest that Mistral consistently produced coherent and clinically meaningful outputs.

Phi-3 demonstrated moderate performance, with a mean score of 3.22 and both the median and mode at 3. Its Fleiss' Kappa (0.179) and Kendall's Tau values (ranging from 0.29 to 0.52) indicated a fair level of agreement among evaluators. While Phi-3 generally adhered to a structured format, it occasionally lacked specificity in the clinical context.

In contrast, Llama exhibited the weakest performance, with a mean score of 2.01, a median of 2, and a mode of 1. The low Fleiss' Kappa value of 0.093 and Tau range of 0.20 to 0.55 reveal minimal agreement among evaluators and point to a lack of reliability in its outputs, which were often irrelevant or off-topic.

Table 1. Benchmark Performance Comparison of Language Models on Anamnesis Formatting Task

Model	Mean	Median	Mode	Std Dev	Fleiss' Kappa
Phi3	3.22	3	3	0.96	0.179
Llama	2.01	2	1	1.03	0.093
Mistral	4.00	4	5	0.93	0.266

A qualitative analysis further clarified these findings. Mistral's outputs, while generally accurate and context-aware, occasionally included excessively detailed or redundant content. Llama frequently introduced content that was unrelated to the input, undermining its usefulness. Phi-3, though typically well-structured, tended to produce overly generic outputs that sometimes lacked clinical depth.

Based on the combination of quantitative scores and qualitative insights, Mistral was selected as the base model for subsequent fine-tuning. Table 2 illustrates representative outputs from each model.

4.2. Fine-tuning Results

Following a single epoch of fine-tuning using a dataset of 100 medical transcripts, the Mistral model exhibited a slight decrease in average rating, from a mean score of 4.00 to 3.89. The median and mode remained unchanged at 4 and 5, respectively, suggesting that the central tendency of evaluations was largely preserved. However, a notable improvement was observed in inter-rater agreement: Fleiss' Kappa increased from 0.266 to 0.38, and Kendall's Tau correlations among evaluators rose as high as 0.78. This indicates that, despite the slight drop in average scores, evaluators were more aligned in their judgments of the fine-tuned outputs.

The modest change in average rating could be attributed to the limitations of the training process. The fine-tuning dataset was generated semi-automatically by another LLM and did not undergo expert medical validation, potentially introducing inconsistencies or stylistic redundancy. Qualitative feedback from evaluators highlighted a tendency

Table 2. Model Output Comparison Example

Input (PT/EN)	PT: tosse produtiva há 10 dias, mais coriza e alina, febre (1 episódio há 9 dias), sem outras queixas. EN: productive cough for 10 days, plus runny nose, fever (1 episode 9 days ago), no other complaints.
Phi3 Score: 2.66	PT: Queixa principal: - Motivo da consulta: Sem informações. EN: Main complaint: - Reason for consultation: No information.
Llama Score: 1	PT: **Queixa Principal:** A queixa principal é a dor de cabeça, que começa na região do pescoço e se espalha para a região da testa. EN: **Main Complaint:** The main complaint is headache, which starts in the neck region and spreads to the forehead.
Mistral Score: 4.66	PT: **Queixa principal:** - Coriza e alina há 10 dias, com 1 episódio de febre há 9 dias. EN: **Main complaint:** - Runny nose for 10 days, with 1 episode of fever 9 days ago.

toward more verbose and repetitive outputs, which sometimes compromised clarity or objectivity. Despite these issues, the increase in evaluator agreement suggests that the fine-tuning process led to more consistent formatting and structure, likely making the model outputs easier to interpret. These findings underscore the importance of high-quality, expert-reviewed datasets and longer or more targeted fine-tuning. A comparison of the outputs before and after fine-tuning is presented in Table 4.

5. Conclusion and Future Works

Our study demonstrates that while current LLMs show promise for medical history formatting, significant challenges remain. The benchmark evaluation revealed Mistral as the strongest baseline model, though its outputs sometimes contained excessive clinical detail. Subsequent fine-tuning improved inter-rater agreement but yielded slightly lower overall scores, likely due to limitations in our training process, particularly the single-epoch training duration and reliance on unvalidated LLM-generated training data. These constraints highlight the importance of both sufficient training time and high-quality, expert-reviewed datasets for medical applications.

As future works, we identify three key directions for improvement: developing comprehensive, clinician-validated training corpora, implementing extended training regimens with careful hyperparameter tuning, and exploring hybrid approaches that combine LLMs with structured medical knowledge bases. Such advances could help achieve the

Table 3. Pre and Post-Fine-Tuning Metrics for Mistral

Metric	Before FT	After FT
Mean	4.00	3.89
Median	4	4
Mode	5	5
Fleiss' Kappa	0.266	0.38

Table 4. Pre and Post-Fine-Tuning Outputs for Mistral

Input (PT/EN)	PT: tosse produtiva há 10 dias, mais coriza e alina, febre (1 episódio há 9 dias), sem outras queixas. EN: productive cough for 10 days, plus runny nose, fever (1 episode 9 days ago), no other complaints.
Mistral Score: 4.66	PT: **Queixa principal:** - Coriza e alina há 10 dias, com 1 episódio de febre há 9 dias. EN: **Main complaint:** - Runny nose for 10 days, with 1 episode of fever 9 days ago.
Fine-tuned Mistral Score: 4.66	PT: **Queixa principal:** Coriza e alina há 10 dias, febre em 1 episódio há 9 dias, sem outras queixas. EN: **Main complaint:** Runny nose for 10 days, fever in 1 episode 9 days ago, no other complaints.

crucial balance between clinical precision and workflow efficiency necessary for real-world adoption, while addressing current limitations related to output consistency and the inclusion of appropriate details. Future work should also investigate the potential of specialized prompt engineering techniques and frameworks, such as Multi-Agent Debate, to further enhance model performance without extensive retraining.

Acknowledgements. This research was conducted as part of the CEREIA project and was supported by multiple institutions and partners. We gratefully acknowledge the contributions from Hapvida NotreDame Intermédica, the Federal University of Ceará (UFC), and the São Paulo Research Foundation (FAPESP). This research was supported by grant 2020/09706-7, São Paulo Research Foundation (FAPESP). We extend our gratitude to all the institutions, partners, and funders involved in this project.

References

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gonçalves, Y., Alves, J., Sá, B., Silva, L., Macedo, J., and da Silva, T. C. (2024). Speech recognition models in assisting medical history. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 485–497, Porto Alegre, RS, Brasil. SBC.

Gür, B. (2012). *Improving speech recognition accuracy for clinical conversations*. PhD thesis, Massachusetts Institute of Technology.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., and Mihalcea, R. (2022). Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Lai, W., Hangya, V., and Fraser, A. (2024). Style-specific neurons for steering llms in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Liu, Q., Qin, J., Ye, W., Mou, H., He, Y., and Wang, K. (2024). Adaptive prompt routing for arbitrary text style transfer with pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18689–18697.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Mukherjee, S. and Dušek, O. (2024). Text style transfer: An introductory overview.

Mukherjee, S., Ojha, A. K., and Dušek, O. (2024). Are large language models actually good at text style transfer? *arXiv preprint arXiv:2406.05885*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, page 311–318, USA. Association for Computational Linguistics.

Reddy, D. (1976). Speech recognition by machine: A review. *Proceedings of the IEEE*, 64(4):501–531.

Soares, M. O. M., Higa, E. d. F. R., Gomes, L. F., Marvã, J. P. Q., da Fonseca Gomes, A. I., and Gonçalves, A. H. C. (2016). Impacto da anamnese para o cuidado integral: visão dos estudantes portugueses. *Revista Brasileira em Promoção da Saúde*, 29:66–75.

Yehia, A. C., Viana, P. R. L., Macedo, M. V. M., de Souza Dias, N. C., Campos, C. C., Jardim, S. N., and de Almeida Garcia, J. N. A. (2024). Anamnese na prática clínica: uma revisão sobre suas aplicações e importância. *Revista da Sociedade Brasileira de Clínica Médica*, 22(2):116–120.