

Automated Report Generation With Chat GPT API For Cardiology: Performance Evaluation And Impact On Physician Burnout

Keywords: Digital health, Automatic report, Cardiology, Chat GPT

- HE-Xperts Consulting LLC, Miami, FL, USA.
- Universitat de Barcelona, Barcelona, Spain

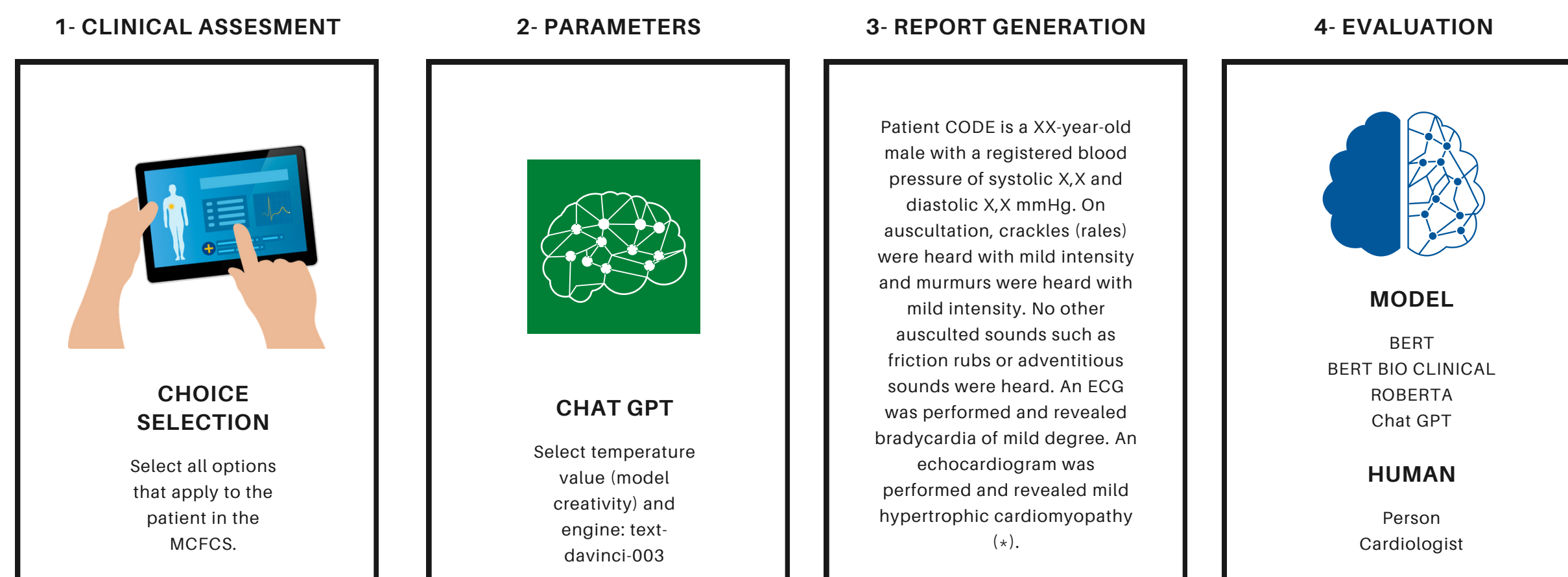


Figure 1. Schematic representation of the automated report generation process. The Python-developed Guiding Multiple Choice and Field Completing System (MCFCS) is depicted, showing a segment of the generated report (*).

01. Introduction

In the realm of modern medicine, the alleviation of administrative burdens on healthcare professionals is paramount. Physician burnout, exacerbated by the strain of paperwork, jeopardizes both the well-being of medical practitioners and the efficiency of healthcare systems. In this pursuit of relief, the integration of cutting-edge technologies becomes indispensable. One such transformative tool is the Chat GPT API, distinguished for its ability to meticulously craft medical notes with utmost clarity and precision (1,2). This study delves into the innovative realm of automated report generation in cardiology, employing the Chat GPT API as a fundamental building block. Through meticulous evaluation, this research illuminates not only the technical prowess of this language model but also its potential to curtail physician burnout significantly.

02. Objective

This study aimed to construct an automated report generator for a cardiology service using a language model and evaluate its performance using algorithmic and human metrics.

References

- Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., ... & Wheless, L. (2022). Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Research square.
- Garg, R. K., Urs, V. L., Agarwal, A. A., Chaudhary, S. K., Paliwal, V., & Kar, S. K. (2023). Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. Health Promotion Perspectives, 13(3), 183.

03. Methods

Developed in Python 3.9, the Guiding Multiple Choice and Field Completing System (MCFCS) integrated patient data, including identification, gender, age, blood pressure, and cardiac results (breath and heart sounds, and ECG/echocardiogram). Using the Chat GPT API (text-davinci-003), it generated plain text reports with varied temperature values. Word count, generation time, and model evaluations (BERT, BIO CLINICAL BERT, ROBERTA, Chat GPT) were measured (Fig. 1). Evaluation scores were set to 1-10 and times for generation and evaluation in seconds.

04. Results

The analysis of generated reports revealed significant findings. Reports produced through the Chat GPT API demonstrated notable success, boasting an average generation time of 8.7 seconds (Fig. 4) and an average word count of 70 (Fig. 3). In comparison, cardiologists took an average of 22 seconds to evaluate reports (Fig. 4), yielding an average score of 8. Remarkably, Chat GPT closely mirrored cardiologist evaluations, scoring an average of 7.45 with an evaluation time of 0.13 seconds (Fig. 4). Conversely, alternative models such as BERT, BIO CLINICAL BERT, and ROBERTA exhibited prolonged evaluation times (averaging above 1.6 seconds) and inferior performance, scoring 3.2, 2.5, and 2.5, respectively. Noteworthy was the performance of BIO CLINICAL BERT, which, being trained with clinical notes, stood as the second-best performer after Chat GPT (Fig. 2). Importantly, the highest score awarded by physicians was achieved with a temperature value of 0.5 in the Chat GPT API (Fig. 2). These outcomes underscore Chat GPT's efficiency and effectiveness in medical report generation and evaluation.

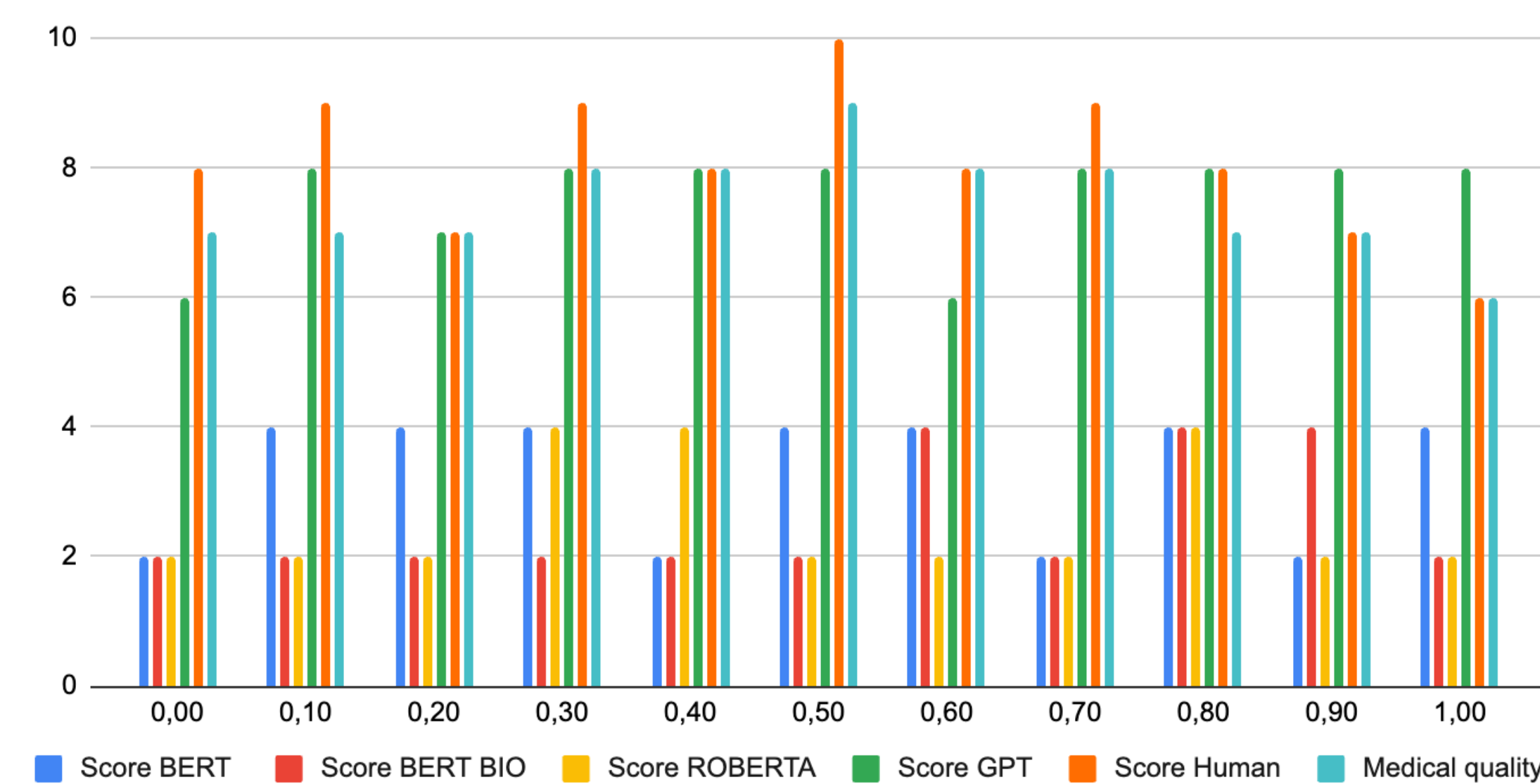


Figure 2. Assessment of the medical report quality. Language models including BERT, BERT BIO, ROBERTA, and Chat GPT were employed for evaluating report quality. Evaluation encompassed assessments by a non-expert in cardiology for general understanding, and a cardiologist for medical accuracy.

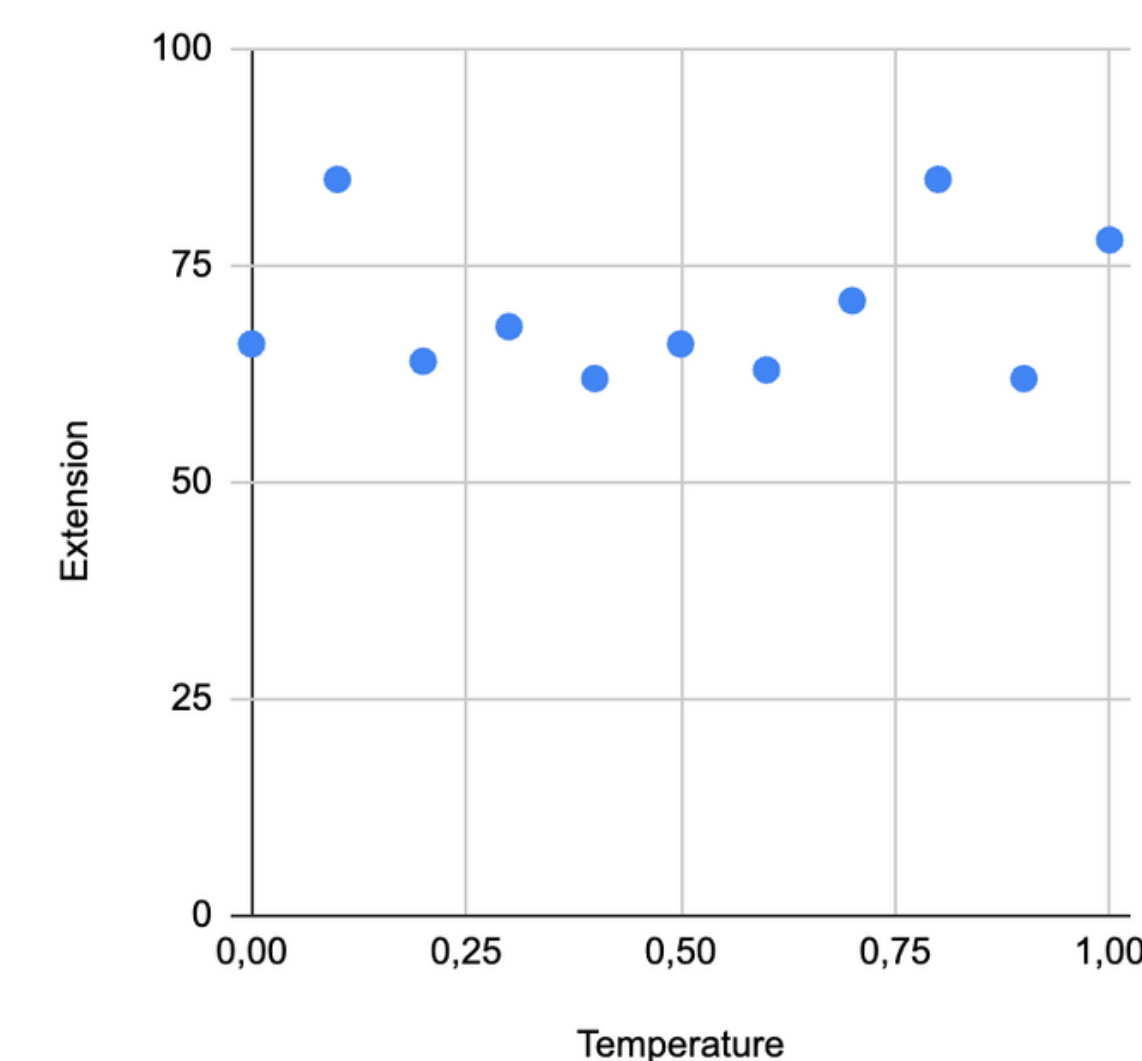


Figure 3. Effects of temperature parameter on the extension of the medical report

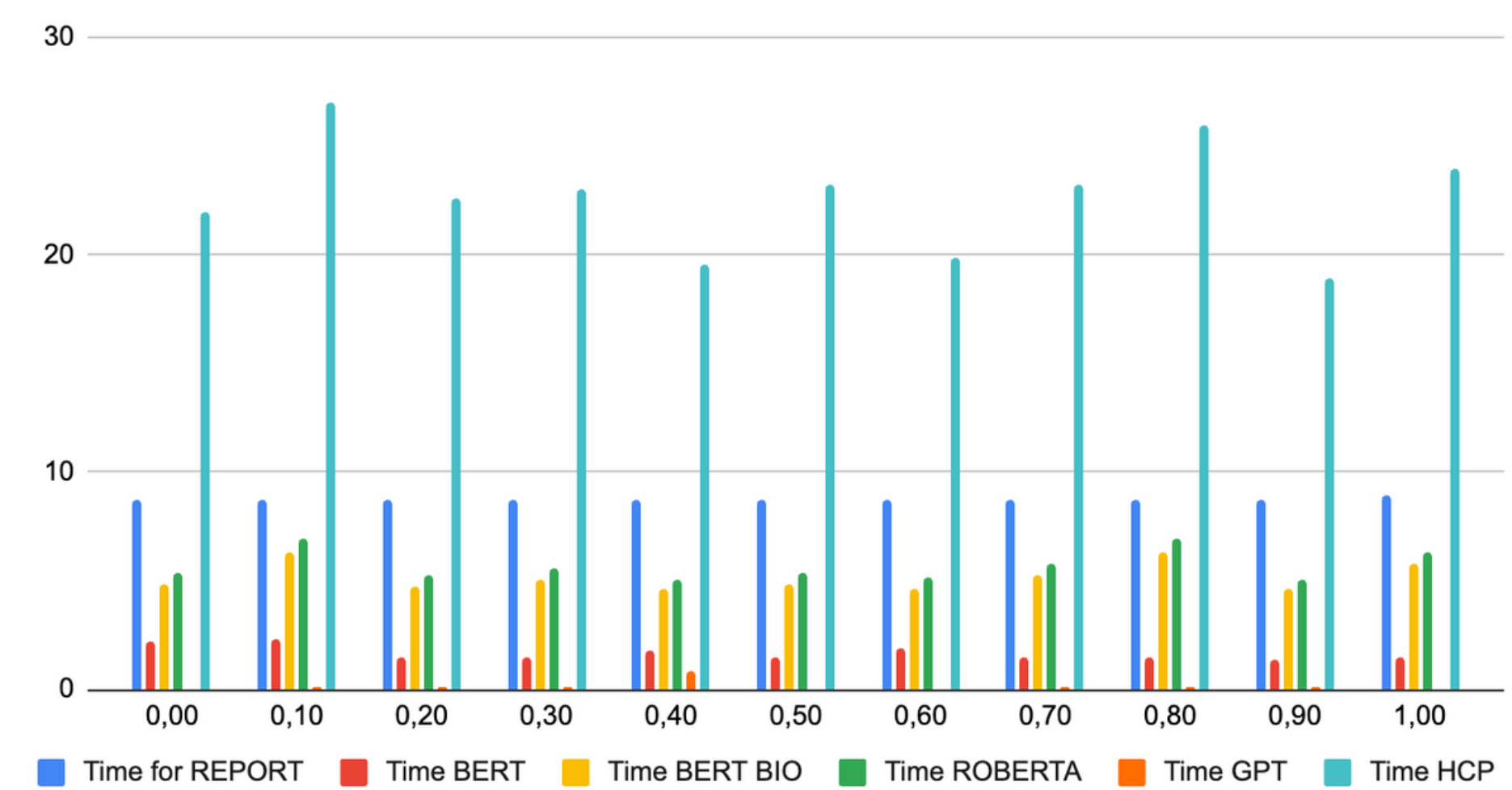


Figure 4. Comparative analysis of report generation and evaluation times for different temperature values. 'Report time' represents the duration Chat GPT took to compose the report, while evaluation times for language models (BERT, BERT BIO, ROBERTA, Chat GPT) and a cardiologist (HCP) are illustrated.

05. Conclusion

The implementation of an automated reporting application enables the rapid generation of high-quality medical reports. Further advancements and refinements hold significant potential for alleviating the burdensome administrative tasks faced by physicians, thus mitigating burnout and optimizing overall healthcare system efficiency.