AI–Driven Clinical Reporting: A Case Study on IQLINIQ

Seyedeh Leili Mirtaheri[1,*], Reza Shahbazian[1], Narges Movahedkor[2], Irina Trubitsyna[1] and Sergio Greco[1]

[1]*Department of Informatics, Modeling, Electronics and System Engineering (DIMES), University of Calabria, Italy*
[2]*Department of Mechanical, Energy, and Management Engineering Department (DIMEG), University of Calabria, Italy*

Abstract

Leveraging the power of generative artificial intelligence (AI) may assist the clinicians in providing prompt care, as well as reducing the financial pressure on patients. Although current generative AI technologies present a possible route for automation, they often lack in report accuracy, quality, and addressing privacy concerns. This paper presents a novel framework for the automatic creation of high-quality reports from clinical meeting audio transcripts. This framework is a part of our clinic management platform (IQLINIQ) and benefits from a three-stage process, including precise audio-to-text transcription, internal anonymization, and a refining phase to ensure consistency and conformity to clinical standards. Using both quantitative measures, including cost and time analysis, and qualitative evaluations, we compare the AI-driven reports against expert-generated ones, individual large language models (LLMs), and a state-of-the-art baseline model, GPT-4-o. Our findings show that our framework noticeably enhances the report preparation process by significantly reducing the time and cost of generating a report from expert-surveyed 3 hours and 750 US-dollars for a complete report to less than 5 minutes and 1 US-dollar, respectively. Improving the report quality by over 10 points compared to existing techniques also underlines the effectiveness of proposed solution.

Keywords

Clinical Reports, Generative AI Report Generation, Large Language Models, Mental Health

1. Introduction

Generative AI is one of the revolutionary trends in technology, quickly becoming mainstream across multiple use cases. The big potential of this disruptive technology can overhaul the nature of industries across multiple sectors, including healthcare [1, 2] and cybersecurity [3]. The prospect of generative AI making processes more efficient and streamlined with recent applications in the generation of clinical reports [4]. Physicians and clinicians today are devoting a significant amount of their time to prepare medical reports and paperwork. It is stated that physicians spend on-average 2 to 6 hours per day dealing with this activity, a substantial amount of time which can be put to better use in more important tasks like patient care and patient throughput enhancement. This administrative time is not only decreasing face-to-face time with patients; it can be a further cause of burnout among physicians, and possibly has an impact

SEBD 2025: 33rd Symposium on Advanced Database Systems, June 16-19, 2025, Ischia, Italy

[*] Corresponding author.

leili.mirtaheri@dimes.unical.it (S. L. Mirtaheri); reza.shahbazian@unical.it (R. Shahbazian); narges.movahed@unical.it (N. Movahedkor); i.trubitsyna@dimes.unical.it (I. Trubitsyna); greco@dimes.unical.it (S. Greco)

on the overall quality of the healthcare system. So surely freeing this wasted time would allow practitioners to better focus on their core competency, which is diagnosis and treatment. For example, ecological momentary assessment (EMA) deployed via a smartphone application [5] has been used to capture the working life of 61 U.S. physicians across 28 ambulatory practice sites. Physician utilization of electronic health record (EHR) consisted of 66.5% of their time spent on direct clinical contact; 20.7% on EHR alone; 7.7% on administrative functions; and 5% on other activities. Most shocking was EHR time: a staggering 44.9% of total physician time spent on EHR. The research finds the relatively low figure on direct patient care to be a concern on efficiency and points toward the imperative need for redesign of EHR and workflow.

The unparalleled success of generative AI models in text generation has made them prevalent in a wide range of applications, and healthcare is no different [6, 4]. The ability of these models to learn and transfer knowledge has created huge interest in how they can revolutionize medical practice. In medicine generally, and most especially in the delicate field of mental health, there have been investigations into using generative AI to speed the creation of clinical reports after psychiatric assessments and clinical visits. Although these early forays into the subject offer us reason to be hopeful regarding the healthcare record of the future, there are still many obstacles. While existing generative AI models have shown powerful text generation capabilities, they still lack the consistency of correct responses and fine-grained comprehension required to generate truly high-quality clinical reports. Their use also raises critical ethical and practical questions about the privacy and security of highly sensitive patient information. Consequently, the consistency gap and the privacy issues clearly represent major obstacles to their broad clinical use, and their use in critical applications requires overcoming such problems.

1.1. Problem Statement

Typically, clinicians are weighed down with the overabundance of administrative tasks, especially clinical report preparation, which takes an average time of 2 to 6 hours a day and significantly restricts time spend on direct patient care. In this respect, EHR systems have overloaded physicians enormously well above spending up to nearly 45% time on EHR-related activities, and thus requires the refurbishment of workflows. With respect to generative AI models in recent years, there has been a current wave of interest in them from academic and clinic applications. Unfortunately, despite strokes of formulating effective means of automatically generating clinical reports, their deficiency in fine-grained understanding coupled with inconsistencies in accuracy hinders their quality. Also, they usually overlook privacy and security concerns regarding the way sensitive patient information is handled. Thus, these limitations present obstacles to the widespread clinical adoption of these models and requires solutions to improve:

- The quality and accuracy of AI-based generated reports.
- Ensuring the privacy and security of patient data.
- Reduced time between use of EHR and report preparation.

1.2. Contributions

To overcome such challenges in report generation, this study presents a novel high-quality production architecture to build reports in clinical setting, giving emphasis to the mental health domain here. We focus on audio transcripts from patient meetings, utilizing multiple unique generative models. By combining rapid generation and moderate level generative models from diverse vendors, we harness their strengths and compensate for their limitations. Therefore, our multimodel approach supports generating high-quality initial reports. Most critically, we offer a novel refinement step performed on these initial reports, where these are combined and polished. The refinement step brings the resulting outcome in accordance with experts written reports, achieving professional precision and maximizing the appropriate contents. We comprehensively assess our system in performance on key performance metrics, such as quality in reports, expenses and costs, and process times. Moreover, we compare consistency in produced reports with professional clinician-produced reports, direct evaluation in reference to humans. Lastly, we compare with state-of-the-art models, comparing performance with high-powered generative model GPT4-o, where we emphasize the performance benefits in our system. The following includes our summarized contributions:

- We present a new multimodel architecture for generating clinical reports, with a refinement step for improved quality and report-appropriate content.
- Rigorous quality, cost, and time evaluations allowing extensive performance evaluations
- Clinical applicability of the generated reports is validated by similarity with expert clinician reports.
- Evaluating with GPT4-o proves the effectiveness and benefits of our proposed method.

We arrange the presented paper in the following structure: initially, sections 2 explores the related literature and existing clinical applications of generative AI models and LLMs. Then, we present detailed description of the proposed architecture in section 3, and evaluation results of the conducted comprehensive assessments, along with detailed description of the utilized dataset and official clinical assessment procedure, are presented in section 4. Finally, we conclude the study with a discussion 4.4 and conclusion 5.

2. Related Works

The requirement for automated clinical report generation have been noticeably noted, and actively researched over the past several years [7]. This is due to sensing the increasing demand for reducing the administrative burden on healthcare professionals and improving efficiency in clinical documentation. In this section, we present the relevant literature in a general sense, on the applications of generative AI in healthcare, across the intersection of natural language processing (NLP), automating report creation, and the use of LLMs.

Automating clinical report generation has been a focus due to its potential to improve efficiency [8, 9, 10]. Transformer models like Biobart-V2 have been used for radiology report summarization, demonstrating their effectiveness in medical text processing [8]. LLMs also offer powerful tools for healthcare tasks such as clinical documentation and diagnosis, although concerns about trust and safety exist [11, 6]. To address these challenges, we utilize multimodel approaches and data anonymization. However, more research into ethical considerations like privacy and bias in LLM use is required [6, 4, 12]. Moreover, they should still be comprehensively evaluated for their effectiveness in specific contexts [13, 14], particularly for medical purposes [15]. Multimodal learning across different data modalities have also been studied for medical imaging applications [16, 17, 18]. The implementation of LLMs in healthcare necessitates robust privacy measures and regulatory frameworks to ensure responsible and secure deployment [6].

Concentrating on clinical reports, the literature has witnessed a growing interest in utilization of LLMs in the medical field. For instance, Google's Med-PaLM 2 has introduced their model, fine-tuned on medical data for medical tasks [19, 20]. Additionally, AI-based systems like *Nabla*[1], *Nannonets* [2], *Notable* [3], *Amelia* [4], *Cognigy*[5], and the *AI for Health* research conducted by Stanford University have introduced new areas for AI to be integrated into and assist the professionals in the healthcare field. Researchers have also conducted studies on evaluating and fine-tuning models like ChatGPT and InstructGPT for specific medical applications. BioGPT [21], pre-trained on PubMed abstracts, has demonstrated superior performance in question answering, relation extraction, and document classification. Similarly, BioMedLM 2.7B and GPT-4-based multi-modal LLMs showcase advancements in this area [22, 23]. Domain-specific versions of BERT, such as BioBERT, PubMedBERT, ClinicalBERT, and BioLinkBERT, have also been developed for scientific and clinical text, demonstrating the adaptability of BERT architectures for medical tasks [24]. Google's PaLM, fine-tuned as Flan-PaLM and Med-PaLM, has achieved state-of-the-art results in medical question answering and clinical reasoning [19]. Additionally, proprietary and open-source medical LLMs like GatorTron [25], Claude [26], and PMC-LLaMA [27] are emerging, contributing to the field's growth. These models, including DRAGON [28], Megatron [29], and Vicuna [30], are also enabling the development of multi-modal LLMs [19, 31].

We recognize that two key components of guaranteeing the quality and dependability of automated systems are the refinement of produced text and the assessment of clinical reports. We tackle this by including a special refinement phase whereby the results of many LLMs are synthesized and polished to meet clinical standards. To give a thorough analysis of our recommended approach, moreover, we combine quantitative measures (cost, time, tokens) with qualitative assessments (quality, clarity, alignment).

3. Proposed Architecture

This section explores the architecture of our approach, proposed for generating clinical reports from meeting audio transcripts. As it is shown in Figure 1, our approach comprises three main stages, including components for real-time transcription, independent and individual report generation models, and a high-level report integration and refinement stage to enhance the report's quality and clarity. The following will go through each of these components in details. Using this architecture, we aim to provide clinicians with precise and detailed clinical

[1] https://www.nabla.com/

[2] https://www.nanonets.health/

[3] https://www.notablehealth.com/

[4] https://amelia.ai/solutions/healthcare/

[5] https://www.cognigy.com/solutions/healthcare

reports, following specified formats. By this multi-stage architecture, we can ensure the quality, consistency, and privacy of sensitive patient data while also reducing the administrative burden on clinicians.

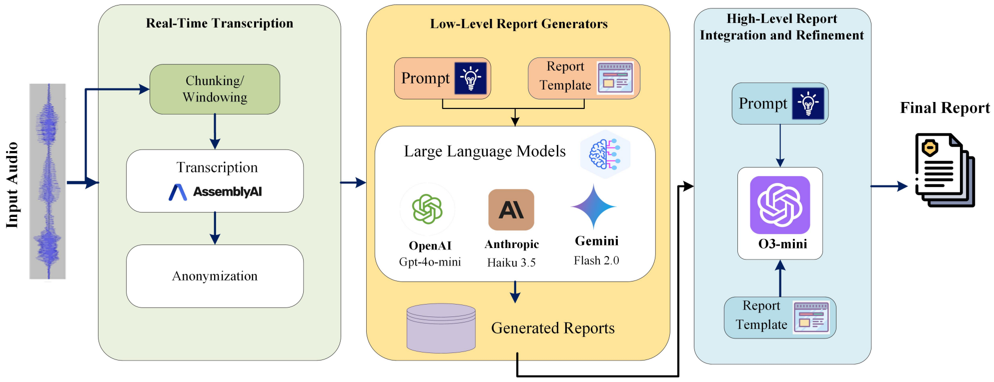


Figure 1: Architecture of our proposed approach. We consider three distinct steps to generate high-quality clinical reports, including real-time audio transcription, independent report generators, and refinement and quality enhancement

Real-time Transcription The initial component of our proposed architecture concentrates on transcribing the input clinical meeting audio into a written format. Various models, either offline or online, can be used as audio transcription tool in the proposed architecture, including *OpenAI Whisper* [32], *Deepgram Nova*, *Google speech to text*, *AssemblyAI*, *Nabla*, *Amazon Transcribe*, and *Azure AI Speech* models. There are several features of audio transcription tools that should be taken into account, among which pricing, processing time, and accuracy among can be considered the most important ones. Therefore, selecting an appropriate tool is all about balancing these features. Some tools may be more expensive to run or take longer to deliver, but richer in features or greater in accuracy. While, others may be quick to respond but inaccurate. In this regard, the choice is one of compromise between speed and accuracy, regarding what best fits the task at hand. Accordingly, we have analyzed and evaluated various audio transcription tools with respect to user needs and resources, and utilized the *Assembly AI* real-time transcription API, as this model is highly accurate and can deal with long audio files. Greater details of the findings can be found in the Section 4.2.

In terms of the real-time transcription, the input audio is required to initially be divided into small, manageable pieces by a chunking process. This enables faster processing and supports real-time transcriptions. The transcript so generated by the transcription module is then run through an Anonymization module to maintain patient confidentiality by eliminating any personally identifiable information. This anonymized transcript then forms the input for the further report generation processes.

Independent Report Generators We utilize a range of LLMs from different sources to create the initial draft reports. We recommend the use of mid-range LLMs, to make it affordable yet still maintain quality, such as Google's Gemini Flash 2.0, Anthropic's Haiku 3.5, and OpenAI's

GPT4-0-mini. Once anonymized, the audio transcripts are fed into every LLM as input, along with a prompt that is designed to guide the generative models towards producing clinical reports. We use official designed questionnaires along with the prompt message to extract relevant information from the audio transcript. Moreover, to provide more accurate analysis by LLMs, we divide the audio script and questionnaires into sections with focused subject, such as history of present illness, concerns, anxiety, allergies and as such.

A specific report template can also be given to every LLM according to the report type to be employed, i.e., psychiatric evaluation or treatment session note, to ensure that the reported output bears a standard form. Therefore, each LLM is processing the provided transcript, clinical questionnaires, designed system prompt, and template, producing an initial report. Multiple LLMs are executing reports simultaneously, which enables us to leverage each model's inherent advantages and procure a diversity of viewpoints. In addition, since each model may have its own weaknesses and strengths, we can benefit from utilizing their natural strengths and compensate for their potential weaknesses. We have used the following system prompt:

> **Report Generator System Prompt:** You are given a list of psychiatric questions and the transcript of a session with the patient. According to the given questions and answers, write the following information as a paragraph in the format of a psychiatric report. Don't add any further information that is not included in the answers.
> **Content:**
> Questions: {*Questionnaire*}
> Transcript: {*Audio Transcript*}

Report Integration and Refinement The final component focuses on unifying the individually generated reports into one consistent, polished, and high-quality final report. For this integration and refinement operation, we utilized the Gemini 03-mini model. We use a particular prompt, meant to steer the integration and refining process by highlighting consistency, thoroughness, and fidelity to medical norms, and the report template is also used in the prior stage guarantees format and structure into consistency. Following the provided medical template, the refined clinical report benefits from the information provided in individual reports, as well as being well organized and thorough, presenting the significant findings and notes from the clinical meeting. We have used the following system prompt for the refinement:

> **Refinement Prompt:** Please combine the following three psychiatric reports for a single patient into one comprehensive and professionally written report, presented in paragraph form. Maintain accuracy and consistency across all information, and use appropriate medical terminology.
>
> **First Report**: {*Content of the first report*}
> **Second Report**: {*Content of the second report*}
> **Third Report**: {*Content of the third report*}

4. Evaluation Results

To evaluate the performance of the proposed architecture, we have conducted an extensive assessment of both cost and time analysis, and the quality of the generated reports. We assess our proposed architecture against several LLMs, the state-of-the-art GPT 4-o model, which is currently the most powerful OpenAI model, and the reports written by professional clinicians. We have specified the following criteria for the qualitative assessments. By *Quality*, we mean the overall quality of the report, including thoroughness, organization, and use of appropriate terminology. *Similarity* refers to how close the report is to the expert report with respect to diagnoses, treatment recommendations, and general evaluation. Ultimately, *Clarity* assesses its writing style and logical flow, how straightforward the report is to grasp.

We surveyed mental health clinicians [6] to establish a baseline for the time and cost linked with conventional report creation, to assess the practical effect of our suggested design. The average time spent and costs accrued by practitioners when manually generating clinical reports were provided in this survey. Using the survey results as a guide to evaluate the possible increases in efficiency and cost-effectiveness that our system might provide in a medical environment, we then measured the performance of our proposed design against these real-world benchmarks.

4.1. Dataset

We used a dataset including five real-case clinical sessions focused on Autism Spectrum Disorder (ASD) evaluations for children aged 6-17 to assess the performance of our proposed architecture [7]. With related records such as patient demographics, clinical notes, and clinician notes, each session featured audio recordings of the medical interaction. It is noteworthy that we have obtained appropriate ethical approvals and consents for using this dataset. Examining our system's performance on this dataset offers a sensible and medically applicable starting point. Emphasis on ASD evaluations makes possible a particular and thorough examination of the capacity of the system to produce within a specific clinical setting precise and complete reports.

4.1.1. Expert Assessment Procedure

Tailored to several age groups, the considered use case clinic uses a thorough, multi-stage process for autism spectrum disorder (ASD) evaluations. Every patient first has a compulsory online interview and autism diagnostic interview-revised (ADI-R) session via *Microsoft Teams* administered by a psychologist. Lasting up to three hours, this session comprises a clinical assistant and psychiatrist; the ADI-R component varies somewhat depending on the psychiatrist used. Clients might then be given further evaluations depending on their age and particular requirements. Autism diagnostic observation schedule (ADOS), cognitive assessment tests, motor evaluations, school well-being assessments, and speech-language pathology (SLP) evaluations are among the tools. While the others are face-to-face, the SLP and school well-being assessments could be carried out on internet. Once all evaluations have been finished, a phone call feedback meeting

[6] The medical data is processed through a clinic management software called *IQLiniQ*, www.iqliniq.com

[7] This study utilizes anonymized medical data provided by clinicians from a mental health clinic in North America. Due to privacy and confidentiality considerations, the name of the clinic cannot be disclosed

is held to go over the results. The facility also has support groups including parent training and resource referral. Important features of this approach are its flexibility across age groups, the compulsory first interview and ADI-R session, the range of extra evaluations available, and the organization of feedback and assistance meetings.

4.2. Audio Transcription

Considering cost, processing time, and word error rate (WER), Table 1 presents a comparative assessment of different audio transcription tools for a 60 minutes audio. For a clear comparison of the models, they are further split into real-time and offline modes. The findings indicate that the *Assembly AI Real-time* model has the lowest price among the real-time transcription tools with merely $0.58, while also resulting in a higher WER. This is while *Speechmatics* has a higher price but less error rate, and *Nabla* is free for merely 30 consultations per month, and then demands $120 per month. Thus, despite its slightly higher WER, the *Assembly AI Real-time* tool seems more cost-efficient and reasonably accurate for large-scale AI-based report generation.

Turning to offline transcription tools, the *Deepgram Enhanced Model* is the cheapest model at $0.1, and took around 0.5 minutes to process. This is while having a high WER of 43, which means that it has the lowest accuracy among offline models. Similarly, other offline models also exhibit a high error rate of over 30 to 43, except for *Assembly AI* tool that showed a remarkably lower WER of merely 9. This model, together with *OpenAI Whisper* and *Deepgram* tools also require a reasonably low price of below 1 dollar, which shows their cost-effectiveness, compared to *Google speech to text*, *Amazon Transcribe*, and *Azure AI Speech.* Despite this, *Deepgram* tools, including both *Nova* and the *Enhanced* model, experienced a significantly lower processing time, requiring merely 10 to 50 seconds for a 60-minute audio. This is particularly important when one demands prompt report generation after the psychiatric session with a patient.

Selecting an appropriate audio transcription model is all about balancing cost, precision, and processing rate. Although processing times range greatly and *Deepgram Nova* provides the fastest offline processing, generally offline models like *Assembly AI* show better accuracy (lower WER) than real-time alternatives. Cost also has a key influence since budget-friendly alternatives frequently sacrifice accuracy whereas more costly models tend to be more accurate. Particularly with its high-performance offline model, *Assembly AI* offers a balanced cost-to-accuracy ratio; however, its real-time service falls in accuracy. Ultimately, the decision hinging on the demands of the particular application will depend on whether critical tasks need more accuracy or time-sensitive activities call for more velocity. Budget limits must also be taken into account. All things considered, and noting that we focused on real-time transcription in our application, we selected the real-time *Assembly AI* tool to provide audio transcriptions, as inputs fed to the report generator stage of our architecture.

4.3. Report Generation

This section discusses the performance analysis of the proposed report generation architecture, regarding both the quantity (cost and time), and the quality based evaluation criteria, including quality, clarity, and similarity to real clinical reports.

Table 1
Cost and Accuracy Performance Evaluation of Different Audio Transcription Models

Model	Price ($)	Duration (Min)	Word Error Rate (WER)
→ Real-Time			
Assembly AI Real-time	**0.58**	-	45
Speechmatics Real-time	1.35	-	36
Nabla Real-time	-	-	36
→ Offline			
AssemblyAI	0.3834	5.16	**9**
Openai Whisper	0.38	5.5	32
Deepgram - Nova	0.25	**0.1**	39
Deepgram - Enhanced Model	**0.1**	0.5	43
Google speech to text	1.5	10	41
Amazon Transcribe	1.5	7	34
Azure AI Speech	1.3	-	27

4.3.1. Cost and Time Analysis

Table 2 shows a contrast of the cost, time, and output tokens among various language models used in our report generation system, comparing individual generators, our suggested refinement-including approach, and the GPT4-o baseline. For a simple comparison of performance across several stages and models, each part of this table is devoted to one of these models. The results are illustrated on average for a complete clinical report. The findings from surveyed clinicians indicate that it could take around 3 hours for an expert to prepare the clinical assessment report. Moreover, the report provision process would cost the patients on average about $750[8], which is noticeably higher than the average expense of AI-generated reports.

Regarding the production of individual reports, three different systems Haiku, GPT4-o-mini, and Gemini Flash 2.0 are used. While GPT4-o-mini produced the most output tokens (meaning longest responses), it imposes the second-lowest cost of $0.003. Still, this model required more time nearly as much as Haiku, and as twice as that of the Gemini. By contrast, the Haiku has a somewhat faster processing time (52.518 seconds), and it produced fewer tokens (2174) than the GPT4-o-mini model. However, it had a greater cost of $0. 028 for a complete clinical report. With 1675 tokens, Gemini Flash 2 yielded the fewest amount of output tokens, having the fastest processing rate (28.165 seconds) and the lowest cost ($0.003). This can imply that Gemini Flash 2. 0 is the most cost-effective alternative for producing first drafts in terms of both speed and expense. Although it is noteworthy that since the token count is lower, it might affect the thoroughness of the report.

The refinement stage (with the o3-mini model) required the most processing time (68.169 seconds) and a medium cost of $0.025, producing the most output tokens (3,853) much like GPT4-o-mini. The main goal for this stage is to integrate the information from individual records, and improve the quality of clinical reports. Hence, it would require longer processing time and heavier output token weights because of attempting to expand the content and thoroughness

[8] The reported value is based on 3+ hour evaluation of ASD in North America and does not necessarily reflects the values in EU.

of the reports. Compared to that, the GPT4o baseline's average processing time was about 53 seconds, a time that was in accordance with the Haiku and GPT-4o-mini. Even so, the cost of a complete report ($0.077) was higher than the personalized report-generator models. At the same time, this model was the one with the least token count (2069), which means that it leads to more brief report texts.

From the findings, there exists a trade-off between processing speed and cost. The processor that was both the fastest and the cheapest is Gemini Flash 2.0, while the most expensive one was GPT4-o, and Haiku and GPT4-o-mini models fell in between. Looking at the output token counts may give us a clue about the length as well as the quantity of details of the reports generated. Regarding this, Gemini Flash 2.0 made the most concise reports, while the o3-mini generated the most verbose. As for GPT4-o, it favored conciseness as well. With regard to the refinement stage, it is more likely to synthesize the reports and expand the initial information from the individual ones, as can be seen in the increased token count. Even though this can mean a longer time of processing and extra cost. The most expensive one of the models was the baseline, meaning that it may not be the most cost-effective model when the purpose is creating numerous reports.

Table 2
Cost and Time Evaluation of the Utilized Models

Provider	Model	Time (s)	Cost ($)	Output Tokens
Individual Report Generation				
OpenAI	GPT4-o-mini	55.644	0.005	3326
Anthropic	Haiku	52.518	0.028	2174
Google	Gemini Flash 2.0	**28.165**	**0.003**	**1675**
Refinement				
OpenAI	o3-mini	68.169	0.025	3853
Baseline				
OpenAI	GPT4-o	52.539	0.077	2069
Clinical	Human Experts	3 hours	750	-

The table outlined in Table 3 is an illustration of the contrasts in report generation time and cost between the proposed multimodel approach, human experts, and the baseline of GPT4-o. The table very clearly highlights the differences between the two methods in terms of cost-effectiveness as well as the measure for time saving. The proposed approach is evaluated in two scenarios: Minimum (Min) and Maximum (Max) of the required time. These account for differences in processing the audio transcripts. In this case, min and max refer to the minimum and the maximum required time taken by the models, both in individual stages and in refinement. The Max scenario refers to the worst case (upper-bound) of sequential implementation of the models.

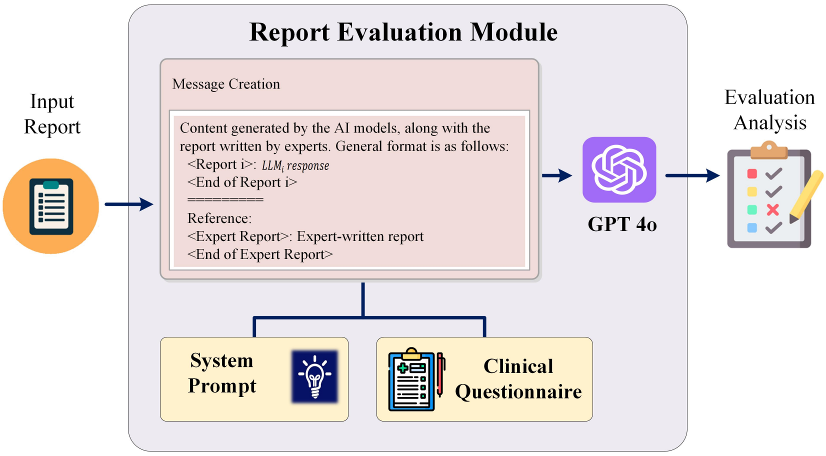
The proposed approach in generating reports showed a minimum of 96 seconds in processing time and a maximum of 204.495 seconds, incurring an overall cost of $0.061. For the GPT4-o baseline, the average time spent per report was 52.5 seconds and the average cost per report was $0.077. We can see that although more time is incurred using the proposed approach, the time is still comparable to the baseline model, and it yields lower cost than the GPT4-o baseline. Notably, both generative AI based approaches markedly surpass the human-based one, which requires significantly more time and cost. Therefore, for high-scale reporting, the proposed approach could be more cost-effective, and be more beneficial at a large scale. In brief, the decision to select one over the other between the proposed technique and GPT4-o baseline will depend on the exact need in the particular scenario. Speed-wise, it can be convenient to consider that GPT4-o succeeds; however, if the dominant factor is cost, the proposed multimodel technique becomes a better alternative notwithstanding more prolonged processing times.

Table 3
Cost and Time Evaluation of the Utilized Models

Method	Scenario	Time (s)	Cost ($)
Proposed	Min	123.812	
	Max	204.495	0.061
GPT4-o	Average	52.539	0.077
Human Experts	Average	3 hours	750

4.3.2. Quality Evaluation Module

We developed a report evaluation module to thoroughly evaluate the performance of our suggested approach using the GPT4-o AI model. Shown in Figure 2, this module helps one systematically evaluate the qualitative components of the produced clinical reports. The first stage of the assessment is the message creation of the input report, which covers compilation of the AI generated report and the expert-authored report that provide the gold standard of reference.



Content generated by the AI models, along with the report written by experts. General format is as follows:
<Report i>: LLM_i *response*
<End of Report i>
==========
Reference:
<Expert Report>: Expert-written report
<End of Expert Report>

Figure 2: Report Quality Evaluation Module

At the core of our evaluation method lies the structural comparison of the generated documents against the expert-written ones. We create a consistent message format for this that includes all relevant information for this analysis. This message contains the generated content, clearly delimited for each model, and the reference point to specify the appropriate expert report and differentiating between professional and generated material. Using the following system prompt, we guide the model through the review process to thoroughly analyze the generated contents on several quality measures. First, the quality of the generated reports is assessed based on thoroughness and organization, along with their clarity of language and structure, emphasizing on readability and simplicity of grasp. At last, the produced material is also assessed based on similarity and alignment with the professional-written reports; thus, by means of this criterion, we can show how much the generated reports represent the expert's patient assessment and advice.

> **Evaluation System Prompt:** I have four distinct psychiatric reports and an expert report for the same patient. Please evaluate each of the four reports individually and assign scores based on the following criteria:
> **Quality**: Overall quality of the report, including thoroughness, organization, and use of appropriate terminology. (Scale: 1-100, 100 being the highest)
> **Clarity**: How easy the report is to understand, including its writing style and logical flow. (Scale: 1-100, 100 being the highest)
> **Similarity**: How closely the report aligns with the expert report in terms of diagnoses, treatment recommendations, and overall assessment. (Scale: 1-100, 100 being the most similar)
> Present your scores in the following format for each report: <Report Number>:<Q:Quality Score>; <C:Clarity Score>; <S:Similarity Score>. For example: <R1>:<Q:7>; <C:8>; <S:6>.

A pre-defined clinical questionnaire can also be included to steer the evaluation process and give particular attention to major elements of report quality and content. This analysis and assessment can guarantee a systematic and uniform review throughout all documents.

4.3.3. Quality Analysis

Comparison of different report generation approaches is given through an evaluation criterion by Quality, clarity, and alignment with expert-written reports. There are various critical aspects of the quality evaluation of the generated clinical reports. Quality measures on the completeness of the report in examining thoroughness and organization, logical flow, and structure, as well as representation of using proper medical terms — ensuring that the clinical findings or recommendations are expressed or articulated accurately. Clarity measures how easily the report can be understood by its intended audience, considering various aspects like writing style, sentence construction, and logical idea flow. Thus, the clear report is conceived concisely without jargon wherever possible and organizes the information competently. Finally, alignment with expert-written reports assesses how closely the AI-generated report matches the content and conclusions of reports written by experienced clinicians. It is, thus, an appraisal of how well the AI has assessed an issue, how good it is at capturing essential points regarding that issue, and how closely its recommendations resemble those given by its human counterparts. This ensures that what the generative models produce accords with established clinical best practices. This section considers all the individual models, the proposed multimodel approach and the baseline GPT4-o regarding their quality performance. The findings indicate that the quality of the GPT4-o-mini model results in a quality score of 81, clarity of 76, and alignment of 71, like the baseline of GPT4o. More improvement in performance is noted in Haiku and then higher in Gemini Flash 2.0 in all metrics by around 5 scores. From the findings, it can also be said that the reports produced by Google's Gemini Flash 2.0 ranked the highest among the individual generator models. The proposed method, on the other hand, with its enhancement step outperformed all the independent report generator and gave an overall high score on all metrics. It could achieve 91.8, 88, and 83 on quality, clarity and similarity to the expert-written report, respectively. Therefore, using multiple medium-level models combined with the refinement can yield a clear integration of different independent reports and enhance the depth, clarity of the reports, and be more aligned with the reports written by experts.

Table 4
Quality Evaluation of the Utilized Models

Provider	Model	Quality	Clarity	Expert-written Report Alignment
Individual Report Generation				
OpenAI	GPT4-o-mini	81	76	71
Anthropic	Haiku	84	79	74
Google	Gemini Flash 2.0	87.2	81.8	77.2
Proposed				
	Multimodel	**91.8**	**88**	**83**
Baseline				
OpenAI	GPT4-o	81	76	80

4.4. Discussion

In this study, we applied quantitative and qualitative evaluation measures to different generative AI models for the generation of clinical reports and reveal time, cost, and quality differences attributed to different models. The findings show that the individual report generator models provided rather quick processing time with the reasonably low cost, attracting interest for the rapid and cheap generation of reports. Among them, Gemini Flash 2.0 was particularly notable, having the fastest processing time (28.165 seconds) and lowest per-report cost ($0.003), although it also yielded the fewest tokens (1675) and maybe less thorough reports. Their speed and cheapness, as shown by the qualitative analysis, came at the cost of report quality and alignment. In comparison, while the multimodel approach was slower in terms of processing time, it has proved to be more cost-effective with respect to total cost across multiple reports when compared to the GPT4-o baseline. Given a processing time about equal to that of personal models (about 53 seconds), the GPT4-o baseline cost $0.077 per report. This is an indication that large-scale deployment would offer a more sustainable solution through the use of the proposed architecture.

Regarding the output tokens produced, the proposed approach, being with a refinement stage, naturally produced the longest reports, averaging 3,853 tokens, suggesting more comprehensive synthesis of information. On the other hand, Gemini Flash 2.0 and GPT4-o were more likely to generate short reports, with Gemini Flash 2.0 at 1675 tokens and GPT4-o at 2069 tokens. The token counts, along with the qualitative analysis, do help in understanding the level of detail and elaboration provided by each model. The qualitative assessment confirmed that the proposed multimodel system was many steps ahead in terms of the quality, clarity, and alignment with expert reports. Particularly, significantly surpassing single models like Gemini Flash 2.0, which scored about 87, 81, and 77, and the GPT4-o baseline, which scored 81, 76, and 80, the suggested approach got ratings of 91.8 for quality, 88 for clarity, and 83 for alignment. Therefore, by merging the outputs of several generative models and then fine-tuning them through a second dedicated stage, our approach managed to achieve quite substantial improvements on all three quality metrics.

The implications are relevant for real-life applications of generative AI within the healthcare industry. A conceivable method for using the proposed multimodel architecture is the automation of the production of high-quality clinical reports that may relieve clinicians of some administrative burden so they can focus on patient care. This is especially relevant given that expert-written reports are said to require about three hours and run patients nearly $750 well more than the expenses linked with artificial intelligence-generate reports. Although, regarding the limitations, further studies are required to optimize the proposed architecture in terms of speed and cost, trying out various combinations of LLMs, and making further improvements to the refinement step. User studies with clinicians can also be beneficial, to test the practical usefulness and acceptance of generated reports in real clinical settings.

We acknowledge the importance of data privacy and in particular in the context of General Data Protection Regulation (GDPR). While we utilized cloud-based services for transcription and LLM inference, our system is modular and portable to offline deployments. Transcription can be handled locally (e.g. Whisper), and LLM inference can be executed on-premise (e.g. LLaMA2). It is also important to stress that cloud-based LLM services are not necessarily incompatible with GDPR. For example, as demonstrated in [33], hybrid approaches can ensure compliance by filtering sensitive content locally before using external services. Our architecture is capable to support strict anonymization pipelines to ensure its compliance with GDPR principles such as data minimization, purpose limitation, and integrity/confidentiality (Articles 5 and 32 GDPR). The anonymization module employs Named Entity Recognition (NER)-based redaction and pseudonymization.

5. Conclusion

This paper attempted to address the urgent requirement for AI-driven high-efficiency and high-quality automation in clinical report generation. This is because we recognize the necessary requirement for streamlining the clinical report production, and reducing the pressure on the clinicians. Noting the limitations of the present technologies, and many automated systems failing with the optimal combination of being accurate, high-quality, and trustworthy, we proposed our application with a three-stage architecture. This includes real-time transcription, low-level report generation, and high-level report integration and refinement. Our architecture exploits the best of various generative AI models, with an added refinement to synthesize information derived from other modalities and align with expert-identified templates. At the same time, this power-packed architecture promotes complete report generation while protecting patient data through anonymization approaches.

The effectiveness of our approach toward significant improvements in report quality, clarity, and alignment with expert-written reports has been demonstrated in evaluations of the proposed system as compared to individual AI report generators (GPT4-o-mini, Gemini Flash 2.0, and Haiku) and the GPT4-o baseline. Processing time is found to be longer for our approach, but it is proved to be more cost-effective, in particular in large-scale application, and probably offers the most detailed reports. Regarding qualitative evaluation, the findings show that the proposed approach showed the highest quality ratings of over 83 in all metrics. These results underline how our design might help to free doctors from administrative burden and therefore raise healthcare.

As for future directions, we aim to pay attention to optimizing the proposed architecture in terms of speed and cost, and examine sophisticated model selection and integration techniques. We will carry out thorough evaluations of its performance under real-world clinical criteria in the following study. Furthermore, we also aim to present top consideration to resolving privacy concerns spawned from these systems by means of techniques including differential privacy and federated learning, therefore guaranteeing responsible and safe deployment. Evaluating the clinical usefulness and acceptance of the produced reports will be best achieved by user studies with professionals, therefore hastening their integration into medical operations.

Acknowledgments

This work was partially supported by PNRR MUR project FAIR(PE0000013) and project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

Declaration on Generative AI

During the preparation of this work, the authors used LanguageTool and QuillBot AI tools in order to grammar and spelling check, and text paraphrasing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos Jr, C. Xiong, Z. Z. Sun, R. Socher, et al., Large language models generate functional protein sequences across diverse families, Nature biotechnology 41 (2023) 1099–1106.

[2] J. Meyers, B. Fabian, N. Brown, De novo molecular design and generative models, Drug discovery today 26 (2021) 2707–2715.

[3] S. L. Mirtaheri, A. Pugliese, Leveraging generative ai to enhance automated vulnerability scoring, in: 2024 IEEE Conference on Dependable, Autonomic and Secure Computing (DASC), 2024, pp. 57–64. doi:10.1109/DASC64200.2024.00014.

[4] H. H. Rashidi, J. Pantanowitz, A. Chamanzar, B. Fennell, Y. Wang, R. R. Gullapalli, A. Tafti, M. Deebajah, S. Albahra, E. Glassy, et al., Generative artificial intelligence in pathology and medicine: A deeper dive, Modern Pathology 38 (2025) 100687.

[5] F. Toscano, E. O'Donnell, J. E. Broderick, M. May, P. Tucker, M. A. Unruh, G. Messina, L. P. Casalino, How physicians spend their work time: an ecological momentary assessment, Journal of General Internal Medicine 35 (2020) 3166–3172.

[6] P. Zhang, M. N. Kamel Boulos, Generative ai in medicine and healthcare: promises, opportunities and challenges, Future Internet 15 (2023) 286.

[7] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, Large language models in medicine, Nature medicine 29 (2023) 1930–1940.

[8] B. A. K. Balouch, F. Hussain, A transformer based approach for abstractive text summarization of radiology reports, in: International Conference on Applied Engineering and Natural Sciences, volume 1, 2023, pp. 476–86.

[9] J. Li, Y. Lin, P. Zhao, W. Liu, L. Cai, J. Sun, L. Zhao, Z. Yang, H. Song, H. Lv, et al., Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer (bert) and in-domain pre-training (idpt), BMC Medical Informatics and Decision Making 22 (2022) 200.

[10] P. Sloan, P. Clatworthy, E. Simpson, M. Mirmehdi, Automated radiology report generation: A review of recent advances, IEEE Reviews in Biomedical Engineering (2024).

[11] Y. Shokrollahi, S. Yarmohammadtoosky, M. M. Nikahd, P. Dong, X. Li, L. Gu, A comprehensive review of generative ai in healthcare, arXiv preprint arXiv:2310.00795 (2023).

[12] A. J. Thirunavukarasu, R. Hassan, S. Mahmood, R. Sanghera, K. Barzangi, M. El Mukashfi, S. Shah, Trialling a large language model (chatgpt) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care, JMIR Medical Education 9 (2023) e46599.

[13] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[14] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al., Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models, PLoS digital health 2 (2023) e0000198.

[15] J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, et al., Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum, JAMA internal medicine 183 (2023) 589–596.

[16] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, M. Zhao, A. K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel, et al., A multimodal generative ai copilot for human pathology, Nature 634 (2024) 466–473.

[17] A. Y. Wang, S. Lin, C. Tran, R. J. Homer, D. Wilsdon, J. C. Walsh, E. A. Goebel, I. Sansano, S. Sonawane, V. Cockenpot, et al., Assessment of pathology domain-specific knowledge of chatgpt and comparison to human performance, Archives of pathology & laboratory medicine 148 (2024) 1152–1158.

[18] S. Chen, Z. Wu, M. Li, Y. Zhu, H. Xie, P. Yang, C. Zhao, Y. Zhang, S. Zhang, X. Zhao, et al., Fit-net: Feature interaction transformer network for pathologic myopia diagnosis, IEEE Transactions on Medical Imaging 42 (2023) 2524–2538.

[19] Y. Matias, G. Corrado, Our latest health ai research updates, Google [Internet] 14 (2023).

[20] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, arXiv preprint arXiv:2212.13138 (2022).

[21] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, Briefings in bioinformatics 23 (2022) bbac409.

[22] D. Zhu, J. Chen, X. Shen, X. Li, M. Elhoseiny, Minigpt-4: Enhancing vision-language understanding with advanced large language models, arXiv preprint arXiv:2304.10592 (2023).

[23] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, X. Gao, Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4, medRxiv (2023) 2023–06.

[24] S. L. Mirtaheri, S. Greco, R. Shahbazian, A self-attention tcn-based model for suicidal ideation detection from social media posts, Expert Systems with Applications 255 (2024) 124855.

[25] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, et al., A large language model for electronic health records, NPJ digital medicine 5 (2022) 194.

[26] J. A. Omiye, J. Lester, S. Spichak, V. Rotemberg, R. Daneshjou, Beyond the hype: Large language models propagate race-based medicine, medRxiv (2023) 2023–07.

[27] C. Wu, X. Zhang, Y. Zhang, Y. Wang, W. Xie, Pmc-llama: Further finetuning llama on medical papers, arXiv preprint arXiv:2304.14454 2 (2023) 6.

[28] M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. S. Liang, J. Leskovec, Deep bidirectional language-knowledge graph pretraining, Advances in Neural Information Processing Systems 35 (2022) 37309–37323.

[29] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, B. Catanzaro, Megatron-lm: Training multi-billion parameter language models using model parallelism, arXiv preprint arXiv:1909.08053 (2019).

[30] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al., Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, See https://vicuna. lmsys. org (accessed 14 April 2023) 2 (2023) 6.

[31] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, R. Daneshjou, Large language models in medicine: the potentials and pitfalls: a narrative review, Annals of internal medicine 177 (2024) 210–220.

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html.

[33] S. Montagna, S. Ferretti, L. C. Klopfenstein, M. Ungolo, M. F. Pengo, G. Aguzzi, M. Magnini, Privacy-preserving llm-based chatbots for hypertensive patient self-management, Smart Health 36 (2025) 100552.