Assessment of Small-Scale Large Language Models for Portuguese-Language Patient Triage and Risk Referral

Luiz R. Faccio[1], **Samuel Feitosa**[1]

[1]Universidade Federal da Fronteira Sul (UFFS)
SC-484, Km 02 - Fronteira Sul, Chapecó - SC, 89815-899

Abstract. *This study examines the use of Large Language Models in patient triage and risk classification. The main objective is to determine whether smaller language models can perform triage tasks effectively. Four small-scale LLMs — gpt-oss:20b, llama3.1:8b, gemma3:12b, and deepseek-r1:14b — were evaluated using 39 fictional test cases to assess their performance, consistency, and reliability. Each case was tested with three different prompts and three validation rounds per prompt. The results show that, although their performance aligns with their model sizes, these LLMs are not yet reliable enough for direct use in clinical workflows. Nonetheless, the study highlights behavioral patterns and potential directions for improving the application of such technologies.*

Resumo. *Este estudo investiga a aplicação de Modelos de Linguagem de Grande Escala no processo de triagem e classificação de risco de pacientes. O objetivo principal é avaliar se modelos menores podem executar essa tarefa de forma eficiente. Foram analisados quatro LLMs de pequeno porte — gpt-oss:20b, llama3.1:8b, gemma3:12b e deepseek-r1:14b — utilizando 39 casos de teste fictícios para medir desempenho, consistência e confiabilidade. Cada caso foi testado com três prompts distintos e três validações por prompt. Os resultados indicam que, embora apresentem desempenho compatível com seus tamanhos, os modelos avaliados ainda não oferecem confiabilidade suficiente para aplicação direta em contextos clínicos. Apesar disso, o estudo permite identificar padrões de comportamento e possíveis caminhos para aprimorar o uso dessas tecnologias.*

1. Introduction

The use of patient triage techniques in emergency care services can be traced back to the early 1980s, and since then, these methods have undergone continuous improvement and widespread implementation [Fry and Burr 2002]. The triage procedure plays a pivotal role within the emergency care system by ensuring that every patient receives timely and appropriate care. It does so by prioritizing patients based on their health indicators and effectively allocating the department's resources [Fekonja et al. 2023, Gerdtz and Bucknall 2001, Andersson et al. 2006, Wireklint et al. 2018]. This process has acquired large influence over the patients outcomes and the overall emergency service flow and quality.

The evolution in the field of Artificial Intelligence (AI) has greatly accelerated in recent times, particularly in techniques related to natural language processing (NLP). Generative Artificial Intelligence (GenAI) is an AI category focused on content generation

(photos, videos, texts, etc.). Within this category are Large Language Models (LLMs), which are AI models specialized in processing human language texts. These models are trained on massive datasets and often use "transformer" architectures, which allows them to draw connections within textual information and generate responses based on context, among other functionalities. Companies in the field have increasingly sought to improve their language models by leveraging chat interfaces, commonly known as "chatbots" [Sarbay et al. 2023]. The intuitive and simplified interaction with models, capable of understanding human language, has enabled the use of this technology to perform various tasks.

In this context, it may be possible to extract significant value from the LLM's capabilities. Given their ability to understand and respond to natural language text, they can analyze symptoms, list potential illnesses that a patient might be experiencing, and ask relevant follow-up questions [Lansiaux et al. 2024]. Thus, when provided with the correct information, language models could assist professionals in the patient triage process, enabling more efficient care.

Accordingly, this research aims to evaluate the efficiency and reliability of four different large language models applied to patient triage in Portuguese language, using the Manchester Triage System (MTS). The benchmarked LLMs are "gpt-oss:20b", "llama3.1:8b", "gemma3:12b" and "deepseek-r1:14b", all freely available at Ollama (https://ollama.com). The study encompasses the development of a model workflow, the assembly of a patient-story repository, an investigation into prompt engineering, and a discussion of the results.

This paper is sectioned to provide context on the patient triage process as well on the LLMs technologies and to discuss the different approaches on combining both elements. Section 2 provides information on key aspects of the patient triage process and Large Language Models. Section 3 covers the related studies on the field. Section 4 elucidates the methodology used for this research, followed by sections 5 and 6 that expose and discuss the results.

2. Theoretical framework

GenAI tools development has largely increased in the recent past. The intuitive and natural interaction enabled by these systems, capable of understanding human language, allows this technology to be applied across a wide range of tasks.

As a subfield of computer science, Machine Learning (ML) focuses on developing computer programs capable of learning and autonomously improving through experience [Mitchell 1997]. Fundamentally, by leveraging representative data from a given environment, well-designed ML algorithms can generalize information and draw inferences from observed examples [Domingos 2012, Duarte and Ståhl 2019, Lorena et al. 2021], thereby enabling them to perform tasks for which they were not explicitly programmed.

2.1. Large Language Models

A large language model (LLM) is built upon a neural network — a powerful machine learning technique. Such networks perform numerous statistical computations to predict the next word in a sequence [Fraser et al. 2023], capturing linguistic and semantic

patterns. Extensive and repetitive training on massive text bodies enables these systems to achieve near-human levels of natural language understanding [Yazaki et al. 2024]. The intensive training — typically using internet-based datasets — enables the adjustment of the neural network parameters, maximizing model coherence and performance. Among LLMs, the most widely adopted architecture is the Transformer, introduced by [Vaswani et al. 2023] in their seminal paper "Attention is All You Need". This architecture leverages a self-attention mechanism that allows the network to process an entire sequence in parallel, enhancing its efficiency in capturing contextual and intratextual relationships.

Although this technology yields great potential, much concern in shown about the quality of the model's answers, questioning the explainability and reliability of the generated text [Lee et al. 2024, Yazaki et al. 2024, Franc et al. 2024a, Franc et al. 2024b, Fraser et al. 2023]. LLMs excel in creating syntactically correct text, that reads superficially plausible, yet, at times, the text is factually incorrect with fully made up information, the so called "hallucinations" [Franc et al. 2024b, Franc et al. 2024a]. In contexts where the factuality of information is essential, such as clinical assessments, hallucinations are extremely dangerous and should be handled with attention.

Seeking to bypass this weaknesses, there are several techniques that can be applied to improve a model's performance on a set of specific tasks. To that end, a model can be fine-tuned: further trained on a pre-defined dataset to give it more information about a desired subject. Along with fine-tuning, the workflow where the model is employed can also be improved to better suit the purpose, allowing the model to work with more precision (e.g. using a Retrieval Augmented Generation (RAG) approach).

2.2. Patient Triage

In emergency healthcare settings, patient triage plays a crucial role in managing patient flow, ensuring effective resource allocation and appropriate service prioritization [Fekonja et al. 2023, Gerdtz and Bucknall 2001, Andersson et al. 2006, Wireklint et al. 2018]. Various triage systems are employed worldwide [Alumran et al. 2020, Fekonja et al. 2023, Wireklint et al. 2018], each with distinct characteristics but sharing the same objective: to provide effective, properly prioritized care while optimizing resource utilization and reducing waiting times. Typically, triage protocols establish algorithms followed by nurses to assess a patient's condition upon initial contact — either within the emergency department or at the scene of an incident. Guided by the selected protocol, healthcare professionals evaluate vital signs such as heart rate, respiratory rate, and the presence of bleeding, and may also record subjective patient reports and other relevant observations in a triage note or anamnesis [Sarbay et al. 2023].

Triage represents a critical component of emergency healthcare, bearing substantial responsibility for patient safety and clinical outcomes. However, its accuracy may vary depending on situational conditions [Gerdtz and Bucknall 2001], being influenced by factors such as professional fatigue, stress levels, and cognitive biases. Moreover, the healthcare provider's ability to effectively interview and accurately assess a patient's clinical status can further affect the triage outcome. Consequently, the triage process is shaped by both individual traits and competencies of the professional, as well as by contextual factors within the work environment [Fekonja et al. 2023].

The Manchester Triage System (MTS) classifies patients into five color-coded priority levels based on clinical urgency: Red for life-threatening conditions; Orange for high-risk cases; Yellow for stable patients with significant symptoms; Green for minor conditions; and Blue for non-urgent cases. Each category specifies a maximum recommended waiting time for medical care.

Incorrect execution of the triage process can lead to a range of issues, potentially resulting in the deterioration of patients' health conditions. Instances of undertriage and overtriage are undesirable outcomes of triage [Fekonja et al. 2023] and should be minimized. During triage, a hierarchical value or color code is assigned to each patient to indicate the urgency of care. When this level is lower than required, the patient may fail to receive timely treatment, constituting undertriage. Conversely, assigning a higher urgency level than necessary results in overtriage, where a patient receives unwarranted priority at the expense of other patients and available resources. Emphasis should be placed on reducing undertriage cases, as these pose greater risk of clinical deterioration of the patient's condition [Franc et al. 2024b].

3. Related Works

Although not yet extensive — particularly in Brazil — the literature on the application of Large Language Models (LLMs) in patient triage is nonetheless significant. Various approaches have been employed to address and enhance different stages of the triage process. Most studies aim to reduce waiting times and improve classification accuracy, while also suggesting that the integration of LLMs can optimize the allocation of human and financial resources and mitigate the negative impact of human bias in triage.

[Franc et al. 2024a] employed the base version of GPT-3.5-turbo in a Repeatability and Reproducibility (R&R) study aimed at evaluating the reliability of the model in patient triage tasks. The authors analyzed 61 prevalidated test cases, each associated with six distinct prompts, which were executed 30 times via the model's API. The results showed accuracy scores ranging from 42.7% to 50.1%, suggesting that GPT-3.5-turbo is not suitable for clinical triage applications.

Similarly, [Franc et al. 2024b] assessed GPT-4 using 391 simulated patient scenarios under the START (*Simple Triage And Rapid Treatment*) protocol. For each scenario, nine prompt designs were created and each was executed ten times through the API. The overall accuracy achieved was 63.9%, with values ranging from 46.7% to 71.8% (SD = 0.7%). The findings indicate that GPT-4, despite its improvements, still demonstrates substantial variability and limited precision.

[Sarbay et al. 2023] evaluated the performance of ChatGPT-4 Turbo in assigning triage levels to 50 simulated patient scenarios, developed using input from emergency care specialists and technical reference materials. Without fine-tuning or advanced prompt engineering, the model was instructed to classify the cases according to the Emergency Severity Index (ESI) protocol, and its outputs were compared with expert-defined reference standards. Overall, ChatGPT demonstrated superior performance in identifying the most critical cases (ESI Level 1), achieving a sensitivity of 88.9% and an F1-score of 0.842, highlighting the potential of large language models as clinical decision-support tools to enhance triage accuracy and healthcare efficiency.

[Gaber et al. 2024] employed language models from the Claude family in clinical

tasks, enhancing one of them using the Retrieval-Augmented Generation (RAG) technique. The models were evaluated on their ability to classify patients according to triage levels, suggest medical specialties, and generate diagnoses, through prompts that represented different scenarios: either patients at home or healthcare professionals. The RAG-enhanced model retrieved abstracts of relevant materials from the PubMed database to enrich its outputs. Model performance was assessed by comparing predictions to expert-defined reference levels and, although none of the models achieved high accuracy in severe cases, none misclassified critical cases as minor ones or vice versa. The models reached up to 65.8% exact accuracy and 82.8% accuracy within one higher level, demonstrating the potential of such tools for clinical decision support.

Other researches also refer to this subject. [Lee et al. 2024] developed a BERT-based system to extract symptoms and medical histories from nurse–patient triage conversations; [Masanneck et al. 2024] evaluated LLMs as second-opinion tools to medical residents; [Patel et al. 2024] examined the trade-off between accuracy and computational cost of Bio-Clinical-BERT versus traditional NLP methods; and [Fraser et al. 2023] compared GPT-4 and GPT-3.5 to Symptom Checker applications, finding greater agreement between GPT-4 and experts.

4. Methodology

The paper's methodology is structured into sequential phases that have been designed to achieve the research goals. The phases are described bellow and also depicted on the Figure 1.

- Phase 1: Literature Review. This phase encompasses the bibliographic review on the topic of LLMs applied to patient triage. The central objective is to verify the existence of similar works and analyze the approaches already proposed and results already obtained.
- Phase 2: Test Cases Repository Creation, LLM Workflow Implementation, and Prompt Engineering. This phase consists of three essential activities carried out in parallel, which provide the input for test execution:
  - Test Cases Repository Creation: This focuses on creating a bank of test cases that can be used for model evaluation. That is, assembling a repository of valid patient scenarios.
  - LLM Workflow Implementation: The large language models that will be part of the experiment will be selected. Furthermore, the workflow for these models to perform patient triage must be implemented and validated.
  - Prompt Engineering: The main objective is to investigate, create, and test different prompt engineering approaches to feed the LLMs. Varying writing, objective, and emphasis, the goal is to create distinct Portuguese prompts and analyze them for use in test execution.
- Phase 3: Tests Execution. In this step, the models will be exposed to the test cases using the defined prompts. A series of iterations with the models will be performed, documenting the responses for subsequent evaluation.
- Phase 4: Evaluation and Discussion. The final phase of the project where all documented responses will be reviewed and will serve as material for the final evaluation of the models' efficiency.

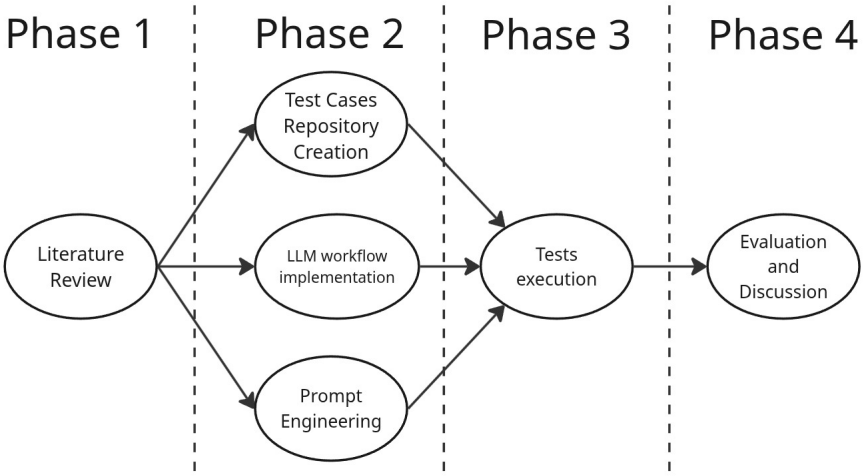

Figure 1. Depiction of the adopted methodology

5. Results and discussion

All the used prompts follow the same structure: first, the task description, after that, the patient information followed by the answering rules. After the prompt engineering phase, there were defined three different prompts, only varying their task description text, keeping the patient information structure and answering rules identical for all. All of the prompts share the same goal: classifying patients according to the MTS. All three prompts keep clarity and precision and are formulated as direct instructions. Besides that, two prompts employ specific prompt engineering techniques: *Role-Prompting* (prompts 1 and 3 use this technique) and *Chain-of-thought prompting* (prompt 3 uses this technique) [Chen et al. 2025]. The final three task descriptions for the prompts, translated from the Portuguese original text, were:

- **Prompt 1:** "You are an intelligent clinical decision support system. Your task is to classify patients according to the Manchester Triage System Protocol (MTS), based on the information provided regarding the main complaint, symptoms, clinical signs, vital signs, and possibly a brief anamnesis. MTS organizes patients into priority categories based on standardized clinical criteria, and each category sets a recommended maximum waiting time for care. Using the patient data below, identify the most appropriate triage category (e.g., Red, Orange, Yellow, Green, or Blue) and justify your choice based on MTS criteria. Here is the patient information:"
- **Prompt 2:** "Classify patients according to the Manchester Triage System Protocol (MTS) based on the information provided. Indicate the most appropriate triage category (Red, Orange, Yellow, Green, or Blue) and justify your choice."
- **Prompt 3:** "You are an intelligent clinical decision-support system with expertise in the Manchester Triage System (MTS). Your task is to classify a patient, following a three-step logical reasoning process (Chain-of-Thought), before providing the final classification. Instructions: 1.Analysis of the Main Complaint and Symptoms: Identify the main complaint, symptoms, and any critical vital signs provided, correlating them with the MTS risk discriminators. State which initial risk category you are considering at this point (e.g., Immediate Risk, Very

Urgent, Urgent). 2. Determination of the Flowchart and Key Discriminator: State which MTS flowchart is most appropriate for the complaint (e.g., Pain, Respiratory Problems, Trauma). Then identify the Key Discriminator (the question or criterion that defines the category) that leads to your choice. 3. Justified Conclusion (MTS): Based on steps 1 and 2, define the priority category (Red, Orange, Yellow, Green, or Blue) and the recommended maximum time to treatment. Final Answer: Present the three reasoning steps (CoT) and, finally, the final category. The patient information follows:"

The final test case repository contains 39 simulated patient histories. This dataset was fabricated by the GPT-5 model, varying age, sex, symptoms and vital signs as well as assigning a correct classification. For each of the test cases combined with each of the prompts, there were made three attempts on each model. Totaling 1404 zero-shot queries.

The models were run locally to avoid the costs associated with using cloud machines, at the expense of longer processing times. Efforts were made to select freely available and relevant models with comparable parameter counts. Ideally, all models would have the same number of parameters; however, this was not feasible. The standard deviation in parameter count (in billions) across the models is 4.3, considering gpt-oss (20B), llama3.1 (8B), gemma3 (12B), and deepseek-r1 (14B).

As part of the main task assigned to the models, they should give an explanation along with the MTS classification. This reasoning serves as material to justify the given color and can be analyzed for further understanding of the models' responses.

Most metrics are divided into two categories: general and mode. The general metrics evaluate each individual model response, whereas the mode metrics evaluate the majority (mode) response across the three validation answers. The mean accuracy across all models and prompts — defined as the proportion of cases in which an LLM produced the exact correct answer — is 33.55% when considering mode responses and 32.91% when considering general responses. Table 1 displays the summary of the results for all models and prompts. And Figure 2 visually displays the comparison between the models' performances.

Model		deepseek-r1-14b	gemma3-12b	gpt-oss-20b	llama3.1-8b	All models (Mean)
Prompt 1	(mode)	35.90%	17.95%	41.03%	30.77%	31.41%
Accuracy	(general)	29.91%	19.66%	43.59%	29.06%	30.56%
Prompt 2	(mode)	43.59%	33.33%	51.28%	23.08%	37.82%
Accuracy	(general)	39.32%	33.33%	46.15%	25.64%	36.11%
Prompt 3	(mode)	28.21%	30.77%	46.15%	20.51%	31.41%
Accuracy	(general)	29.06%	31.62%	46.15%	21.37%	32.05%
All prompts	(mode)	35.90%	27.35%	46.15%	24.79%	33.55%
(Mean)	(general)	32.76%	28.21%	45.30%	25.36%	32.91%

Table 1. Accuracy of all models and prompts

Notably, gpt-oss:20b achieved the highest accuracy, when queried with prompt 2 the model got 51,28% accuracy on the mode metric and 46,15% on the general metric. Overall, the results suggest that model accuracy increases slightly when using the mode of multiple responses produced by the same model (a 0.64-percentage-point increase, 1.95%

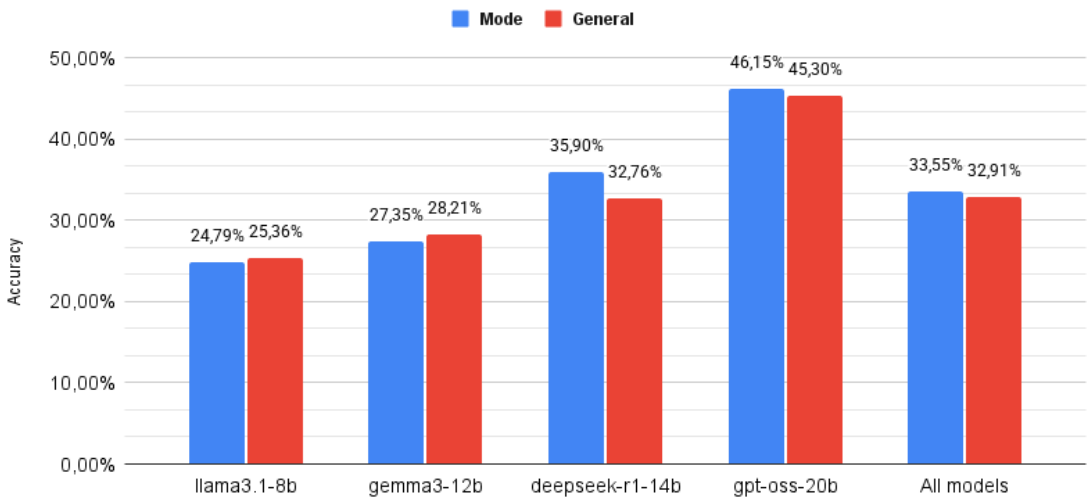

Figure 2. Chart of model accuracies

improvement). Although the improvement is minimal, it indicates that aggregating several answers can reduce randomness ('noise') and better capture the model's central tendency. Additional testing with a larger number of validation steps is needed to confirm whether this internal answer-voting mechanism consistently provides benefits.

By applying the validation mechanism, it is possible to assess the intra-model consistency — or the reproducibility — of the generated answers. To quantify this consistency, we compute an agreement rate, which measures how frequently the three validation responses for a given prompt and patient case converge to the same classification. For each instance, the agreement rate is calculated as the proportion of responses that correspond to the most frequent label among the three validations (i.e., the mode). An agreement rate of 1 indicates complete reproducibility across the three runs, while lower values reflect increasing levels of variability. Considering all four models and all prompts, the mean agreement rate was 82.19%, indicating that the models generally produced highly consistent classifications across repeated runs of the same prompt. Table 2 presents the agreement rates for each model and prompt. Table 2 also shows the repeatability rate for each model, that is, the rate which classifications for the same patient case remain stable despite variations in the prompt.

Model	Prompt 1	Prompt 2	Prompt 3	Repeatability Rate
deepseek-r1-14b	70.94%	64.96%	75.21%	50.43%
gemma3-12b	96.58%	100.00%	96.58%	81.20%
gpt-oss-20b	78.63%	82.91%	74.36%	60.68%
llama3.1-8b	77.78%	76.07%	92.31%	58.12%
All Models (Mean)	80.98%	80.89%	84.62%	62,61%

Table 2. Agreement (reproducibility) rate within the three validation responses for each model and prompt

Notably, *gemma3-12b* showed the highest repeatability on its responses, achieving a rate of 81.20% when considering the mode answer for each of the three prompts. This

shows that although the prompt may vary, the model is able keep the same response for a given patient scenario. At the opposite end is *deepseek-r1-14b*, with only 50.43% of repeatability rate, pointing it has higher sensitivity to prompt phrasing.

An important aspect of model evaluation is understanding how the models fail. As previously discussed, both undertriage and overtriage are undesirable triage outcomes that should be minimized, especially undertriage, which may lead to a deterioration in the patient's clinical condition[Fekonja et al. 2023]. Table 3 presents the undertriage and overtriage rates, calculated based on the misclassifications produced by each model.

Model	deepseek-r1-14b	gemma3-12b	gpt-oss-20b	llama3.1-8b	Mean
Overtriage	77.93%	100.00%	81.68%	94.48%	88.52%
Undertriage	22.07%	0.00%	18.32%	5.52%	11.47%

Table 3. Overtriage and undertriage rates per model, considering mode responses

All models consistently tended to over-classify patient severity, assigning higher urgency levels than appropriate and thus indicating a lower risk to patient safety. When considering all misclassifications across all models, the overtriage rate reached 88.52%, whereas only 11.47% of errors were cases of undertriage. Notably, the *gemma3-12b* model exhibited no undertriage cases, meaning that whenever it produced an incorrect classification, it always assigned at least one level higher than the correct triage category.

Prompt	Accuracy		Undertriage		Overtriage		Agreement Rate
	(mode)	(general)	(mode)	(general)	(mode)	(general)	
prompt 1	31.41%	30.56%	6.41%	10.21%	93.59%	89.79%	63.46%
prompt 2	37.82%	36.11%	19.00%	16.02%	81.00%	83.98%	64.74%
prompt 3	31.41%	32.05%	9.02%	8.67%	90.98%	91.33%	70.30%

Table 4. Metrics for each prompt

Finally, the three prompts used in this study had noticeably different impacts on the models' outcomes. Table 4 summarizes the main statistics for each prompt. In terms of accuracy, prompt 2, despite being the simplest formulation, achieved the highest scores (37.82% in mode accuracy and 36.11% in general accuracy), indicating that more concise or direct instructions may help models with this profile converge toward more accurate answers. However, prompt 2 also exhibited the highest undertriage rates among the three (19.00% mode; 16.02% general), suggesting that this simplicity may come at the cost of increased clinical risk, as the models were more likely to underestimate patient severity. Figure 3 visually displays the mode metrics for each prompt, simplifying the comparison.

Prompt 3 presented a balanced behavior relative to the others: while its accuracy was lower (31.41% mode; 32.05% general), it achieved the highest agreement rate among all prompts (70.30%). This indicates that its structure may promote greater intra-model consistency, even if the answers are not necessarily correct.

6. Conclusion

This research sought to benchmark four different small-scale freely available LLMs. The results point that these models are not fitting for the process on their own, clarifying some

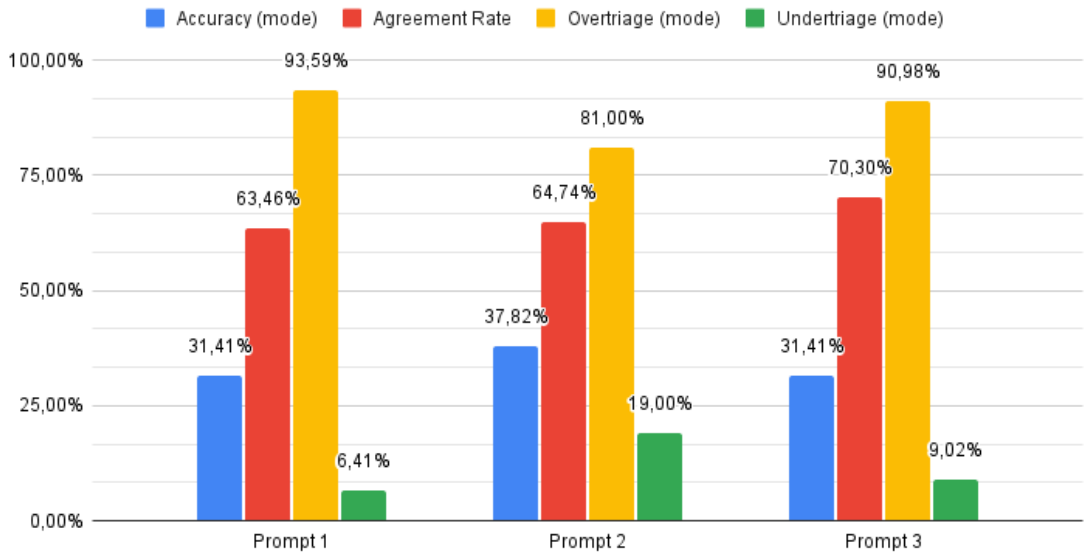

Figure 3. Chart of prompt metrics, considering the mode

of their strengths and fragilities.

The largest model employed in our evaluation was gpt-oss:20b, with approximately 20 billion parameters, it was also the one with better performance metrics. Although parameter count alone does not determine an LLM's effectiveness, it provides a useful point of comparison with prior studies. The models analyzed in related work — such as GPT-3.5-turbo and GPT-4 in [Franc et al. 2024b, Franc et al. 2024a], ChatGPT-4 Turbo in [Sarbay et al. 2023], and the Claude-based models examined by [Gaber et al. 2024] — do not disclose their exact sizes, but are generally understood to operate at scales substantially larger than 20B. Despite this presumed increase in computational capacity, those studies consistently report variability, limited reproducibility, and accuracy constraints in clinical triage tasks. This comparison indicates that, in this context, model accuracy does not appear to scale linearly with model size.

Despite their reduced parameter counts, smaller open models can still demonstrate notable capability in clinical triage tasks, particularly when evaluated under solid prompting and validation strategies. An important advantage of these models is their minimal hardware requirements, which enable fully local execution without dependence on external servers or proprietary APIs and reduces operational costs. And because the models used in this study are freely available, they also allow other researchers to replicate, extend, and compare results.

Looking ahead, several opportunities for future work can help to improve the understanding of LLM behavior in the clinical triage contexts. One possible direction is an analysis of the explanations produced by the models — evaluating not only the final classification, but also the reasoning quality, coherence, and medical plausibility of the explanations across prompts and models.

Future studies could also expand the validation protocol, increasing the number

of validation queries to better assess intra-model consistency and sensitivity to prompt variation. Alongside this, extending the range of evaluated models, including both larger systems and newly released small-scale open models.

Finally, applying the same evaluation framework to different datasets, including real-world triage records or expert-reviewed case sets. That would allow researchers to test the reliability of the LLM tools across different scenarios. These efforts would contribute to a more comprehensive understanding of LLM performance in clinical triage and guide the development of safer and more reliable decision-support tools.

References

Alumran, A., Alkhaldi, O., Aldroorah, Z., Alsayegh, Z., Alsafwani, F., and and, N. A. (2020). Utilization of an electronic triage system by emergency department nurses. *Journal of Multidisciplinary Healthcare*, 13:339–344.

Andersson, A.-K., Omberg, M., and Svedlund, M. (2006). Triage in the emergency department – a qualitative study of the factors which nurses consider when making decisions. *Nursing in Critical Care*, 11(3):136–145.

Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6).

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, 55:78–87.

Duarte, D. and Ståhl, N. (2019). Machine learning: A concise overview. In Said, A. and Torra, V., editors, *Data Science in Practice*, pages 27–58. Springer International Publishing.

Fekonja, Z., Kmetec, S., Fekonja, U., Mlinar Reljić, N., Pajnkihar, M., and Strnad, M. (2023). Factors contributing to patient safety during triage process in the emergency department: A systematic review. *Journal of Clinical Nursing*, 32(17-18):5461–5477.

Franc, J. M., Cheng, L., Hart, A., Hata, R., and Hertelendy, A. (2024a). Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. *Canadian Journal of Emergency Medicine*, 26(1):40–46.

Franc, J. M., Hertelendy, A. J., Cheng, L., Hata, R., and Verde, M. (2024b). Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage of Simulated Patients Using the Simple Triage and Rapid Treatment (START) Protocol: Gage Repeatability and Reproducibility Study. *Journal of Medical Internet Research*, 26:e55648. Publisher: JMIR Publications Toronto, Canada.

Fraser, H., Crossland, D., Bacher, I., Ranney, M., Madsen, T., and Hilliard, R. (2023). Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and Physicians for Patients in an Emergency Department: Clinical Data Analysis Study. *JMIR mHealth and uHealth*, 11:e49995.

Fry, M. and Burr, G. (2002). Review of the triage literature: Past, present, future? *Australian Emergency Nursing Journal*, 5(2):33–38.

Gaber, F., Shaik, M., Franke, V., and Akalin, A. (2024). Evaluating large language model workflows in clinical decision support: referral, triage, and diagnosis. *medRxiv*, pages 2024–09. Publisher: Cold Spring Harbor Laboratory Press.

Gerdtz, M. F. and Bucknall, T. K. (2001). Triage nurses' clinical decision making. an observational study of urgency assessment. *Journal of Advanced Nursing*, 35(4):550–561.

Lansiaux, E., Baron, M.-A., and Vromant, A. (2024). Navigating the landscape of medical triage: Unveiling the potential and challenges of large language models and beyond. *The American Journal of Emergency Medicine*, 78:224.

Lee, S., Lee, J., Park, J., Park, J., Kim, D., Lee, J., and Oh, J. (2024). Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. *The American Journal of Emergency Medicine*, 77:29–38. Publisher: Elsevier.

Lorena, A., Faceli, K., Almeida, T., de Carvalho, A., and Gama, J. (2021). *Inteligência Artificial: uma abordagem de Aprendizado de Máquina (2a edição)*.

Masanneck, L., Schmidt, L., Seifert, A., Kölsche, T., Huntemann, N., Jansen, R., Mehsin, M., Bernhard, M., Meuth, S. G., and Böhm, L. (2024). Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study. *Journal of Medical Internet Research*, 26:e53297. Publisher: JMIR Publications Toronto, Canada.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill.

Patel, D., Timsina, P., Gorenstein, L., Glicksberg, B. S., Raut, G., Cheetirala, S. N., Santana, F., Tamegue, J., Kia, A., and Zimlichman, E. (2024). Traditional Machine Learning, Deep Learning, and BERT (Large Language Model) Approaches for Predicting Hospitalizations From Nurse Triage Notes: Comparative Evaluation of Resource Management. *JMIR AI*, 3(1):e52190. Publisher: JMIR Publications Inc., Toronto, Canada.

Sarbay, , Berikol, G. B., and Özturan, U. (2023). Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): A preliminary, scenario-based cross-sectional study. *Turkish Journal of Emergency Medicine*, 23(3):156–161.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Wireklint, S. C., Elmqvist, C., Parenti, N., and Göransson, K. E. (2018). A descriptive study of registered nurses' application of the triage scale retts©; a swedish reliability study. *International Emergency Nursing*, 38:21–28.

Yazaki, M., Maki, S., Furuya, T., Inoue, K., Nagai, K., Nagashima, Y., Maruyama, J., Toki, Y., Kitagawa, K., Iwata, S., Kitamura, T., Gushiken, S., Noguchi, Y., Inoue, M., Shiga, Y., Inage, K., Orita, S., Nakada, T., and Ohtori, S. (2024). Emergency Patient Triage Improvement through a Retrieval-Augmented Generation Enhanced Large-Scale Language Model. *Prehospital Emergency Care*, pages 1–7.