

Evaluation of Language Models in the Medical Context Under Resource-Constrained Settings

Andrea Posada, Daniel Rueckert, Felix Meissen, and Philip Müller

June 25, 2024

Abstract

Since the emergence of the Transformer architecture, language model development has increased, driven by their promising potential. However, releasing these models into production requires properly understanding their behavior, particularly in sensitive domains such as medicine. Despite this need, the medical literature still lacks technical assessments of pre-trained language models, which are especially valuable in resource-constrained settings in terms of computational power or limited budget. To address this gap, we provide a comprehensive survey of language models in the medical domain. In addition, we selected a subset of these models for thorough evaluation, focusing on classification and text generation tasks. Our subset encompasses 53 models, ranging from 110 million to 13 billion parameters, spanning the three families of Transformer-based models and from diverse knowledge domains. This study employs a series of approaches for text classification together with zero-shot prompting instead of model training or fine-tuning, which closely resembles the limited resource setting in which many users of language models find themselves. Encouragingly, our findings reveal remarkable performance across various tasks and datasets, underscoring the latent potential of certain models to contain medical knowledge, even without domain specialization. Consequently, our study advocates for further exploration of model applications in medical contexts, particularly in resource-constrained settings. The code is available on <https://github.com/anpoc/Language-models-in-medicine>.

1 Introduction

Natural Language Processing (NLP) holds immense promise in the medical domain. Recently, the medical community has shown substantial interest in leveraging state-of-the-art language models to address various medical challenges [1, 2]. In particular, generative large language models (LLMs) have showcased emergent abilities beyond their original training objectives, such as text summarization and question-answering [3]. These newfound abilities have enabled LLMs to perform tasks of significant clinical importance, including passing medical examinations, summarizing clinical and radiological reports, as well as medical dialogues, extracting drug names from medical notes, responding to patient inquiries, and writing medical histories and physical assessments [2, 4].

The versatility of language models can be attributed to a convergence of factors [2, 4, 5]. The first factor is their ability to learn valuable patterns within large amounts of unlabeled data via self-supervision. The second factor revolves around the Transformer architecture [6] and its suitability for efficient parallel processing on modern computing hardware. Lastly, the third factor encompasses the crucial process of fine-tuning language models to align their responses with human expectations through instruction tuning.

Integration of language models in medical settings is becoming a reality as partnerships between developers and healthcare systems continue to grow [7]. The potential benefits are significant, as they can derive broadly applicable representations from extensive medical corpora at scale. Recent research indicates that LLMs with more than 100 billion parameters can encapsulate clinical knowledge [8]. Nevertheless, it is essential to recognize that our understanding of the behavior of LLMs and language models in general still needs to be completed [4]. Moreover, implementing these models also carries risks, such as the generation of inaccurate results, a phenomenon known as hallucinations, and the potential amplification of existing biases [1, 4].

Hence, the deployment of language models in sensitive fields, such as healthcare, should be approached with the utmost care [5]. To further understand and assess language models' performance in clinical

scenarios, we conducted comprehensive evaluations using models of varying sizes, ranging from 110M to 13B parameters, and encompassing different knowledge domains. We focus on text classification, using three different approaches and conditional text generation tasks. The approaches employed for text classification and the application of zero-shot prompting alongside allow performance improvement without the need for training or fine-tuning. For its part, the selection of the model size range considers resource constraints, in terms of limited budgets or computational power, encountered not only by many research groups but also by other actors and institutions, such as hospitals, where handling models with more than 110 billion parameters is not feasible.

Moreover, the use of small language models aligns with the crucial imperative of environmental sustainability, as larger models often require significant energy and computational resources. It also opens up possibilities for organizations to develop applications that can run directly on commodity hardware and small devices rather than relying on cloud-based services. In fact, the prospect of deploying these language models on smartphones and other mobile devices operating at the edge without cloud dependence is promising [9]. Therefore, the use of language models in resource-constrained settings not only addresses practical challenges but also has great potential in edge and local computing.

The paper is structured as follows. Section 1 introduces the study and provides an overview of its objectives. Section 2 elaborates on the preliminaries, establishing foundational concepts relevant to the investigation. Related work in the existing literature is explored in Section 3. Section 4 details the methodology adopted, and Section 5 provides the most significant results and analysis. Finally, Section 6 contains the conclusions drawn from the study, summarizing key insights and implications for future research directions.

2 Preliminaries

The evolution of natural language processing can be condensed into four major groups of models: (1) statistical models, (2) neural language models, (3) pre-trained language models, and (4) large language models [10]. Each of these groups represents a paradigm shift in natural language modeling and has contributed significantly to the conception of language models as we know them today.

The first transition, from statistical to neural language models, marked a shift from predicting words based on minimal local context to probabilistically assessing word sequences using neural networks. Additionally, this transition introduced the concept of representing words as low-dimensional continuous embeddings based on their contextual usage (distributional semantics). The second transition, from neural language models to pre-trained language models, represented a move from task-specific model development towards pre-training and fine-tuning methodologies. Lastly, the third transition to large language models constituted a transformation from discriminative AI to generative AI, from model-centric to data-centric approaches, and from fine-tuning to prompt-tuning [10–12]. These advances have paved the way for more sophisticated language models with broader applications and improved capabilities.

2.1 From task-specific model development towards pre-training and fine-tuning: pre-trained language models

The emergence of pre-trained language models gave relevance to pre-training and fine-tuning methodologies, which was a paradigm shift, as mentioned above. This shift has driven research toward designing more efficient architectures and refinement of pre-training strategies, demonstrating significant advantages in enriching language understanding and improving model performance in various tasks [10, 11]. One of the first attempts at pre-trained language models is ELMo [13]. ELMo was developed to capture context-aware word representations by pre-training a bidirectional Long Short-Term Memory (biLSTM) network and fine-tuning it for subsequent downstream tasks. However, the introduction of the Transformer architecture represented a qualitative leap in NLP, offering highly parallelizable structures and self-attention mechanisms.

The Transformer [6] follows the autoencoder archetype, giving rise to three families of models: (1) BERT-family or encoder-only models, (2) GPT-family or decoder-only models, and (3) text-to-text or encoder-decoder models. Their graphical representations are shown in Fig. 1.

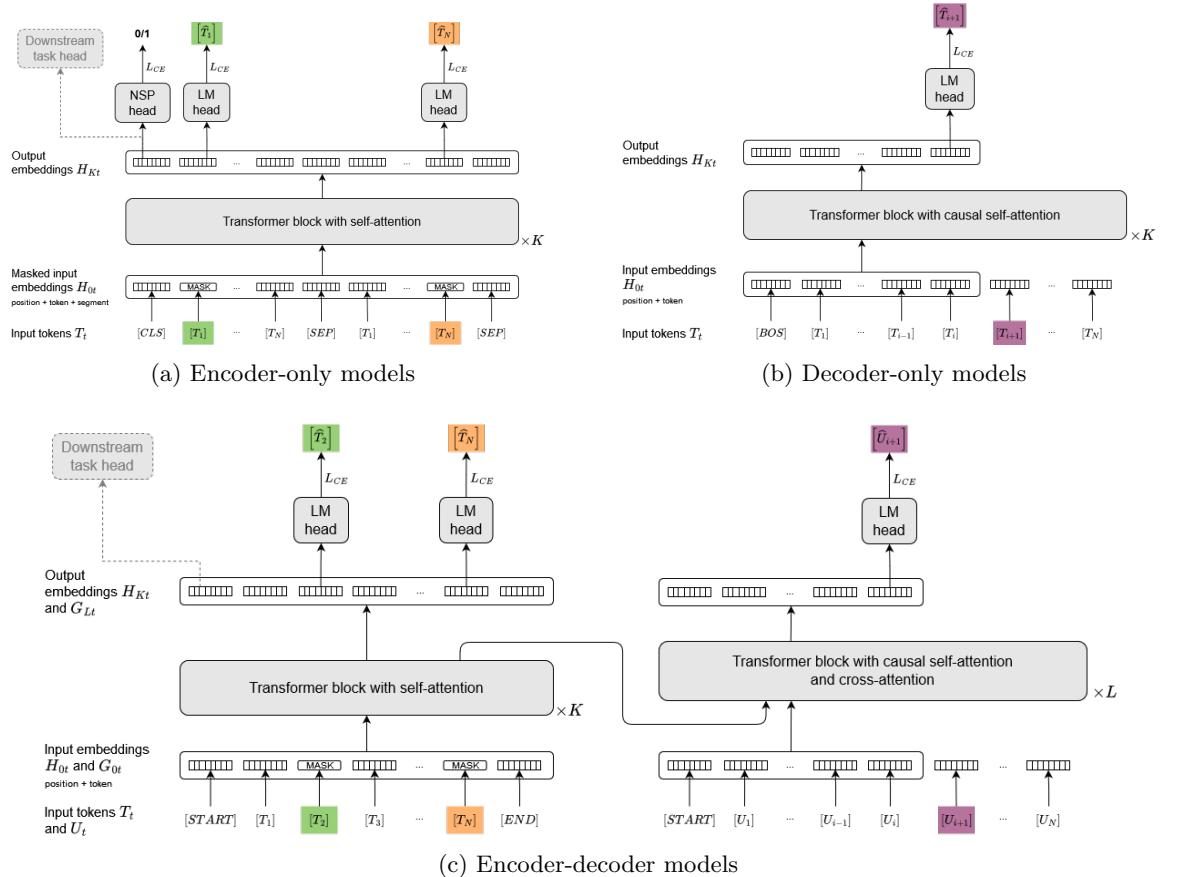


Figure 1: Graphical representation of the three families of Transformer-based models: encoder-only, decoder-only, and encoder-decoder models. The correspondence between outputs and targets is signaled by colors. In the encoder-only models, the input is tokenized and some of the tokens are masked. They are then fed into Transformer blocks with self-attention to obtain contextualized output embeddings, which are further processed by next sentence prediction (NSP) and language model (LM) heads or used by downstream task-specific heads. The NSP head may or may not be necessary, depending on the training objective. In the decoder-only models, the input is tokenized and fed to Transformer blocks with causal self-attention. The causal self-attention ensures that the information flows unidirectionally, namely from left to right. The encoder-decoder models are used for text-to-text tasks. The input text is processed by the encoder, which is similar to that of encoder-only models, excluding the NSP head, and which flows information to the decoder via the cross-attention mechanism. This information is used together with the target output so that the decoder learns to produce the latter generatively.

2.1.1 Encoder-only models

Encoder-only models, exemplified by BERT [14], are based on masked language modeling (MLM), where parts of the input are masked to encourage the model to reconstruct the original sequence, leveraging contextual information bidirectionally. These models can be stated as $v_{1:n} \rightarrow \phi(v_{1:n})$. In particular, their contextual embeddings have been proven highly effective as general-purpose semantic features, significantly boosting performance in discriminative NLP tasks.

2.1.2 Decoder-only models

Decoder-only models focus on autoregressive language modeling, that is, predicting the next token in a sequence based on previous tokens. These models produce contextual embeddings and a distribution over the next tokens v_{i+1} , which can be stated as $v_{1:i} \rightarrow \phi(v_{1:i}), \mathbb{P}(v_{i+1}|v_{1:i})$. However, the contextual embeddings they generate depend solely on the left context. Currently, most research efforts are directed toward decoder-only models due to their exceptional performance in conditional generation tasks and their demonstrated emergent capabilities.

2.1.3 Encoder-decoder models

Text-to-text models, or encoder-decoder models, are trained to learn the correspondence between a pair of texts and can be stated as $v_{1:n} \rightarrow \phi(v_{1:n}), \mathbb{P}(w_{1:m}|\phi(v_{1:n}))$. These models combine bidirectional contextual embeddings with the capability to generate output sequences, making them versatile in various text-to-text tasks without requiring additional heads for fine-tuning. Moreover, by having a wide spectrum of language tasks that can be translated into text-to-text representation, these models can potentially be used for a wide range of applications.

2.2 From discriminative AI to generative AI: large language models

More recently, it has been observed that increasing the size of language models, whether in terms of model size, dataset size, or computational resources, often results in improved model capabilities in various tasks [15–21], including those that require specialized scientific knowledge and reasoning [22]. Known as the scaling law [23], this phenomenon posits a power-law relationship between model performance, measured by the loss, and the factors mentioned above.

Several studies have delved into this relationship by progressively training larger language models, up to hundreds of billion parameters, such as GPT-3 [17], PaLM [19], Galactica [24], LLaMA models [25, 26], Claude [27], Gemini 1.5 [28], and Mistral [29]. Although scaling primarily involves increasing model size while retaining similar architectures and pre-training tasks as smaller models [10], large language models (LLMs) have exhibited unexpected emergent capabilities [3, 30], enabling them to address a wide range of complex tasks. GPT-3 and ChatGPT serve as precursors to these large language models [10, 11].

A particularly intriguing aspect of LLMs is their few-shot and zero-shot capabilities [31], enabling them to adapt to various tasks without requiring gradient-based parameter updates [8, 12, 17, 21, 32–35]. Through these and other prompting techniques, emerging capabilities have been observed, enabling LLMs to exhibit apparent reasoning abilities [8]. Nonetheless, prompting techniques are not exclusive to LLMs but are also applicable, for example, to pre-trained language models. Similarly, numerous studies highlight the competence of LLMs as implicit knowledge bases [8, 22, 24, 36]. As for prompt engineering, its role is highly significant due to its benefits, which have proven to be more effective in large or instruction-tuned models. [8, 21, 32]. In the medical domain, for example, Chain of Thought (CoT) has been used to address explainability [37] and in-context learning to mitigate the need for costly medical annotations [11].

Among LLMs, GPT-4 stands out for its exceptional performance, often matching or surpassing human performance in various tasks [38–40], including in scientific domains such as biology, chemistry, and medicine [41]. In the general domain, extensive evaluations of this model have been conducted [22, 42–44], even exploring the path toward Artificial General Intelligence (AGI) [39]. In the medical domain, recent studies have begun to investigate its utility in various applications, ranging from medical chatbots [45] to medical competency exams [40] and its applications in radiology [38, 46–51], among others [52–55].

Nevertheless, there are other relevant models in the medical field, and unlike GPT-4, most of them are domain-specific. Generally, these models are derived from general-purpose LLMs, with exceptions like

GatorTron [56], Galactica, and GatorTronGPT [57]. The most common training technique is instruction fine-tuning, as seen in Galactica, MedAlpaca [58], PMC-LLaMA [59], Med-PaLM 2 [60], GatorTronGPT, and ClinicalGPT [61]. However, although less frequently, reinforcement learning from human feedback (RLHF) and reinforcement learning from AI feedback (RLAIF) techniques are also adopted, HuatuoGPT [62] being an example. Recent research studies also indicate a multimodal trend that supports various types of healthcare data, including electronic health records (EHR), medical images, and medical sequence signals. Examples of these developments include LLaVAMed [63], MedAGI [64], OphGLM [65], Visual Med-Alpaca [66], and MedFlamingo [67], CheXzero [68].

Despite their advantages, LLMs also have limitations. One of these are computational challenges due to their high computational resource requirements [8, 10]. Conducting repetitive and ablation studies to investigate the impact of training strategies for LLMs is costly and time-consuming. Therefore, in certain cases, LLMs may not represent a cost-effective solution. Since smaller language models have also shown the effectiveness of using curated scientific and biomedical corpora for both discriminative and generative language modeling [38], they represent an alternative to LLMs. Such models include BioBERT [69], SciBERT [70], PubMedBERT [71], BioMegatron [72], ScholarBERT [73], BioGPT [74], and ClinicalBERT [75]. However, the current literature still lacks comparative studies on these smaller language models.

3 Related Work

Comparative studies investigating language models are crucial in advancing our understanding of them, shedding light on their functionalities, and pinpointing their constraints. Despite previous research, a notable gap persists in the literature. This gap is particularly significant in fields that require heightened sensitivity, such as medicine, where a thorough understanding of models is imperative. Although existing research in this domain focuses primarily on specific tasks, datasets, or models [5, 8, 12, 32, 38], there are exceptions. Some of these studies are detailed below.

In [11], the authors conduct a descriptive study delineating the capabilities and limitations of language models within the healthcare domain, encompassing pre-trained and large language models. They explain the development process, provide insight into training data, methodologies, and optimization strategies, and investigate concerns about integrating LLMs in healthcare. These concerns focus on fairness, accountability, transparency, and ethics. Likewise, in the context of integration concerns, the faithfulness problem within existing research on artificial intelligence in healthcare and medicine is explored in [76]. This analysis involves identifying the causes of nonfaithful results, possible evaluation metrics for faithfulness, and potential mitigation strategies for this issue.

Regarding studies closely aligned with the present work, notable contributions include those by Soni *et al.* [77], Lehman *et al.* [78], and Zhou *et al.* [79]. In [77], the authors assess the cost-effectiveness of pre-training and fine-tuning in Transformer language models, specifically in BERT, BioBERT [69], Clinical BERT [75], and XLNet [80]. These models underwent pre-training and fine-tuning on general, biomedical, and clinical-domain data combinations for medical question-answering (QA) tasks. Their findings revealed that BERT-based models exhibit superior performance when fine-tuned on multiple datasets, incorporating both general and clinical domain data. This observation suggests a gap in well-generalizable medical QA datasets. Furthermore, their results underscored the positive impact of initial fine-tuning on general-domain datasets, such as SQuAD, before fine-tuning on clinical datasets, highlighting the quality of the large, manually constructed SQuAD dataset. Overall, it was observed that initial fine-tuning improves performance in most cases. It is important to note that this study focused solely on question-answering and did not evaluate prompting techniques.

On its side, in [78], the authors' research delves into whether LLMs, primarily trained on general web text, are suitable for highly specialized, safety-critical domains such as medicine or if domain-specialized models offer better performance. The authors evaluated 12 language models, ranging from 220 million to 175 billion parameters, measuring their performance in three clinical tasks. As part of their experiments, T5 models were trained from scratch using MIMIC-III and MIMIC-IV clinical notes to investigate the efficiency of clinical tokens. Their findings suggest that relatively small specialized clinical models significantly outperform all in-context learning approaches, even when fine-tuned on limited annotated data. However, this study does not assess the ability of models to handle long text and considered neither

decoder-only models nor instruction-tuned ones.

Lastly, [79] provides a comprehensive overview of the development and deployment of LLMs in medicine, including the challenges and opportunities that they face. The main highlight of this study is that it is not only a discursive but also a practical assessment. The authors begin with an introduction to the principles of existing medical LLMs, including their basic model structures, the number of parameters, and the sources and scales of data used for model development. A comparative evaluation of the performance of different LLMs in various medical tasks is then provided, comparing them with state-of-the-art lightweight models. This review is framed by the research questions: 1) What are the practices for developing medical LLMs? 2) How to measure the medical task performance of LLMs in a medical setting? 3) How have medical LLMs been used in actual practice? 4) What are the challenges in using medical LLMs; and 5) How to develop and deploy medical LLMs more effectively?

4 Methodology

The experiments involve two target tasks: text classification and conditional text generation. Specifically, text classification is addressed through three different approaches: (i) contextual embedding similarity, (ii) natural language inference (NLI), and (iii) multiple-choice question-answering (MCQA). These are explained in Section 4.1.2. As detailed below, each task is evaluated using diverse datasets, models, and performance metrics.

The evaluation process comprises 53 models, which exhibit a size spectrum ranging from 110 million to 13 billion parameters. These models belong to three knowledge domains: general, biomedical or scientific, and clinical. Their characterization can be found in Table 1. **OpenLLaMA 13B**, **Flan-T5-XXL**, **T5-V1.1-11B**, and **T0++** are run using float16 precision.

In order to conduct a comprehensive analysis, we recognize the importance of evaluating the chosen models across a broad spectrum of medical specialties, ultimately aiming at a general assessment of medical knowledge. To achieve this, we used the transcriptions dataset. Nevertheless, also recognizing the potential of artificial intelligence in particular medical fields, we narrowed our focus in addition to radiology, one of the most promising fields for AI integration. Consequently, we opt for the MIMIC-CXR dataset along with its labeled counterpart, MS-CXR.

4.1 Text classification

For the text classification task, we evaluate two datasets using three approaches and their corresponding models. It is essential to mention that fine-tuning is not performed; however, we use zero-shot prompt-engineering. In the following, we provide a concise overview of the datasets, followed by an explanation of the approaches employed. We refer the reader to Appendix A for more details on the preprocessing steps and the data characteristics of the datasets.

4.1.1 Datasets

Transcriptions is a multi-label collection of electronic health records covering several medical specialties. The preprocessing procedure for this dataset involves removing null entries, modifying the report’s format, and selecting and renaming labels. After preprocessing, the resultant dataset consists of 2074 samples and 29 labels. Model performance is evaluated using the AUC score.

In particular, since the EHRs are used in their entirety, certain samples contain token vectors that exceed the maximum input length allowed for multiple models. There are even samples whose token vectors exceed 2048 tokens, the maximum input length supported among the models. Consequently, different strategies have been adopted for each approach to effectively address this challenge. The adopted strategies, which require splitting the sentence into parts, will produce different results than processing the entire sample at once due to the attention mechanisms. These approaches are detailed in Section 4.1.2. **MS-CXR** is a multi-class dataset comprising sections of X-ray reports, each accompanied by annotations made by a radiologist [100–102]. The number of classes is eight, well distributed and represented in 718 unique text samples. The preprocessing procedure is more straightforward than for the other datasets

Table 1: The models used in this study are categorized by their type, domain, and size. Each model is presented with its number of parameters and may have one or more superscripts. They are 0: model used for contextual embedding similarity, 1: model used for natural language inference (NLI), 2: model used for multiple-choice questions, 3: model used for text generation, †: instruction-tuned model, ‡: cross-encoder model.

and includes removing samples with missing information and duplicates. In contrast to transcriptions, there is no evidence of any sample exceeding the maximum input length allowed for any of the models.

Model performance is evaluated using a set of metrics in their macro-averaged version to ensure a comprehensive assessment. The metrics are accuracy, F1-score, precision, and recall.

4.1.2 Approaches

Contextual embedding similarity is grounded in the cosine similarity between the contextual embeddings of the text under consideration and its corresponding label. The models employed are exclusively encoder-only, comprising 11 models subjected to evaluation. The final contextual or sentence embedding is determined using three distinct pooling strategies: the CLS-token embedding, average pooling of the token-level embedding, and maximum pooling of the token-level embedding. In addition, prompt-engineering is implemented for the label. The specific details of the proposed prompts can be found in Appendix B.

In cases where the input tokens exceed the maximum input size defined by the model, the samples are split into smaller parts, processing each individually. The embedding of the sample varies depending on the pooling strategy implemented, as described below. In particular, the models used in this approach have a maximum input token size of 512 tokens.

- *CLS pooling:* The contextual embeddings of the constituent parts of a sample are obtained by CLS pooling. The contextual embedding of the sample is calculated as the element-wise average among the contextual embeddings of the constituent parts.
- *Maximum pooling:* The contextual embedding of each sample is determined as the element-wise maximum of the output token-level embeddings of the constituent parts.
- *Average pooling:* The average operator considers the number of non-padding tokens of the parts into which a sample's text is segmented. Thus, the contextual embedding of each sample is calculated as the element-wise weighted average among the token-level embeddings of its constitutive parts.

Natural language inference for text classification involves representing the text to be classified as a premise and the labels as hypotheses. Subsequently, a multi-classification task with three classes - entailment, neutral, and contradiction- is performed. Conceptually, this approach can be thought of as having n classifiers, where n corresponds to the number of labels or classes. The models employed are encoder-only (cross-encoders) or encoder-decoder architectures.

How the output is processed varies depending on whether the dataset is multi-label or multi-class. The entailment and contradiction outputs are transformed into binary probabilities in the multi-label scenario. This transformation effectively translates the problem into n binary text classifiers, each classifier for one of the labels. In contrast, in the multi-class scenario, only the entailment logits are used to predict the class associated with a sample. When the length of the input tokens exceeds the maximum input size defined by the model, the scores of these samples are calculated by aggregating the scores obtained from their constituent parts. Aggregation measures that can be considered include maximum, average, or median.

Furthermore, like the previous approach, the influence of prompts is considered during the model evaluation. The prompts used in this context are identical to those employed in the contextual embedding similarity approach, and their details can be found in Appendix B.

Multiple-choice question-answering framework enables encoder-decoder and decoder-only architectures to perform text classification. A total of 27 models are assessed in this approach. Table 1, which provides details of the models used, shows the inclusion of pre-trained models and their instruction-tuned counterparts.

For this approach to work, a prompt is required. Therefore, the impact of the prompt is measured in the analysis. The prompts are formulated based on those most commonly used in instruction-tuning models for multiple-choice question-answering tasks. However, in these prompts, the question format is not restrained. For this reason, this study also considers the impact of the question itself. Additional information on the prompts used in this approach can be found in Appendix B.

It is essential to mention that the logit space has been constrained to align with the set of viable alternatives presented in a multiple-choice scenario. For this, the token identifiers associated with the feasible response options are determined and used to filter the logit space.

4.2 Conditional text generation

We evaluate one dataset for the conditional text generation task, using perplexity as the evaluation metric. Perplexity is a measure of uncertainty on the value of a sample from a discrete probability distribution. If we have a tokenized sequence $X = (x_0, x_1, \dots, x_T)$, then

$$\text{PPL}(X) = \exp\left\{-\frac{1}{T} \sum_{t=1}^T \log p_\theta(x_t | x_{<t})\right\}$$

where $\log p_\theta(x_t | x_{<t})$ is the log-likelihood of the t -th token conditioned on the preceding tokens $x_{<t}$ according to our model.

Below, a concise overview of the dataset employed is provided, followed by an explanation of the approach adopted. A more detailed characterization of MIMIC-CXR and its preprocessing procedure is presented in Appendix A.

4.2.1 Datasets

MIMIC-CXR is composed of X-ray reports [102–104]. As part of the preprocessing procedure, the relevant sections are extracted using the code provided by Johnson *et al.* [105, 106]. Following this extraction process and the subsequent removal of null and duplicated samples, the resulting dataset contains 57,711 samples. Furthermore, none of the samples exceed the maximum input size allowed for the proposed models.

4.2.2 Approaches

For this task, 20 decoder-only models are evaluated, as detailed in Table 1. Among these models, we find Galactica, whose tokenizer lacks special tokens. As a result, the analysis is conducted in two parts. In the first part, the Beginning of Sequence (BOS) token is included during tokenization, with perplexity calculated from the first token of the texts. The BOS token is a special token usually employed by generative models to indicate the start of a text. In cases where a model’s tokenizer does not have the BOS token predefined, such as the Falcon-7B, it is defined as its first special token. The second part of the analysis proceeds without the BOS token, and perplexity is determined from the second token of the texts, and so on.

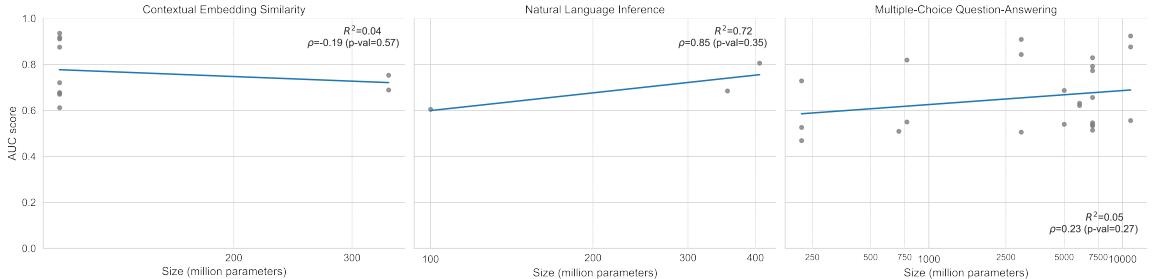
5 Results and Discussion

The main findings are outlined below, each supported by the figures and tables provided. These findings can be further verified by the figures in Appendix C.

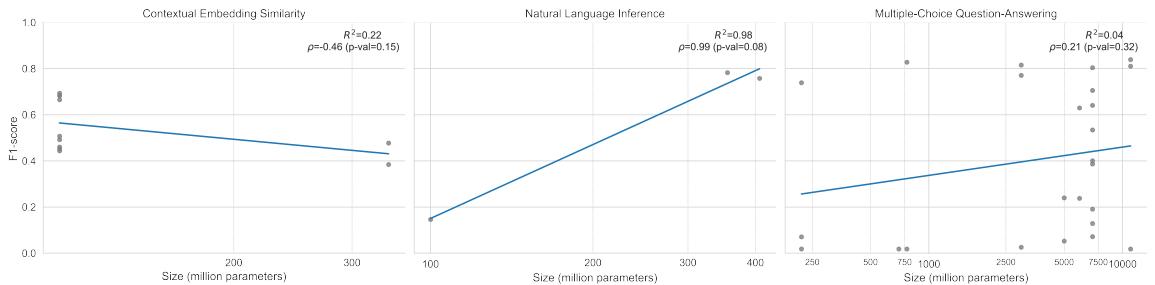
5.1 Model size: more parameters do not generally translate into better performance

The experiments conducted for text classification yield findings that challenge the assertion that larger models consistently deliver superior performance. An analysis of the impact of the logarithm of size on model performance indicates insufficient evidence to conclude that the Pearson correlation between size and performance is statistically significant in all the cases (refer to Fig. 2). This observed discrepancy may be attributed to several factors, including the relatively small size of the models considered, with a number of parameters up to billions rather than tens or hundreds of billions.

Specifically, in contextual embedding similarity, the trend of performance improvement with increasing size is almost nonexistent. As depicted in Fig. 3, for instance, models such as SapBERT (m06)



(a) Classification results for transcriptions dataset



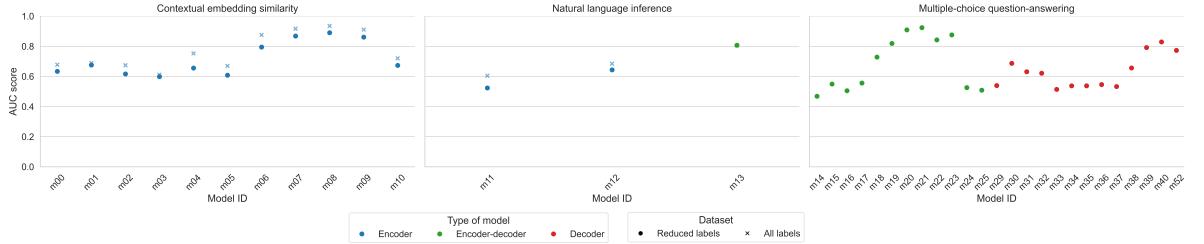
(b) Classification results for the MS-CXR dataset

Figure 2: Analysis of the impact of the logarithm of size on model performance. Model performance is defined as the highest achieved performance per model. An analysis of the Pearson coefficient suggests that there is not enough evidence to establish the statistical significance of the correlation, as reflected by the p-values. However, it is important to expand the sample size to validate this observation, considering a minimum of 30 or 35 models per approach.

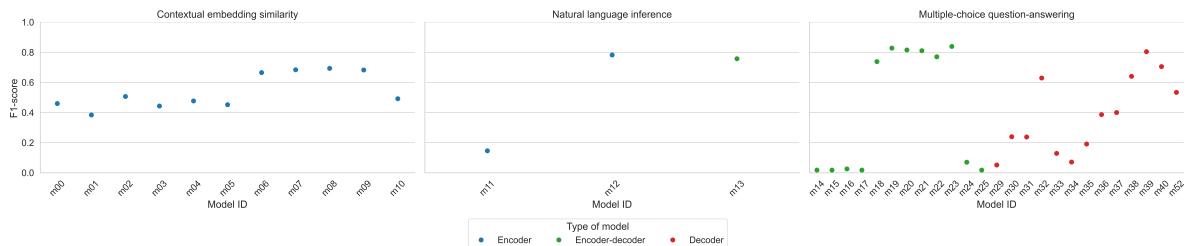
and BioLORD (m07-m09), which excel in this approach, outperform models even three times larger in both datasets. When solely the performance deltas associated with increasing the number of parameters for the same type of model are examined, they are both positive and negative. For the transcriptions dataset, as indicated by the lighter crosses in Fig. 3, BERT_{LARGE} (m01) marginally outperforms BERT_{BASE} (m00), similar to the case of BiomedBERT-large (abstracts only) (m04) versus BioMedBERT (abstracts only) (m03) where positive deltas are also obtained. However, for the MS-CXR dataset, BERT_{BASE} (m00) surpasses BERT_{LARGE} (m01) in all metrics, and BioMedBERT (abstracts only) (m03) outperforms BiomedBERT-large (abstracts only) (m04) in precision. Now, by comparing BioMedBERT (abstracts only) (m03) and BiomedBERT (abstracts + full text) (m02), performance gains are obtained when more training data are used.

In natural language inference, while the results align with the expectation that larger models tend to lead to better performance, more experiments are needed to draw a solid conclusion. In the MS-CXR dataset, the difference between the most effective NLI-DeBERTa_{base} model and the least effective RoBERTa_{LARGE}-MNLI and BART Large-MNLI models ranges from [36.32, 51.59] across the metrics considered. These differences suggest that larger models among the models used always lead to performance improvement. However, reaching the same conclusion for the transcriptions dataset is not straightforward, given the results especially obtained for RoBERTa_{LARGE}-MNLI (m12), which, as depicted in Fig. 3, is closer to the performance of NLI-DeBERTa_{base} (m11) than to that of BART Large-MNLI (m13).

When it comes to multiple-choice question-answering, the impact of the increase in size on performance is also unclear. While positive Pearson correlations are obtained, there is insufficient evidence to deem them statistically significant. When analyzing this impact within the T5 model family, as shown in Fig. 3, the effect is minimal or inconsistent when considering their non-instruction-tuned versions (m14-m17, m24-m25). Conversely, when examining the instruction-tuned versions of T5 (m18-m23), a consistent positive effect of size on performance is observed for the transcriptions dataset. Although a positive effect is also present for the MS-CXR dataset, it is only for T0 models (m22-m23).



(a) Classification results for transcriptions dataset



(b) Classification results for the MS-CXR dataset

Figure 3: The highest classification scores achieved by the models per approach for the evaluated datasets. In the scatter plots for the transcriptions dataset, lighter crosses represent scores considering the full set of labels, while darker points correspond to the reduced set of labels for comparison with the multiple-choice QA approach. Across both datasets, the highest-performing models remain the same: for contextual embedding similarity, the BioLORD models (m07-m09); for NLI, MNLI fine-tuned RoBERTa and BART (m12-m13); and for multiple-choice QA, the larger instruction-tuned models within the T5 family (m20-m23) and the instruction-tuned models within the LLaMA family (m39-m40, m52). Overall, the larger instruction-tuned T5 models emerge as the top performers. The correspondence between the model and its ID is found in Table 1.

Overall, the results of the text classification task suggest that increasing model size only sometimes leads to improved performance, whether comparing different models or evaluating the same models. Therefore, while model size may be a relevant factor in determining performance, training data and objectives may be more decisive. Expanding the sample size could be essential to validate this observation, considering a minimum of 30 or 35 models per approach. The previous insights can be verified in more detail in Figs. 13 to 18 in Appendix C.

In the context of the conditional text generation task, as shown in Fig. 4, increasing the size of the model contributes to higher performances, regardless of whether the BOS token is included. Also, an improvement in the performance of the second versions (m42, m44) versus the first versions (m41, m43) of OpenLLaMA is observed. Considering that the difference between these OpenLLaMA versions is the dataset used for pre-training, the conditional text generation task results do not contradict the above conclusion for text classification. Furthermore, our findings do not conflict with existing scaling laws, usually associated with LLMs. First, because we have no control over training, i.e., we cannot guarantee that the model is not exposed to variations other than the number of parameters (size). Second, Chinchilla’s law does not consider inter-model comparisons.

5.2 Model domain: more than a specialized domain; model architecture, training data, and training objective

Current medical datasets remain relatively small compared to those of the general domain, covering only a tiny region of the medical knowledge space [79]. Since some of the available domain-specific models are specialized by continuous pre-training on only one of these datasets, their generalization ability may be hampered, counter to the goal of creating well-generalizable domain-specific models and even achieving the ideal of artificial general intelligence (AGI). As discussed previously, the results suggest

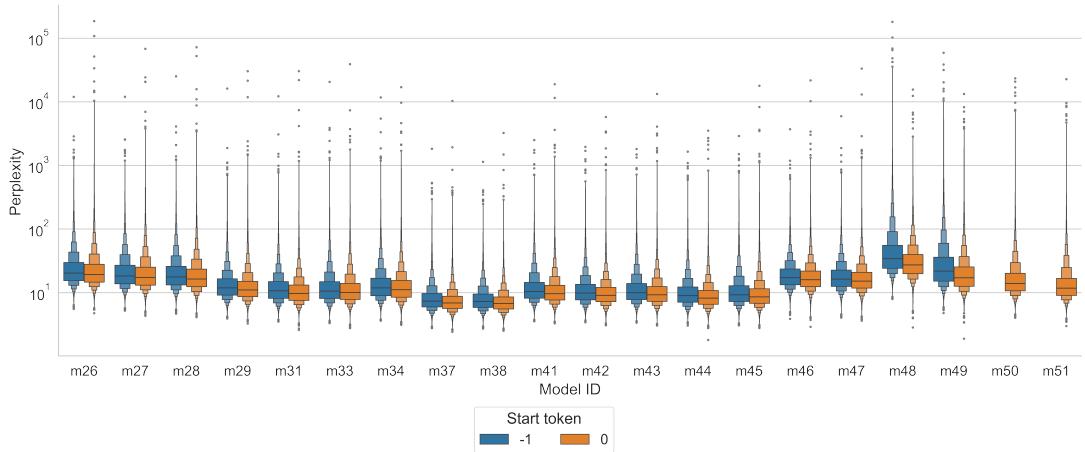


Figure 4: Performance scores for the MIMIC-CXR dataset, disaggregated by BOS token usage. The perplexities are displayed in logarithmic scale. Not using the BOS token is beneficial for 77.78% (14 out of 18) of the models, with the exceptions of the GPT2 models and **Palmyra Base 5B**. Concerning outliers, their presence is quite strong. Moderate outliers, above quantile 0.75 by 1.5 times the IQR, represent between 7% and 11% of the data, with BioGPT models having the highest percentages. Extreme outliers, above quantile 0.75 by three times the IQR, make up between 4% and 7% of the data, with most models exhibiting percentages around 4% and 5%. The correspondence between model and its ID is found in Table 1.

that architecture, training data, and training objectives are crucial in determining model generalization, possibly outweighing model size.

For the classification task through contextual embedding similarity, the effectiveness of domain specialization in improving performance is not evident, as displayed in Fig. 3. For the transcriptions dataset, refer to the lighter blue crosses in the figure. For example, **Bio+Clinical BERT** (m10) achieves lower results than expected, being around the middle of the performance ranking. Even in biomedical domain specialization, **BERT_{BASE}** (m00), a general domain model, outperforms some of the BiomedBERT models (m02-m04), challenging the superiority of domain-specific models. These observations are valid for both datasets. In particular, the latter observation also extends to **BERT_{LARGE}** (m01).

Two comparisons arise when employing multiple-choice question-answering: T5 models (m14-m15) versus their clinical specialized versions (m24-m25) and **Alpaca** (m39) versus **MedAlpaca** (m52). However, the evidence supporting the hypothesis that domain specialization significantly improves performance is limited, as displayed in Fig. 3. Specifically, in the transcriptions dataset, differences in AUC scores between ClinicalT5 (m24-m25) and T5 (m14-m15) are 5.75 and -4.11 for the base and large sizes, respectively. This trend persists across all prompts. Similarly, the difference in AUC scores between **MedAlpaca** (m52) and **Alpaca** (m39) is -1.86 , negatively affecting performance. In the MS-CXR dataset, the difference in F1-scores between ClinicalT5 (m24-m25) and T5 (m14-m15) is 5.24 and 0 for the base and large sizes, respectively. For **MedAlpaca** (m52) versus **Alpaca** (m39), the F1-score deteriorates even by 26.97. Thus, the evidence suggests that domain specialization provides either minimal or nonexistent benefits for this task.

Similar patterns to those discussed previously are identified in the text generation task, as observed in Fig. 4. Remarkably, the second version of LLaMA, **LLaMA 2-7B** (m38), stands out as the model with the highest predictive capacity for the dataset, with an average perplexity of 9.12. In contrast, **BioGPT** (m48) presents the greatest difficulty in comprehending the dataset, with an average perplexity of 38.72. These results are paradoxical, especially considering that **BioGPT** (m48) is domain-specialized while **LLaMA 2-7B** (m38) is not. A more detailed analysis by size reveals that

1. for the medium size, **BioGPT** (m48) does not outperform any of the general domain models, whereas **GPT-2-PubMed Medium** (m46) does;

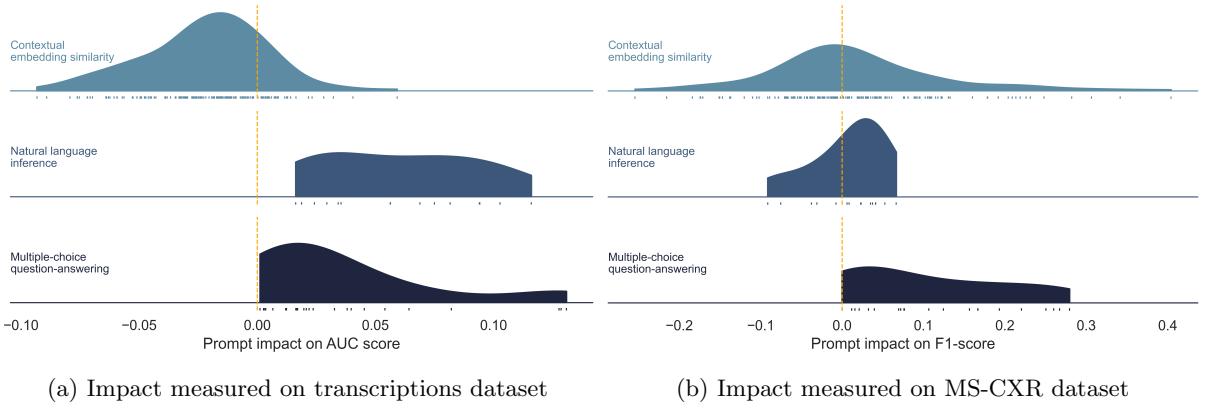


Figure 5: Prompt impact distributions on model performance. The prompt impact on contextual embedding similarity and NLI is quantified as the difference in performance resulting from prompt usage, with positive values indicating improvement. As the distributions revealed, its usage only sometimes enhances performance. In the case of multiple-choice QA, the prompt impact is calculated as the variation in performance, expressed in standard deviations, across different prompts. Optimal scenarios entail non-extreme values, suggesting neither nonexistent prompt influence on performance nor significant performance dependency on prompt wording. The distributions unveil some prompt-dependent models.

2. for the large size, domain specialization proves beneficial; and
3. for the XL and XXL sizes, neither Galactica (m50-m51) nor **BioGPT-Large** (m49) clearly outperforms general domain models.

Consequently, the only specialized models that demonstrate an advantage are GPT-2-PubMed (m46-m47).

Considering the insights discussed and the remarkable performance of **SapBERT** (m06) and **BioLORD** (m07-m09) models in the contextual embedding similarity approach, as well as Flan-T5 (m18-m21) and T0 (m22-m23) in the multiple-choice question-answering approach, the fundamental role played by training data, training objectives, and model architectures is underscored. Continual pre-training for named entity recognition or medical entity linkage using contrastive learning on UMLS data might contribute to the success of SapBERT and BioLORD models in the contextual embedding similarity task. On the other hand, employing instruction-tuned text-to-text models represents a compelling approach to achieving high performance in multiple-choice question-answering. Additionally, exploring not only the hypothesis but also the impact of domain instruction fine-tuning could provide further insights into model performance and effectiveness in specialized domains. Given that it is impossible to conclude the impact of the model domain on NLI, expanding the analysis to incorporate domain-specialized NLI models in biomedical and clinical domains could be valuable.

5.3 Prompting and instruction-tuned models: key in determining model performance

One of the central points of the study is to analyze the influence of prompts on the models and text classification approaches under investigation. As shown in Fig. 5, the results underscore the fundamental role of the prompt and its wording in the performance of the model, demonstrating both positive and negative impacts. Consequently, we advocate using prompts and advanced prompt engineering techniques to guide the model toward better results. Nevertheless, this process should not be limited to a single prompt due to the observed and well-known phenomenon of prompt brittleness [11]. As well, instruction fine-tuning proves advantageous for the models. Detailed results for each classification approach are presented below.

In contextual embedding similarity, using a prompt does not always confer benefits. For the transcripts dataset, the mean impact on the AUC score is -2.25 points, with differences ranging from -9.32

to 5.91 points. Overall, using any of the proposed prompts improves performance for 45.45% of the model + pooling strategy combinations. As for the BioLORD-PMB, BiomedBERT family, BERT_{BASE}, and SciBERT models, none of the suggested prompts led to AUC score improvements. Inversely, for the MS-CXR dataset, the prompt impact on performance is generally more positive, albeit with more variability. The mean impact, along with its range, is 1.94 (−25.17, 37.62) for accuracy, 1.30 (−25.43, 40.40) for F1-score, −2.28 (−43.91, 28.43) for precision, and 1.74 (−17.79, 32.53) for recall. Overall, employing any of the proposed prompts represents benefits for 69.70% to 84.85% of the model + pooling strategy combinations, depending on the metric. As for the BioMedBERT (*abstracts only*) and BiomedBERT-large (*abstracts only*), their performance scores, except for precision, improve with any of the proposed prompts, while for the BioLORD family and Bio+Clinical BERT, their scores are always hindered.

When examining the prompt impact on NLI, more consistent benefits are observed than for contextual embedding similarity. Using any of the proposed prompts yields improvements in the transcriptions dataset. The mean impact is 8.03 for BART Large-MNLI, 2.15 for NLI-DeBERTa_{base}, and 4.87 for RoBERTa_{LARGE}-MNLI, with larger models profiting the most from its usage. On its side, in the MS-CXR dataset, using a prompt only sometimes results in gains, particularly for NLI-DeBERTa_{base}. For this model, the impact on the F1-score is −3.42, while for BART Large-MNLI and RoBERTa_{LARGE}-MNLI, it is 2.18 and 2.78, respectively. For NLI-DeBERTa_{base}, positive prompt impacts are only observed for precision. For NLI, there are certain prompts with a high positive impact, whereas others do not, mostly independent of the model.

Regarding multiple-choice question-answering, the importance of prompts is evident, as they can significantly influence model performance. For example, in the transcriptions dataset, the proportion of models performing better than random guessing (AUC scores greater than 50) can improve from 52% to 96% with appropriate prompts. Similarly, in the MS-CXR dataset, this proportion can change from 25% to 85% when measured by the F1-score (scores greater than 1.25). Thus, its importance is highlighted not only by the high performance obtained but also by its variability, as reflected in Fig. 5. This is further supported by Figs. 17 and 18 in Appendix C. When analyzing the impact of templates versus questions, both components of the prompt (see Appendix B), it is found that the wording of templates has more influence on performance than the wording of questions, and no single prompt (template + question) works universally well for all models.

More specifically, in the transcriptions dataset, the performance ranking between the T5 models and between the Flan-T5 and T0 models remains consistent independent of the prompt. Moreover, Flan-T5-XXL (m21) and T0++ (m23) consistently outperform the other models regardless of the prompt. On average, the impact of the template, measured by the standard deviation of the AUC score per model, is 3.85, while the effect of the question within a template is 0.80. For the MS-CXR dataset, it becomes evident that some models, such as T5 and Falcon-7B, are unsuitable for this approach and dataset. The prompt plays an even more critical role than in the transcriptions dataset, affecting the ranking of models. The sensitivity is such that, with certain prompts, instruction-tuned models yield similar results to their base counterparts. On average, the impact of the template is 11.02, while the effect of the question within a template is 1.74. Interestingly, one of the proposed role-based prompts works remarkably well for the instruction-tuned versions of the T5 family despite not being the optimal choice overall.

Concerning instruction-tuning, these models generally outperform their non instruction-tuned counterparts. The instruction-tuned T5 versions, whether T0 or Flan-T5, in any size considered, exhibit superior performance than their base counterparts. In the transcriptions dataset, Flan-T5 either performs better or is on par with T0. However, in the MS-CXR dataset, T0 outperforms Flan-T5 in the 11B parameter version only when comparing the highest performances due to the considerable variability of T0++. Also, instruction-tuning improves performance consistently for the LLaMA group, whereas this is not always the case for the other generative models: MPT and GPT-J are exceptions in the transcriptions dataset and Falcon in the MS-CXR dataset. Overall, instruction-tuning represents a gain, with mean improvements of 21.45 in AUC-score and 43.55 in F1 score between the highest performances per model and 16.70 in AUC-score and 37.61 in F1 score between the mean performances for the transcriptions dataset and the MS-CXR dataset, respectively.

More details on the prompt impact can be found in Fig. 20 in Appendix C.

Table 2: Highest performing models for text classification per approach and metric. The approaches are encoded in the following way: CES stands for contextual embedding similarity, NLI for natural language inference, and MCQA for multiple-choice question-answering.

Dataset	Metric	CES		NLI		MCQA	
		Model (pooling prompt)	Score	Model (prompt)	Score	Model (prompt)	Score
MS-CXR	Accuracy	BioLORD-STAMB2-v1-STS2 (avg 5)	69.65	RoBERTa _{LARGE} -MNLI (5)	76.50	T0++ (t03+q04, t03+q05)	81.68
	F1-score	BioLORD-STAMB2-v1-STS2 (avg)	69.34	RoBERTa _{LARGE} -MNLI (5)	78.25	T0++ (t03+q05)	83.90
	Precision	BioLORD-PMB (CLS)	83.34	RoBERTa _{LARGE} -MNLI (5)	80.74	Alpaca 7B (t02+q03)	85.85
	Recall	BioLORD-STAMB2-v1-STS2 (avg)	72.58	RoBERTa _{LARGE} -MNLI (5)	82.28	T0++ (t03+q04)	89.18
Transcriptions	AUC score	BioLORD-STAMB2-v1-STS2 (avg 1)	89.01	BART Large-MNLI (4)	80.75	Flan-T5-XXL (t01+q09)	92.38

5.4 Text classification: instruction-tuned T5: models to go for

As shown in Table 2 and Fig. 3, for both datasets, the most favorable results are obtained with larger instruction-tuned T5 models, namely Flan-T5 (m18-m21) and T0 (m22-m23). The latter highlights the possibility of representing discriminative tasks as generative ones by framing them as instructions. It is also emphasized that generative tasks are not exclusive to decoder-only models, and text-to-text models might be an architecture that needs to be explored more and potentially promising. For example, instruction-tuned versions of T5 with 3B parameters (m20, m22) provide higher results than decoder-only models, almost three times larger, applying to both datasets.

When explicitly comparing precision and recall values, which can be critical depending on the application, the optimal choice of models may vary. Alpaca (m39) and LLaMA 2-CHAT-7B (m40) demonstrate proficiency in precision, while Flan-T5-XXL (m21) and T0++ (m23) excel in recall. Also, BioLORD (m07-m09) models, with 110 million parameters, exhibit precision and AUC score differences equal to or less than 3 points compared to models with a number of parameters between 7 billion and 11 billion.

With respect to the datasets, the MS-CXR dataset poses more challenges for the models than the transcriptions dataset. This difference in difficulty could be attributed to factors such as (i) the format of the data, (ii) the fact that the reports are from different medical subfields, and (iii) the length of the reports. Nevertheless, the highest performers per approach consistently show results indicative of clinical knowledge or clinical notions.

5.5 The groups of generative models: LLaMA and GPT

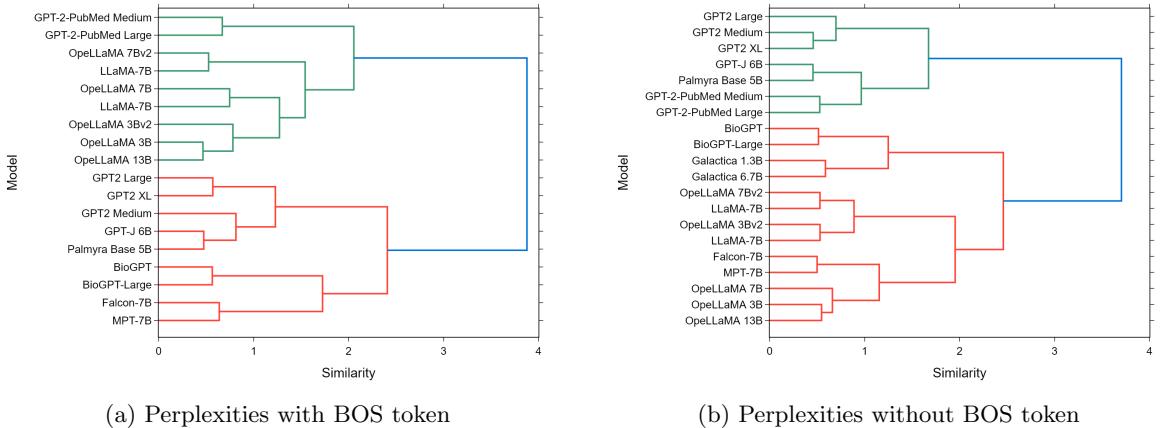


Figure 6: Dendrograms of the principal components of UMAP after being applied to the perplexities. Two major clusters of models are observed: the GPT-2 models and the LLaMA models.

Two procedures are carried out to gain insight into whether the models exhibit similar perplexity behavior and to identify potential clusters among them. The first procedure involves calculating the correlations between the models. Spearman’s and Pearson’s correlations are considered, assessing monotonic

and linear relationships. The second procedure involves dimensionality reduction via UMAP, followed by hierarchical clustering, represented in Fig. 6 using dendrograms.

The correlation analysis reveals that, in general, all models are positively related. This positive relation indicates that most samples have a similar relative difficulty level for the models. However, the BioGPT models are the only exceptions, showing low Spearman correlations compared to the other models and positioning them considerably far from the rest of the evaluated models. The observed behavior in terms of monotonic relationships remains consistent regardless of the use of the BOS token. Regarding linear relationships, groups of models are clearly observed, including LLaMA models, OpenLLaMA models, and GPT-2 models. Although the previous groups are somewhat expected, some unexpected associations are also evident, such as between **Falcon**-7B and **MPT**-7B and between **Palmyra Base** 5B and **GPT-J** 6B. These two groups have notably weak intergroup correlations.

With respect to the LLaMA and OpenLLaMA models, their linear relationship weakens when the BOS token is used, indicating more pronounced performance disparities. Possibly, training data plays an important role in this context, as it is essentially their main difference [92]. Further insights, provided by the correlations and supported by the dendrograms in Fig. 6, reveal the existence of two main groups of models: the GPT-2 models and the LLaMA models.

The analysis of the standard deviations reveals the presence of exceptionally challenging samples for the models, that is, outliers. This phenomenon is visually depicted in Fig. 4, which illustrates the distribution of perplexity between different models. Overall, the perplexity distributions have considerable variance, with the LLaMA models demonstrating the slightest variation. More detailed analyses are required to estimate the importance of these outliers and possible patterns among them.

6 Conclusion

This study comprehensively explored a range of language models with varying sizes and domain fields. These models, encompassing 52 in total, were tested in two fundamental natural language processing tasks: text classification and text generation. The size range of these models goes from 110 million to 13 billion parameters, which is relatively small compared to recent language models. Our findings, based on these specific models, have significant implications, particularly for researchers and organizations operating under resource-constrained contexts.

In text classification, our research has explored the effectiveness of three distinct approaches: context embedding similarity, natural language inference, and multiple-choice question-answering. BioLORD and SapBERT models have demonstrated remarkable performance in text classification via contextual embedding similarity. Additionally, the instruction-tuned versions of T5, Flan-T5 and T0, followed by the instruction-tuned versions of LLaMA, have exhibited good results in the multiple-choice question-answering approach. However, for the NLI approach, we have identified the need for extended exploration, particularly in the context of specialized domain models. This implies that further research is needed to fully understand the potential of NLI in text classification, especially in specialized domains.

A common thread running through our findings is the significance of the prompt in improving text classification performance across different datasets and approaches. The significance of prompts extends beyond mere performance gains; they present a viable alternative to the resource-intensive processes of training and fine-tuning language models, which are often associated with substantial financial and environmental costs, especially for large-scale models. Moreover, effective prompt engineering is essential to mitigate prompt brittleness, ensuring more robust and reliable outcomes.

Since, in general, medical datasets remain relatively small and cover only a small region of the medical knowledge space [79], domain-specific models, specialized by continuous pre-training on only one of these datasets, might see their generalization ability hindered. This practice could be one of the explanations for the results obtained. The results also suggest that the architecture, training data, and training objectives are crucial in determining the model’s generalization abilities, possibly outweighing model size.

Regarding the text generation task, two groups of models are identified based on the perplexities obtained on MIMIC-CXR. These groups consist of the GPT-2 models and the LLaMA models. Notably, the LLaMA models stand out due to their low perplexities with minimal variation. Further research is needed to identify and understand the outliers within these results, as they could hold important insights.

In conclusion, this research highlights the critical role of prompts in language model inference and reaffirms the effectiveness of instruction-tuned generative models in addressing downstream tasks. It also highlights the significance of model architecture, training data, and training objectives, potentially even more so than model size, in its generalization capacity. We advocate for further investigations into topics such as model calibration, i.e., how certain the model is about output, prompt tuning, and performance concerning issues like hallucinations and biases, among others. Such studies can lead to more effective and ethical applications of language models, particularly in sensitive fields like healthcare.

References

- [1] W. Wiggins and A. Tejani, “On the opportunities and risks of foundation models for natural language processing in radiology,” *Radiology. Artificial Intelligence*, vol. 4, no. 4, p. e220119, 07 2022.
- [2] N. H. Shah, D. Entwistle, and M. A. Pfeffer, “Creation and adoption of large language models in medicine,” *JAMA*, vol. 330, no. 9, pp. 866–869, 09 2023. [Online]. Available: <https://doi.org/10.1001/jama.2023.14217>
- [3] J. Wei *et al.*, “Emergent abilities of large language models,” *ArXiv*, vol. abs/2206.07682, 2022.
- [4] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, “Large language models in medicine: The potentials and pitfalls : A narrative review,” *Ann. Intern. Med.*, vol. 177, no. 2, pp. 210–220, 02 2024.
- [5] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, “Can large language models reason about medical questions?” *Patterns*, vol. 5, no. 3, p. 100943, 2024.
- [6] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fdbd053c1c4a845aa-Paper.pdf
- [7] G. Kuling, B. Curpen, and A. L. Martel, “Bi-rads bert and using section segmentation to understand radiology reports,” *Journal of Imaging*, vol. 8, no. 5, 2022. [Online]. Available: <https://www.mdpi.com/2313-433X/8/5/131>
- [8] K. Singhal *et al.*, “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7972, pp. 172–180, 09 2023.
- [9] M. Abdin *et al.*, “Phi-3 technical report: A highly capable language model locally on your phone,” *ArXiv*, vol. abs/2404.14219, 2024.
- [10] W. X. Zhao *et al.*, “A survey of large language models,” *ArXiv*, vol. abs/2303.18223, 2023.
- [11] K. He *et al.*, “A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics,” *ArXiv*, vol. abs/2310.05694, 2023.
- [12] L. Tang *et al.*, “Evaluating large language models on medical evidence summarization,” *npj Digital Medicine*, vol. 6, no. 1, p. 158, 08 2023. [Online]. Available: <http://dx.doi.org/10.1038/s41746-023-00896-7>
- [13] M. E. Peters *et al.*, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: <https://aclanthology.org/N18-1202>
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>

- [15] H. W. Chung *et al.*, “Scaling instruction-finetuned language models,” *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024. [Online]. Available: <http://jmlr.org/papers/v25/23-0870.html>
- [16] J. Hoffmann *et al.*, “An empirical analysis of compute-optimal large language model training,” in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 30 016–30 030. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf
- [17] T. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418fb8ac142f64a-Paper.pdf
- [18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” OpenAI, Tech. Rep., 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:160025533>
- [19] A. Chowdhery *et al.*, “Palm: Scaling language modeling with pathways,” *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023. [Online]. Available: <http://jmlr.org/papers/v24/22-1144.html>
- [20] J. W. Rae *et al.*, “Scaling language models: Methods, analysis & insights from training gopher,” *ArXiv*, vol. abs/2112.11446, 2022.
- [21] J. Wei *et al.*, “Finetuned language models are zero-shot learners,” *ArXiv*, vol. abs/2109.01652, 2022.
- [22] D. Hendrycks *et al.*, “Measuring massive multitask language understanding,” *ArXiv*, vol. abs/2009.03300, 2021.
- [23] J. Kaplan *et al.*, “Scaling laws for neural language models,” *ArXiv*, vol. abs/2001.08361, 2020.
- [24] R. Taylor *et al.*, “Galactica: A large language model for science,” *ArXiv*, vol. abs/2211.09085, 2022.
- [25] H. Touvron *et al.*, “Llama: Open and efficient foundation language models,” *ArXiv*, vol. abs/2302.13971, 2023.
- [26] ——, “Llama 2: Open foundation and fine-tuned chat models,” *ArXiv*, vol. abs/2307.09288, 2023.
- [27] Anthropic, “Introducing the next generation of Claude,” <https://www.anthropic.com/news/clause-3-family>, 03 2024.
- [28] M. Reid *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *ArXiv*, vol. abs/2403.05530, 2024.
- [29] A. Q. Jiang *et al.*, “Mistral 7b,” *ArXiv*, vol. abs/2310.06825, 2023.
- [30] R. Schaeffer, B. Miranda, and S. Koyejo, “Are emergent abilities of large language models a mirage?” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 55 565–55 581. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf
- [31] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Comput. Surv.*, vol. 55, no. 9, 01 2023. [Online]. Available: <https://doi.org/10.1145/3560815>

- [32] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, “Large language models are few-shot clinical information extractors,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Dec. 2022, pp. 1998–2022. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.130>
- [33] V. Sanh *et al.*, “Multitask prompted training enables zero-shot task generalization,” *ArXiv*, vol. abs/2110.08207, 2022.
- [34] A. Lampinen *et al.*, “Can language models learn from explanations in context?” in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Dec. 2022, pp. 537–563. [Online]. Available: <https://aclanthology.org/2022.findings-emnlp.38>
- [35] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022, pp. 22199–22213. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf
- [36] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, “TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jul. 2017, pp. 1601–1611. [Online]. Available: <https://aclanthology.org/P17-1147>
- [37] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *ArXiv*, vol. abs/2201.11903, 2023.
- [38] Q. Liu *et al.*, “Exploring the boundaries of GPT-4 in radiology,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Dec. 2023, pp. 14414–14445. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.891>
- [39] S. Bubeck *et al.*, “Sparks of artificial general intelligence: Early experiments with gpt-4,” *ArXiv*, vol. abs/2303.12712, 2023.
- [40] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, “Capabilities of gpt-4 on medical challenge problems,” *ArXiv*, vol. abs/2303.13375, 2023.
- [41] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, “GPTEval: A survey on assessments of ChatGPT and GPT-4,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, May 2024, pp. 7844–7866. [Online]. Available: <https://aclanthology.org/2024.lrec-main.693>
- [42] J. López Espejel, E. H. Ettifouri, M. S. Yahaya Alassan, E. M. Chouham, and W. Dahhane, “Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts,” *Natural Language Processing Journal*, vol. 5, p. 100032, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2949719123000298>
- [43] R. Bommasani, P. Liang, and T. Lee, “Holistic evaluation of language models,” *Annals of the New York Academy of Sciences*, vol. 1525, no. 1, pp. 140–146, 2023. [Online]. Available: <https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.15007>
- [44] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, and Y. Zhang, “Evaluating the logical reasoning ability of chatgpt and gpt-4,” *ArXiv*, vol. abs/2304.03439, 2023.
- [45] P. Lee, S. Bubeck, and J. Petro, “Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine,” *New England Journal of Medicine*, vol. 388, no. 13, pp. 1233–1239, 2023, pMID: 36988602. [Online]. Available: <https://doi.org/10.1056/NEJMsr2214184>

- [46] M. A. Fink, “Large language models such as chatgpt and gpt-4 for patient-centered care in radiology; [große sprachmodelle wie chatgpt und gpt-4 für eine patientenzentrierte radiologie],” *Radiologie*, vol. 63, no. 9, pp. 665–671, 2023. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168660418&doi=10.1007%2fs00117-023-01187-8&partnerID=40&md5=2c353e9e7688dcdde1033218f733ebc3>
- [47] Q. Lyu *et al.*, “Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential,” *Visual Computing for Industry, Biomedicine, and Art*, vol. 6, no. 1, p. 9, 2023. [Online]. Available: <http://dx.doi.org/10.1186/s42492-023-00136-5>
- [48] L. C. Adams *et al.*, “Leveraging gpt-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study,” *Radiology*, vol. 307, no. 4, p. e230725, 2023, pMID: 37014240. [Online]. Available: <https://doi.org/10.1148/radiol.230725>
- [49] R. Bhayana, R. R. Bleakney, and S. Krishna, “Gpt-4 in radiology: Improvements in advanced reasoning,” *Radiology*, vol. 307, no. 5, p. e230987, 2023, pMID: 37191491. [Online]. Available: <https://doi.org/10.1148/radiol.230987>
- [50] Z. Wu *et al.*, “Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task,” *ArXiv*, vol. abs/2304.09138, 2023.
- [51] M. Ranjit, G. Ganapathy, R. Manuel, and T. Ganu, “Retrieval augmented chest x-ray report generation using openai gpt models,” *ArXiv*, vol. abs/2305.03660, 2023.
- [52] B. Meskó and E. J. Topol, “The imperative for regulatory oversight of large language models (or generative ai) in healthcare,” *npj Digital Medicine*, vol. 6, no. 1, p. 120, 07 2023. [Online]. Available: <http://dx.doi.org/10.1038/s41746-023-00873-0>
- [53] D. Gala and A. N. Makaryus, “The utility of language models in cardiology: A narrative review of the benefits and concerns of chatgpt-4,” *International Journal of Environmental Research and Public Health*, vol. 20, no. 15, 2023. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85167748409&doi=10.3390%2ffijerph20156438&partnerID=40&md5=42bf6f987c91c22783b0b329917043d1>
- [54] S. B. Atallah, N. R. Banda, A. Banda, and N. A. Roeck, “How large language models including generative pre-trained transformer (gpt) 3 and 4 will impact medicine and surgery,” *Techniques in Coloproctology*, vol. 27, no. 8, pp. 609—614, 07 2023. [Online]. Available: <http://dx.doi.org/10.1007/s10151-023-02837-8>
- [55] K. Cheng, Q. Guo, Y. He, Y. Lu, S. Gu, and H. Wu, “Exploring the potential of gpt-4 in biomedical engineering: The dawn of a new era,” *Annals of biomedical engineering*, vol. 51, no. 8, pp. 1645–1653, 08 2023. [Online]. Available: <https://doi.org/10.1007/s10439-023-03221-1>
- [56] X. Yang *et al.*, “A large language model for electronic health records,” *npj Digital Medicine*, vol. 5, no. 1, p. 194, 12 2022. [Online]. Available: <http://dx.doi.org/10.1038/s41746-022-00742-2>
- [57] C. Peng *et al.*, “A study of generative large language model for medical research and healthcare,” *npj Digital Medicine*, vol. 6, no. 1, p. 210, 11 2023. [Online]. Available: <http://dx.doi.org/10.1038/s41746-023-00958-w>
- [58] T. Han *et al.*, “Medalpaca – an open-source collection of medical conversational ai models and training data,” *ArXiv*, vol. abs/2304.08247, 2023.
- [59] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, “PMC-LLaMA: toward building open-source language models for medicine,” *Journal of the American Medical Informatics Association*, p. ocae045, 04 2024. [Online]. Available: <https://doi.org/10.1093/jamia/ocae045>
- [60] K. Singhal *et al.*, “Towards expert-level medical question answering with large language models,” *ArXiv*, vol. abs/2305.09617, 2023.

- [61] G. Wang, G. Yang, Z. Du, L. Fan, and X. Li, “Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation,” *ArXiv*, vol. abs/2306.09968, 2023.
- [62] H. Zhang *et al.*, “HuatuoGPT, towards taming language model to be a doctor,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Dec. 2023, pp. 10859–10885. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.725>
- [63] C. Li *et al.*, “Llava-med: Training a large language-and-vision assistant for biomedicine in one day,” in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 28541–28564. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/5abcd8ecdacba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf
- [64] J. Zhou, X. Chen, and X. Gao, “Path to medical agi: Unify domain-specific medical llms with the lowest cost,” *ArXiv*, vol. abs/2306.10765, 2023.
- [65] W. Gao *et al.*, “Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue,” *ArXiv*, vol. abs/2306.12174, 2023.
- [66] C. Shu, B. Chen, F. Liu, Z. Fu, E. Shareghi, and N. Collier, “Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities,” <https://github.com/cambridgetl/visual-med-alpaca>, 2013.
- [67] M. Moor *et al.*, “Med-flamingo: a multimodal medical few-shot learner,” in *Proceedings of the 3rd Machine Learning for Health Symposium*, ser. Proceedings of Machine Learning Research, vol. 225. PMLR, 10 Dec 2023, pp. 353–367. [Online]. Available: <https://proceedings.mlr.press/v225/moor23a.html>
- [68] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, “Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning,” *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1399–1406, 09 2022. [Online]. Available: <http://dx.doi.org/10.1038/s41551-022-00936-9>
- [69] J. Lee *et al.*, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 09 2019. [Online]. Available: <https://doi.org/10.1093/bioinformatics/btz682>
- [70] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. [Online]. Available: <https://aclanthology.org/D19-1371>
- [71] Y. Gu *et al.*, “Domain-specific language model pretraining for biomedical natural language processing,” *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, 10 2021. [Online]. Available: <https://doi.org/10.1145/3458754>
- [72] H.-C. Shin *et al.*, “BioMegatron: Larger biomedical domain language model,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Nov. 2020, pp. 4700–4706. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.379>
- [73] Z. Hong *et al.*, “Scholarbert: Bigger is not always better,” *ArXiv*, vol. abs/2205.11342, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:248987342>
- [74] R. Luo *et al.*, “BioGPT: generative pre-trained transformer for biomedical text generation and mining,” *Briefings in Bioinformatics*, vol. 23, no. 6, 09 2022, bbac409. [Online]. Available: <https://doi.org/10.1093/bib/bbac409>

- [75] E. Alsentzer *et al.*, “Publicly available clinical BERT embeddings,” in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Jun. 2019, pp. 72–78. [Online]. Available: <https://www.aclweb.org/anthology/W19-1909>
- [76] Q. Xie, E. J. Schenck, H. S. Yang, Y. Chen, Y. Peng, and F. Wang, “Faithful ai in medicine: A systematic review with large language models and beyond,” *medRxiv*, 2023. [Online]. Available: <https://www.medrxiv.org/content/early/2023/07/01/2023.04.18.23288752>
- [77] S. Soni and K. Roberts, “Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, May 2020, pp. 5532–5538. [Online]. Available: <https://aclanthology.org/2020.lrec-1.679>
- [78] E. Lehman *et al.*, “Do we still need clinical language models?” in *Proceedings of the Conference on Health, Inference, and Learning*, ser. Proceedings of Machine Learning Research, vol. 209. PMLR, 06 2023, pp. 578–597. [Online]. Available: <https://proceedings.mlr.press/v209/eric23a.html>
- [79] H. Zhou *et al.*, “A survey of large language models in medicine: Progress, application, and challenge,” *ArXiv*, vol. abs/2311.05112, 2024.
- [80] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf
- [81] Y. Liu *et al.*, “Roberta: A robustly optimized bert pretraining approach,” *ArXiv*, vol. abs/1907.11692, 2019.
- [82] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, “Self-alignment pretraining for biomedical entity representations,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Jun. 2021, pp. 4228–4238. [Online]. Available: <https://aclanthology.org/2021.naacl-main.334>
- [83] F. Remy, K. Demuynck, and T. Demeester, “BioLORD: Learning ontological representations from definitions for biomedical concepts and their textual descriptions,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Dec. 2022, pp. 1454–1465. [Online]. Available: <https://aclanthology.org/2022.findings-emnlp.104>
- [84] Sentence Transformers - Cross-Encoders. (2021) cross-encoder/nli-deberta-base. [Online]. Available: <https://huggingface.co/cross-encoder/nli-deberta-base>
- [85] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: <http://jmlr.org/papers/v21/20-074.html>
- [86] Google. (2023) google/t5-v1.1. [Online]. Available: <https://huggingface.co/google>
- [87] AI at Meta. (2023) facebook/bart-large-mnli. [Online]. Available: <https://huggingface.co/facebook/bart-large-mnli>
- [88] V. Sanh *et al.*, “Multitask prompted training enables zero-shot task generalization,” *ArXiv*, vol. abs/2110.08207, 2021.
- [89] Q. Lu, D. Dou, and T. Nguyen, “ClinicalT5: A generative language model for clinical text,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Dec. 2022, pp. 5436–5443. [Online]. Available: <https://aclanthology.org/2022.findings-emnlp.398>

- [90] A. Radford *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [91] W. E. team, “Palmyra-base Parameter Autoregressive Language Model,” <https://dev.writer.com>, 01 2023.
- [92] X. Geng and H. Liu, “Openllama: An open reproduction of llama,” May 2023. [Online]. Available: https://github.com/openlm-research/open_llama
- [93] W. E. team, “Camel-5B InstructGPT,” <https://dev.writer.com>, 04 2023.
- [94] B. Wang and A. Komatsuzaki, “GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model,” <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- [95] NLP Cloud, “nlpcloud/instruct-gpt-j-fp16,” 2023. [Online]. Available: <https://huggingface.co/nlpcloud/instruct-gpt-j-fp16>
- [96] E. Almazrouei *et al.*, “The falcon series of open language models,” *ArXiv*, vol. abs/2311.16867, 2023.
- [97] MosaicML NLP Team. (2023) Introducing mpt-7b: A new standard for open-source, commercially usable llms. [Online]. Available: www.mosaicml.com/blog/mpt-7b
- [98] R. Taori *et al.*, “Stanford alpaca: An instruction-following llama model,” https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [99] Y. Papanikolaou. (2020) healx/gpt-2-pubmed. [Online]. Available: <https://huggingface.co/healx>
- [100] B. Boecking *et al.*, “MS-CXR: Making the most of text semantics to improve biomedical vision–language processing (version 0.1),” 2022. [Online]. Available: <https://doi.org/10.13026/b90j-vb87>
- [101] ——, “Making the most of text semantics to improve biomedical vision–language processing,” in *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022, pp. 1–21. [Online]. Available: https://doi.org/10.1007/978-3-031-20059-5_1
- [102] A. L. Goldberger *et al.*, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), circulation Electronic Pages: <http://circ.ahajournals.org/content/101/23/e215.full> PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.
- [103] A. E. W. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng, “The MIMIC-CXR database,” 2019. [Online]. Available: <https://doi.org/10.13026/C2JT1Q>
- [104] A. E. W. Johnson *et al.*, “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” *Sci Data*, vol. 6, no. 1, p. 317, 2019. [Online]. Available: <https://doi.org/10.1038/s41597-019-0322-0>
- [105] A. E. W. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, “The mimic code repository: enabling reproducibility in critical care research,” *Journal of the American Medical Informatics Association*, vol. 25, no. 1, pp. 32–39, 2018.
- [106] A. Johnson *et al.*, “Mit-lcp/mimic-code: Mimic code v2.2.1,” Jul. 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6818823>

A Data

This section describes the data employed and outlines the corresponding preprocessing procedure.

A.1 Transcriptions

Transcriptions is a multi-label dataset with 40 different labels and 2,358 data samples. The data were extracted from Kaggle, and additional information about the labels can be found in MTSamples.com.

A.1.1 Preprocessing

The preprocessing procedure involves the removal of samples that lack associated reports, adjusting the formatting of the report, and selecting and renaming labels. Formatting adjustments are necessary because line breaks are encoded as comma patterns. To ascertain the final format, we considered the original data source MTSamples.com and the results generated by ChatGPT as a guide to knowledge of language models.

In terms of labels, less relevant categories were excluded due to their broad level of generality or lack of association with a specific medical specialty. Precisely, the eliminated labels are: “Consult - History and Phy.”, “Discharge Summary”, “Emergency Room Reports”, “General Medicine”, “Hospice - Palliative Care”, “IME-QME-Work Comp etc.”, “Letters”, “Office Notes”, “Pain Management”, “SOAP / Chart / Progress Notes”. Additionally, several labels contained the “/” character, indicating “or”, which we explicitly replaced with the latter. For example, “Allergy / Immunology” was transformed into “Allergy or Immunology”. Subsequently, the labels “Chiropractic” and “Physical Medicine - Rehab” were merged into a unified category called “Physical Medicine and Rehabilitation, or Chiropractic”. Other modifications include transforming “ENT - Otolaryngology” into “Otolaryngology”, “Hematology - Oncology” into “Hematology or Oncology”, “Lab Medicine - Pathology” into “Laboratory Medicine or Clinical Pathology”, “Pediatrics - Neonatal” into “Pediatrics or Neonatal”, and “Speech - Language” into “Speech and Language”.

Upon completion of the preprocessing, the initial count of 40 different labels is reduced to 29, and the number of samples to consider is 2,074.

A.1.2 Description

The class distribution is visualized in Fig. 7. Surgery is the most prevalent category (in 52.46% of the samples), followed by Cardiovascular or Pulmonary (17.89%) and Orthopedic (17.11%). On the other hand, Allergy or Immunology (0.33%), preceded by Autopsy (0.38%) and Laboratory Medicine or Clinical Pathology (0.38%), are the least frequent categories. The number of labels per sample ranges from 1 to 4, with an average of 2 labels per sample. Additionally, some labels never co-occur within the same sample.

The results of analyzing label leakage, which refers to whether a label appears explicitly in the text to be classified, are shown in Fig. 8. For most labels, label leakage is minimal, except for Autopsy (62.50%), Rheumatology (40.00%), Speech and Language (33.33%), and Surgery (29.50%). Labels without label leakage are Allergy or Immunology, Cardiovascular or Pulmonary, Cosmetic or Plastic Surgery, Diets and Nutritions, Hematology or Oncology, Laboratory Medicine or Clinical Pathology, Obstetrics or Gynecology, Pediatrics or Neonatal, Physical Medicine and Rehabilitation, or Chiropractic, Psychiatry or Psychology, and Sleep Medicine. The presence of labels in texts of other labels is not considered, given that this is a multi-label dataset, and the analysis and interpretation of such occurrences are inherently complex.

A.2 MS-CXR

MS-CXR [100–102] is a multi-class dataset with 8 different classes and a corpus of 1,448 data samples, comprising 718 unique samples. The data can be obtained from [102].

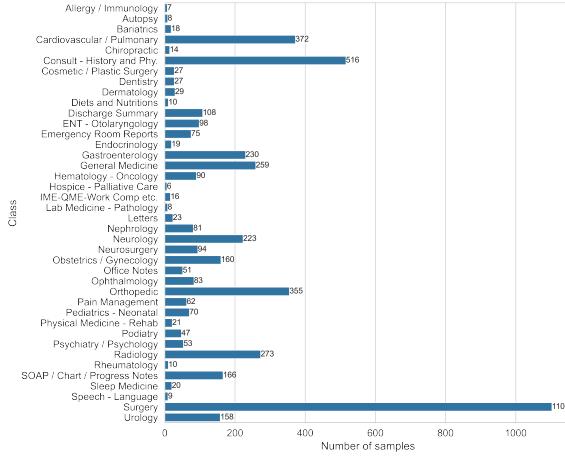


Figure 7: Label distribution of the transcriptions dataset after preprocessing

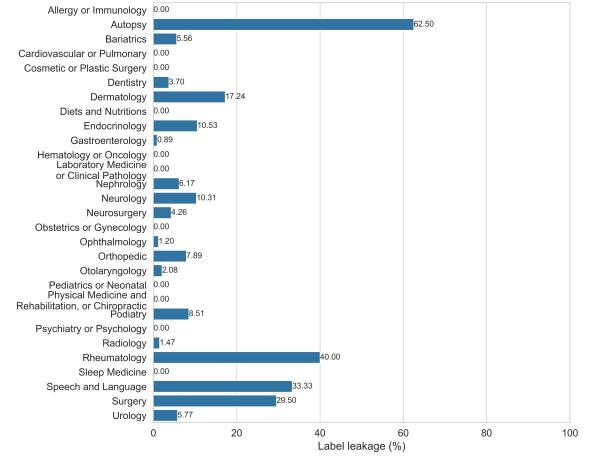


Figure 8: Label leakage for the transcriptions dataset

A.2.1 Preprocessing

The preprocessing procedure involves removing instances without associated reports and eliminating duplicates. To be precise, 730 samples (50.41%) were identified as duplicates, with a maximum of 82 and an average of 3 duplicates, considering only repeated reports. In addition, when duplicate reports do not agree with the assigned label, either of these labels is evaluated as the true one.

A.2.2 Description

The class distribution is depicted in Fig. 9. Overall, the dataset does not exhibit class imbalance. The most frequent classes are Pneumonia (24.37%), closely followed by Pneumothorax (21.17%), while the less frequent classes are Cardiomegaly (5.15%), preceded by Edema (5.43%).

Upon analysis of label leakage, as presented in Fig. 10, a high label leakage is observed, except for Lung Opacity, which has a low leakage rate of 1.33%. In particular, Consolidation, Edema, and Pneumothorax exhibit leakage rates that exceed 90%. Classes with leakage rates below 50% include Pneumonia, Cardiomegaly, and Lung Opacity, as mentioned earlier. Regarding the presence of labels in text from other labels, notable occurrences include Consolidation in the classes of Edema (12.82%) and Pneumonia (24.00%), and Pleural Effusion in the Atelectasis class (21.43%).

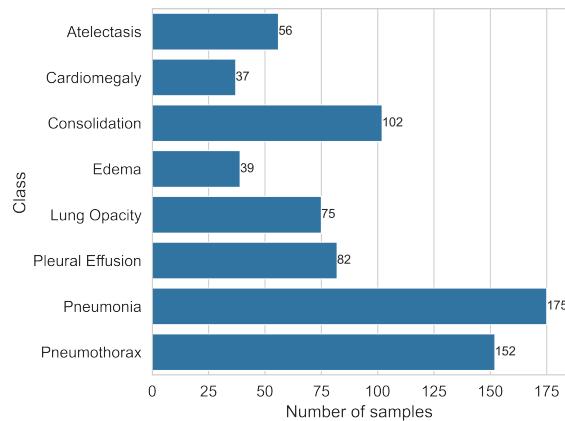


Figure 9: Label distribution of the MS-CXR dataset

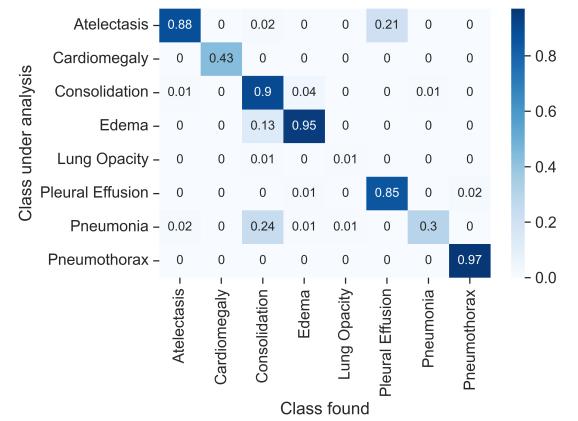


Figure 10: Label leakage and presence of class label on other classes for MS-CXR dataset

To conclude, each class’s word count per text is measured, and their distributions are presented in Fig. 11. Classes with shorter texts include Cardiomegaly and Pneumothorax. Although classes with longer texts are not observed, there are flatter distributions with heavy tails, suggesting that the length of texts in these classes is less concentrated around a specific value.

A.3 MIMIC-CXR

MIMIC-CXR [102–104] is a dataset of radiographic reports that encompasses 78,584 samples. After extracting the most pertinent sections, 75,029 samples are identified as informative. This dataset is accessible through [102].

A.3.1 Preprocessing

The preprocessing procedure involves extracting the most relevant sections from chest X-ray reports using the codes [105] designed for this purpose and publicly available on GitHub [106]. In addition, texts lacking content and duplicate samples are removed. Texts lacking information are defined as those that are empty or match one of the following: “.”, “As above”, “As above.”, “As above..”, “None.”, “See above.”, “No changes.”, “__”, “__ earlier”, “__”, or “__”. Those mentioned above were identified after meticulously examining texts with a maximum length of two words. In total, these non-informative texts represent merely 0.26% of the dataset. Regarding duplicates, 1.69% of the total samples are duplicated, comprising 23.07% of the dataset. The text with the most duplicates is ”No acute cardiopulmonary process.” representing 7.88% of the samples. On average, each text appears twice in the dataset.

Upon completion of the preprocessing steps, the dataset results in 57,711 samples, composed mainly of impressions (81.92%) and findings (17.48%).

A.3.2 Description

Considering the nature of this dataset, its description focuses mainly on the distribution of the number of words per sample, as shown in Fig. 12. This distribution is left-skewed, with a peak of around 10 words per sample. Moreover, there is a significant plateau between 20 and 40 words per sample. Interestingly, the distribution’s right tail extends beyond 150 words per sample. In summary, most texts (75%) contain at most 51 words, with a pronounced peak of around 10 words per sample. However, this dataset also includes longer texts, some reaching up to 307 words.

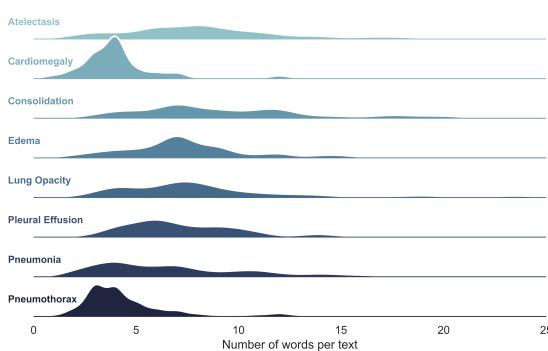


Figure 11: Length distribution of reports, in number of words, for MS-CXR dataset

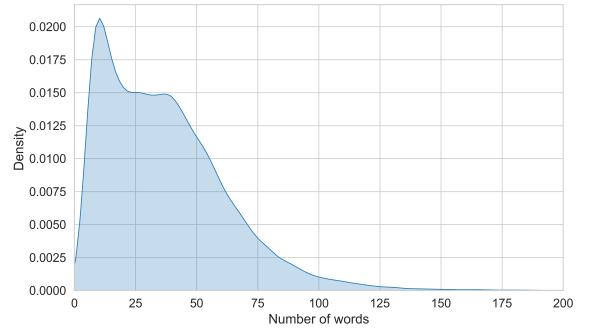


Figure 12: Length distribution of reports in number of words for MIMIC-CXR dataset

B Prompts

The prompts used for the text classification task via contextual embedding similarity, natural language inference (NLI), and multiple-choice question answering (QA) are presented.

B.1 Prompts for text classification via contextual embedding similarity and NLI

The prompts proposed for text classification using contextual embedding similarity and Natural Language Inference (NLI) are exclusively applied to the label (in the case of NLI, to the hypothesis). Table 3 lists the prompts used. Template ID 0 is the default to generate the hypothesis in the zero-shot text classification using the NLI setting, as documented in HuggingFace.

Table 3: Templates to be used as contextual embedding similarity and NLI prompts. The column “Label prompt” refers to the template, and the column “Dataset” specifies the target dataset.

ID	Label prompt	Dataset
0	This example is {label}.	Transcriptions, MS-CXR
1	This is an example of {label}.	Transcriptions, MS-CXR
2	This report belongs to the category {label}.	Transcriptions
3	This report belongs to the medical speciality {label}.	Transcriptions
4	This report belongs to the medical speciality: {label}.	Transcriptions
5	The diagnosis is {label}.	MS-CXR
6	There is evidence of {label}.	MS-CXR
7	These findings are consistent with {label}.	MS-CXR

B.2 Prompts for text classification via multiple-choice QA

The proposed prompts for text classification via multiple-choice question answering are based on the default templates specific to various of the considered instruction-tuned models. These templates are systematically assessed using a set of questions, enabling us to quantify the influence of the question wording. For the MS-CXR dataset, we also incorporate role-based questions. The prompts, their corresponding datasets, and specific requirements are summarized in Table 5 and Table 4.

Each class or label is encoded with an uppercase letter denoting the option, followed by its name. For instance, if the first label is “ y_1 ”, it is represented as “(A) y_1 ” within the prompt. In the context of the transcriptions dataset, there are 29 distinct labels. However, due to their large number, we include the top 10 most frequent labels and categorize the remaining labels under an additional “Other” option. Specifically, for the transcriptions dataset, we employ templates t01, t02, t03, t04, and t07 along with questions q07, q08, and q09. Whereas for MS-CXR dataset, we employ templates t01, t02, t03, t04, t07, t11, and t13, and questions q03, q04, and q05.

Table 4: Questions to be used for the multiple-choice QA templates. The column “Dataset” specifies the target dataset.

ID	Question	Dataset
q01	What is the most plausible diagnosis?	MS-CXR
q02	What is the patient’s diagnosis?	MS-CXR
q03	What is the diagnosis?	MS-CXR
q04	Which one of the following is the diagnosis?	MS-CXR
q05	Which one is the patient’s diagnosis?	MS-CXR
q06	Which of the options is the most likely to be the diagnosis?	MS-CXR
q07	Which category does the report belong to?	Transcriptions
q08	What is the field that best suits the report?	Transcriptions
q09	Which one is the topic of the report?	Transcriptions

Table 5: Templates to be used as multiple-choice QA prompts. Regarding the column “Requirements”, “report” refers to the text sample, “options” to the labels provided as choices, and “question” to the question itself (see Table 4). Note that the term “question” sometimes appears capitalized, indicating that the question begins with an uppercase letter when integrated into the template.

ID	Prompt	Requirements	Dataset
	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.		
t01	<pre>#### Instruction: {question} Select one of the following options: {options}</pre> <p>#### Input: {report}</p> <p>#### Response: ()</p> <p>Context: {report}</p> <p>Question: {question}</p>	report, options, QUESTION	Transcriptions, MS-CXR
t02	<p>Options: {options}</p> <p>Answer: (</p> <p>Context: {report}</p> <p>Question: Based on the context, {question}</p>	report, options, QUESTION	Transcriptions, MS-CXR
t03	<p>Options: {options}</p> <p>Answer: (</p>	report, options, question	Transcriptions, MS-CXR
t04	<p>{report}. Which one of the following, if true, most strengthens the argument? {options}. (</p> <p>Read the following and answer the question.</p> <p>{report}</p>	report, options	Transcriptions, MS-CXR
t05	<p>{question} {options}</p> <p>(</p> <p>{report}</p>	report, options, QUESTION	Transcriptions, MS-CXR
t06	<p>What's the best answer to this question: {question}</p> <p>{options}</p> <p>(</p> <p>{report}</p>	report, options, QUESTION	Transcriptions, MS-CXR
t07	<p>{question} {options}</p> <p>(</p> <p>Read this chest x-ray report: "{report}"</p>	report, options, QUESTION	Transcriptions, MS-CXR
t08	<p>Now answer this question: "{question}"</p> <p>{options}</p> <p>(</p>	report, options, QUESTION	MS-CXR
t09	<p>Knowing that "{report}", how would one answer "{question}"</p> <p>{options}</p> <p>(</p> <p>{report}</p> <p>Based on the above text, what's the best answer to this question: {question}</p>	report, options, QUESTION	Transcriptions, MS-CXR
t10	<p>{options}</p> <p>(</p> <p>{report}</p> <p>Based on the above text, what's the best answer to this question: {question}</p>	report, options, QUESTION	Transcriptions, MS-CXR
t11	<p>You are a doctor and have the following information about a patient from a chest x-ray: {report}. Which one of the following, if true, most strengthens the argument? {options}. (</p>	report, options	MS-CXR
t12	<p>You are a doctor and have the following information about a patient from a chest x-ray: {report}. {question} {options}. (</p>	report, options, QUESTION	MS-CXR
t13	<p>I want you to act as a virtual doctor. I will describe my symptoms and you will choose the most probable diagnosis among the following: {options}. You should only reply with the chosen diagnosis, and nothing else. My request is "{report}". (</p>	report, options	MS-CXR
t14	<p>I want you to act as a virtual doctor. I will describe my symptoms and you will choose a diagnosis among the possible diagnoses. You should only reply with the chosen diagnosis, and nothing else. Do not write explanations. The possible diagnoses are: {options}. My request is "{report}". (</p>	report, options	MS-CXR

C Supplementary results

This appendix presents supplementary figures and tables to support the results presented. They are displayed first by dataset and then by task or approach.

C.1 Classification task

C.1.1 Contextual embedding similarity

Results are depicted in Figs. 3, 13 and 14. The mapping between the models and their ID is

m00: BERT _{BASE}	m06: SapBERT
m01: BERT _{LARGE}	m07: BioLORD-STAMB2-v1
m02: BiomedBERT (abstracts + full text)	m08: BioLORD-STAMB2-v1-STS2
m03: BiomedBERT (abstracts only)	m09: BioLORD-PMB
m04: BiomedBERT-large (abstracts only)	m10: Bio+Clinical BERT
m05: SciBERT	

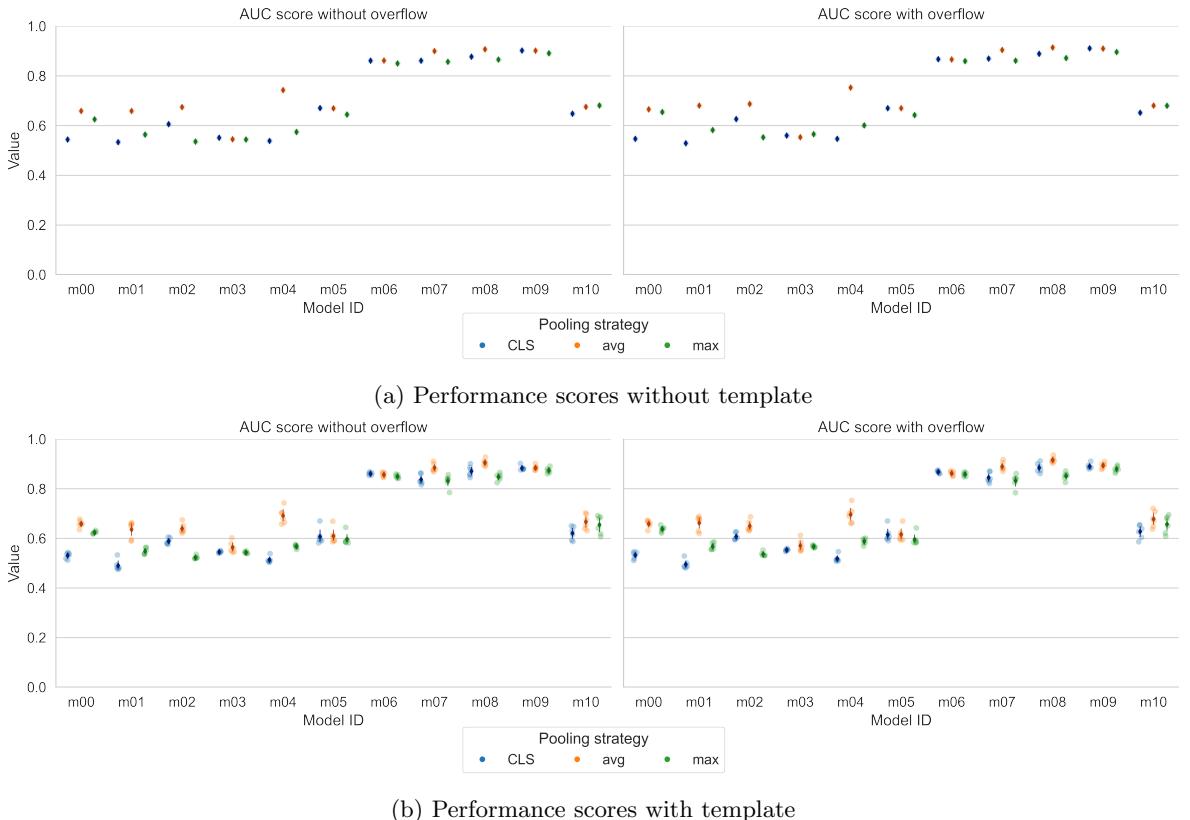
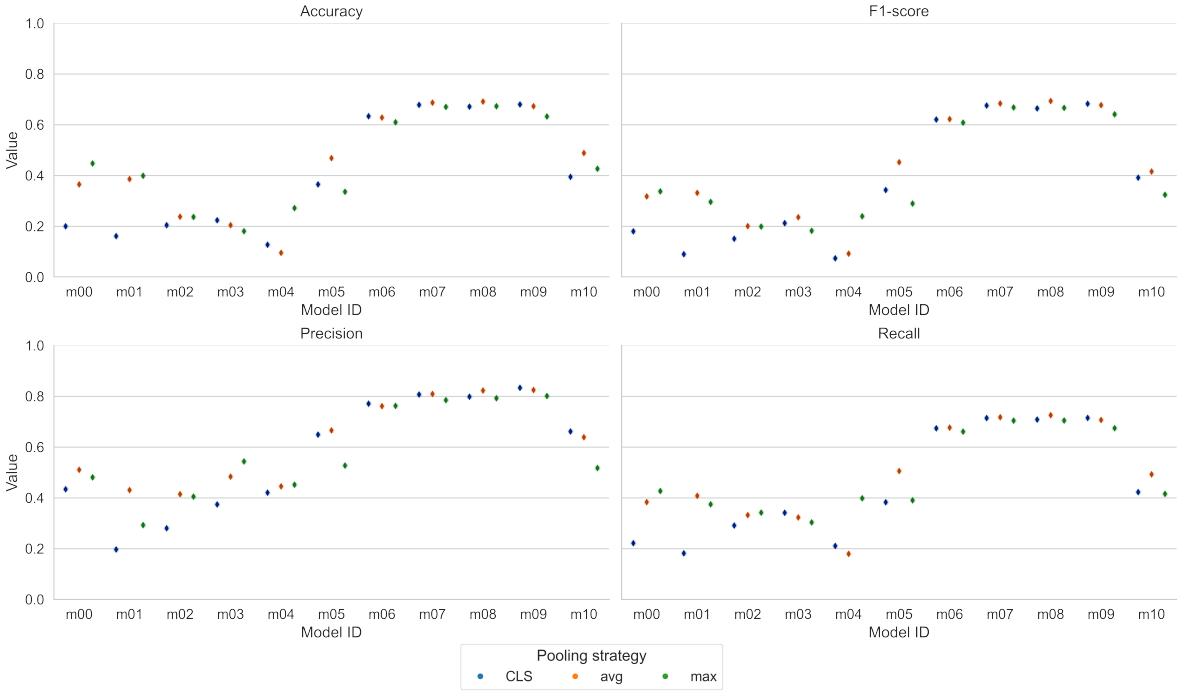
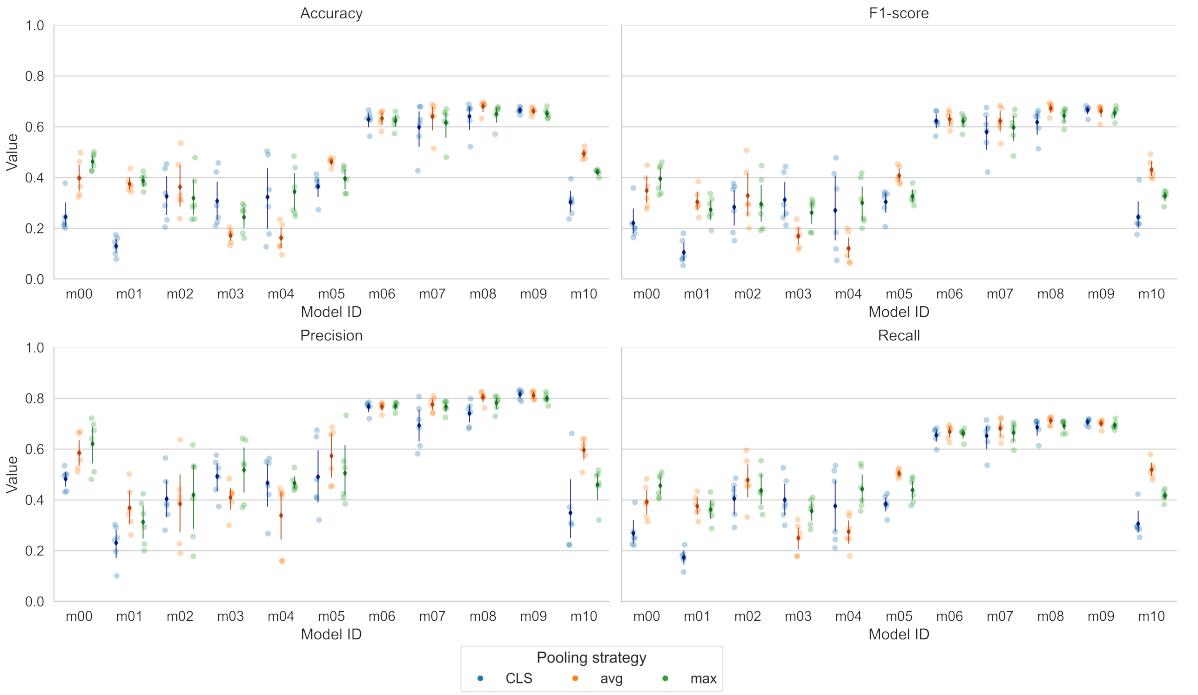


Figure 13: Performance scores for transcriptions dataset using contextual embedding similarity, disaggregated by template and overflow usage, and pooling strategy. Using overflow leads to improvement on 87.88% of the cases, being a case a model + pool strategy. The impact of using overflow is, on average, 0.90 points of the AUC score. Thus, most of the trends observed when not using overflow are kept. Regarding the pooling strategy, when using overflow, average pooling produces the best results (8 of 11 models), followed by CLS pooling. SapBERT and BioLORD models stand out due to their performance, having more than a 10-point difference in the AUC score with the rest of the models when the best configurations are compared.



(a) Performance scores without template



(b) Performance scores with template

Figure 14: Performance scores for MS-CXR dataset using contextual embedding similarity, disaggregated by template usage and pooling strategy. SapBERT and BioLORD models stand out due to their performance, having at least a difference of 12, 16, 10, and 16 points in accuracy, F1-score, precision, and recall, respectively, with the rest of the models when the best configuration is compared. However, this gap no longer exists when templates are allowed to be used. That means that templates have a primordial role in determining model performance, both to boost or hinder it. Specifically, it boosts the performance on between 69.70% to 84.85% of the cases depending on the performance metric, being a case a model + pool strategy. Regarding the pooling strategy, average pooling produces the best results (5-8 of 11 models, depending on the metric).

C.1.2 Natural language inference

Results are depicted in Figs. 15 and 16. The mapping between the models and their ID is

m11: NLI-DeBERTa_{base} m12: RoBERTa_{LARGE}-MNLI m13: BART Large-MNLI

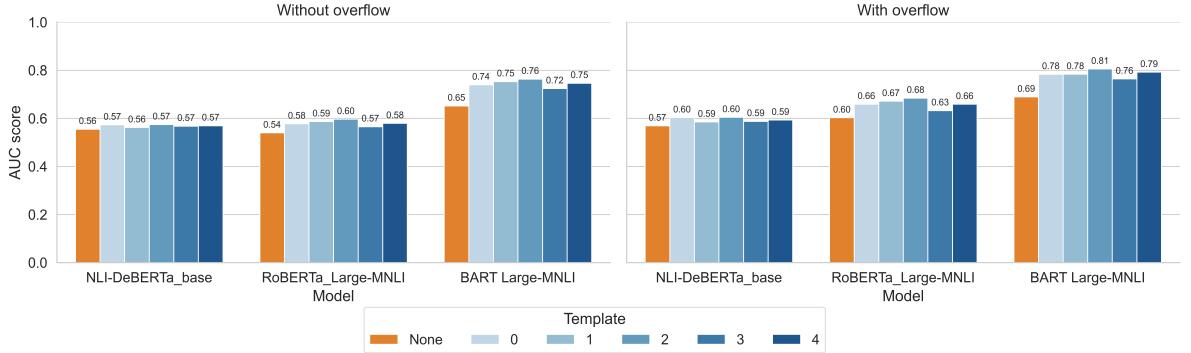


Figure 15: Performance scores for transcriptions dataset using natural language inference, disaggregated by template and overflow usage. Using overflow leads to improvement on all three models, having an impact of 2.15, 4.87, 8.03 on NLI-DeBERTa_{base}, RoBERTa_{LARGE}-MNLI, and BART Large-MNLI respectively. Noteworthy to observe is the benefit that using overflow represents for RoBERTa_{LARGE}-MNLI: from performing lower than NLI-DeBERTa_{base} to outperforming it when using overflow. Using a template consistently improves performance, particularly template 2, resulting in AUC scores of up to 80.54.

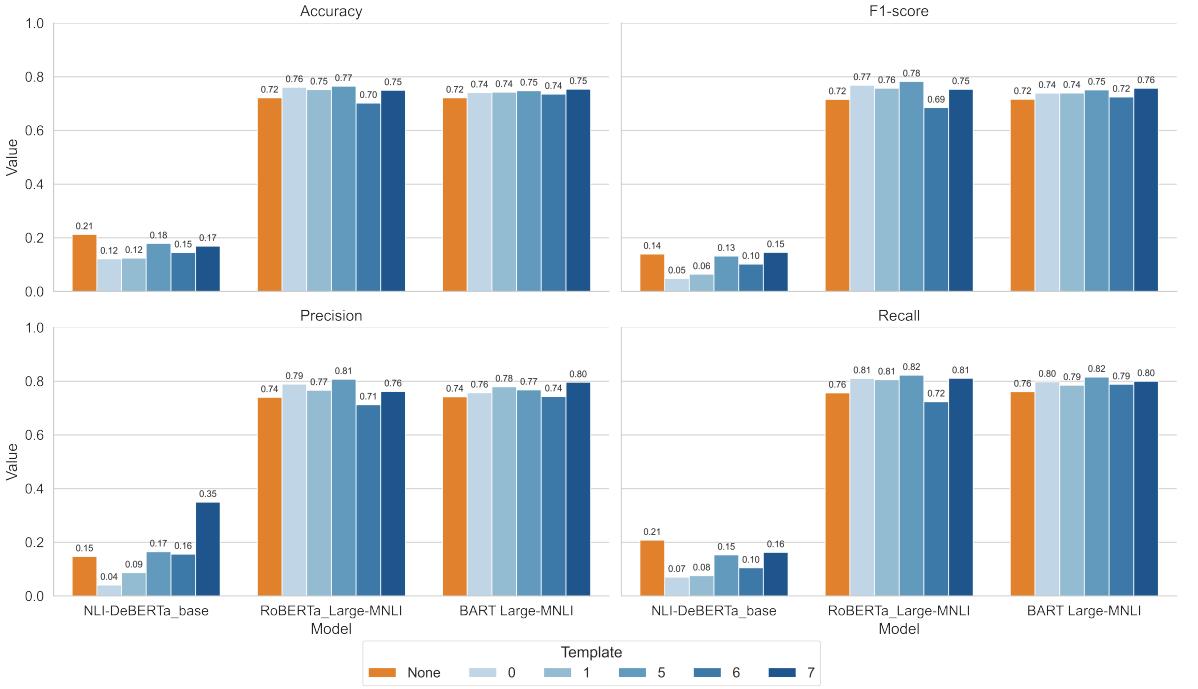


Figure 16: Performance scores for MS-CXR dataset using natural language inference, disaggregated by template. Using overflow improves the large models, impacting the F1-score on -3.43 , 2.78 , 2.18 on NLI-DeBERTa_{base}, RoBERTa_{LARGE-MNLI}, and BART Large-MNLI, respectively. Template 5 is a good choice for the large models (6.65 for RoBERTa_{LARGE-MNLI} and 3.49 for BART Large-MNLI), while template 6 is not (-3.09 for RoBERTa_{LARGE-MNLI} and 0.86 for BART Large-MNLI). The choice of template can represent that RoBERTa_{LARGE-MNLI} is better than BART Large-MNLI, as they both have performances quite similar.

C.1.3 Multiple choice question answering

Results are depicted in Figs. 17 to 20. The mapping between the models and their ID is

m14: T5-V1.1-Base	m21: Flan-T5-XXL	m31: GPT-J 6B	m38: LLaMA 2-7B
m15: T5-V1.1-Large	m22: T0-3B	m32: Instruct GPT-J	m39: Alpaca 7B
m16: T5-V1.1-3B	m23: T0++	m33: Falcon-7B	m40: LLaMA 2-CHAT-7B
m17: T5-V1.1-11B	m24: ClinicalT5-base	m34: Falcon-7B-Instruct	m52: MedAlpaca 7b
m18: Flan-T5-Base	m25: ClinicalT5-large	m35: MPT-7B	
m19: Flan-T5-Large	m29: Palmyra Base 5B	m36: MPT-7B-Instruct	
m20: Flan-T5-XL	m30: Camel 5B	m37: LLaMA-7B	

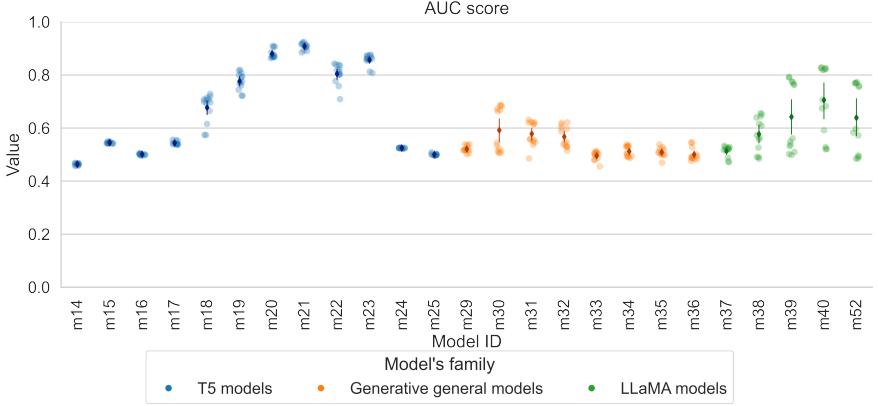


Figure 17: Performance scores for transcriptions dataset using multiple-choice question answering. The T5 family of models represents text-to-text models, whereas the rest of the models represent autoregressive models. Instruction tuning usually leads to a performance increase, with an impact of 21.45 on the AUC score when comparing the best performance per model. Considering the model size, text-to-text models perform similarly or better than their autoregressive counterparts, and **Flan-T5-XXL** is the best-performing model of all. Regarding the sensitivity of the models to the prompt used, the high sensitivity of the models is reflected in the clustering of the intra-model performance together with the visible variability of the latter.

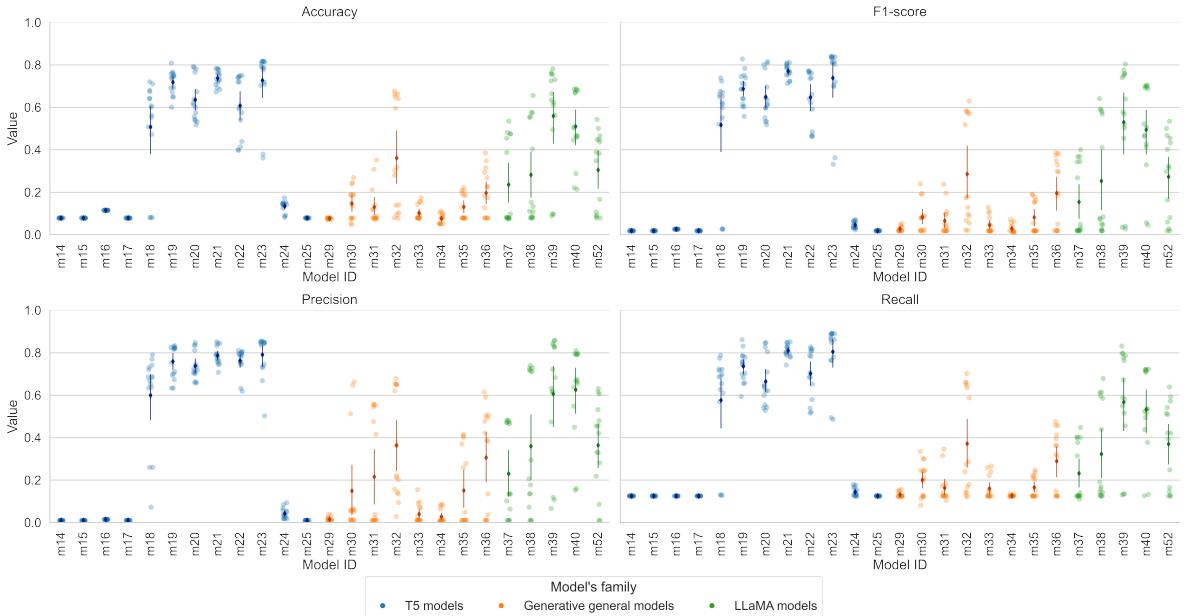


Figure 18: Performance scores for MS-CXR dataset using multiple-choice question answering. The T5 family of models represents text-to-text models, whereas the rest of the models represent autoregressive models. Instruction-tuning leads to a performance increase, except for **Falcon-7B**, with an impact of 43.55 on the AUC score when comparing the best performance per model. Taking into account the model size, text-to-text models perform similarly or better than their autoregressive counterparts, having that T0++ achieves the best scores for accuracy, F1-score, and recall. However, **Alpaca 7B** does it for precision. On the other hand, models that are not suitable for this task are T5 models, in all their size versions, and **ClinicalT5-large**. Regarding the sensitivity of the models to the prompt used, the high sensitivity of the models is reflected in the clustering of the intra-model performance together with the visible variability of the latter.

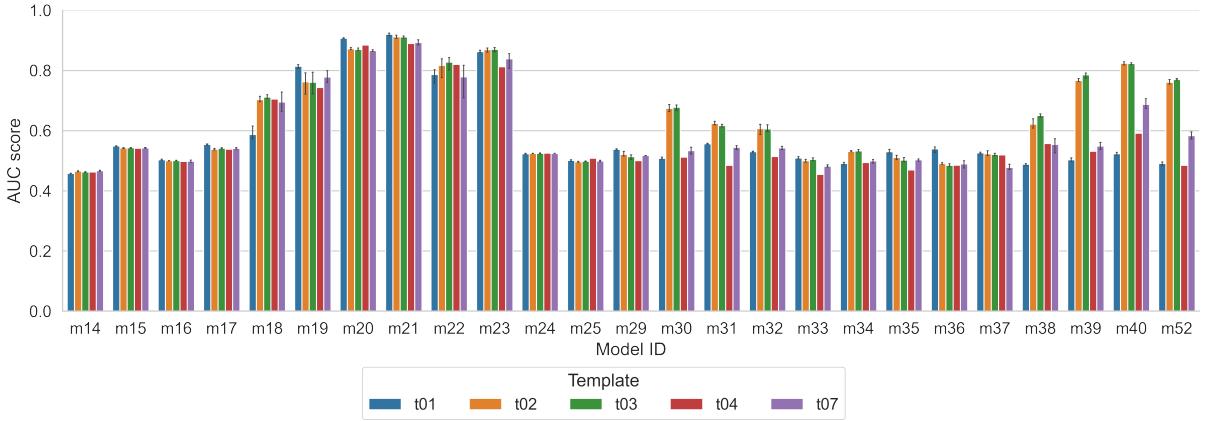


Figure 19: Prompt analysis for the transcriptions dataset using multiple-choice question answering, disaggregated by template. Neither a template nor a question works best for all the models. Regarding the impact measured in terms of standard deviations, templates have an average impact of 3.84. On its side, questions have an impact of 0.80 on a template. Thus, the template’s wording plays a more important role than the questions itself. In general, prompting has a great impact on performance.

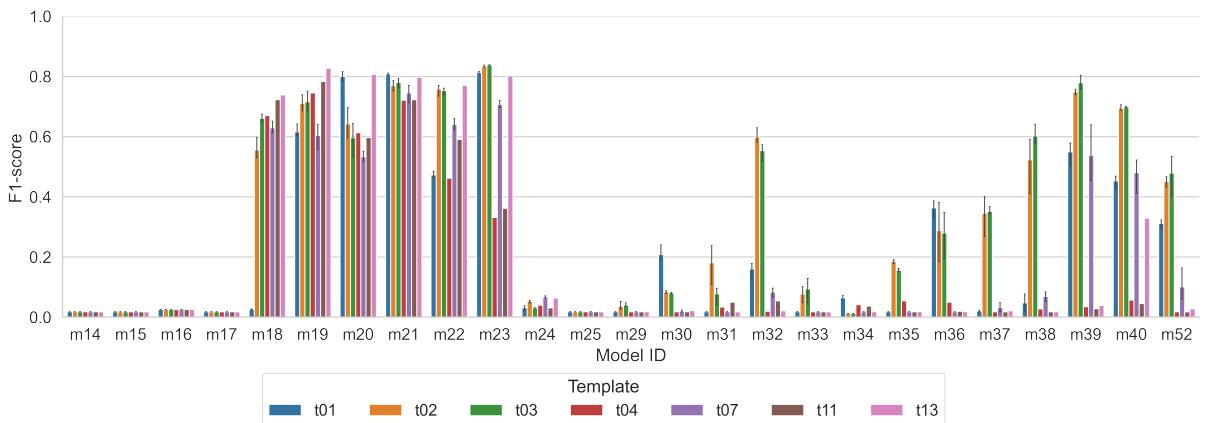


Figure 20: Prompt analysis for the MS-CXR dataset using multiple-choice question answering, disaggregated by template. Neither a template nor a question works best for all the models. However, some templates are least suitable for this task and dataset. For example, template 4 is not good for the T0 and LLaMA family; template 7 is not for Flan-T5 models, and some generative no LLaMA models. Also, for some models, the role prompting strategy does not give good results, with emphasis in template 11. Regarding the impact, measured in terms of F1-score standard deviations, templates have an average impact of 11.02. On its side, questions have an impact of 1.74 on a template. Thus, the template’s wording plays a more important role than the questions itself. In general, prompting greatly impacts performance, even decisive in terms of the ranking of models.