Towards Improved Recall in Medical Document Summarization: The GoLLIE Approach

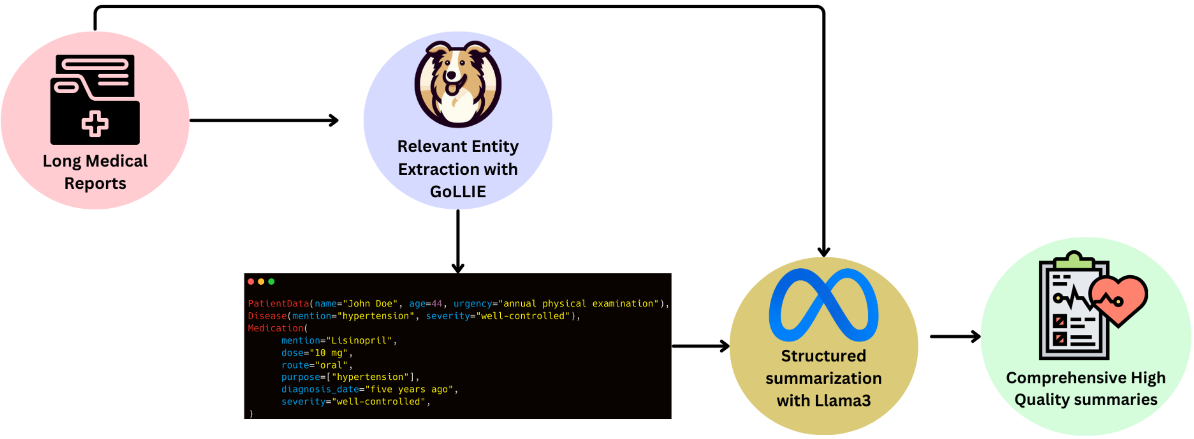
Neil De La Fuente[1,2], Joan Samper[1,2], and Daniel Vidal[1,2]

[1]Computer Vision Center
[2]Universitat Autònoma de Barcelona

May 28, 2024

Abstract

Maintaining the accuracy of extracted information in medical document summarization is crucial due to the potential consequences of errors. Errors such as false positives, where incorrect information is included, false negatives, where important information is omitted, and hallucinations, where the system generates information that was not present in the original text, can lead to significant issues, such as misdiagnoses, inappropriate treatments, and overall compromised patient safety. This project leverages GoLLIE [9], a Guideline-following Large Language Model for Information Extraction, to enhance recall by identifying key entities and essential details in medical texts. GoLLIE uses specific guidelines to ensure no critical information is omitted. The extracted entities and details are used to generate structured summaries using a few-shot learning approach on Llama3 [2]. This method eliminates the need for extensive retraining of the summarizing LLM. Code and demo are publicly available: *github.com/Neilus03/recsum*.



Long Medical Reports
Relevant Entity Extraction with GoLLIE
Structured summarization with Llama3
Comprehensive High Quality summaries

```
PatientData(name="John Doe", age=44, urgency="annual physical examination"),
Disease(mention="hypertension", severity="well-controlled"),
Medication(
      mention="Lisinopril",
      dose="10 mg",
      route="oral",
      purpose=["hypertension"],
      diagnosis_date="five years ago",
      severity="well-controlled",
)
```

Figure 1: Summarization Pipeline.

1 Motivation

In the healthcare landscape, the huge amounts of data generated through patient interactions, diagnostics, and treatments present both opportunities and challenges. One critical challenge is the effective and efficient communication of essential information. Medical reports, document patient histories, diagnostic results, and treatment plans, are essential to communicate the state and diseases to the patients. However, their complexity and length can present a difficulty in understanding and accessibility for patients, doctors, and other stakeholders in the medical sector.

Condensing medical reports into clear, concise summaries offers significant benefits across the healthcare system. For patients, this enhances comprehension, enabling them to better understand their health conditions, treatment options, and necessary follow-up care. Improved understanding leads to higher patient satisfaction and better adherence to treatment plans, ultimately resulting in better health outcomes.

For doctors, summarized reports allow quick assimilation of critical information, saving time and reducing workload. This efficiency enables more focused patient interactions and more accurate clinical decision-making.

Health institutions benefit from improved information flow, enhancing operational efficiency, reducing costs associated with miscommunication and redundant procedures, and improving overall quality of care. Furthermore, summarized data is valuable for medical research, public health monitoring, and policy-making, driving improvements in the healthcare system as a whole.

In this paper, we present our approach to leveraging GoLLIE for extracting and summarizing essential information from medical documents to address these challenges effectively.

2 Background

Recently, significant advancements have been made in the field of text summarization, driven largely by the emergence of powerful Language Models. In the context of healthcare, these advancements hold particular promise for addressing the challenge of efficiently and accurately summarizing complex medical documents.

A common approach to this task is *extractive summarization*, which involves selecting key sentences from the original text. Traditional extractive methods have relied on simpler techniques, but recent advancements leverage the power of contextual embeddings. For example, BERTSUM [7], an adaptation of the BERT [4] model, uses its contextual understanding to identify and extract the most important sentences from medical documents.

While extractive methods are effective in maintaining factual accuracy, they may not always produce the most coherent or human-like summaries. *Abstractive summarization*, on the other hand, aims to generate new sentences that capture the essence of the original text, leading to more fluent and readable summaries. Powerful language models like Pegasus [11], BART [6], and T5 [8] have demonstrated strong performance in abstractive summarization tasks by being fine-tuned on medical datasets. However, ensuring that these abstractive summaries remain faithful to the original text and avoid generating inaccurate information is an ongoing challenge. Efforts like the FaMeSumm [12] framework address this by fine-tuning pre-trained language models on domain-specific data to improve faithfulness in medical summaries.

The emergence of Large Language Models (LLMs) like *Gemini 1.5* and *GPT-4*, and their open-source counterparts, such as *Mistral* and *Llama-2*, has significantly advanced text summarization capabilities. These models exhibit remarkable accuracy and generate contextually relevant summaries.

To effectively utilize these LLMs, several techniques have been developed. In-context learning provides the model with examples within the prompt itself, allowing it to learn patterns and generate summaries tailored to the desired output without requiring modifications to the model's weights. This approach is highly efficient and leverages the extensive pre-trained knowledge of these LLMs.

Another powerful technique is low-rank adaptation (LoRA) [5], a fine-tuning method that focuses on adjusting a small subset of model weights, making it computationally efficient while still significantly enhancing performance on specific tasks. The introduction of QLoRA [3], which incorporates 4-bit quantization, further optimizes this process, enabling the efficient fine-tuning of even larger models.

Despite the advancements and capabilities of LLMs, they still face significant challenges, particularly with hallucinations and recall accuracy. Hallucinations occur when models produce information that sounds plausible but is incorrect or misleading. This is especially problematic in the medical field, where accuracy is crucial. Various methods to mitigate hallucinations have been explored, as detailed in Hallucination is Inevitable [10], but they often fall short due to computational limitations and the complexity of real-world data. These limitations mean that hallucinations cannot be entirely avoided, requiring careful oversight when using these models.

Additionally, LLMs can struggle with recall, sometimes failing to retrieve all relevant information accurately. This can compromise the completeness and reliability of the summaries they generate.

To address the challenges aforementioned in medical summarization, our approach combines the strengths of powerful LLMs with domain-specific models. Specifically, our method incorporates GoLLIE, a guideline-following LLM for precise information extraction, and Llama3, which enhances the model's ability to produce accurate and comprehensive summaries. By integrating these techniques, our approach aims to significantly improve the reliability and fidelity of medical document summarization, ensuring both accuracy and completeness.

3 Proposed Approach

This paper introduces a novel approach to improve recall in medical document summarization, addressing the critical need for accurate and complete information extraction in clinical settings. Our method, leverages the strengths of two powerful language models: GoLLIE (Guideline-following Large Language Model for Information Extraction) and Llama3. The approach consists of two main stages:

- **Information Extraction with GoLLIE:** GoLLIE, guided by a set of predefined guidelines tailored for medical documents, identifies and extracts key entities (e.g., patient demographics, diagnoses, medications) and essential details (e.g., symptoms, treatment plans, test results). Further detailes on guidelines are given in the Appendix A.1.

- **Structured Summarization with Llama3:** The extracted information is then structured into a predefined format along with the text to summarize and fed to Llama3, which generates a concise and informative summary using a few-shot learning approach.

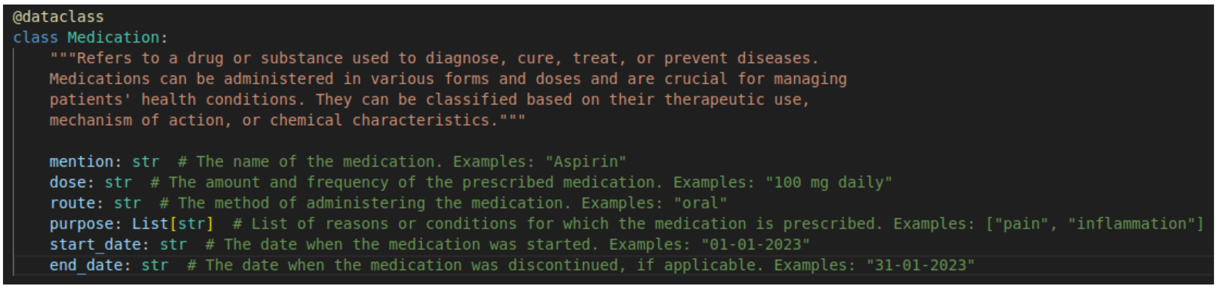
By combining GoLLIE's precision in information extraction with Llama3's ability to generate coherent summaries, our method aims to improve recall and ensure that no crucial information is omitted in the summarization process. A diagram of the summary generation pipeline is shown in Figure 1.

3.1 GoLLIE for Information Extraction

As mentioned, GoLLIE is a large language model specifically designed for information extraction, particularly in domains where accuracy and completeness are very important. Unlike traditional information extraction techniques that rely heavily on rule-based systems or require extensive labeled data for supervised learning, GoLLIE utilizes a guideline-following approach, allowing it to be applied in a zero-shot fashion on domains where it wasn't explicitly trained on.

At its core, GoLLIE is a finetuned version of CodeLlama [1], which is trained on a massive dataset of text and code, enabling it to understand natural language and code-like instructions. This allows us to provide GoLLIE with specific guidelines that outline the key entities and details to extract from medical documents. These guidelines are defined using a Python library called 'Data Classes' that allows to define data structures in a clear and concise way. Using this library, we can express the guidelines in a structured format.

An example of a guideline used to extract medication information can be seen below:



```
@dataclass
class Medication:
    """Refers to a drug or substance used to diagnose, cure, treat, or prevent diseases.
    Medications can be administered in various forms and doses and are crucial for managing
    patients' health conditions. They can be classified based on their therapeutic use,
    mechanism of action, or chemical characteristics."""

    mention: str  # The name of the medication. Examples: "Aspirin"
    dose: str  # The amount and frequency of the prescribed medication. Examples: "100 mg daily"
    route: str  # The method of administering the medication. Examples: "oral"
    purpose: List[str]  # List of reasons or conditions for which the medication is prescribed. Examples: ["pain", "inflammation"]
    start_date: str  # The date when the medication was started. Examples: "01-01-2023"
    end_date: str  # The date when the medication was discontinued, if applicable. Examples: "31-01-2023"
```

Figure 2: Data Class definition for Medication extraction

In the particular case shown in Figure 2, we are providing GoLLIE instructions to extract all mentions of drugs that follow this structure. Each data class is composed by an arbitrary number of attributes, in this case, if GoLLIE detects a mention of a drug, it would proceed to extract, if available, its dosage, administration route and purpose, along with its start and end date. This process would be repeated for every data class defined. A total of 9 data classes were defined (see APPENDIX for more detailed information).

Using this approach, we can efficiently extract a wide variety of structured information from medical texts, even if GoLLIE was not specifically trained on those specific entities. Additionally, we have developed an equivalent set of guidelines in Spanish, enabling the extraction of information from Spanish medical reports as well.

3.2 Llama3 for Structured Summarization

Once GoLLIE completes the information extraction phase, the extracted data is organized into a structured format. This structure can be tailored to the specific requirements of medical summarization, ensuring that the information is presented clearly and logically.

We utilize a one-shot learning approach with Llama3-70b to generate the final summary. This involves providing Llama3 with a single example that demonstrates the desired input and output format. The input consists on the extracted entities and attributes along with the original text to be summarized, while the output is variable, if we want an schematic summary, it would be a bullet-point based summary outlining main concepts, while if we prefer a prose-like summary, the output would be a concise paragraph summarizing the report. For a more in depth understanding of the prompt please refere to Appendix A.2.

Leveraging Llama3's powerful language generation capabilities in this way eliminates the need for extensive fine-tuning or retraining of the model. The one-shot learning paradigm allows Llama3 to quickly adapt to the task of generating structured summaries from the extracted medical information.

4 Experiments and Results

In this section, we present the evaluation of our proposed method using a dataset of medical reports. We compare the performance of several models, including Mistral, Llama3-8b, Gemma-7b-it, and Llama3-70b, both with and without the GoLLIE framework in a one-shot learning set up where we provide each of the LLMs an example. It is worth to note that evaluating summaries is inherently challenging due to the subjective nature of summarization and the difficulty in defining objective metrics for quality. Thus, our evaluation employs both quantitative and qualitative metrics to provide a more comprehensive analysis of the models' performance.

4.1 Evaluation Dataset

The evaluation dataset was provided by Asho Corporation. From these medical reports, we selected 10 that included ground truth summaries to compute ROUGE and BERT scores. Additionally, we manually extracted keywords from 12 medical reports to evaluate the Keyword Density and Keyword Recall metrics. This dual approach ensures a robust evaluation, capturing both the fidelity and the completeness of the summaries generated by the models. The selected reports include a diverse set of medical cases, ensuring that the evaluation covers a wide range of clinical scenarios and terminologies.

4.2 Keyword Density

Keyword Density measures the proportion of important keywords in the summary relative to the total number of words. This metric ensures that the summary is concise while retaining essential information. A higher keyword density indicates a more information-rich summary, which is crucial for the clarity and usefulness of medical document summarization. A detailed explanation of the metric is provided in A.4.2

For instance, in a medical context, keywords could include critical terms such as diagnoses, symptoms, medications, and treatment plans. Ensuring these keywords are present in the summary is vital for maintaining the informational integrity of the document. A summary with high keyword density is not only brief but also highly informative, enabling medical professionals to quickly grasp the essential details.

Figure 3 illustrates the keyword density for different models with and without the GoLLIE framework. Models integrated with GoLLIE consistently exhibit higher keyword density compared to raw models, indicating that GoLLIE helps generate more concise summaries without sacrificing critical information. This improvement can be attributed to GoLLIE's ability to follow predefined guidelines that prioritize the extraction and retention of important medical terms.

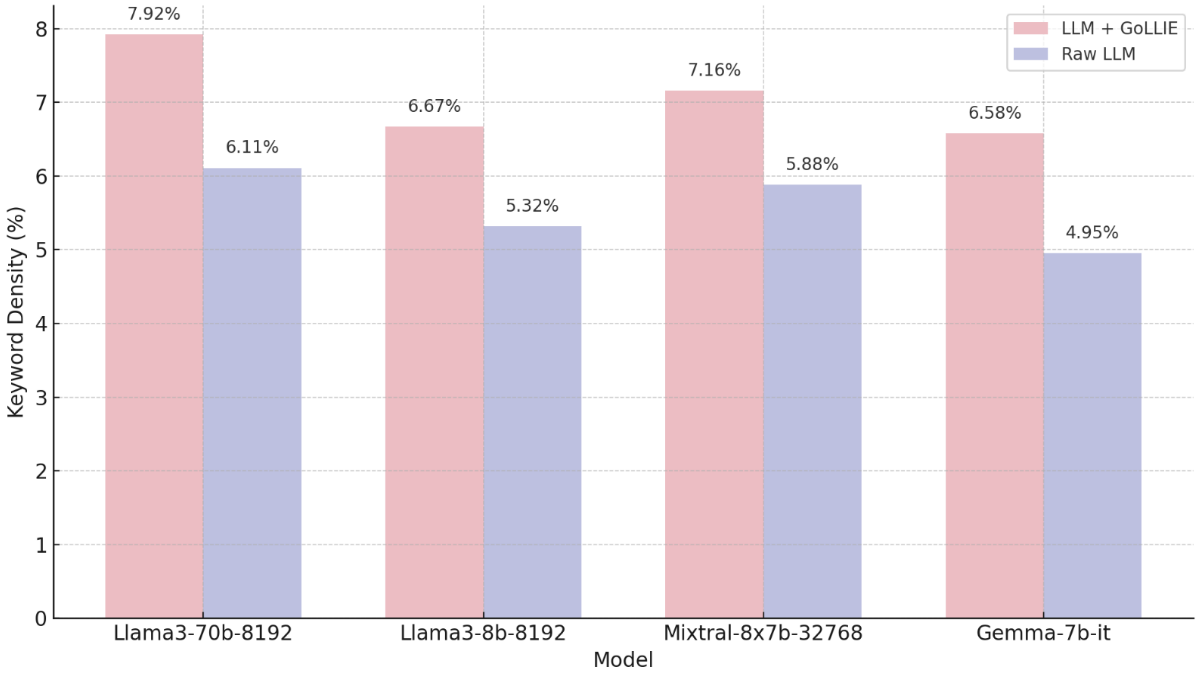


Figure 3: Keyword Density for Different Models With and Without GoLLIE

4.3 Keyword Recall

Keyword Recall measures the proportion of essential keywords retained in the summary compared to the ground truth. It ensures that no important information is omitted. This metric is particularly vital in the medical field, where the omission of critical information can lead to significant negative outcomes, such as incorrect diagnoses or inappropriate treatment plans.

To calculate keyword recall, we compare the keywords present in the generated summary against those identified by human experts in the ground truth. A higher keyword recall indicates that the summary retains most of the crucial information, making it reliable for clinical use. A deeper explanation of the metric can be found in A.4.2.

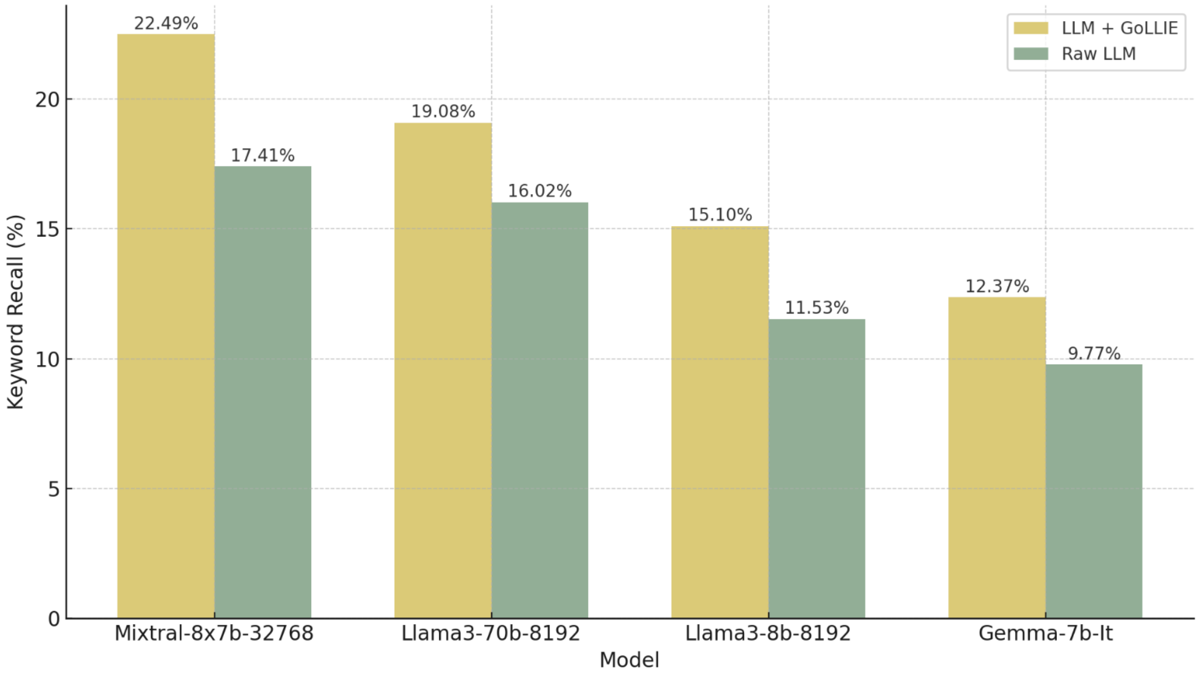


Figure 4: Keyword Recall for Different Models With and Without GoLLIE

Figure 4 shows that GoLLIE-enhanced models outperform raw models in keyword recall, demonstrating better retention of critical information. This improvement suggests that GoLLIE effectively captures and preserves essential details in the summaries, making them more useful and accurate for medical purposes.

4.4 ROUGE Scores

ROUGE scores evaluate the quality of the generated summaries by comparing them to ground truth summaries. We compute ROUGE-1, ROUGE-2, and ROUGE-L scores to assess the overlap of unigrams, bigrams, and longest common subsequence respectively. These scores are essential for evaluating the fluency and readability of the summaries.

ROUGE-1 measures the overlap of individual words, ROUGE-2 measures the overlap of two consecutive words, and ROUGE-L considers the longest common subsequence, which helps evaluate the syntactic structure and coherence of the summary. Higher ROUGE scores indicate better quality summaries that are closer to human-written ground truth.

Figure 5 shows the average ROUGE scores for the Llama3-70b model with and without GoLLIE. The results indicate that GoLLIE does not significantly affect ROUGE scores but maintains the same quality as raw models. This demonstrates that the guidelines provided by GoLLIE do not degrade the quality of the summaries, ensuring that the generated text remains coherent and fluent.

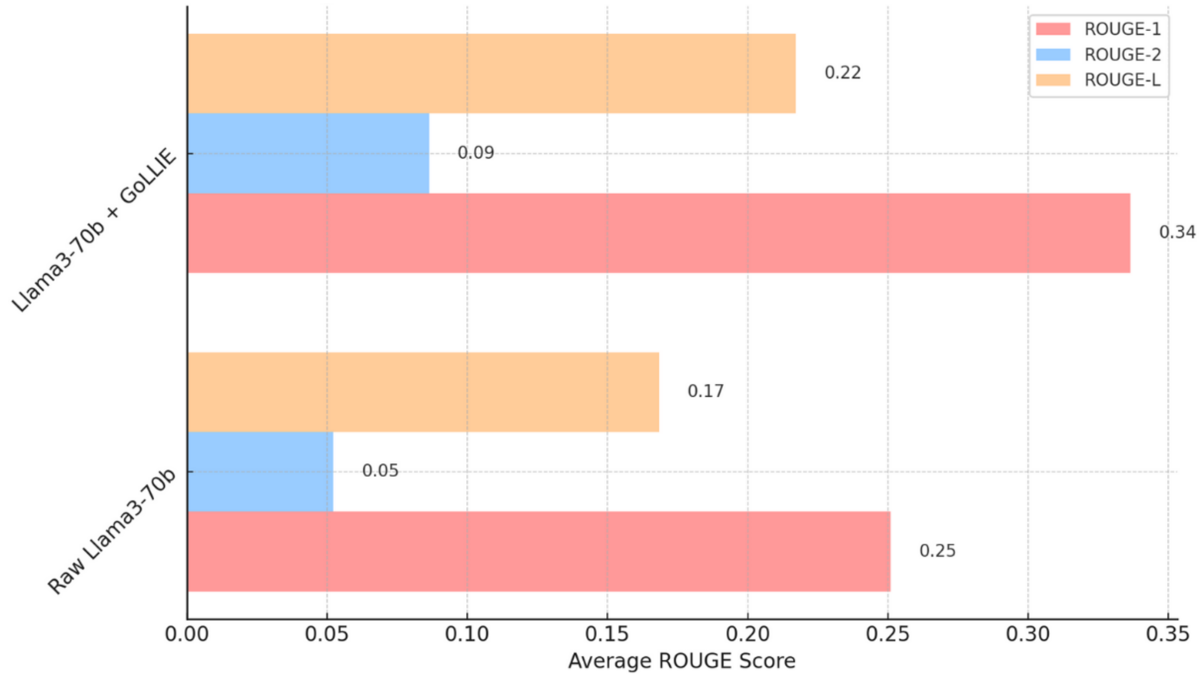


Figure 5: Average ROUGE Scores for Llama3-70b With and Without GoLLIE

4.5 BERT Scores

BERT scores measure the semantic similarity between the generated summaries and ground truth, considering precision, recall, and F1 score. These metrics provide a deeper insight into how well the summaries capture the meaning and important details of the original text. Higher BERT scores indicate summaries that are semantically similar to the ground truth, ensuring that the meaning and key information are preserved.

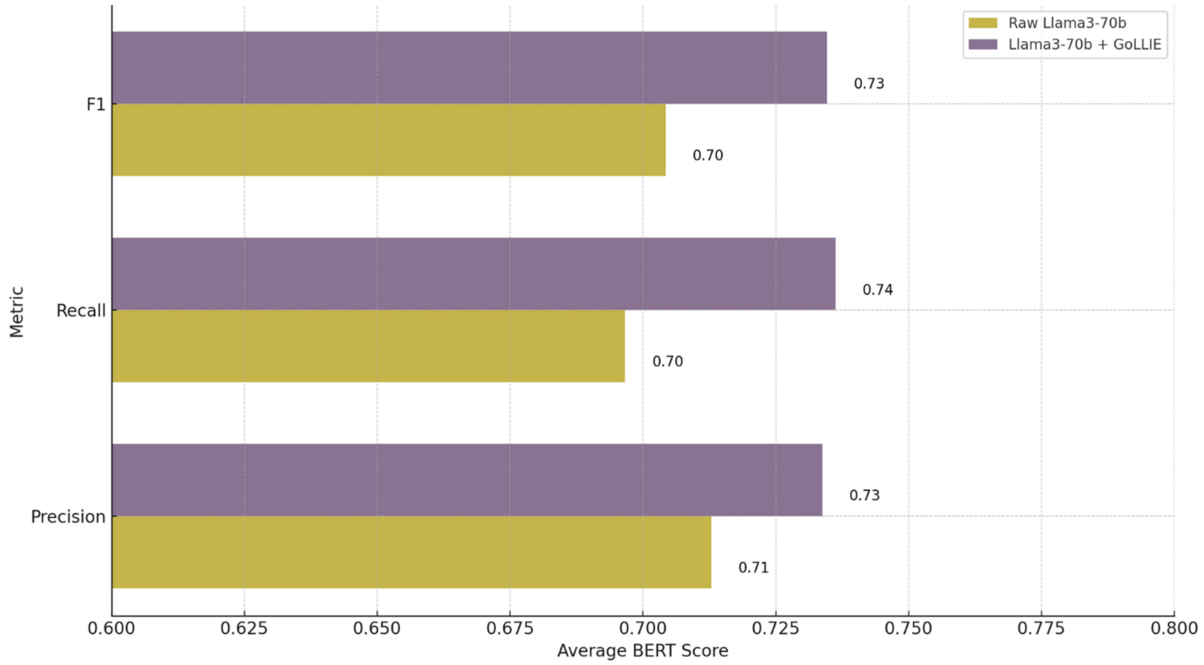


Figure 6: Average BERT Scores for Llama3-70b With and Without GoLLIE

Figure 6 presents the average BERT scores for the Llama3-70b model with and without

GoLLIE. The results indicate that GoLLIE slightly improves recall and F1 score while maintaining high precision. This ensures that the summaries are not only accurate but also comprehensive, capturing a broader range of relevant information.

4.6 Qualitative Analysis

From a qualitative perspective, GoLLIE-enhanced summaries include more relevant keywords and critical information with fewer words, as evidenced by the keyword density and recall metrics. This enhancement leads to summaries that are not only shorter but also more information-dense and useful for medical professionals. The qualitative analysis involves examining specific examples of summaries generated by the models to assess their practical utility and informativeness, Some examples are provided in A.4.3. Additionally, rapid experiments can be made using the web-app demo available in the GitHub repository.

The examples highlight the differences in the summaries generated with and without GoLLIE. The GoLLIE-enhanced summary is generally more concise and focused on essential details, demonstrating the effectiveness of our approach in extracting and retaining critical information while maintaining brevity.

5 Discussion

This research introduces a novel approach to addressing the persistent challenges of recall and accuracy in medical document summarization, leveraging the strengths of GoLLIE and Llama3. Existing methods, such as LLMs, despite advancements in abstractive summarization, often struggle with issues like faithfulness, hallucinations, and accurate capture of critical information. Our proposed method aims to mitigate these limitations through a two-stage approach that prioritizes information extraction and structured summarization.

The core contribution of this research lies in the integration of GoLLIE, a guideline-following LLM specifically designed for precise information extraction. By utilizing predefined guidelines tailored to medical documents, GoLLIE extracts key entities and details, ensuring no critical information is overlooked. This approach, unlike traditional methods reliant on rule-based systems or extensive labeled data, enables zero-shot application across diverse medical domains, enhancing the model's versatility and efficiency.

Furthermore, the integration of Llama3, a powerful 70-billion-parameter LLM for natural language generation, facilitates the production of structured summaries based on the extracted information. This approach utilizes a one-shot learning paradigm, eliminating the need for extensive model retraining and allowing flexible adaptability to different summarization formats (e.g., bullet points, concise paragraphs). This dual-stage process ensures that the final summaries are both comprehensive, coherent and structured in the desired way.

The potential implications of this approach are significant. By enhancing recall and accuracy in medical summarization, our method can contribute to several key areas:

- **Improved patient comprehension:** Concise and comprehensive summaries can empower patients to better understand their diagnoses, treatments, and follow-up care. This leads to increased satisfaction and improved adherence to medical plans, ultimately enhancing health outcomes.

- **Enhanced clinical decision-making:** Healthcare professionals benefit from efficient access to critical information, enabling more accurate assessments, optimized treatment decisions, and improved patient outcomes. Quick assimilation of critical data allows for more focused patient interactions and effective clinical workflows.
- **Streamlined information flow:** Efficient summarization optimizes information flow within healthcare institutions, reducing costs associated with miscommunication and redundant procedures. Improved information flow enhances operational efficiency and overall quality of care.

While the proposed approach shows promising advancements, it is crucial to acknowledge potential limitations, some of them are stated below:

- **Data availability and quality:** The effectiveness of GoLLIE's information extraction relies on the availability of comprehensive and well-structured medical data. Addressing data biases and limitations remains a crucial aspect of future research. Efforts should focus on expanding datasets to cover diverse medical scenarios and ensuring data quality.
- **Guideline development and maintenance:** Defining and maintaining comprehensive guidelines for GoLLIE, covering a wide range of medical domains and evolving clinical practices, is an ongoing challenge. Research into automated guideline generation and validation is necessary to ensure ongoing adaptability and effectiveness. Developing adaptive guidelines that evolve with medical advancements will be crucial.

Future research directions will focus on further enhancing the capabilities of our approach. Some potential improvements include the development of automated generation and validation pipelines to enhance adaptability and reduce human effort. This could involve Reinforcement Learning from Human Feedback approaches to continuously update and refine guidelines based on new data and further optimize Llama3's summarization capabilities and improve alignment of generated summaries with clinical needs. This approach can help the model learn from practical applications and continuously improve its performance. Finally, a more rigorous and extended evaluation and benchmarking of our approach against existing summarization methods is essential. Benchmarking on standardized medical summarization datasets will be crucial for validating the effectiveness and robustness of our approach as well.

As presented in the results section, the GoLLIE approach surpasses the raw LLM approach across all evaluated metrics. These results indicate that our method serves as a foundation for advancing the field of medical document summarization. By addressing the limitations of current LLMs and introducing a novel approach that prioritizes recall and accuracy, our method has the potential to significantly improve the communication and utilization of critical medical information.

In conclusion, this research demonstrates a significant step forward in medical document summarization by combining precise information extraction with coherent summary generation. The integration of GoLLIE and Llama3 addresses key challenges in the field, providing a robust solution that enhances patient comprehension, clinical decision-making, and information flow within healthcare institutions. Future work will continue to build on these findings, exploring new methodologies and technologies to further improve the effectiveness and scalability of our approach.

6 Conclusion

This research presents a method to improve recall in medical document summarization using GoLLIE for precise information extraction and Llama3 for structured summarization. Our approach enhances patient comprehension, clinical decision-making, and information flow in healthcare. Future work will focus on addressing data and guideline challenges to further refine our method and validate its effectiveness.

References

[1] Meta AI. Code llama: Open foundation models for code. `https://ai.meta.com/research/publications/code-llama-open-foundation-models-for-code/`, 2023.

[2] Meta AI. Introducing meta llama 3: The most capable openly available llm to date. `https://ai.meta.com/blog/meta-llama-3/`, 2024.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[7] Yang Liu. Fine-tune bert for extractive summarization, 2019.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[9] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*, 2024.

[10] Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *ArXiv*, abs/2401.11817, 2024.

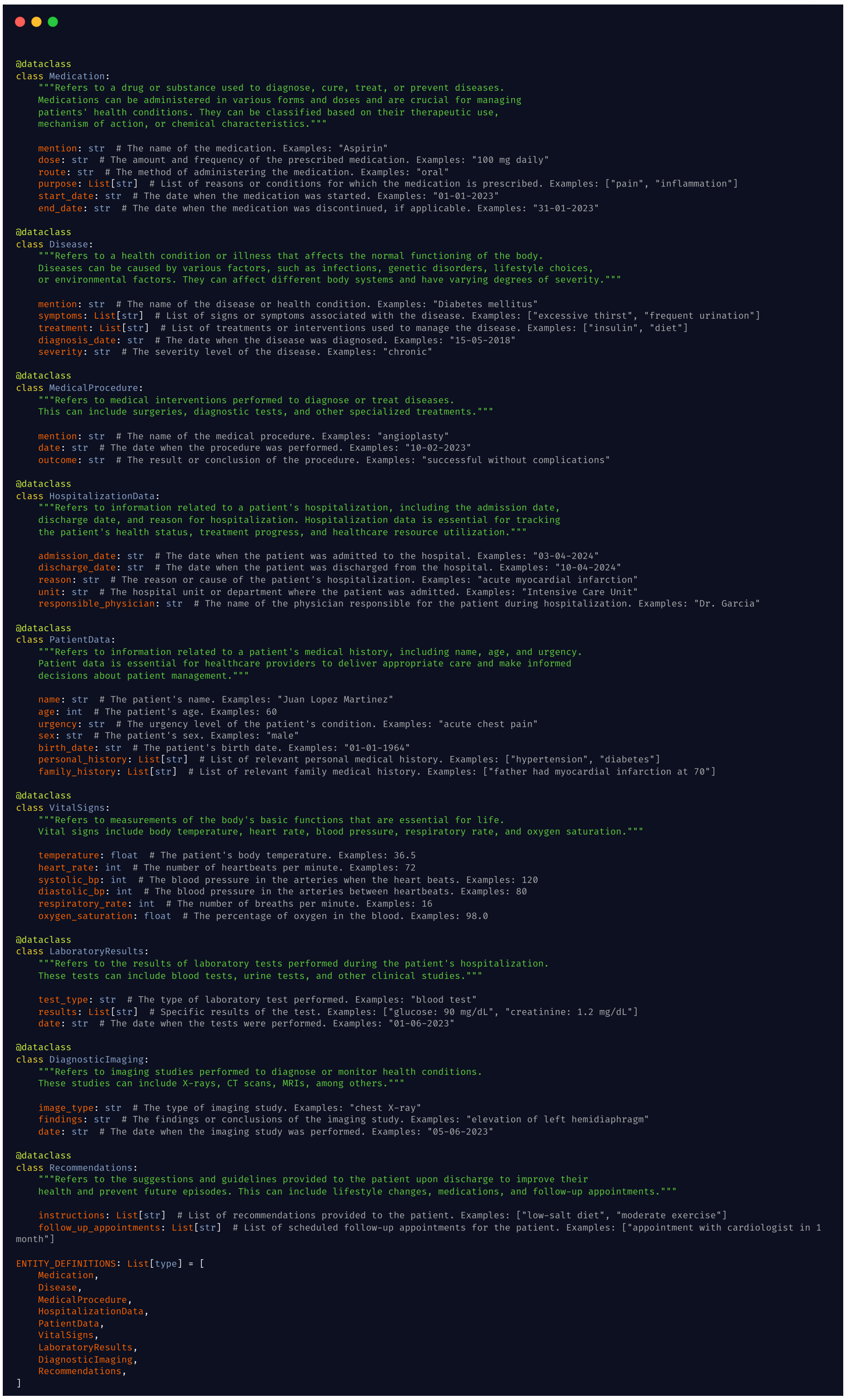
[11] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.

[12] Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra, and Rui Zhang. FaMeSumm: Investigating and improving faithfulness of medical summarization. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10915–10931, Singapore, December 2023. Association for Computational Linguistics.

A Appendix

A.1 GoLLIE Guidelines

The following data classes were used to define the extraction guidelines for GoLLIE:



```python
@dataclass
class Medication:
    """Refers to a drug or substance used to diagnose, cure, treat, or prevent diseases.
    Medications can be administered in various forms and doses and are crucial for managing
    patients' health conditions. They can be classified based on their therapeutic use,
    mechanism of action, or chemical characteristics."""

    mention: str  # The name of the medication. Examples: "Aspirin"
    dose: str  # The amount and frequency of the prescribed medication. Examples: "100 mg daily"
    route: str  # The method of administering the medication. Examples: "oral"
    purpose: List[str]  # List of reasons or conditions for which the medication is prescribed. Examples: ["pain", "inflammation"]
    start_date: str  # The date when the medication was started. Examples: "01-01-2023"
    end_date: str  # The date when the medication was discontinued, if applicable. Examples: "31-01-2023"

@dataclass
class Disease:
    """Refers to a health condition or illness that affects the normal functioning of the body.
    Diseases can be caused by various factors, such as infections, genetic disorders, lifestyle choices,
    or environmental factors. They can affect different body systems and have varying degrees of severity."""

    mention: str  # The name of the disease or health condition. Examples: "Diabetes mellitus"
    symptoms: List[str]  # List of signs or symptoms associated with the disease. Examples: ["excessive thirst", "frequent urination"]
    treatment: List[str]  # List of treatments or interventions used to manage the disease. Examples: ["insulin", "diet"]
    diagnosis_date: str  # The date when the disease was diagnosed. Examples: "15-05-2018"
    severity: str  # The severity level of the disease. Examples: "chronic"

@dataclass
class MedicalProcedure:
    """Refers to medical interventions performed to diagnose or treat diseases.
    This can include surgeries, diagnostic tests, and other specialized treatments."""

    mention: str  # The name of the medical procedure. Examples: "angioplasty"
    date: str  # The date when the procedure was performed. Examples: "10-02-2023"
    outcome: str  # The result or conclusion of the procedure. Examples: "successful without complications"

@dataclass
class HospitalizationData:
    """Refers to information related to a patient's hospitalization, including the admission date,
    discharge date, and reason for hospitalization. Hospitalization data is essential for tracking
    the patient's health status, treatment progress, and healthcare resource utilization."""

    admission_date: str  # The date when the patient was admitted to the hospital. Examples: "03-04-2024"
    discharge_date: str  # The date when the patient was discharged from the hospital. Examples: "10-04-2024"
    reason: str  # The reason or cause of the patient's hospitalization. Examples: "acute myocardial infarction"
    unit: str  # The hospital unit or department where the patient was admitted. Examples: "Intensive Care Unit"
    responsible_physician: str  # The name of the physician responsible for the patient during hospitalization. Examples: "Dr. Garcia"

@dataclass
class PatientData:
    """Refers to information related to a patient's medical history, including name, age, and urgency.
    Patient data is essential for healthcare providers to deliver appropriate care and make informed
    decisions about patient management."""

    name: str  # The patient's name. Examples: "Juan Lopez Martinez"
    age: int  # The patient's age. Examples: 60
    urgency: str  # The urgency level of the patient's condition. Examples: "acute chest pain"
    sex: str  # The patient's sex. Examples: "male"
    birth_date: str  # The patient's birth date. Examples: "01-01-1964"
    personal_history: List[str]  # List of relevant personal medical history. Examples: ["hypertension", "diabetes"]
    family_history: List[str]  # List of relevant family medical history. Examples: ["father had myocardial infarction at 70"]

@dataclass
class VitalSigns:
    """Refers to measurements of the body's basic functions that are essential for life.
    Vital signs include body temperature, heart rate, blood pressure, respiratory rate, and oxygen saturation."""

    temperature: float  # The patient's body temperature. Examples: 36.5
    heart_rate: int  # The number of heartbeats per minute. Examples: 72
    systolic_bp: int  # The blood pressure in the arteries when the heart beats. Examples: 120
    diastolic_bp: int  # The blood pressure in the arteries between heartbeats. Examples: 80
    respiratory_rate: int  # The number of breaths per minute. Examples: 16
    oxygen_saturation: float  # The percentage of oxygen in the blood. Examples: 98.0

@dataclass
class LaboratoryResults:
    """Refers to the results of laboratory tests performed during the patient's hospitalization.
    These tests can include blood tests, urine tests, and other clinical studies."""

    test_type: str  # The type of laboratory test performed. Examples: "blood test"
    results: List[str]  # Specific results of the test. Examples: ["glucose: 90 mg/dL", "creatinine: 1.2 mg/dL"]
    date: str  # The date when the tests were performed. Examples: "01-06-2023"

@dataclass
class DiagnosticImaging:
    """Refers to imaging studies performed to diagnose or monitor health conditions.
    These studies can include X-rays, CT scans, MRIs, among others."""

    image_type: str  # The type of imaging study. Examples: "chest X-ray"
    findings: str  # The findings or conclusions of the imaging study. Examples: "elevation of left hemidiaphragm"
    date: str  # The date when the imaging study was performed. Examples: "05-06-2023"

@dataclass
class Recommendations:
    """Refers to the suggestions and guidelines provided to the patient upon discharge to improve their
    health and prevent future episodes. This can include lifestyle changes, medications, and follow-up appointments."""

    instructions: List[str]  # List of recommendations provided to the patient. Examples: ["low-salt diet", "moderate exercise"]
    follow_up_appointments: List[str]  # List of scheduled follow-up appointments for the patient. Examples: ["appointment with cardiologist in 1
month"]

ENTITY_DEFINITIONS: List[type] = [
    Medication,
    Disease,
    MedicalProcedure,
    HospitalizationData,
    PatientData,
    VitalSigns,
    LaboratoryResults,
    DiagnosticImaging,
    Recommendations,
]
```

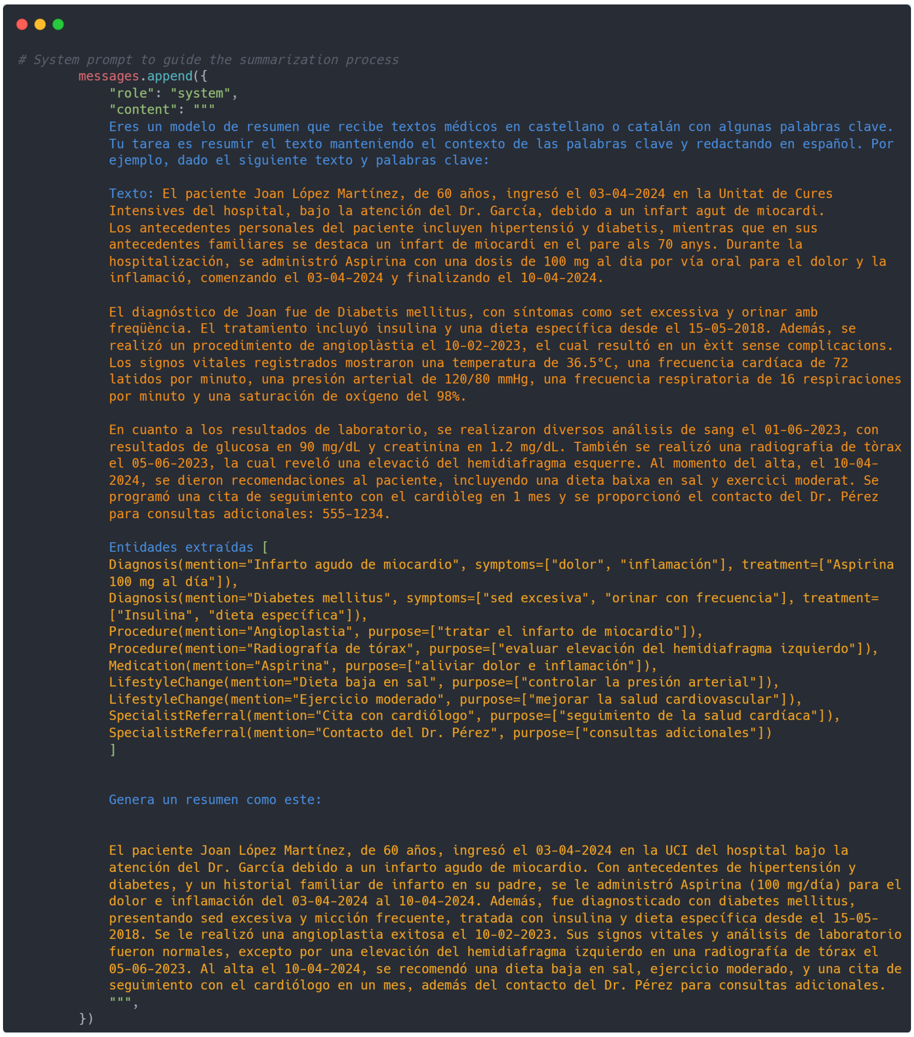
Figure 7: Medical Entities and Extraction Guidelines for GoLLIE

As it can be seen in Figure 7, each data class includes a set of attributes that specify the information to be extracted. These data classes are used to define the extraction guidelines for GoLLIE, ensuring that the model captures essential information from medical documents.

A.2 Llama3-70b on Groq

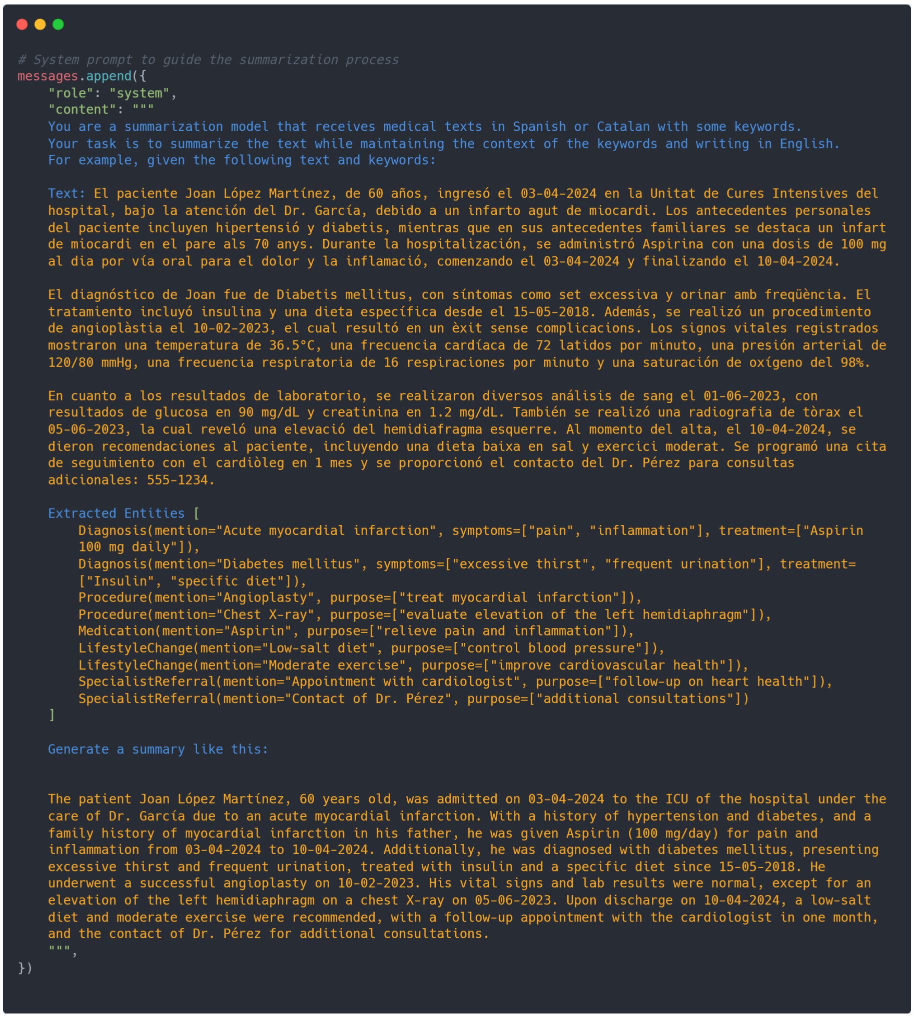
We use the Llama3-70b model on Groq, which is an advanced platform utilizing LPUs (Language Processing Units). Groq's architecture, based on LPUs, is particularly well-suited for speeding up LLM inference.

The Llama3-70b model is chosen as the most capable open-source model, making it an excellent fit for tasks requiring high accuracy and fluency in natural language processing, such as text summarization.



```
# System prompt to guide the summarization process
        messages.append({
            "role": "system",
            "content": """
            Eres un modelo de resumen que recibe textos médicos en castellano o catalán con algunas palabras clave.
            Tu tarea es resumir el texto manteniendo el contexto de las palabras clave y redactando en español. Por
            ejemplo, dado el siguiente texto y palabras clave:

            Texto: El paciente Joan López Martínez, de 60 años, ingresó el 03-04-2024 en la Unitat de Cures
            Intensives del hospital, bajo la atención del Dr. García, debido a un infart agut de miocardi.
            Los antecedentes personales del paciente incluyen hipertensió y diabetis, mientras que en sus
            antecedentes familiares se destaca un infart de miocardi en el pare als 70 anys. Durante la
            hospitalización, se administró Aspirina con una dosis de 100 mg al dia por vía oral para el dolor y la
            inflamació, comenzando el 03-04-2024 y finalizando el 10-04-2024.

            El diagnóstico de Joan fue de Diabetis mellitus, con síntomas como set excessiva y orinar amb
            freqüència. El tratamiento incluyó insulina y una dieta específica desde el 15-05-2018. Además, se
            realizó un procedimiento de angioplàstia el 10-02-2023, el cual resultó en un èxit sense complicacions.
            Los signos vitales registrados mostraron una temperatura de 36.5°C, una frecuencia cardíaca de 72
            latidos por minuto, una presión arterial de 120/80 mmHg, una frecuencia respiratoria de 16 respiraciones
            por minuto y una saturación de oxígeno del 98%.

            En cuanto a los resultados de laboratorio, se realizaron diversos análisis de sang el 01-06-2023, con
            resultados de glucosa en 90 mg/dL y creatinina en 1.2 mg/dL. También se realizó una radiografia de tòrax
            el 05-06-2023, la cual reveló una elevació del hemidiafragma esquerre. Al momento del alta, el 10-04-
            2024, se dieron recomendaciones al paciente, incluyendo una dieta baixa en sal y exercici moderat. Se
            programó una cita de seguimiento con el cardiòleg en 1 mes y se proporcionó el contacto del Dr. Pérez
            para consultas adicionales: 555-1234.

            Entidades extraídas [
            Diagnosis(mention="Infarto agudo de miocardio", symptoms=["dolor", "inflamación"], treatment=["Aspirina
            100 mg al día"]),
            Diagnosis(mention="Diabetes mellitus", symptoms=["sed excesiva", "orinar con frecuencia"], treatment=
            ["Insulina", "dieta específica"]),
            Procedure(mention="Angioplastia", purpose=["tratar el infarto de miocardio"]),
            Procedure(mention="Radiografía de tórax", purpose=["evaluar elevación del hemidiafragma izquierdo"]),
            Medication(mention="Aspirina", purpose=["aliviar dolor e inflamación"]),
            LifestyleChange(mention="Dieta baja en sal", purpose=["controlar la presión arterial"]),
            LifestyleChange(mention="Ejercicio moderado", purpose=["mejorar la salud cardiovascular"]),
            SpecialistReferral(mention="Cita con cardiólogo", purpose=["seguimiento de la salud cardíaca"]),
            SpecialistReferral(mention="Contacto del Dr. Pérez", purpose=["consultas adicionales"])
            ]


            Genera un resumen como este:


            El paciente Joan López Martínez, de 60 años, ingresó el 03-04-2024 en la UCI del hospital bajo la
            atención del Dr. García debido a un infarto agudo de miocardio. Con antecedentes de hipertensión y
            diabetes, y un historial familiar de infarto en su padre, se le administró Aspirina (100 mg/día) para el
            dolor e inflamación del 03-04-2024 al 10-04-2024. Además, fue diagnosticado con diabetes mellitus,
            presentando sed excesiva y micción frecuente, tratada con insulina y dieta específica desde el 15-05-
            2018. Se le realizó una angioplastia exitosa el 10-02-2023. Sus signos vitales y análisis de laboratorio
            fueron normales, excepto por una elevación del hemidiafragma izquierdo en una radiografía de tórax el
            05-06-2023. Al alta el 10-04-2024, se recomendó una dieta baja en sal, ejercicio moderado, y una cita de
            seguimiento con el cardiólogo en un mes, además del contacto del Dr. Pérez para consultas adicionales.
            """,
        })
```

Figure 8: System prompt in Spanish that Llama 3 receives to generate the summary



```
# System prompt to guide the summarization process
messages.append({
    "role": "system",
    "content": """
    You are a summarization model that receives medical texts in Spanish or Catalan with some keywords.
    Your task is to summarize the text while maintaining the context of the keywords and writing in English.
    For example, given the following text and keywords:

    Text: El paciente Joan López Martínez, de 60 años, ingresó el 03-04-2024 en la Unitat de Cures Intensives del
    hospital, bajo la atención del Dr. García, debido a un infarto agut de miocardi. Los antecedentes personales
    del paciente incluyen hipertensió y diabetis, mientras que en sus antecedentes familiares se destaca un infart
    de miocardi en el pare als 70 anys. Durante la hospitalización, se administró Aspirina con una dosis de 100 mg
    al dia por vía oral para el dolor y la inflamació, comenzando el 03-04-2024 y finalizando el 10-04-2024.

    El diagnóstico de Joan fue de Diabetis mellitus, con síntomas como set excessiva y orinar amb freqüència. El
    tratamiento incluyó insulina y una dieta específica desde el 15-05-2018. Además, se realizó un procedimiento
    de angioplàstia el 10-02-2023, el cual resultó en un èxit sense complicacions. Los signos vitales registrados
    mostraron una temperatura de 36.5°C, una frecuencia cardíaca de 72 latidos por minuto, una presión arterial de
    120/80 mmHg, una frecuencia respiratoria de 16 respiraciones por minuto y una saturación de oxígeno del 98%.

    En cuanto a los resultados de laboratorio, se realizaron diversos análisis de sang el 01-06-2023, con
    resultados de glucosa en 90 mg/dL y creatinina en 1.2 mg/dL. También se realizó una radiografia de tòrax el
    05-06-2023, la cual reveló una elevació del hemidiafragma esquerre. Al momento del alta, el 10-04-2024, se
    dieron recomendaciones al paciente, incluyendo una dieta baixa en sal y exercici moderat. Se programó una cita
    de seguimiento con el cardiòleg en 1 mes y se proporcionó el contacto del Dr. Pérez para consultas
    adicionales: 555-1234.

    Extracted Entities [
        Diagnosis(mention="Acute myocardial infarction", symptoms=["pain", "inflammation"], treatment=["Aspirin
        100 mg daily"]),
        Diagnosis(mention="Diabetes mellitus", symptoms=["excessive thirst", "frequent urination"], treatment=
        ["Insulin", "specific diet"]),
        Procedure(mention="Angioplasty", purpose=["treat myocardial infarction"]),
        Procedure(mention="Chest X-ray", purpose=["evaluate elevation of the left hemidiaphragm"]),
        Medication(mention="Aspirin", purpose=["relieve pain and inflammation"]),
        LifestyleChange(mention="Low-salt diet", purpose=["control blood pressure"]),
        LifestyleChange(mention="Moderate exercise", purpose=["improve cardiovascular health"]),
        SpecialistReferral(mention="Appointment with cardiologist", purpose=["follow-up on heart health"]),
        SpecialistReferral(mention="Contact of Dr. Pérez", purpose=["additional consultations"])
    ]

    Generate a summary like this:


    The patient Joan López Martínez, 60 years old, was admitted on 03-04-2024 to the ICU of the hospital under the
    care of Dr. García due to an acute myocardial infarction. With a history of hypertension and diabetes, and a
    family history of myocardial infarction in his father, he was given Aspirin (100 mg/day) for pain and
    inflammation from 03-04-2024 to 10-04-2024. Additionally, he was diagnosed with diabetes mellitus, presenting
    excessive thirst and frequent urination, treated with insulin and a specific diet since 15-05-2018. He
    underwent a successful angioplasty on 10-02-2023. His vital signs and lab results were normal, except for an
    elevation of the left hemidiaphragm on a chest X-ray on 05-06-2023. Upon discharge on 10-04-2024, a low-salt
    diet and moderate exercise were recommended, with a follow-up appointment with the cardiologist in one month,
    and the contact of Dr. Pérez for additional consultations.
    """,
})
```

Figure 9: System prompt in English that Llama 3 receives to generate the summary

In Figure 8 and Figure 9, we show the primary prompts used to summarize texts in Spanish and English. These prompts provide the LLM with basic instructions and an example that includes text, extracted entities, and a corresponding summary to facilitate one-shot learning, also known as in-context learning. This method allows the model to generate accurate summaries by learning from a single example provided in the prompt.

Additionally, we include other system prompts to adjust the output format, making the summary more schematic or more compact based on user preferences. There is also a prompt specifically for summarizing the text without using the Gollie-extracted entities, which involves a simple instruction for the model to summarize the text without any extra information. Finally, the user prompt contains the actual text to be summarized.

A.3 Implementation Details

We used Python for all the inference of the models, data cleaning, the prompt system, and the overall logic of our project. For the web interface, we utilized Flask along with HTML, CSS, and JavaScript. The models were loaded and inference was performed on an NVIDIA A40 GPU with 48GB of RAM memory.

A.4 Evaluation

Evaluating the performance of text summarization models is a challenging task. While quantitative metrics such as ROUGE scores provide some insight, they often fall short in capturing the actual summary quality, coherence, and relevance as. This is why a comprehensive evaluation approach is necessary. To select the ground truth and understand the dataset to design the right metrics an exploratory analysis of the dataset has been done, this can be found in the GitHubGitHub repository of the project.

A.4.1 Density

For the evaluation of the model the used approach relies on regex. By the use of regular expressions the exact match of keywords with the summary text is computed, by obtaining the keyword count in the summary denoted with K. This keyword count is divided by the total number of words in the summary, denoted by LS, to obtain the keyword density of the summary.

Density formula :
D = Density
K = Total number of keywords in the summary
LS = Total number of words in the summary

$$D = K/LS$$

A.4.2 Keywords Recall

For the keyword recall the used approach relies on regex. By the use of regular expressions the exact match of keywords with the summary text is computed, by obtaining the keyword count in the summary denoted with K. The number of keywords found in the summary, denoted by LK , is then divided by the total number of keywords in the ground truth, by obtaining a recall metric.

KR = Keyword Recall
K = Total number of keywords in the summary
LK = Total number of Keywords in the ground truth

$$KR = K/LK$$

A.4.3 Qualitative analysis samples

In this appendix section you can find samples of the models. This results can change for new generations of text, as the LLMs that generate the text have some probabilistic techniques to generate new text despite of giving always the same input.

Llama3-70b with GoLLIE:

> Paciente de 27 años, **natural** de **Bolivia**, ingresada por insuficiencia suprarrenal primaria, diagnosticada con crisis de Addison. Presentaba síntomas de diarrea, vómito y debilidad general. Se realizaron varias pruebas diagnósticas, incluyendo gastroscopia, colonoscopia, tomografía abdominal y broncoscopia, que revelaron gastritis crónica activa, **lesiones superficiales** en la **mucosa** gástrica, un nódulo hipodenso en el segmento VI del hígado y **lesiones fibrosas cicatriciales** en el ápex del pulmón izquierdo. Se diagnosticó también

tuberculosis pulmonar y probablemente adrenitis inflamatoria/infecciosa. Se inició tratamiento con varios medicamentos, incluyendo **antibióticos** y corticosteroides.

Llama3-70b without GoLLIE:

El paciente de 27 años ingresó con síntomas de diarrea, vómitos y debilidad generalizada. La exploración física objetivó una pigmentación cutánea intensa, palidez y dolor abdominal en epigastrio. Los análisis de sangre mostraron hiponatremia, hipocortisolismo y anticuerpos anti-suprarrenales negativos. La tomografía abdominal reveló un engrosamiento suprarrenal bilateral con mínima trabeculación de la grasa periadrenal, sugestivo de adrenalitis inflamatoria o infecciosa. La tomografía de tórax mostró lesiones fibrocísticas cicatriciales en el ápex del pulmón izquierdo y bronquiectasias bilaterales en bases pulmonares. Se sospecha la presencia de tuberculosis pulmonar. Se instauró tratamiento con hidrocortisona y se le dio de alta con recomendaciones para seguir tratamiento en domicilio. Se solicitaron pruebas adicionales, como la broncoscopia, para evaluar la sospecha de tuberculosis pulmonar.

The highlighted words denote those that were forgotten or not retrieved by the raw LLM while being correctly retrieved by the LLM with GoLLIE.