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 26, 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*.

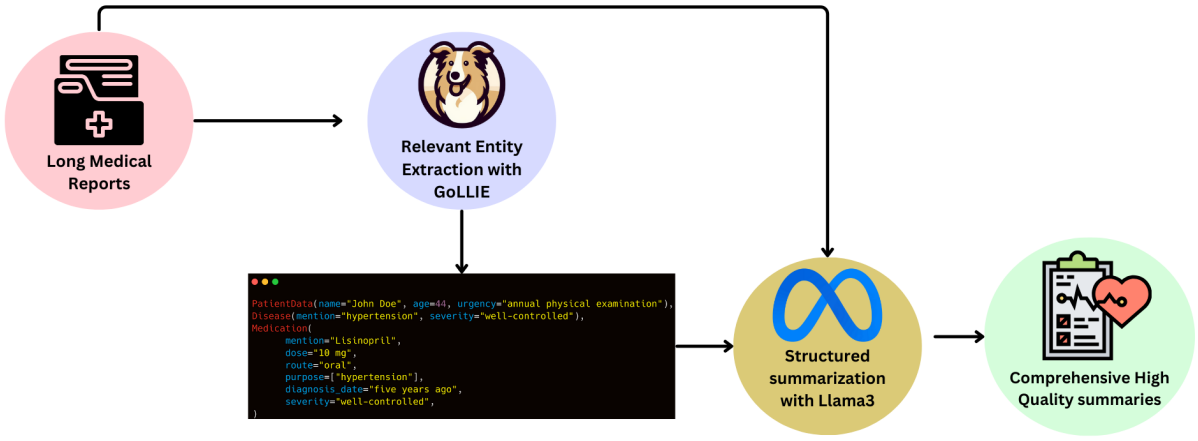

Figure 1: Summarization Pipeline.

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 few-shot learning approach with Llama3-70b to generate the final summary. This involves providing Llama3 with a small number of examples that demonstrate 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.

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

4 Experiments and Results

5 Discussion

This research presents a novel approach to addressing the persistent challenges of recall and accuracy in medical document summarization, particularly in the context of limitations posed by current LLMs. Existing methods, while demonstrating advancements in abstractive summarization, often struggle with faithfulness, hallucination, and the accurate capture of critical information. Our proposed method, leveraging GoLLIE and Llama3, aims to mitigate these limitations by introducing 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 information extraction, to enhance recall. By utilizing predefined guidelines tailored to medical documents, GoLLIE aims to extract key entities and details, ensuring that no critical information is overlooked and missed. This approach, unlike traditional methods reliant on rule-based systems or extensive labeled data, allows for zero-shot application across diverse medical domains.

Furthermore, the integration of Llama3, a powerful 70 billion parameter LLM for natural language generation, enables the production of structured summaries based on the extracted information. This approach utilizes a few-shot learning paradigm, eliminating the need for extensive model retraining and allowing for flexible adaptability to different summarization formats (e.g., bullet points, concise paragraphs).

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

- **Improved patient comprehension:** Concise and comprehensive summaries can empower patients to better understand their diagnoses, treatments, and follow-up care, leading to increased satisfaction and improved adherence to medical plans.
- **Enhanced clinical decision-making:** Doctors and healthcare professionals can benefit from efficient access to critical information, enabling more accurate assessments, optimized treatment decisions, and ultimately improved patient outcomes.
- **Streamlined information flow:** Efficient summarization can optimize information flow within healthcare institutions, leading to enhanced operational efficiency, reduced costs associated with miscommunication, and improved overall quality of care.

While the proposed approach offers promising advancements, it's crucial to acknowledge potential limitations and avenues for future research:

- **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 for further research.
- **Guideline development and maintenance:** Defining and maintaining a comprehensive set of 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.

Future research directions will focus on further enhancing the capabilities of our approach. Some of these improvements may include the development of automated guideline generation techniques for generating and validating GoLLIE's guidelines, enhancing adaptability and reduce human effort. Incorporating reinforcement learning techniques, particularly from human feedback, can further optimize Llama3's summarization capabilities and improve the alignment of generated summaries with clinical needs. In addition to these, a more rigorous and extended evaluation and benchmarking of our approach is essential to establish its performance compared to existing summarization methods. Benchmarking against standardized medical summarization datasets will be crucial for validating the effectiveness of our approach.

This research 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 holds the potential to significantly improve the communication and utilization of critical medical information. Ongoing research efforts will focus on further development and refinement to maximize the impact and clinical utility of this 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.

7 Acknowledgements

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 3: Medical Entities and Extraction Guidelines for GoLLIE

As it can be seen in Figure 3, 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

A.3 Implementation Details