Titre: Title:	Ontology-Constrained Generation of Domain-Specific Clinical Summaries
Auteur: Author:	Gaya Mehenni
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Mehenni, G. (2025). Ontology-Constrained Generation of Domain-Specific Clinical Summaries [Master's thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/67792/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/67792/
Directeurs de recherche: Advisors:	Amal Zouaq
Programme: Program:	GÉNIE INFORMATIQUE

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ontology-Constrained Generation of Domain-Specific Clinical Summaries

GAYA MEHENNI
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Août 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Ontology-Constrained Generation of Domain-Specific Clinical Summaries

présenté par **Gaya MEHENNI**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Farida CHERIET, présidente
Amal ZOUAQ, membre et directrice de recherche
Gilles PESANT, membre

DEDICATION

À ma famille, mes amis et tout ceux que j'ai rencontrés qui ont fait de moi la personne que je suis aujourd'hui...

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my research director, Professor Amal Zouaq. Her constant support and invaluable guidance over the past few years have been extremely formative and have allowed me to grow both professionally and personally.

I am immensely grateful to my parents, my brother, and my sister. I am fortunate to be able to count on them, as their support has been the condition for my success.

A big thank you as well to my girlfriend and to all my friends at Polytechnique. I am thinking in particular of my colleagues at the LAMA-WeST laboratory, whose constructive feedback pushed me to surpass myself as a researcher, as well as all those I have met over these past few years who have made my journey unforgettable.

RÉSUMÉ

Ce mémoire propose une approche novatrice pour relever les défis critiques de la génération de résumés cliniques qui soient à la fois précis, pertinents et adaptés au domaine médical. Utilisant les grands modèles de langue (LLMs), cette méthode vise également à réduire les "hallucinations", c'est-à-dire la production d'informations factuellement incorrectes. Le point de départ de cette recherche est le constat que les note cliniques actuelles de patients contiennent une grande quantité de données que les médecins doivent examiner en détail. L'objectif est d'automatiser ce processus afin de diminuer l'épuisement professionnel des médecins et d'améliorer l'efficacité des soins de santé.

Bien que les LLM présentent un potentiel considérable pour la synthèse de textes, leur tendance à générer des informations erronées et leurs difficultés à traiter des données spécifiques à un domaine, surtout dans le contexte médical où la confidentialité est primordiale, posent des risques importants. L'objectif central de cette recherche est d'utiliser les ontologies médicales, des représentations structurées de connaissances factuelles, pour guider les LLM afin de produire des résumés plus fiables et spécifiques à une spécialité. L'hypothèse est que l'intégration de ces ontologies dans le processus de génération des LLM améliorera la précision et la pertinence des résumés cliniques.

La méthodologie s'articule autour de plusieurs axes, en commençant par une analyse d'adaptation au domaine médical via une annotation basée sur les ontologies pour identifier les concepts clés. Par la suite, une stratégie d'extraction d'informations s'appuie sur ces ontologies pour créer une représentation structurée des notes cliniques. Finalement, un nouveau processus de décodage contraint et guidé par l'ontologie est appliqué, utilisant une approche qui favorise le contenu aligné sur les relations ontologiques et permet de minimiser les incohérences. Les résultats expérimentaux, obtenus notamment avec le jeu de données MIMIC-III, montrent des améliorations significatives dans la génération de résumés adaptés et une réduction des hallucinations. Ces conclusions indiquent que le fait de contraindre la génération des LLMs à l'aide d'ontologies diminue efficacement la génération d'informations erronées.

De plus, ce mémoire introduit aussi MedHal, un nouvel ensemble de données conçu spécifiquement pour l'évaluation de la détection des hallucinations dans les textes médicaux. MedHal surmonte les limites des jeux de données actuels en intégrant diverses sources et tâches médicales et en fournissant un volume important d'exemples. Ces exemples sont également annotés avec des explications indiquant les incohérences factuelles. Ceci permet un entraînement et une évaluation plus robustes des modèles de détection d'hallucination.

Les retombées de cette recherche pour le secteur de la santé sont considérables. Ce travail contribue de manière significative à l'avancement de la synthèse de textes assistée par les modèles de langue dans le domaine médical en apportant des solutions concrètes aux défis de la factualité et de l'adaptation au domaine. Enfin, nous apportons également une solution pour alléger la charge de travail des docteurs, améliorer la qualité des soins, accélérer la recherche en IA médicale et faciliter un déploiement plus sûr des LLMs dans le domain médical.

ABSTRACT

This thesis presents an innovative approach to address the critical challenges of generating accurate, relevant, and domain-adapted clinical summaries using Large Language Models (LLMs), while simultaneously mitigating hallucinations. Recognizing that Electronic Health Records (EHRs) contain vast amounts of structured and unstructured data, which clinicians must review thoroughly, this research aims to automate this process to reduce burnout and improve healthcare efficiency.

While LLMs offer significant potential for summarization, their inherent tendencies to hallucinate and their limitations with out-of-distribution data, particularly in the privacy-sensitive medical domain, pose substantial risks. The core objective of this research is to leverage medical ontologies, structured representations of factual domain knowledge, to guide LLMs towards generating more grounded, domain-relevant, and specialty-specific summaries (e.g., tailored for radiologists versus oncologists). The underlying hypothesis is that integrating ontologies into the LLM generation process will enhance the factual accuracy and relevance of clinical summaries.

The methodology encompasses several key components: an initial domain adaptation analysis using ontology-based annotation to identify and prioritize relevant concepts; an ontology-based prompting strategy for information extraction, leading to a Concept-Structured Representation (CSR) of clinical notes; and a novel ontology-guided constrained decoding process. This decoding mechanism utilizes a beam search approach, incorporating hierarchy, property, and similarity scores to favour content that aligns with ontological relationships and reduces factual inconsistencies. The experimental results demonstrate significant improvements in generating domain-adapted summaries of clinical notes and in hallucination reduction, particularly through the application of the proposed methods on the MIMIC-III dataset. The findings indicate that constraining LLM output with ontological knowledge effectively reduces the generation of erroneous information.

Furthermore, the thesis introduces MedHal, a new large-scale dataset specifically designed for evaluating hallucination detection in medical texts. MedHal addresses the limitations of existing smaller, single-task datasets by incorporating diverse medical text sources and tasks, providing a substantial volume of annotated samples with explanations for factual inconsistencies. This allows for more robust training and evaluation of medical hallucination detection models. The MedHal dataset proves valuable for developing more effective medical hallucination detection systems.

The implications of this research are considerable for healthcare, as it offers a way to reduce clinician workload by providing precise and relevant information, enhance the quality of patient care through domain-adapted data, accelerate medical AI research by providing a standardized evaluation framework, and facilitate the safer deployment of LLMs in clinical settings by mitigating the critical issue of hallucinations. This work makes a contribution to advancing LLM-assisted text summarization in the medical field by offering concrete solutions to challenges of factuality and domain adaptation through ontology integration and robust evaluation tools.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

Electronic Health Records (EHRs) document every aspect of a patient's stay during a hospital admission. Containing an overwhelming amount of unstructured (e.g. lab reports, progress notes) and structured (vital signs, lab results) information, these documents must be reviewed by clinicians prior to the point of care, since a thorough analysis of all this data needs to be made before making a formal decision about a patient's diagnosis. This process is time-consuming and can lead to clinicians burnout [1]. Large Language Models (LLMs), which have shown major improvements in language understanding in recent years, can automate this process [2] and help reduce clinician burnout. Automating the summarization process of clinical notes could not only ease the burden currently placed on healthcare systems, but also improve the work quality of doctors. However, this process hinges on two critical requirements. The summary must be factually grounded, containing only information explicitly present in the original clinical note. Additionally, the summarization process must be tailored to the doctor's specialty as it must only retain the relevant information needed by the clinician reading the summary. Since the information needed by radiologists differs significantly from that needed by oncologists, different summaries should be generated for different areas of focus. However, while LLMs have improved a lot on tasks like summarization and information extraction since the original Transformer paper [3], they still show certain limitations in some areas. Indeed, major limitations of LLMs include their inherent tendency to hallucinate information [4–6] and their inability to handle out-of-distribution data. These limitations are particularly important as a single value or word hallucinated can lead to disastrous consequences for the patient. Plus, classical LLMs have limited exposure to medical data during training due to privacy restrictions, which further diminishes their performance when applied to healthcare domains. To address these issues, this research aims to leverage ontologies in conjunction with LLMs to create domain-adapted summaries of clinical notes. Ontologies are a structured representation of the knowledge of a specific domain (in our case, the medical domain). The information present in ontologies is known to be factual and relevant to the domain and thus can be used to guide LLMs towards more grounded and domain-relevant generations.

1.1 Definitions

1.1.1 Ontologies

Definition : *An ontology is a formal, explicit specification of a shared conceptualisation* [7].

More precisely, an ontology is a structured representation of the knowledge linked to a specific field. This field can be as precise as quantum mechanics or as broad as the medical domain. This structure is created through three types of objects with specific goals :

- Classes: Abstract representations of domain-specific concepts that constitute the fundamental categorical units within the ontological framework
- Properties: Attributes that characterize and define the intrinsic nature of classes, facilitating their formal specification within the knowledge domain
- Relationships: Formal associations that establish semantic connections between classes, enabling the expression of complex inter-dependencies within the ontological structure

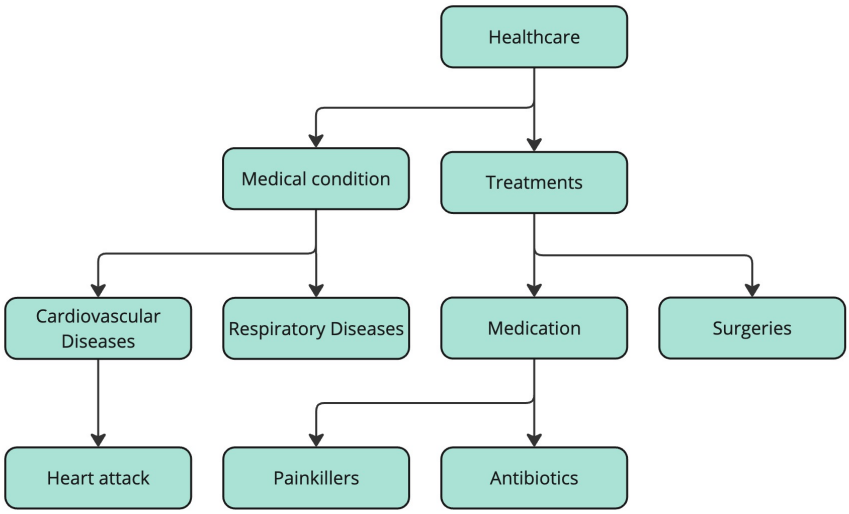


Figure 1.1 Example of ontology representing a small portion of the medical domain

For instance, an ontology of the medical domain could contain classes representing concepts like "Treatments", "Medication" and "Medical condition". As shown in Figure 1.1 , the "Treatments" class can be linked to the "Medication" class through a relation of type *subClassOf* (shown by the arrows in Figure 1.1). This research is focused on ontologies representing the

medical domain as it aims to use them on clinical notes. More precisely, it focuses on medical ontologies that encompass all aspects of healthcare and medicine. This preference for general medical ontologies originates from the need to consolidate all possible information relevant to process clinical notes in a single structure for simplicity.

1.1.2 Large Language Models

Large Language Models are neural architectures based on the Transformer architecture [3]. They work by utilizing the attention mechanism which allows them to attend to every part of a sequence at once. Contrary to the original Transformer architecture which depicted an "encoder-decoder" architecture, recent LLMs usually follow a "decoder-only" architecture. Encoder-decoder models compress the input sequence using an encoder and use a decoder to generate the output sequence. However, decoder-only models directly predict the next sub-word (called token) that should follow the input sequence. Through rigorous training, these models achieve remarkable performance only using this single paradigm. For example, given the input sequence "*The capital of France is* ", the next tokens generated for a well trained model would be *Pa-ris* (the dash separating the tokens generated). More precisely, for a single inference step, the model not only outputs the next token, but also the probability distribution of all tokens in its vocabulary. While the most probable token is usually chosen as the next token, some algorithms prefer to rely on another metric to choose which token should be selected in the distribution. More information is given about this in Section 2.4.

1.2 Medical Domain

This section examines the unique constraints within the medical field. It provides a brief overview of patient admission procedures and explores the difficulties encountered when using LLMs to analyze clinical notes.

1.2.1 Admission Procedure

An admission is the acceptance of a patient to a hospital's care unit for the purpose of receiving medical care. It is associated with a single patient and encompasses a lot of information required by healthcare providers. While an admission corresponds to a single patient, a patient may have multiple admissions to a hospital over time, reflecting different episodes of care. During the point of stay, multiple clinical notes can be authored by various healthcare professionals about the patient including nurses, radiologists, oncologists, and other specialists. These notes detail information like medical history, diagnostics and treatment

plans. When the patient is discharged from the hospital, a comprehensive document called the discharge summary is written detailing every aspect of the patient's stay.

1.2.2 Clinical notes

Clinical notes contain extensive patient information, but present significant challenges in interpretation due to their heterogeneous format and content. They range from structured laboratory results in tabular format to detailed narrative reports. The variability in these notes stems from multiple sources: different specialists have distinct documentation requirements, and even within the same specialty, clinicians may employ different documentation styles based on their training and preferences. This structural heterogeneity persists despite attempts at standardization through frameworks such as SOAP (Subjective, Objective, Assessment, Plan) [8] and PICO (Population, Intervention, Comparison, Outcome) [9]. While some healthcare institutions implement standardized documentation protocols, the challenge of format variability remains significant. This is particularly evident in the context of Large Language Models (LLMs), where models trained on clinical notes from one institution often demonstrate poor generalization when applied to notes from different institutions.

1.3 Problem Statement

This section details the different aspects of the problem of clinical summarization as tackled by this research.

1.3.1 Summarization

In Natural Language Processing (NLP), the summarization task is defined to be the process of transforming an input sequence (or source) I of arbitrary size into a shorter sequence of text O [10]. This transformation $T(x)$ is guided by a criterion $C(I, O)$ which measures the mutual information that needs to be conserved between I and O. This criterion is maximized to obtain O. More formally,

$$O = \max_{T} \; C(I, T(I)) \tag{1.1}$$

such that $\text{Length}(O) < \text{Length}(I)$. Following this definition, we aim to reduce $\text{Length}(O)$ as much as possible, leading to more dense summaries, to save doctors as much time as possible. Moreover, we aim to generate a structured representation that encompasses all the essential details needed by a specialist for easy querying.

1.3.2 Domain Adaptation

As mentioned in Section 1, the content required by different medical specialties varies significantly. Given multiple clinical notes associated to a patient, it is important to carefully select the information that is the most relevant to the specialist reading the final summary. While a nurse might not look at MRI results when analyzing electronic health records, they will carefully analyze patient responses to treatments in the clinical notes as that information is important for them. This aspect of domain adaptation is particularly crucial, as the effectiveness of a summarization system cannot be measured solely by its ability to preserve information. Even if a summary is grounded by its source, it becomes ineffective if it fails to prioritize the information relevant to the healthcare professional reading it. Following Equation 1.1, we adjust the definition of C to incorporate the medical specialty D in the criterion leading to $C(x_i, x_o, D)$. The ability to precisely determine and control what information is included in summaries for different medical specialties is a central component of this research. In this sense, we define a domain as a set of ontology classes of interest related to a specific medical field. This definition operates under the assumption that a comprehensive medical ontology that encompasses the broader medical field exists.

Definition: *Let O be an ontology that encompasses an entire field (e.g. medicine). A domain is defined to be a set of ontology classes S of interested related to a specific sub-field (radiology, nursing, etc).*

Therefore, we define domains specifically in relation to medical fields, rather than using the traditional definition of domains (law, medicine, mathematics, etc). This domain-specific approach directly influences a fundamental aspect of clinical summarization that is often overlooked in the literature : summary relevance.

Definition: *Summary relevance measures the degree to which the information provided in a summary is important to the reader.*

This aspect is particularly important when the ground truth summary is not accessible as we have no way of measuring if the information that was used to generate the summary is the one that was crucial given the clinical notes. We propose measuring summary relevance by assessing how well a summary covers the most important concepts to the reader. In our case, the reader is linked to a specific domain (radiology, nursing, etc) and we measure if the concepts covered in the summary are consistent with those of the domain.

1.3.3 Groundedness

While the domain-adaptation part is important, it is also crucial that summaries generated from clinical notes remain grounded. By grounded, we mean that every single piece of information mentioned in the summary must be backed by a statement in the original clinical notes.

Definition: *A summary is grounded if and only if each statement it contains can be directly traced back to and verified against the source text.*

The importance of groundedness in clinical summarization cannot be overstated. Incorrect information in summaries of clinical notes can lead to serious medical errors and potentially compromise patient safety. These summaries may propagate misunderstandings or introduce artificial facts that could influence clinical decision-making. The literature often uses "factuality" interchangeably with "groundedness," but we distinguish them here. Groundedness is context-dependent, while factuality is not.

Definition: *A text is factual if and only if each statement it contains can be directly traced back to and verified against a source text (context) or general knowledge.*

Both groundedness and factuality can be indirectly measured through hallucination metrics. Further details regarding these two terms and their measurement will be provided in Sections 2.3 and 2.5.2.

1.3.4 Hallucinations in the Medical Domain

While in certain domains, hallucinations of LLMs do not have a direct impact on people, this risk is particularly concerning in healthcare settings as it directly impacts patient outcomes. Recent research has made progress in summary groundedness by reducing LLM hallucinations through various approaches [11, 12] and evaluation metrics [13, 14], but significant challenges remain. The fundamental difficulty lies in the lack of a formal, consistent methodology for detecting ungrounded statements as it depends heavily on concept definitions and linguistic formulations. While this research does not propose a method for completely preventing hallucinations, it aims to improve the groundedness of clinical text summarization, thus reducing LLM hallucinations. Finally, this research also focuses on creating a benchmark for evaluating different methods on factuality in medical text generation. The current state-of-the art methods for evaluating summary groundedness mainly rely on large-scale language models, which require substantial computation resources and infrastructure - a significant barrier to accessibility for many researchers and institutions. Plus, these models often lack specialty in the medical domain, which hurts their performance when given clinical documentation.

1.3.5 Structured Summarization

While textual summaries effectively reduce the mental load of reviewing multiple documents, healthcare professionals might require specific data points rather than a comprehensive overview of the patient's stay. A nurse, for instance, may need to quickly access a patient's weight to determine appropriate medication dosage - a task that still remains time-consuming even with well-crafted summaries, as it requires parsing through narrative text to locate the specific information. This limitation originates from the underlying nature of textual summaries : they cannot be queried easily. While current state-of-the-art methods [2, 15] for clinical summarization mainly focus on narrative generation, few have tried structuring the summaries using certain formats [16–19]. However, these summaries can usually only be queried based on the clinical section, not by medical concepts. To address this constraint, we aim to incorporate the structure of ontologies into the summarization process. This structure could help organize information according to medical concepts or properties (e.g. weight, procedures, treatments) while maintaining semantic relationship with the clinical note. This structured format offers multiple benefits: it enhances information accessibility through direct queries, improves document readability, and provides a standardized interface between different clinical note formats. Since note formats such as SOAP and PICO are defined based on sections containing different medical concepts, an easy mapping can be defined if the granularity of the method is at the concept level. This standardization could facilitate information exchange across different medical specialties and healthcare institutions, effectively bridging documentation gaps in clinical communication.

1.3.6 Infrastructure Constraints

Healthcare institutions face significant constraints that must be considered in the development of clinical summarization systems. While large language models with hundreds of billions of parameters demonstrate superior performance, their deployment in healthcare settings is often impractical. First, most healthcare institutions lack the specialized hardware infrastructure necessary to host and operate such large models effectively, and the substantial operational costs (energy consumption, maintenance, etc) make these solutions financially unsustainable. Plus, healthcare providers are bound by strict privacy regulations and data protection requirements. The transmission of sensitive clinical documentation to external data centers thus becomes impossible. This problem is even more complicated by data locality requirements, as many governmental institutions require patient data to be processed and stored within the same region as the healthcare institution, some even requiring that the data remains in the physical premise of the facility. This makes cloud-based solutions

unsuitable regardless of their performance.

1.4 Research Questions

Our study aims to enhance multiple aspects of automatic clinical text generation. We focus on two primary areas: improving the generation process itself and developing better evaluation methods for clinical text.

1.4.1 Improving Generation Process

We investigate the following research questions:

1. How can large language model (LLM) generated clinical summaries be effectively adapted across different medical domains ?
2. What strategies enable the integration of medical ontologies to constrain and guide LLM text generation ?
3. What mechanisms can be implemented to reduce hallucinations in LLM-generated clinical content ?

1.4.2 Advancing Evaluation Methods

Recognizing the limitations of current evaluation metrics in clinical settings, particularly for hallucination detection, we also explore these questions:

1. What constitutes an efficient and reliable hallucination detection metric specifically designed for clinical text ?
2. How can hallucination detection approaches be unified across diverse medical NLP tasks, including question answering, summarization, and information extraction ?

1.5 Thesis Outline

This thesis is structured into five chapters, each addressing a critical aspect of our research. Chapter 2 lays the groundwork with a detailed literature review, exploring five essential topics: the fundamentals of summarization, the specific challenges of clinical text summarization, the phenomenon of hallucinations in generative models, various decoding strategies, and relevant evaluation metrics. Chapter 3 introduces our novel approach for generating

domain-adapted clinical summaries, a method initially presented at the EKAW conference. This chapter highlights our key innovations: an ontology-constrained decoding strategy that enhances relevance and mitigates hallucinations, and a new ontology-driven domain-aware summarization process. Chapter 4 details our work on developing a dedicated large-scale medical hallucination detection dataset. This dataset serves a dual purpose: to evaluate the performance of current models in detecting medical hallucinations and to act as a valuable training resource. Concluding the thesis, Chapter 5 provides a concise summary of our contributions, discusses the inherent limitations of our work, and outlines promising avenues for future research.

CHAPTER 2 LITERATURE REVIEW

This literature review is about summarization, clinical text summarization, domain adaptation, hallucinations, constrained generation, and evaluation metrics.

2.1 Summarization

As mentioned in Chapter 1, summarization is the process of transforming an input sequence into a shorter sequence of text. Its main objective is to grasp as much information as possible contained in the input text in the output text. The input sequence is a text sequence of arbitrary size and the criterion is a measure of the mutual information preserved between the input and output sequences according to a formal definition of what needs to be conserved. This transformation is guided by the type of inputs and processes.

The literature distinguishes three main categories of summarization tasks based on the nature of the input sequence. When processing a single document, the task is referred to *single-document summarization* (SDS) [20, 21]. In scenarios involving multiple documents as input, the task becomes *multi-document summarization* (MDS) [22, 23]. The third category, *query-focused summarization* (QFS) [24–27] extends these approaches by incorporating an additional query component in the input sequence. A notable distinction between these approaches lies in how their information criterion is defined. In SDS and MDS, the criterion for what information should be preserved is typically implicit in the model architecture and training process, leading to potential variations in information selection across different approaches. In contrast, QFS provides an explicit, sample-specific criterion through its query component, offering clearer guidance on what information should be prioritized in each summary.

While these categories define how input is structured and information is selected, the actual mechanism for generating summaries falls into three main paradigms: *extractive*, *abstractive*, and *hybrid* summarization [10, 21].

2.1.1 Extractive Summarization

Extractive summarization relies on directly retrieving parts of the input text to put them into the output text. This means that every part of the summary is directly copied from the input text. Extractive summarization is usually a two step process, where the first step consists of finding the important information and the second step involves joining the

retrieved information to generate the final summary [10]. Different kinds of methods can be used to do these tasks. They are regrouped into two categories : *statistical methods* and *neural methods*

Statistical Methods

Statistical methods in extractive summarization employ various heuristic approaches that leverage both custom-designed features and frequency-based metrics of words and sentences. Fundamental techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) [28–30] evaluate the significance of text segments by analyzing their lexical composition while accounting for the diminishing impact of commonly occurring terms across the document (or multiple documents in the case of MDS). These approaches have then been further enhanced by incorporating structural and semantic features, including sentence position, sentence relevance, and topic coverage metrics mainly through optimization techniques [31–33]. Another branch of statistical methods transforms documents into graph-based representations, where sentences are depicted as nodes and their semantic relationships as weighted edges [34–36]. This transformation enables the application of graph-based algorithms to identify central nodes, which theoretically correspond to the most salient sentences in the source document. These methods vary mainly in how they construct the graph and in what types of graph they construct (directed, undirected).

Neural Methods

Neural methods encompass a diverse range of approaches based on deep neural networks and their variants. The foundation of these methods was established through word embedding techniques like Word2Vec [37] and Glove [38], which enabled the representation of text sequences as dense vector embeddings. These representations demonstrated remarkable effectiveness in sentence classification and facilitated sentence clustering and selection based on relevance and meaning. The integration of these embeddings with sophisticated architectures such as Recurrent Neural Networks [39], Long Short-Term Memory networks [40], Convolutional Neural Networks and Graph Neural Networks [41, 42] further enhanced the capability of extraction summarization systems. However, the field underwent a paradigm shift with the introduction of the Transformer architecture [3]. Due to their attention mechanism, Transformers enabled the creation of contextual representations that surpassed previous methods in terms of accuracy and versatility. This breakthrough led to the development of Pre-trained Language Models (PLMs), including BERT, RoBERTa, Sentence-BERT and DeBERTa. These models have been successfully employed in various approaches, focusing

on inter-sentence relationships [20], sentence hierarchy [43, 44], and graph-based representations [43]. Recent innovations have expanded the methodological landscape, incorporating reinforcement learning by reformulating summarization as a ranking task [45, 46]. Additionally, the emergence of diffusion models has introduced novel approaches to generating summary sentence representations, which are then utilized to extract relevant sentences from the source document [24]

Granularity

The granularity of the extraction must be defined prior to the retrieval. Depending on the size of the input text, multiple granularities can be used, the main one being at the sentence-level [20]. However, other avenues have been explored by researchers like using sections instead of sentences [17, 47, 48] . In the case of MDS, the granularity can even be extended to the document level if the task involves multiple documents of small sizes.

Limitations

Extractive summarization represents the most reliable approach for maintaining complete groundedness, as it constructs summaries exclusively from verbatim excerpts of source documents. However, this methodology presents several challenges. One significant limitation lies in the granularity of extraction—individual sentences, when isolated from their surrounding context, may lead to a change in semantic meaning. In clinical settings, this can be particularly problematic. For example, consider a statement noting "patient's condition has improved" appearing in the source document after documentation of medication A administration. If this statement is extracted and placed in the summary following a mention of medication B administration, it creates a misleading causal relationship. While the original document clearly attributed the improvement to medication A, the restructured summary incorrectly suggests that medication B was responsible for the patient's improvement.

Furthermore, extractive summarization often struggles with maintaining natural flow and cohesion in the summaries. Since sentences are extracted from different sections of the source documents, the final summary may appear fragmented. This can result in summaries that, while factually accurate, are less engaging and potentially more difficult for healthcare providers to quickly comprehend in time-sensitive clinical settings.

2.1.2 Abstractive Summarization

While extractive summarization builds a summary by combining multiple sequences in the input text, abstractive summarization generates summaries by synthesizing new text based on the input information using a language model - the summary is written from scratch. The evolution of abstractive summarization can be traced through distinct technological phases with a significant transformation occurring with the emergence of LLMs.

Before LLMs

The introduction of pre-trained encoder-decoder architectures, notably BART [49] and T5 [50], marked a pivotal advancement in abstractive summarization. These Pre-trained Language Models (PLMs), leveraging extensive pre-training corpora, demonstrated remarkable efficiency in task-specific fine-tuning with minimal data requirements. This capability boosted the development of various enhancement techniques, which can be categorized into three main approaches: architectural modifications, training methods, and processing techniques.

Architectural Methods : Architectural innovations focus on creating summarization-fitted architectures to better suit summarization tasks [51]. Significant advances include enhanced self-attention mechanisms that expand token processing capacity, which is particularly beneficial for Multi-Document Summarization (MDS) [52–54]. Researchers have also incorporated hierarchical structures within attention mechanisms [55–58]. A notable innovation involves the integration of guidance signals into PLM encoders, improving output faithfulness and relevance [2, 59]. These guidance signals effectively direct the model's attention to relevant tokens during inference.

Training Methods : Training-focused approaches aim to optimize the fine-tuning process to enhance models' summarization capabilities. Contrastive learning techniques [60–63] address this by training models to discriminate between multiple candidate summaries, effectively identifying optimal outputs. These methods particularly target exposure bias, a phenomenon where autoregressive models' outputs vary between training and inference phases due to their reliance on ground truth summaries during training [64]. Additional innovations include specialized pre-training strategies, such as importance-based denoising [65] and n-gram prediction, which further refined the summarization process.

Processing methods : The third category encompasses both pre-processing and post-processing techniques. Pre-processing methods [66] focus on input optimization and data filtering, while post-processing approaches [67–69] include error correction and constrained gen-

eration strategies, which ensure higher quality outputs. More details about post-processing techniques are given in section 2.4.

After LLMs

The emergence of instruction-tuned LLMs like GPT-3.5, Claude, and Llama [70], has fundamentally transformed text summarization research. As models grow increasingly large and get computationally intensive to train, the research community has pivoted from traditional fine-tuning approaches to developing techniques that better leverage the knowledge already embedded within these LLMs. While these models already demonstrate strong summarization capabilities in their base form, researchers have developed various approaches to extract and enhance their abilities. These approaches can be broadly categorized into two main directions : prompting techniques, multi-agent systems and constrained generation. While constrained generation is more deeply explained in section 2.4, we detail here the two other directions.

Prompting techniques : Prompting techniques aim to optimize how tasks are presented to LLMs by providing specific instructions, examples or possible reasoning paths. Notable innovations include few-shot prompting, which helps LLMs better understand desired output formats and content through examples, and chain-of-thought (CoT) prompting, which enhances LLMs' reasoning capability leading through better summarization performance. The reasoning process, usually called element-aware summarization, helps the model better distinguish which elements must be present in the final summary [46, 71–73]. Soft prompting, which uses trainable continuous vectors instead of fixed text templates (discrete prompts), has also surfaced as a technique to improve summarization capabilities of LLMs [74, 75].

Multi-agent systems : Multi-agent systems, on the other hand, leverage the interaction between multiple LLMs to improve summary quality. In a vast majority of these systems, some LLMs generate summaries or summary components, while others act as evaluators of the generated content [76]. Through iterative feedback loops, summaries are progressively refined until they meet the evaluators' quality criteria. This approach has proven particularly effective in detecting and correcting inaccurate information [77].

Limitations

Abstractive summarization generates cohesive summaries with natural flow by producing new text rather than extracting existing passages. However, this approach faces a significant challenge: LLMs can hallucinate, introducing facts or details not present in the source text.

This limitation is particularly problematic in production environments, especially in critical domains like medicine where accuracy is crucial. While current LLMs demonstrate strong summarization capabilities due to their extensive training on summarization tasks, their tendency to hallucinate remains a barrier to their deployment in contexts where even a single factual error could have serious consequences.

2.1.3 Hybrid Methods

Hybrid methods combine extractive and abstractive approaches to achieve performance superior to either methods applied independently. These techniques typically implement a multi-stage pipeline [78, 79] where extractive models identify, cluster or filter relevant segments of the input text, and abstractive models process these segments to generate the final summary [80]. The prompting techniques discussed in Section 2.1.2 can be applied to either or both models to further enhance the pipeline's overall performance. This hybrid architecture can be implemented in two main ways. In the first approach, the extractive component is directly integrated into the abstractive model's architecture [81–83]. In the second method, a traditional pipeline approach is adopted where the output of the extractive model is sent to the abstractive model without the two models being interlinked. The most popular method in this sense is based on Retrieval-Augmented Generation (RAG), which has emerged as the predominant hybrid method in recent years. In RAG systems, a retriever model identifies and extracts important portions of the input text, which are then processed by an abstractive model [84, 85]. Modern RAG architectures have been enhanced with additional components such as re-ranking mechanisms [86, 87] and query refinement techniques [88]. The effectiveness of hybrid approaches, particularly RAG, can be attributed to their ability to mitigate the hallucination problems commonly associated with LLMs processing large inputs. More details are mentioned about RAG in section 2.3.3. By using the extractive components to reduce the input size, these systems produce more concise and factually accurate summaries while maintaining the natural flow characteristics of abstractive methods.

2.2 Clinical Text Summarization

Clinical text summarization represents a unique challenge in the broader field of natural language processing. While general text summarization has experienced significant advances in recent years, the development of clinical text summarization has progressed at a slower pace, primarily due to the absence of reference summaries. This limitation stems from the strict privacy regulations surrounding clinical records, which creates a substantial barrier to research advancement in this domain. The research community has found some workarounds with syn-

thetics clinical notes, but these are limited in terms of applicability. It is important to clarify that our focus specifically addresses the summarization of clinical notes to improve efficiency for healthcare professionals. While the broader field of medical summarization encompasses various tasks, including medical dialogue summarization, research paper summarization, and patient health question summarization [89], our research specifically concentrates on clinical note summarization. State-of-the-art methods are detailed in Section 2.2.4.

2.2.1 Challenges

The challenges in clinical text summarization extend beyond bare data accessibility. Several inherent characteristics of medical records make this task particularly complex compared to other domain-specific summarization tasks. These challenges include the considerable length of medical records [90], their hybrid nature requiring both extractive and abstractive summarization approaches [91], and the substantial variation in writing styles across different healthcare providers. Furthermore, the requirement for absolute factual accuracy poses a significant challenges for LLMs, which are prone to hallucinations. All generated statements must be strictly factual and traceable to their original source for healthcare professionals' verification. Additionally, the specialized medical terminology presents a notable challenge [91], as LLMs typically have limited exposure to such domain-specific vocabulary during training.

2.2.2 Generating Discharge Summaries

While the majority of research efforts in medical summarization have centered on research papers and medical dialogue [89, 92], there has been a growing attention to electronic health record summarization. While most clinical note datasets with corresponding summaries remain private [91], some public resources have emerged to support research in this field. Among these is MIMIC-III [93], a semi-public dataset containing over 45,000 de-identified patient admissions, though access requires completion of privacy training. This dataset has spawned several research directions, particularly focusing on discharge summary prediction and its components [94, 95]. The generation of the *Brief Hospital Course* section, which summarizes a patient's entire hospital stay, has received particular attention [2,15,96]. Additional research has explored generating chief complaint sections [97] and histories of present illness [95]. A key challenge with generating discharge summaries is that they tend to be too general. This broad scope often fails to capture the domain-specific information that specialists require, thus limiting this task's application in the real world.

2.2.3 Radiology Reports

A parallel stream of research has emerged around radiology report summarization, utilizing the MIMIC-CXR dataset [98]. These reports typically contain three sections: *Background*, *Findings*, and *Impressions*. The task involves summarizing the detailed *Findings* section, using the *Impressions* section as ground truth, as it represents a condensed version of the findings [99–101].

2.2.4 State-of-the-art Methods

State-of-the-art approaches to clinical note summarization primarily employ hybrid methodologies based on medical entity annotators [2, 15, 102], specialized prompting techniques [15, 100] or reinforcement learning [94, 101]. Our research diverges from conventional approaches that generate unstructured output text. Instead, we aim to produce structure reports customized to specific medical specialists needs. This approach addresses the significant variation in how clinical documents are usually structured and presented [103], enabling clinicians to efficiently query specific patient attributes while maintaining access to comprehensive information when needed. The domain of structured summarization in clinical settings remain relatively unexplored as research mainly focused on generated structured summaries (SOAP/PICO) from patient-doctors conversations [19, 104]. Limited work has addressed the transformation of existing clinical notes into structured, queryable summaries [105]. State-of-the-art methods on structured summarization is mainly based on extractive approaches that classify utterances into sections of the structured summary [17, 19, 104–106]. Plus, our research specifically addresses the need to adapt these summaries for different medical specialties, as different specialists require distinct types of information [107]. This represents a novel direction in the field, as currently, to the best of our knowledge, there exists no dataset containing ground truth summaries tailored to different medical domains based on the same clinical notes.

2.3 Hallucinations

Hallucinations represent a significant concern in LLMs, particularly regarding their application in critical fields like medicine. Following [108], we define a hallucination as a span s of generated tokens $w_i...w_{i+j}, j \geq i$ that is not supported from either user-provided context or factual data. Informally, hallucinations are defined as *generated content that is nonsensical or unfaithful to the provided source content* [108, 109].

2.3.1 Open-Domain vs Closed Domain Hallucinations

The literature distinguishes between hallucinations based on the type of knowledge accessed by LLMs during generation [110]. During pre-training and supervised fine-tuning, knowledge is embedded within the LLM's parameters. This is referred to as *parametric knowledge* [111]. LLMs consistently utilize this knowledge when responding to user queries. The second form, *contextual knowledge*, derives from user input (called prompts) [111]. For example, when prompting a model to summarize clinical notes using domain-specific important elements, parametric knowledge guides the selection of domain-relevant information, while contextual knowledge enables the generation of summaries specific to the provided clinical document. This distinction has led to the categorization of hallucinations based on these two knowledge types [110]. *Open-domain hallucinations* relate to parametric knowledge or training data [112]. If a model has learned during training that bananas are yellow, it should consistently output this information. If it generates that bananas are blue, this constitutes an open-domain hallucination (if the information is not provided in the user prompt). *Closed-domain hallucinations*, on the other side, relate to contextual knowledge [4]. In our context, closed-domain hallucinations are essentially the inverse of groundedness. When a generated text is free from closed-domain hallucinations, it is, by definition, grounded, as per Definition 1.3.3. These types of hallucinations are further categorized into intrinsic and extrinsic types.

2.3.2 Intrinsic vs Extrinsic Hallucinations

The literature defines *intrinsic hallucinations* as outputs that directly contradict the provided context [4,6]. In contrast, *extrinsic hallucinations* are statements in the generation that seem plausible, but cannot be verified by the original context provided [4, 6, 108]. For example, if a source document indicates low blood pressure, but the model generates that the blood pressure is high, this qualifies as an intrinsic hallucination. However, if the source document contains no blood pressure information, but the model generates a statement saying that blood pressure is high, this represents an extrinsic hallucination.

2.3.3 Solutions

Researchers have developed various approaches to mitigate hallucinations in LLMs, which can be systematically categorized into four distinct solution types : design-time solutions involving architectural modifications, training-time solutions that enhance learning paradigms, generation-time solutions that optimize inference procedures, and external tools that augment model capabilities through external systems.

Design-time solutions

Design-time solutions encompass architectural modifications to the original Transformer architecture [3], specifically aimed at reducing hallucinations. These changes primarly focus on three key areas : copy mechanisms, softmax redefinition and parametric memories.

Copy mechanisms enables models to directly duplicate text from the input rather than generating new tokens [113, 114]. This approach has proven effective in tasks requiring minimal closed-domain hallucinations, such as summarization, as it eliminates the need for models to regenerate sequences that already exist in the input document, thereby reducing hallucinated content through direct sequence copying. While these architectural modifications predated the emergence of LLMs and became close to obsolete due to LLMs being better and better at copying text from the input sequence due to their large training phase, recent research has improved factual accuracy by integrating these mechanisms into LLMs without requiring any fine-tuning [115, 116].

Softmax function modification is a also a technique used to prevent hallucinations which focuses on the Softmax function, a fundamental component of LLMs' autoregressive property. Recent work has identified the Softmax function as a significant constraint on LLMs' expressiveness and faithfulness [117, 118]. To address this limitation, researchers have proposed various modifications to the Softmax function, primarily centered around implementing mixtures of Softmax functions [117, 118].

Parametric memories involve modifying the classical architecture of LLMs in order to integrate a new neural memory module whose goal is to store knowledge [119–121]. This memory system can be modified post-training to correct hallucinated content. It primarily targets closed-domain hallucinations by attempting to separate the model's capacity for natural language generation from its learned world knowledge.

Training-time solutions

Training-time solutions aim to modify the traditional training paradigms of LLMs to enhance factuality, though this approach often presents a trade-off where models become overly cautious to avoid potential inaccuracies.

Reinforcement learning has emerged as a prominent technique for improving LLM response accuracy. Specifically, Reinforcement Learning from Human Feedback (RLHF) [122] has shown significant potential in improving factual generation. This approach involves sampling multiple generations for a given prompt, ranking these generations using human evaluators, training a reward model based on these evaluations and training the original model

with the reward model. Established algorithms in this domain include Proximal Policy Optimization (PPO) [122], Direct Preference Optimization (DPO) [123], and Kahneman-Tversky Optimization (KTO) [124]. Recent advancements have introduced variations of reinforcement learning approaches that provide fine-grained feedback on specific elements within generations [125, 126]. While these methods require more extensive annotations, it offers more precise guidance by directly identifying which parts of the generated content are incorrect according to specific criteria such as factuality or toxicity. Additionally, researchers have applied reinforcement learning reasoning hallucinations by penalizing models when reasoning steps lead to incorrect conclusions [127, 128].

Loss function adjustement techniques aim to reduce hallucinations by modifying the convential fine-tuning paradigm. Traditional fine-tuning relies on negative log-likelihood minimization represented by the following loss function :

$$\mathcal{L} = -\sum_{i=0}^{N} \log p_\theta(y_t | x_i y_{<t}) \tag{2.1}$$

Where:

- N represents the number of samples
- y_t is the next token to be considered at time t
- p_θ is the model's probability of generating token y_t
- x_i is the input sequence; $y_{<t}$ is the generated sequence up to time t

This conventional loss function focuses solely on predicting the next token, lacking mechanisms to verify factual accuracy. Given that LLMs are trained on massive internet-sourced corpora, next-token prediction does not inherently prevent hallucinations. Research has shown that this training objective may potentially increase hallucination rates [129–131]. While current advances in the field focus on using LLMs for data filtering when gathering the training data [132], researchers have explored various improvements to this loss function [131, 133], including approaches based on contrastive learning [60, 134, 135].

Latent space understanding methods seek to explore the underlying representational space generated during LLM training. These approaches often leverage weight interpolation [136], based on the observation that the final latent space from optimization exhibits linear connectivity [137]. A key concept in this field is the task vector. Given a pre-trained model

θ_{base} and a fine-tuned model $\theta_{fine-tuned}$ on a specific task t. A task vector is defined as :

$$v = \theta_{fine-tuned} - \theta_{base} \tag{2.2}$$

This approach allows for steering model attention towards specific capabilities [138]. Initially developed for domains outside the field of NLP [139], these techniques have shown effectiveness for LLMs [140–142]. By identifying the task vector associated with hallucinated content, researchers can potentially develop intervention strategies that steer the model away from generating inaccurate information by generating anti-hallucination vectors [143]. It is important to note that this technique is really sensible to the task t and the architecture used as it depends on the latent space generated when training a certain architecture on a task.

Generation-time solutions

Generation-time solutions include techniques that modify the classical inference pipeline to improve task accuracy. These can be categorized into pre-generation and post-generation techniques.

Pre-generation techniques aim to reformulate and augment the context given to LLMs to enhance model accuracy and reduce hallucinations. These techniques include zero-shot prompt engineering, few-shot prompting, and chain-of-thought reasoning. Zero-shot prompt engineering involves formulating the problem with precise context for the LLM. This includes several approaches:

- **Role prompting:** Assigns a specific role to the LLM to establish the perspective from which it should respond ("You are a doctor whose role is", "You are a mathematician", etc) [144, 145].
- **Style prompting:** Specifices the desired output style or tone ("Write in a poetic manner", "Write in a professional manner", etc) [146].
- **Emotion prompting:** Incorporates emotional elements to emphasize the importance of the question ("This is important to me", "My career depends on it") [147, 148].
- **Format prompting:** Defines the required output format (json, yaml, list, etc) [149, 150].

While seemingly straightforward, these techniques have demonstrated improved performance and reduced hallucinations in specific tasks. Few-shot prompting involves presenting the LLM with n examples of a task before requesting it to perform the same task [151]. This technique

is particularly effective when similar examples are available and the same process needs to be applied to new cases. For instance, LLMs can easily transform statements into questions when provided with a few examples without requiring fine-tuning. The main problems with this technique lies in determining the optimal value of n and selecting appropriate examples. While increasing the number of examples generally improves performance [151], it also increases token usage. Several approaches have been proposed for example selection, some based on example similarity to the input [152–154] and other more sophisticated methods based on filtering and generated samples [155, 156]. Chain-of-Thought (CoT) prompting encourages LLMs to articulate their reasoning process before providing an answer [157]. This technique has shown to improve performance and reduce hallucinations in tasks requiring reasoning capabilities. CoT can be combined with few-shot prompting to demonstrate possible reasoning paths for solving problems. More advanced CoT-based techniques have incorporated decoding methods [158, 159] and contrastive learning to validate reasoning paths and enhance performance [160, 161]. More sophisticated pre-generation techniques include problem decomposition [162, 163], which decomposes the problem into smaller simple problems, and ensemble techniques, which use multiple prompts on the same task or sample multiple reasoning paths [164, 165] in order to average their predictions. While these pre-generation techniques can be effective in certain use cases, they all come with a significant trade-off: increase token usage, resulting in longer inference times and higher memory requirements.

Post-generation techniques involve multiple inference passes with the model to verify, modify or critique its initial response. While these additional passes increase the latency until the final output is generated, they enable the model to improve its answer through self-reflection. Methods like Self-Check [166] and Chain-of-Verification (COVE) [167] generate feedback based on the original input and the model's initial answer, using this feedback to identify and correct potential hallucinations. These approaches have demonstrated improvements in summarization tasks, where summaries can be broken down into individual statements for the LLM to validate or invalidate. Alternative methods [168, 169] generate multiple possible outputs and task the LLM with evaluating these alternatives, leading to reduced hallucinations as the model selects the most factually accurate version. The feedback in these approaches can be generated by the LLM itself [170, 171], other LLMs [172, 173], and templated questions [166, 167]. Some researchers have also taken a preventive approach by attempting to stop the model before hallucination occurs. For instance, self-familiarity [174] evaluates the model's familiarity with the concepts in the input instruction and withholds response generation when encountering unfamiliar concepts. Similarly, Self-Ask [175] determines whether additional questions need to be asked before providing an answer. While these methods have demonstrated improvements in text generation tasks, they still face some

limitations:

- **Computational overhead:** The requirement for multiple inference passes significantly increases computational requirements per sample. This creates a trade-off between speed and accuracy: researchers and users must choose between a faster response that might contain hallucinations or a slower, more accurate response. Essentially, factuality must be traded for speed. More recently, work has been done to remove that trade-off [176]
- **Reliability of self-evaluation:** These methods rely on the assumption that LLMs can accurately identify hallucinated content in their own outputs. However, there is no guarantee that LLMs won't hallucinate during their evaluation process. While recent research suggest that LLMs may have some awareness of when they are hallucinating [112], these studies have been limited to open-ended hallucinations and specific use-cases.

External tools

Hallucination mitigation strategies increasingly leverage external tools to ground LLMs' generations. These tools include a broad range of resources external to the LLM's parameters, including databases, web search engines, knowledge bases, and code execution environments.

Retrieval Augmented Generation (RAG) emerges as the most promising approach, utilizing external vector databases to enhance model performance [86, 177]. RAG transforms how LLMs access and utilize information by supplementing parametric knowledge with contextual retrieval. LLMs can thus use their contextual knowledge on par with their parametric knowledge in order to respond to a query. The RAG process involves several key steps:

1. **Document Encoding:** An embedding model transforms documents into vector representations. Documents can include various information types such as text or knowledge graph triplets
2. **Database Creation:** Encoded documents are stored in a vector database.
3. **Query Retrieval:** During inference, the input query is embedded using the same embedding model
4. **Context Augmentation:** The most similar k documents (a configurable hyperparameter) are retrieved and integrated into the prompt.

5. **Generation:** The LLM generates the answer to the query using the provided context.

For example, when queried "What is the age of the patient ?", RAG might retrieve and append the admission note of a patient containing such information to provide contextual grounding. This approach has demonstrated significant hallucination reduction, particularly in question-answering and up-to-date knowledge-dependent tasks [177]. Researchers have developed numerous improved versions of RAG including:

- **Re-ranking mechanisms:** these mechanisms take as an input the k most relevant documents and re-ranks the documents. This step is more accurate than the classical retrieval method as it does not rely on pre-computed embeddings. The query is directly compared to each document through a re-ranking model [86, 87].
- **Query routing and refinement techniques:** These techniques aim to get a better understanding of the query leading to improved vector representation [88, 178, 179].
- **Optimized document chunking:** This steps aims to filter out irrelevant information in documents to optimize how documents are chunked in the vector database [180].
- **Context optimization:** This step aims to reduce the size of the retrieved content while filtering out the irrelevant information through data filtering and summarization [181].

Despite its effectiveness, RAG is not infallible in terms of hallucinations. Two primary challenges persist [6]:

- **Retrieval Failure:** When retrieved content fails to relevantly match the query, it can increase hallucination as it introduces a lot of irrelevant information [182].
- **Generation Bottleneck:** The model may struggle to effectively utilize or extract information from the retrieved context [183]. This is especially the case if the context is noisy or conflicting [184].

Alternative approaches to hallucination mitigation using external tools extend beyond RAG. Web search integration represents a promising avenue, enabling LLMs to query internet resources through APIs and retrieve up-to-date, factual information in real time [185]. Code execution environments offer another strategy for hallucination mitigation, particularly in technical domains like mathematics and software development. By allowing LLMs to execute code and receive immediate runtime feedback, these environments provide a concrete

verification mechanism [185–187]. Researchers have observed significant reductions in hallucinations in problems that require precise calculation or algorithmic problem-solving, where the gap between generated text and executable logic can be bridged.

2.4 Decoding strategies

Decoding strategies represent modifications to the classical inference process that lead to generations satisfying specific constraints. Modern Large Language Models (LLMs) generate text from left to right, one token at a time. At each time step t, the LLM outputs not just the next token, but a probability distribution over its entire vocabulary. This provides access to each token's probability of being the next in the sequence. Formally, given the input X and the generated tokens $y_1...y_t$, LLMs output the probability of token t given the generated sequence and the input :

$$P(y_t|x, y_1...y_{t-1}) = LLM(x, y_1...y_{t-1}) \tag{2.3}$$

Traditionally, the most probable token is selected as LLMs are trained to maximize the probability of the next token in the sequence. This approach is known as *Greedy Decoding*. Using equation 2.3, greedy decoding is defined as :

$$y_t = \underset{j}{\operatorname{argmax}}\, P(y_t|x, y_1...y_{t-1}) \tag{2.4}$$

While this technique can sometime lead to close-to-optimal generations, numerous alternatives have been proposed to modify this decoding process. Notably, the process isn't limited to tracking only one token at a time. It is possible to track multiple tokens simultaneously - for instance the k most probable - leading to multiple possible generations at each time step. This approach, known as *Beam Search* [188], tracks multiple possible generation paths concurrently. These generation paths are referred to as beams. Recent state-of-the-art decoding strategies vary in their application level: either at the token level or the beam level. Token-level strategies modify the model's probability distribution at each time stamp, while beam-level strategies alter beam selection criteria to meet constraints.

2.4.1 Token-level Decoding Strategies

Token-level decoding strategies serve multiple purposes, including generation diversification and structure incorporation. For instance, when requiring the model to output valid JSON, constraints must be applied at the token level. Key token-level algorithms include :

- **Temperature Sampling** adjusts the token probability distribution using a hyper-parameter τ that controls the distribution's skewness [189]. The next token is then sampled from the adjusted probability distribution. In practice, the implementation involves modifying the Softmax function :

$$\frac{e^{y_i}}{\sum_{k=1}^{n} e^{y_k}} \longrightarrow \frac{e^{\frac{y_i}{\tau}}}{\sum_{k=1}^{n} e^{\frac{y_k}{\tau}}} \tag{2.5}$$

- **Top-k Sampling** only considers the k most probable tokens during the sampling process [190].

- **Top-p Sampling** considers only the most probable tokens whose cumulative probability meets or exceeds p during sampling [191].

- **η-Sampling** discards tokens whose entropy falls below a specified threshold, with the next token sampled from the remaining candidates [192].

- **Contrastive Search** adjusts token probability distribution using a degeneration penalty to enhance diversity while maintaining coherence [193].

- **Contrastive Decoding** modifies the probability distribution using the distribution from a smaller LLM to enhance generation quality [194]. It uses the mistakes a smaller LLM would make to improve the generations of a bigger one.

Attribute-based discriminators

Researchers have developed methods to condition generation on desired attributes [195–197]. These attributes could be topics the model should address during the generation or tone that should be incorporated into the output sequence. These approaches employ a Bayesian factorization trick and a discriminator model that predicts attribute satisfaction given the generated sequence.

$$LLM_{\text{fine-tuned on } a}(x, y_1 ... y_t - 1) = P(y_t | x, y_1 ... y_t - 1, a) \tag{2.6}$$

Given the attribute a, rather than conditioning through fine-tuning as in equation 2.6, they utilize this factorization trick :

$$P(y_t | x, y_1 ... y_{t-1}, a) \propto P(a | x, y_1 ... y_t) * P(y_t | x, y_1 ... y_{t-1})$$
$$P(y_t | x, y_1 ... y_{t-1}, a) \propto \text{Discriminator}(a | x, y_1 ... y_t) * LLM(x, y_1 ... y_{t-1})$$

Structure-based discriminator

A significant branch of decoding strategies focuses on imposing format constraints on generation. This is particularly valuable when integrating LLM outputs with APIs or tools (see Section 2.3.3). Researchers have employed finite-state machines to constrain LLM decoding processes, ensuring generations conform to predefined regular expressions [198–202]. [203] extended this approach using formal representations based on logic and code. Further advances have established a general framework allowing any grammars to be included with the input through an incremental token-level parser [204]. These grammars can be input-dependent, allowing per-sample grammar variation, which is particularly useful for tasks like entity linking and extractive summarization. These algorithms utilize token pruning, where certain token probabilities are set to zero to prevent invalid token generation. These methods can be summarized by the algorithm below:

1. Generate probability distribution of tokens :
   $P(y_t|x, y_1...y_{t-1}) = LLM(x, y_1...y_{t-1})$
2. Process the probability distribution through a parser :
   $P^*(y_t|x, y_1...y_{t-1}) = \text{Parser}(P(y_t|x, y_1...y_{t-1}), x, y_1...y_{t-1}, G)$
3. Select the next token based on the adjusted probability distribution :
   $y_t = \underset{j}{\text{argmax}}\, P^*(y_j|x, y_1...y_{t-1})$

2.4.2 Beam-based Decoding Strategies

Beam-based decoding strategies operate at the beam level, maintaining multiple possible generations simultaneously based on the beam search algorithm [188]. This approach mitigates the risk of missing high-probability token sequences that might be obscured by low-probability tokens at intermediate steps. The final output typically selects the most probable beam, though the k most probable sequences could also be considered. Beam search has proven particularly effective in tasks with variable output lengths, such as summarization and translation [205, 206]. Current state-of-the-art beam-based decoding strategies include:

- **Diverse Beam Search** organizes the k most probable sequences into n groups and incorporates a diversity term to generate more diverse beams [207].
- **Grid Beam Search** extends beam search to accommodate lexical constraints in generation, such as required words or text sequences in the final output [208].

- **Best-first Beam Search** enhances beam search efficiency through priority queue implementation [209].
- **Improved Beam Search for Hallucination Mitigation** employs an NLI model to re-rank beam probabilities, reducing hallucinated content in generations [210].

2.5 Evaluation Metrics

Evaluation metrics assess LLMs' performance on specific tasks. In the context of this work, we focus on two categories of metrics: those evaluating summarization capabilities and those measuring hallucination levels in model generations.

2.5.1 Summarization Evaluation Metrics

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [211] stands as the field's standard metric for summarization evaluation. Given a reference summary and the ground truth summary, ROUGE measures the n-gram overlap between the two summaries, where n determines the level of textual similarity required (in practice n is usually 1 or 2). For example, ROUGE-1 and ROUGE-2 measure unigram and bigram overlap respectively. Plus, ROUGE supports multiple ground truth summaries. The metric supports multiple ground truth summaries and is formally defined as:

$$\text{ROUGE-N}(C) = \frac{\sum_{S \in R} |G(C, N) \cap G(S, N)|}{\sum_{S \in R} |G(S, N)|} \tag{2.7}$$

where R represents the set of reference summaries, and $G(S, n)$ returns the set of n-grams in summary S. Variants like ROUGE-L and ROUGE-W incorporate longest common subsequence (LCS) approaches [211]. However, ROUGE's reliance on exact n-gram matching presents a significant limitation. It penalizes semantically equivalent summaries that use different wordings, potentially favoring summaries with matching words but conflicting information over those expressing identical information through different wording. METEOR is another metric that improves ROUGE by considering word stems [212]. It calculates the harmonic mean of unigram precision and recall, incorporating exact word matches and stemming to account for morphological variations. Additionally, METEOR considers synonyms, making it more flexible in assessing the semantic accuracy of translations. However, it is still limited in terms of word formulations. This limitation led to the creation of context similarity metrics.

Context Similarity Metrics. Context similarity metrics leverage pre-trained encoder mod-

els to evaluate summary quality. These metrics utilize the encoder's ability to map text sequences into a high-dimensional latent space where contextually similar content are closer to each other. BERTScore [213], based on the BERT model, has emerged as the standard metric in this category. Alternatives include BARTScore, which evaluates text likelihood using the BART model [49], MoverScore [214], which enhances BERTScore with soft alignments, and SynWMD [215], which incorporates syntactic awareness. While effective at measuring summary relevance (see 1.3.2), these metrics may assign high similarity scores to contradictory statements about the same topic, such as conflicting patient age descriptions. For example, the sentences "The patient is 70 year old" and "The age of the patient is 60 years old" will have a high BERTScore as the sentences detail the same topic (Age). However, they detail contradictory statements.

Reference-free Metrics. A fundamental challenge in summarization evaluation lies in the absence of a definitive "gold" summary. Reference summaries provided in datasets serve as examples of acceptable summaries rather than comprehensive solutions. This inherent limitation makes it difficult to assess similarity between summaries. In addition, this challenge is compounded by the fact that the ground truth used for metric computation represents just one of many possible valid summaries. To address these challenges, researchers have developed reference-free automatic metrics, focusing primarily on factual detection:

- **SummaQA** evaluates summary quality by quantifying its ability to answer questions automatically generated from the source text [216].
- **SUPERT** assesses summaries by measuring their semantic similarity to pseudo-reference summaries, which are constructed from salient sentences in the source text [23].
- **QuestEval** builds upon SummaQA's framework, leveraging pre-trained language models to compute quality scores [217].

2.5.2 Hallucination Evaluation Metrics

While reference-free metrics inherently assess hallucination, researchers have also developed specialized metrics specifically targeting hallucination detection in generated content:

- **SRLScore** evaluates factual consistency by comparing fact tuples extracted from the input document and the generated summary [218].
- **FACTSCORE** decomposes generated content into atomic facts and calculates the percentage supported by reliable knowledge sources [219].

- **FEQA** quantifies hallucinations by assessing the summary's ability to accurately answer questions derived from the source text [220].

Another significant approach utilizes Natural Language Inference (NLI) models to detect factual inconsistencies. These models assess whether a given premise entails a hypothesis by computing an entailment score. Metrics such as SummaC [221], FactCC [222], and DAE [223] leverage this capability by treating the input text as premises and decomposed summary statements as hypotheses. However, these metrics are still pretty limited as the models are not entirely 100% accurate.

2.5.3 LLM-as-a-Judge

A recent development in evaluation methodology known as *LLM-as-a-Judge* involves using LLMs themselves as judges [224, 225]. This approach typically employs larger models to evaluate smaller ones, as using models of comparable size would raise questions about evaluation reliability. Research has shown that LLM-based evaluation correlates more strongly with human judgment, particularly in tasks like summarization where multiple valid answers exist [226]. Most common LLMs used as evaluators are usually general closed large language models like GPT4 [227], Claude and Gemini [228]. However, some alternatives have been proposed by the open source community like the Prometheus models [229] which are specifically fine-tuned for evaluation.

CHAPTER 3 ONTOLOGY-CONSTRAINED GENERATION OF DOMAIN-SPECIFIC CLINICAL SUMMARIES

This chapter is based on the paper *Ontology-Constrained Generation of Domain-Specific Clinical Summaries* by Gaya Mehenni and Amal Zouaq published in the 24th International Conference on Knowledge Engineering and Knowledge Management, on November 26th 2024 [230]. It also incorporates subsequent research and developments stemming from that work.

3.1 Introduction

In the past few years, large language models (LLMs) have demonstrated significant advancements in their extraction and summarization capabilities [75, 89, 95, 231], offering potential to automate the processing of complex medical information, such as Electronic Health Records (EHRs) and clinical notes [2]. These records, regrouping overwhelming amounts of information, are known to significantly contribute to clinician burnout [1]. Generating domain-specific summaries, which would efficiently encapsulate the information needed by each specialist could ease this task. However, applying LLMs to medicine presents several challenges: clinical documentation uses specialized terminology, lacks standardized structure. Plus, LLMs tend to hallucinate content, which, in this setting, is particularly problematic. These issues get even worse when the generated content must be tailored to different medical contexts, as information priorities vary substantially between medical specialties. For instance, oncologists require a different set of information as radiologists. Medical ontologies offer a solution to these problems by providing structured knowledge representations that can be used to identify key concepts and relationships within particular fields aka *domains*. These can guide the extraction process of relevant information from clinical notes, enabling domain-adapted summarization, while reducing hallucinations. A remaining challenge involves constraining LLM generation to specific domain concepts and properties to avoid producing non-factual, ungrounded information.

3.2 Methodology

Our research explores the potential of using ontologies to guide a language model towards relevant information using prompting and constrained generation. By imposing contraints on the generation process using ontological structures, we aim to enhance the extraction and summarization capabilities of language models while mitigating the risk of hallucinations.

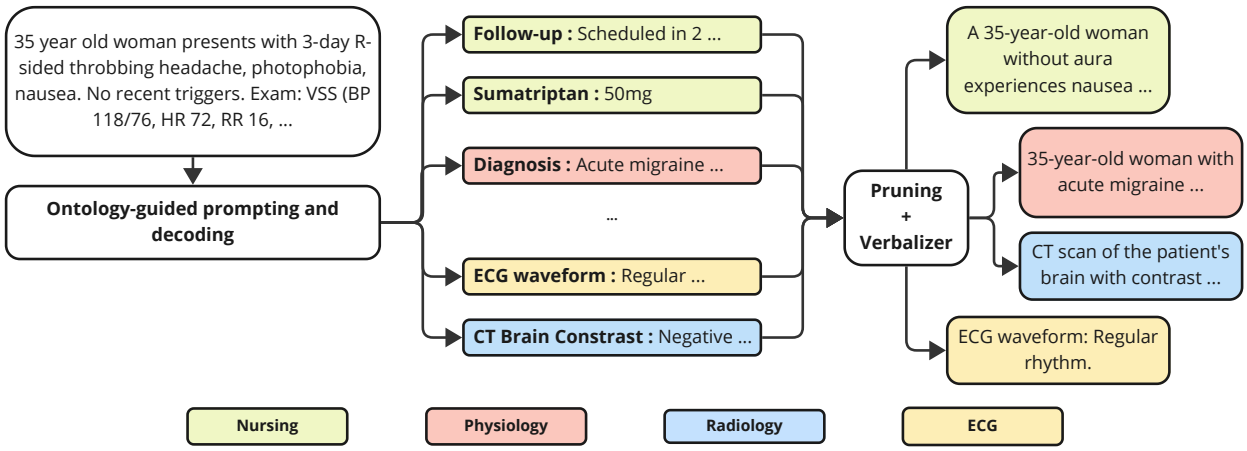


Figure 3.1 General overview of how our method generates domain-adapted clinical summaries

To achieve this, we propose a methodology that leverages the structured knowledge encoded within ontologies to guide models generations in an informative and grounded manner. At the heart of our methodology lies the concept of ontology-guided beam search. We utilize the ontology in conjunction with the diverse beam search algorithm [207] to evaluate the relevance and factual accuracy of potential beam candidates in relation to the input context. By iteratively assessing the alignment of candidate generations with ontological knowledge, we expect to enhance the overall groundedness (see Section 1.3.3) and relevance (see Section 1.3.2) of the generated text, ensuring that it aligns more closely with the knowledge represented in the ontology. To further enhance the groundedness and relevance of generations, we incorporate an evaluation of beam paths based on their resemblance to the clinical note. This ensures that the information extracted by LLMs is not only aligned with the ontological knowledge but also reflective of the specific details and nuances present in the source clinical document. Figure 3.2 illustrates how, from multiple clinical notes for a given patient, our method generates both a textual (unstructured) summary and a structured summary defined by ontology concepts. This structured representation allows for adaptation of the final summary to various medical domains, such as cardiology and oncology. Furthermore, our approach is model-agnostic, requiring only token probabilities for its application.

As shown in figure 3.2, our methodology is based on five main steps: domain adaptation analysis, information extraction, constrained decoding, pruning and verbalization.

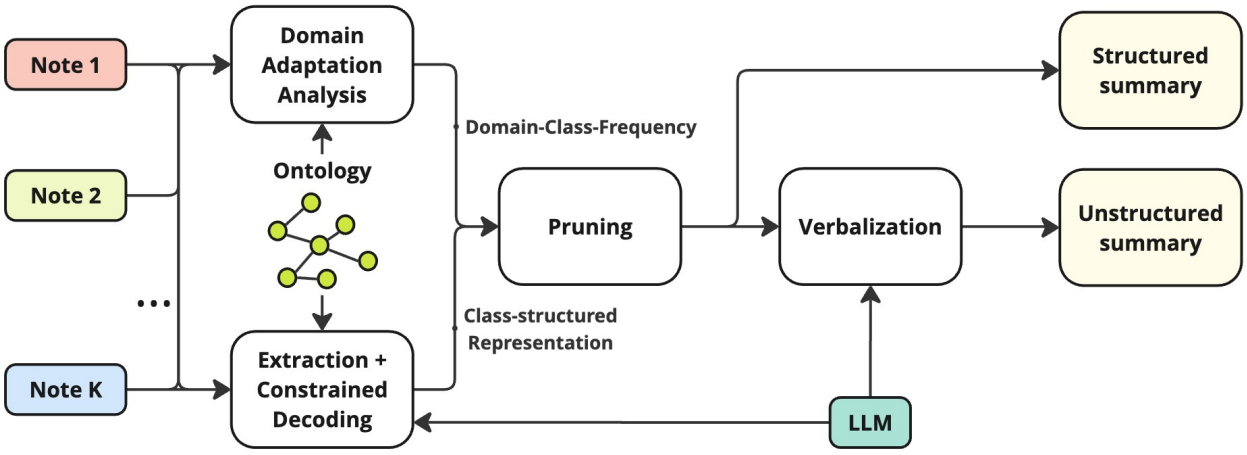


Figure 3.2 **Overall architecture of our method:** Structured and unstructured summaries can be generated from multiple clinical notes about the same patient

3.2.1 Domain Adaptation Analysis

As mentioned in section 1.3.2, we define a domain to be a set of ontology concepts of interest related to a specific medical field. While the domain can be defined by medical experts, we define it based on clinical notes related to specific medical fields. Given texts that are linked to a certain domain D, we aim to identify the most important concepts in each text. We start by annotating each text using an ontology-based annotator. In our case, we utilized the MedCAT annotator [232]. We then create a set S based on a minimum occurrence threshold. We presume that the annotator can detect different formulations of the same concept (abbreviations, plurals, etc). Then, using the ontology, we retrieve all ancestors of each concept and add the ancestors to S. This step aims to filter out overly specific concepts tied to individual patient notes, as our goal is to capture a general overview of domain-relevant concepts. Figure 3.3 provides an example illustrating the importance of this filtering process in achieving such a broad understanding of key domain concepts. Subsequently, the frequency of each class in S is computed and stored in a class-to-frequency dictionary. We define this dictionary as the *Domain-Class-Frequency* (DCF) dictionary. Its goal is to store the most relevant concepts in a domain to later guide the generations towards these concepts. Then, each DCF is normalized according to the average DCF, computed by averaging the class frequencies across all domains. This normalizing step ensures that each DCF only contain relevant concepts to the domain and reduces the weight of general medical concepts that are higher in the ontology hierarchy. Finally, only the top k most frequent concepts are kept. Here, k controls how precise we want our definition of a domain to be. This methodology is

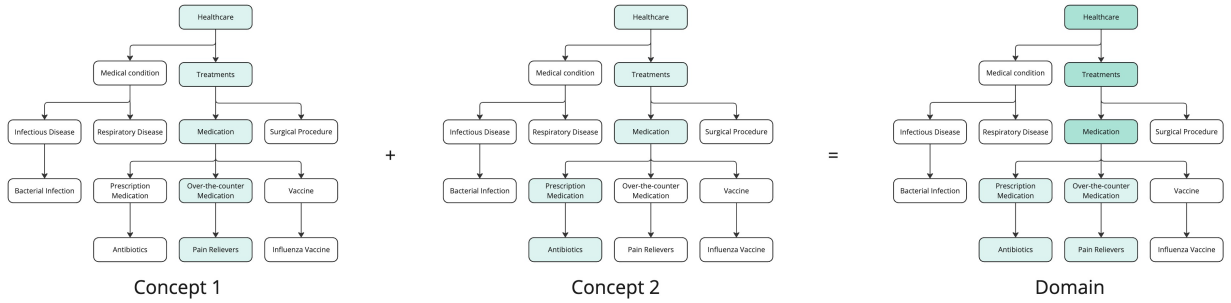


Figure 3.3 **Domain Adaptation Analysis:** By retrieving all ancestors of each concept, we get a broad understanding of general concepts present in the domain. In this case, we only showed two concepts for the domain, but this algorithm should be computed with multiple concepts.

presented in algorithm 1. As shown in Figure 3.2, the DCF dictionary is used in the pruning phase to perform domain adaptation.

3.2.2 Information Extraction using Ontology-based Prompting

Building upon the domain definition through ontological concepts established in the previous step, we now focus on extracting concept-specific information from clinical notes. Drawing inspiration from [71], we introduce a summarization process guided by ontology-based prompting. This approach aims to generate structured representations of clinical notes, enabling doctors to efficiently query concept-based information while facilitating domain-specific summarization through integration with the DCF dictionary (see Section 3.2.1). The process begins with the annotation of multiple clinical notes from a patient using a medical ontology to identify relevant medical concepts. For each clinical note, we identify the k most frequent concepts using the same process defined in Section 3.2.1 minus the normalization stage. These concepts serve as the basis for subsequent steps. Capping the number of concepts to k helps us balance computational efficiency with the amount of information extracted. While including all concepts would yield more exhaustive data, it would incur significant computational overhead, and infrequent concepts are often harder for models to reliably extract due to their limited mentions. The extraction process involves prompting the model to extract information about individual concepts on the same clinical note through multiple inference passes. This architecture allows for efficient parallelization, as each note and concept can be processed independently. Following a RAG-like methodology, we enhance the prompt by incorporating ontological information associated with each concept, specifically derived from the concept's restriction properties. The prompt template and an example

Algorithm 1 Domain Adaptation Analysis

```
function DOMAINADAPTATIONANALYSIS(domainTexts, ontology, annotator, k)
    concepts ← []
    for all text in domainTexts do
        textConcepts ← annotator(text)
        for all textConcept in textConcepts do
            ancestors ← ontology.getAllAncestors(textConcept)
            for all ancestor in ancestors do
                concepts.append(ancestor)
            end for
        end for
    end for
    frequencies ← getFrequencies(concepts)
    averageFrequencies ← getAverageFrequencies(frequencies)
    return getMostCommon(frequencies - averageFrequencies, k)
end function
```

are illustrated in Figure 3.4.

Our hypothesis is that this ontological augmentation provides the model with richer context about the concepts. Including a concept's ontological characteristics improves the model's extraction capabilities by providing a comprehensive understanding of the concept's meaning and significance. The final output for each clinical note is a concept-structured representation (CSR), which maps detected ontology concepts to their corresponding extracted summaries from the notes. These summaries, referred to as "extracted values," are generated by the model using the specified prompt template. An example of a CSR is shown in Table 3.1.

Concept	**Extracted Value**
Laboratory test	Non-diagnostic repolarization abnaormalities on tracing #1.
Pain / sensation finding	The patient has an inferior myocardial infarction of indeterminate age.
Finding of heart rhythm	A regular supraventricular rhythm of indeterminate mechanism.

Table 3.1 Example of concept-structured representation associated to a clinical note

3.2.3 Constrained Decoding

During the information extraction phase, the model is prompted to extract information related to a specific concept. To ensure the relevance and factual accuracy of the generated

Here is a clinical note about a patient

[Clinical Note]

In a short sentence, extract the information that is related to the "[concept]" medical concept from the clinical note. "[concept]" is characterized by "[restriction properties]" If the concept is not mentioned in the note, respond with 'N/A'. Only output the extracted information.

Figure 3.4 Prompt template used to extract information

Here is a clinical note about a patient

John Doe, 68 y/o M, presents with dyspnea and mild chest discomfort. Vitals: BP 110/70, HR 98, SpO2 92%. Lungs show crackles. Electrocardiogram (ECG) performed: Sinus tachycardia with ST depression in leads II, III, aVF. Plan includes labs, CXR, O2, and cardiology consult.

In a short sentence, extract the information that is related to the "electrocardiogram" medical concept from the clinical note. "electrocardiogram" is characterized by "Evaluation - action AND Heart Structure and Electrocardiographic monitor and recorder, device". If the concept is not mentioned in the note, respond with 'N/A'. Only output the extracted information.

Figure 3.5 Example of prompt used in the case of the "Electrocardiogram" concept (as a real note from MIMIC can't be shown, a synthetic note was generated to illustrate the prompt)

responses, a novel decoding strategy is employed. This strategy leverages the knowledge embedded within the ontology to guide the model towards more relevant answers (see Section 1.3.2) and minimize hallucinations, ensuring groundedness with the clinical notes (see Section 1.3.3). Figure 3.6 illustrates the overall decoding process.

Our constrained decoding algorithm utilizes diverse beam search [207], incorporating grouped beam search to generate more diverse results during decoding. The algorithm prioritizes beams that exhibit textual similarity to the clinical note and contain concepts related to the target concept through various ontological relations (hierarchical, restriction properties). However, the challenge of detecting concepts from individual tokens is not trivial as concepts might include several tokens. For instance, in the sentence *The patient displayed signs of acute respiratory distress syndrome*, the respiratory disease mentioned is spread out throughout multiple words (acute respiratory distress syndrome). To address this, the beam score is computed after generating a specific number of tokens, denoted as the *generation window* (W). Every W tokens, the generated tokens are analyzed, and the beams are re-ranked

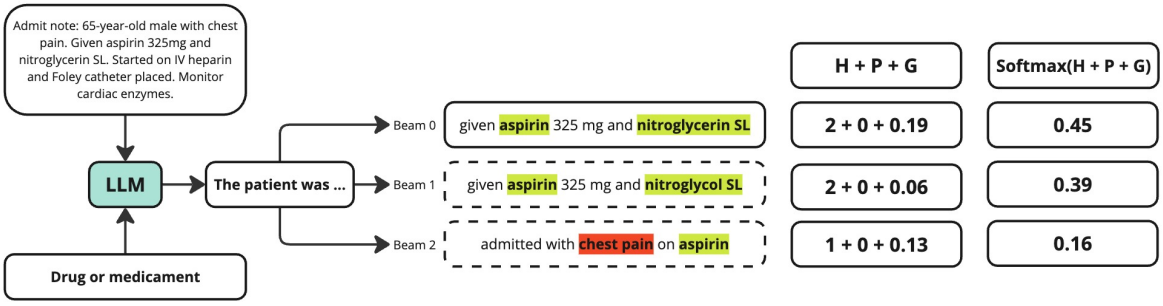


Figure 3.6 **Constrained decoding process:** Each beam (represented by a rectangle) corresponds to a generation window. Concepts highlighted in green indicate membership in a child concept of the base concept (e.g., "Drug or medicament") within the ontology. The presence of such concepts enhances the hierarchy score, increasing the likelihood of the beam being selected as a final output. The similarity score is calculated using the ROUGE-2 score between the generation window and the clinical notes.

based on a new computed score described below. The same annotator used for concept identification during the extraction phase (Section 3.2.2) is employed for the analysis of W.

Our new beam score comprises three sub-scores: hierarchy score (H), property score (P), and similarity score (S). For all scores, we define B to be the base class used to create the prompt. This corresponds to the concept designated by the "[concept]" tag in the prompt template. For instance, in Figure 3.5, the base concept is *electrocardiogram*. T denotes the newly generated tokens within the generation window, C represents the set of concepts detected in T, and $A(c)$ signifies the set of ancestors for concept c according to the ontology. Overall, our objective is to favor beams that include concepts linked to the base concepts through relations in the ontology and that are textually similar to the clinical note. These relationships can be hierarchical or restriction-based.

Hierarchy Score

The hierarchy score (H) quantifies the presence of descendants of the base class within the generated beam:

$$H = H_{bf} \frac{1}{|C|} \sum_{c \in C} \mathbb{1}\{b \in A(c)\} \tag{3.1}$$

Here, H_{bf} represents the hierarchy boost factor, a hyperparameter that controls the influence of the hierarchy score on the final beam score. H is calculated based on the *subClassOf* property of concepts in the ontology. The intuition behind this score is that, when querying about the patient's diseases for instance, we expect the model's generations to contain

ontology concepts which inherit from the "disease" class, like "infectious diseases" or "respiratory diseases". The hierarchy sub-score guides beams towards relevant concepts within the ontology's hierarchy.

Property Score

The property score (P) assesses the relevance of a beam to the base concept based on its associated concepts through restriction properties. While applicable to any class property, we focus on restriction properties in this context. This score enables the decoding process to incorporate knowledge inferred from the ontology. Beams that mention concepts present in the restriction property objects are favored as they theoretically should be more relevant since they are related in the ontology. For example, when prompted about the "Fever" concept, we would prefer the answer to mention that this diagnosis was made because the patient has a body temperature above the normal range. We would then favor beams that mention the "Body Temperature" class. The property score is calculated as:

$$P = \frac{P_{bf} \sum_{c \in C} \mathbb{1}\{c \in P(b)\}}{|C||P(b)|} + \text{R2}(T, P^{'}(b))$$

where P_{bf}, similar to H_{bf}, is the property boost factor and R2 is the ROUGE-2 score. $P(x)$ is the set of concepts related to x through restriction properties and $P^{'}(x)$ is a natural language formulation of $P(x)$. In practice, we only consider *And* and *Or* restrictions. Given a class property restriction of the form { *property1* : *value1*, ... }, to compute $P^{'}(x)$, we simply concatenate all values in the case of an *And* restriction. For example, if P(Fever) = *AND(Interprets: Body Temperature, Has Interpretation: Above Reference Range)*, we have $P^{'}$(Fever) = *Body Temperature Above Reference Range.* In the case of an *Or* restriction, we add *or* between every value. We compute the ROUGE-2 score between the natural language formulation of the restrictions and the newly generated tokens in order to address the annotator's limitations when detecting concepts. This phenomenon mainly happens when the generation window cuts in half certain concepts that contain multiple words. For example, without the ROUGE-2 score, a beam containing *Body temperature above reference* won't be favoured as the annotator might not correctly identify the *Above Reference Range* concept. Adding the ROUGE-2 score allows the beam to still be favoured even though the word *range* was not included in the generation window.

Similarity Score

The similarity score (S) measures the textual similarity between a beam and the clinical notes. Given that model answers for a class should be concise and extract information directly from the notes, we hypothesize that the ROUGE-2 score effectively captures this similarity based on n-gram overlap:

$$S = S_{bf}\ \mathrm{R2}(T, N)$$

where S_{bf} is the similarity boost factor, N is the clinical note and T the tokens generated in that generation window. This score favors beams that closely resemble the clinical note, reducing the risk of hallucinations due to model paraphrasing. Furthermore, the similarity score is crucial for improving factuality in generations that contain specific values, such as a patient's vital signs (e.g., blood pressure, heart rate, temperature) mentioned in a clinical note. In such cases, other scores like the hierarchy or property scores might not help the model identify the correct numerical values, as these are not typically linked to ontology concepts. The similarity score addresses this by favoring beams that contain the accurate n-grams corresponding to these specific values in the original note.

Final Score

A beam's score B_{ours_i} is computed from its hierarchy score H_i, its property score P_i and similarity score S_i and converted to a probability with a softmax function applied across all beam scores.

$$B_{ours_i} = \mathrm{Softmax}(H_i + P_i + S_i) \tag{3.2}$$

To merge each beam scores with the original beam probability B_i, we linearly interpolate between B_{ours_i} and B_i using the following formula

$$B'_i = w_0 * B_i + w_1 * B_{ours_i} \tag{3.3}$$

where $w_0, w_1 \in [0, 1]$ are hyper-parameter controlling how much we want the ontology-based decoding process to have an effect on the generation such that $w_0 + w_1 = 1$. It is important to note that current LLM implementation usually work in log space for the probabilities. Thus, a conversion to linear space is needed before interpolating between the probabilities. Beams are then re-ranked based on this B'_i to control the LLM's generation process.

3.2.4 Pruning

After the information extraction phase (Section 3.2.2), each clinical note is represented by a Concept-structured representation (CSR). To tailor the generated summary to a specific domain, we prune the CSR to retain only the information relevant to that domain. This pruning process leverages the Domain Concept Frequency (DCF) of the target domain (see Section 3.2.1). The pruning step works as follows: Given a domain's DCF and a clinical note's CSR, we iterate through the CSR and, for each ontology concept in the CSR, we check if it is present in the DCF. If that's the case, we keep it and its associated information in the pruned CSR. Additionally, we also consider related concepts. Since concepts in the DCF and CSR might not match perfectly (e.g., the DCF might have "Cardiomyopathy", while the CSR has the more specific concept "Dilated Cardiomyopathy"), we incorporate a level of generalization. We keep a concept from the CSR if it's within α "child" nodes (in the ontology hierarchy) of a concept that is present in the DCF. For instance, if α=1, and the DCF contains "Cardiomyopathy," we would keep "Dilated Cardiomyopathy" (a direct child of "Cardiomyopathy"), even though it is not an exact match. This allows us to capture relevant information even when the extracted concepts are more specific than the domain concepts in the DCF. The result of this pruning process is a domain-specific CSR, containing only the information deemed relevant to the target domain. This pruned CSR then serves as the basis for generating the final domain-adapted summary.

3.2.5 Verbalization

The verbalization stage transforms the structured summary from the pruning phase into a textual format with an LLM. This conversion enables us to evaluate our methods using classical summarization metrics, which are designed to compare texts. Additionally, this step serves clinicians who prefer reading a conventional text rather than a concept-structured summary.

3.3 Experiments

This section details the experiments conducted to evaluate the proposed methodology, including the data used and the preprocessing steps applied. Our evaluation focuses on three key aspects: the effectiveness of generating domain-adapted summaries and the impact of our constrained decoding approach on relevance and groundedness.

3.3.1 Data

Experiments were performed using the Medical Information Mart for Intensive Care (MIMIC-III) dataset [93]. MIMIC-III comprises over 45,000 de-identified patient admissions to critical care units, containing 1.4 million clinical notes spanning 15 medical specialties. Table 3.2 describes the key columns of the dataset relevant to our experiments. The CATEGORY

Column name	Description
ROW_ID	Id of the row
SUBJECT_ID	Id of the patient
HADM_ID	Id of the hospital stay (admission)
CHARTDATE	Date at which the clinical note was written
CHARTTIME	Date and time at which the clinical note was written
CATEGORY	Type of note
DESCRIPTION	Description of the note (report, progress note)
TEXT	Note text

Table 3.2 Description of the main columns in the MIMIC-III dataset

column plays a crucial role in defining our domains, as further elaborated in Section 3.3.1. This column specifies the type of recorded note, enabling a clear classification of the clinical data. Possible values for this column are : *Case Management, Consult, Discharge summary, ECG, Echo, General, Nursing, Nursing/other, Nutrition, Pharmacy, Physician, Radiology, Rehab Services, Respiratory, Social Work.*

Pre-processing

MIMIC-III clinical notes underwent three preprocessing stages to prepare the data for our experiments. These steps were applied consistently across all evaluation tasks to maintain a unified dataset.

BHC Filtering

As the BHC task (see Section 2.2.2) clearly contains a set of clinical notes related to each other through an admission id, we first filter the data to retain only notes suitable for this task. This involves the following steps:

1. Remove all notes that have a NaN value in these columns : TEXT, HADM_ID, CHARTDATE, CATEGORY, DESCRIPTION

2. Remove all admissions that do not contain a discharge summary

We perform this step in order to easily identify clinical notes linked to the same admission. These clinical notes can afterwards be used as context for information extraction and domain-adapted summarization.

Length Filtering

To manage computational resources, we filter admissions based on the length and number of associated clinical notes. Specifically, we remove any admissions that:

- Contain more than 10 clinical notes in total.
- Contain any individual clinical note exceeding 2048 tokens in length (token counts are determined using the Llama-3 tokenizer [132]).

These two criteria are applied in conjunction; an admission is removed if either condition is met.

Subset Selection

Finally, to further control the dataset size for processing, we cap the total number of clinical notes to approximately 5000. We achieve this by iterating through the admissions and keeping all notes associated with each admission until we reach the 5000-note limit. To ensure that complete admissions are retained (as information might be spread across multiple notes), we include all notes from the final (partially processed) admission, even if this results in a total slightly exceeding 5000. This is important because removing notes from the last admission could leave out relevant information for a certain domain.

Data Overview

Following the pre-processing steps described above, we analyze the resulting dataset to understand its characteristics and diversity. Table 3.3 summarizes the impact of each preprocessing step, showing the number of clinical notes and admissions remaining after each stage. We observe a significant reduction in data size, particularly after the length filtering and subset selection stages, which were implemented to manage computational resources. All main experiments are done using the final subset of 5005 clinical notes regrouping 775 admissions. To understand the dataset's composition, we examine the distribution of clinical notes across different medical domains. Figure 3.7 illustrates the distinct values in the CATEGORY column of the MIMIC-III dataset, representing the various medical specialties covered in our study. We observe that the dataset encompasses a wide range of medical domains, including "Nursing", "Radiology", and "ECG" among others. This diversity allows us to evaluate the domain adaptation capabilities of our proposed methodology.

Step	# Clinical notes left after step	# Admissions left after step
Initial data	2,083,180	58,362
BHC Filtering	1,528,004	47,006
Length Filtering	28,551	4,489
Subset Selection	5005	775

Table 3.3 Number of clinical notes and admissions left after each pre-processing steps

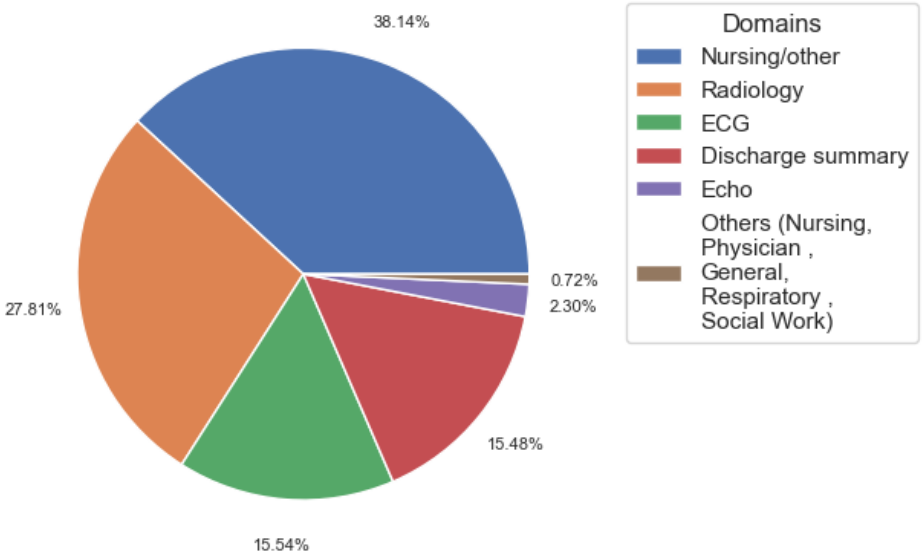
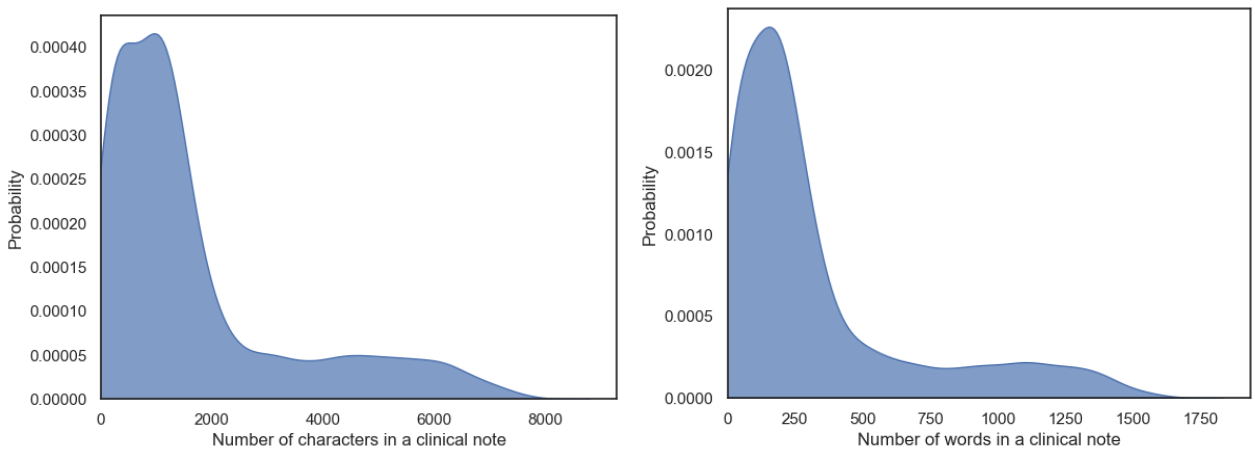


Figure 3.7 Proportions of domains present in our subset of MIMIC-III

Next, we analyze the length of the clinical notes, measured in both words and characters. Figure 3.8 presents the distribution of these lengths. We observe that the majority of clinical notes have less than 2000 characters and 500 words. Figure 3.9 presents these distributions, broken down by medical domain. We find that the length of clinical notes varies considerably across domains. For example, ECG notes tend to be shorter, while radiology reports are generally longer, reflecting the different nature of the information captured in these note types.

Finally, we investigate the medical diversity of the dataset. To quantify this, we calculate the ratio of medical terms to the total number of words in the clinical notes. Medical terms are identified using the same concept annotation tool described in Section 3.2.2. This metric provides an indication of the concentration of medical knowledge within the text. Table 3.4 shows the overall medical diversity of the dataset, as well as the diversity within each medical domain. We find that the average medical diversity is around 0.23, meaning that



(a) Character distribution

(b) Word distribution

Figure 3.8 Distributions of clinical notes in our subset of MIMIC-III

roughly 23% of the words in the dataset are identified as medical terms. Some domains, such as "Echo" or "ECG" might exhibit higher medical diversity (around 0.30), while others, like "Social Work," may have lower diversity (around 0.13).

Domain	**Medical Diversity**
Echo	0.30
ECG	0.28
Radiology	0.26
Respiratory	0.26
Physician	0.24
Discharge Summary	0.23
Nursing	0.21
Nursing/other	0.21
General	0.16
Social Work	0.13
Average	**0.23**

Table 3.4 Medical diversity of clinical notes per domain

Models & Ontology

As we planned to compare the performance of general models and domain-specific models on clinical extraction and summarization, we conducted experiments across models that were

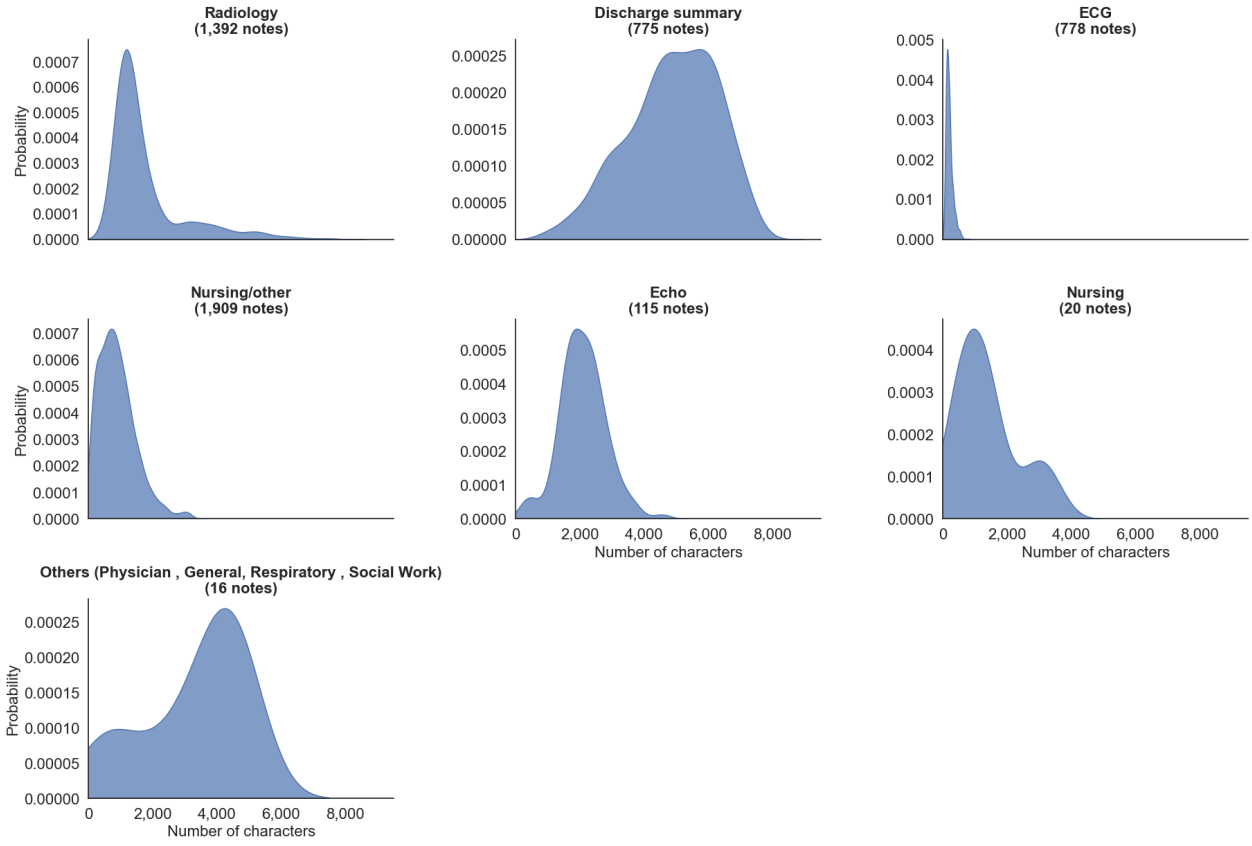


Figure 3.9 Character distribution of clinical notes per domain

fine-tuned on different corpora to evaluate our method's effectiveness. We began with the Llama-3-8B-Instruct model [132], known for its strong performance. We then utilized models trained on medical data like medicine-Llama3-8B [233] and Llama3-OpenBioLLM-8B [234], the latter being one of the best medical models [234], outperforming GPT3.5 Turbo and Meditron-70B [235] on various medical QA datasets. We then evaluated the Llama-3.2-3B model [132], a smaller variant with 3 billion parameters, to understand how model size influences our methodology's effectiveness. As our ontology-based decoding method relies on beams, it might perform worse when beam candidates are bad in general.

For the medical terminology, we leveraged SNOMED-CT [236], a comprehensive ontology spanning diverse medical fields. This ontology contains 374673 concepts with over 1.7 million axioms. We employed the MedCAT annotator [232] to detect SNOMED-CT concepts within the text. It performs Named Entity Recognition (NER) and linking to any concept vocabulary like SNOMED-CT. Trained on electronical health records, it achieves state-of-the-art performance on NER datasets like MedMentions [237].

Domain Adaptation

To adapt our method to specific medical domains, we followed the methodology outlined in Section 3.2.1 and extracted the 100 most frequent concepts for each domain within MIMIC-III (defined by the CATEGORY column). Figure 3.10 displays the top five most frequent concepts for each domain, illustrating the significant variation in concept representation across different medical specialties. Prior to this analysis, we pruned irrelevant branches of the SNOMED-CT ontology (e.g., Linkage concepts, Qualifier values) to focus on clinically relevant concepts.

Domain Adaptation Test Set

While our domain adaptation analysis was performed across all domains, for evaluation, we focused on only four. This selection was based on domains that ensured sufficiently distinct conceptual representations, and maintained a challenging evaluation set. The domain considered are : *Nursing, ECG, Physician* and *Radiology.* As we aim to generate one summary per domain per admission, our original test set of 775 admissions leads to a final 3100 pair dataset (admission, domain-adapted summary). We define this set as the domain adaptation test set.

3.3.2 Evaluation

We evaluate the performance of our method using multiple metrics. First, we assess the impact of our constrained decoding method compared to traditional methods on information extraction using a pairwise comparison metric. We perform this evaluation based on two criteria: groundedness (see Definition 1.3.3) and relevance (see Definition 1.3.2). Moreover, we also perform an extrinsic evaluation of our methodology by generating domain-adapted summaries of admissions and evaluating how clinically relevant the summaries are to each domain.

Groundedness and Relevance of Extracted Information

We evaluate the impact of our proposed constrained decoding process on the groundedness and relevance of generated extractions. Specifically, we assess how constrained decoding enhances the model's ability to extract accurate and relevant information. Since absolute ground truth for extraction from clinical notes is unavailable for the MIMIC dataset, we employ a pairwise comparison approach to assess the relative effectiveness of the different decoding strategies using an evaluator model. A win rate is then computed for each method

Figure 3.10 Top 5 most frequent concepts of each domain in MIMIC-III after performing a domain adaptation analysis

based on how often the evaluator model selects it as the most effective. We compare three distinct generation methods during the extraction phase: (1) standard (greedy) search, (2) diverse beam search, and (3) our constrained decoding approach. We leverage the state-of-the-art Prometheus-8x7b-v2.0 model [229] for evaluation. Prometheus acts as an evaluator, receiving as input: (1) the original prompt; (2) a rubric detailing quality criteria (on a scale of 1 to 5); and (3) two response variants.

Based on the rubric, Prometheus determines the superior response. We evaluate the extractions of the 5 most frequent concepts of clinical notes, leading to 25025 extractions per approach. These concepts are chosen using the same approach described in Section 3.2.2. For every single clinical note and each of the 5 concepts, we obtain three distinct extractions (one from each method). We then form all possible pairs from these three extractions, ensuring each pair consists of extractions generated by two methods from the same clinical note and targeting the same concept. However, we noticed that Prometheus was occasionally inconsistent in its judgments; swapping the order of responses sometimes changed its preference. To mitigate this inconsistency, we compute all permutations of the three methods, leading to six pairwise comparisons per extraction. We send all possible pairs to Prometheus to compute the win rate of each method against each other. Given M the number of samples comparing method A and method B, T the number of ties and E the number of samples where Prometheus' answer could not be parsed, we compute the win rate of method A with the following formula:

$$\text{Win Rate}(A) = \frac{W_A}{M - T - E} \tag{3.4}$$

Where W_A represents the number of samples where method A was judged superior by Prometheus versus method B. In our case, $M = 25025 \times 2 = 50050$.

Groundedness Groundedness accuracy is assessed using a rubric from the original authors of Prometheus presented in Figure 3.12.

Table 3.13 presents the groundedness win rates between all pairs for all models. These win rates correspond to the percentage of samples that a technique has won over another technique. For example, for Llama-3B-Instruct, our constrained decoding approach won 58% of the time against greedy generation. The results show a clear advantage of our constrained decoding approach over greedy generation and diverse beam search for most models tested. For Llama-3B-Instruct, constrained decoding achieved a 3.7% higher win rate against greedy generation compared to diverse beam search. Similarly, for Llama-8B-Instruct, the improvements was 5.8%. Llama-OpenBioLLM-8B showed the most improvement with an 81% win rate against greedy search.

Relevance Similar to the groundedness evaluation, we compute a relevance win rate score in order to evaluate how relevant the extractions are based on the concepts. This score is computed using a custom rubric which is present in Figure 3.14.

Table 3.15 presents the relevance win rates between all pairs for all models. Similar to groundedness evaluation, the results in Table 3.15 show an advantage of our method in the

You are a fair judge assistant tasked with providing clear, objective feedback based on specific criteria, ensuring each assessment reflects the absolute standards set for performance.

###Task Description:
An instruction (might include an Input inside it), a response to evaluate, and a score rubric representing a evaluation criteria are given.
1. Write a detailed feedback that assess the quality of two responses strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, choose a better response between Response A and Response B. You should refer to the score rubric.
3. The output format should look as follows: "(write a feedback for criteria) [RESULT] (A or B)"
4. Please do not generate any other opening, closing, and explanations.

###Instruction:
Here is a clinical note about a patient :
———————-
{Clinical note}
———————-
In a short sentence, extract the information that is related to the "Social context" medical concept from the clinical note. If the concept is not mentioned in the note, respond with 'N/A'. Only output the extracted information.

###Response A:
Family declined to pronounce.

###Response B:
Family will not revisit.

###Score Rubric:
Are the model's responses factually correct and well-supported by evidence?
Score 1: The model's responses are mostly incorrect or based on unfounded information.
Score 2: The model sometimes provides factually correct responses, but inaccuracies are common.
Score 3: The model generally provides factually correct information, though some errors occur.
Score 4: The model often provides factually accurate information with only occasional minor errors.
Score 5: The model consistently provides responses that are factually correct and well-supported by evidence.

###Feedback:
Both Response A and Response B accurately convey the social context according to the given clinical note. However, Response B, "Family will not revisit.", captures the intent of the family better than Response A's "Family declined to pronounce.". Response A can be misinterpreted as the family taking an active role in the death declaration, when in reality they wish to avoid any further contact with the medical facility. Therefore, Response B provides a more comprehensive understanding of the family's situation. In conclusion, both responses are factually correct, but Response B provides a clearer and more accurate representation of the social context.
[RESULT] B

Figure 3.11 Example of prompt given to Prometheus using the factuality rubric (the original clinical note from MIMIC was omitted)

[Are the model's responses factually correct and well-supported by evidence?]
Score 1: The model's responses are mostly incorrect or based on unfounded information.
Score 2: The model sometimes provides factually correct responses, but inaccuracies are common.
Score 3: The model generally provides factually correct information, though some errors occur.
Score 4: The model often provides factually accurate information with only occasional minor errors.
Score 5: The model consistently provides responses that are factually correct and well-supported by evidence.

Figure 3.12 Groundedness rubric used to compare extracted values with Prometheus-2

	GS	DBS	OCD
GS	-	45	42
DBS	55	-	46
OCD	**58**	**54**	-

(a) Llama-Instruct-3B

	GS	DBS	OCD
GS	-	46	40
DBS	54	-	43
OCD	**60**	**57**	-

(b) Llama-3-8B-Instruct

	GS	DBS	OCD
GS	-	38	40
DBS	**62**	-	**52**
OCD	60	48	-

(c) medicine-Llama3-8B

	GS	DBS	OCD
GS	-	27	19
DBS	73	-	37
OCD	**81**	**63**	-

(d) Llama-OpenBioLLM-8B

Figure 3.13 Win rates of each model on groundedness (GS: Greedy search, DBS: Diverse Beam Search, OCD: Ontology-Constrained Decoding) using $H_{bf} = 3, P_{bf} = 1$ and $S_{bf} = 10$. The number of ties and parsing errors are indicated in A.1

majority of cases.

Analysis Our results shown in Figures 3.13 and 3.15 demonstrate the effectiveness of constrained decoding in improving both groundedness and relevance across models. Interestingly, the improvement appears to be more pronounced in the larger Llama-3-8B-Instruct model, potentially due to the richer set of beam candidates available in larger models, which our constrained decoding can effectively leverage. A similar trend was observed for relevance. Furthermore, we noted that the magnitude of improvement was remarkably similar for both groundedness and relevance. This suggests a possible correlation between the two, where focusing on relevant information may inherently come with more grounded extractions, and

[Are the model's responses relevant to the medical concept mentioned?]
Score 1: The model's answer is irrelevant to the medical concept and completely misses information that is related to the medical concept.
Score 2: The model's short summary is mainly irrelevant, but mentions one or two things related to the medical concept mentioned.
Score 3: The model's short summary is somewhat irrelevant, but contains key elements related to the concept mentioned.
Score 4: The model's short summary is mainly relevant, but contains some elements that are not linked to the medical concept.
Score 5: The model's short summary mentions everything related the the medical concept perfectly without missing any detail.

Figure 3.14 Relevance rubric used to compare extracted values with Prometheus-2

	GS	DBS	CD
GS	-	45	42
DBS	55	-	46
CD	**58**	**54**	-

(a) Llama-Instruct-3B

	GS	DBS	CD
GS	-	47	40
DBS	53	-	43
CD	**60**	**57**	-

(b) Llama-3-8B-Instruct

	GS	DBS	CD
GS	-	38	39
DBS	**62**	-	**52**
CD	61	48	-

(c) medicine-Llama3-8B

	GS	DBS	CD
GS	-	26	20
DBS	74	-	37
CD	**80**	**63**	-

(d) Llama-OpenBioLLM-8B

Figure 3.15 Win rates of each model on relevance (GS: Greedy search, DBS: Diverse Beam Search, OCD: Ontology-Constrained Decoding) using $H_{bf} = 3, P_{bf} = 1$ and $S_{bf} = 10$. The number of ties and parsing errors are indicated in A.1

vice-versa.

When examining medical models, we observe seemingly contradictory performance patterns. For Llama-Medicine-8B, our proposed method unexpectedly decreases the performance of diverse beam search in both groundedness and relevance. In contrast, OpenBioLLM demonstrates dramatically different results. With this model, our method substantially outperforms both alternative decoding strategies with a 63% win rate against diverse beam search. This strong improvement suggests that the model's beam candidates are inherently good but often hidden because the model prioritizes less optimal choices with higher probabilities. The contradictory performance of Llama-Medicine-8B presents an interesting case for analysis as

it is the only model that lead to worse performance after applying our decoding strategy. We hypothesize that these results may be attributed to the unique training paradigm employed for this model, which may have led to less optimal beam candidates. Another possible explanation might be that the new training paradigm allows the model to generate accurate beam likelihoods leading to optimal beam selection by the model itself. As for non-medical models, we see consistent improvements of our method. When analyzing extractions across various models, it is surprising to find that medical models aren't inherently better at extracting medical information compared to their non-medical counterparts.

Concept	**Extractions**
Clinical Finding	**GS :** No acute intracranial hemorrhage or mass effect **DBS :** No acute intracranial hemorrhage or mass effect **OCD :** No acute intracranial hemorrhage or mass effect, no significant interval change since the prior exam, and no sign of fracture or bone destruction.
Electrocardiogram abnormal	**GS :** Non-specific lateral ST segment depression compatible with ischemia. **DBS :** Non-specific lateral ST segment depression compatible with ischemia. **OCD :** Normal sinus rhythm, rate 75. Non-specific lateral ST segment depression compatible with ischemia.
Disease	**GS :** N/A **DBS :** N/A **OCD :** Low-grade fever, incision site infections (CDI) at right buttock and right breast
Physical Object	**GS :** A stent to the left anterior descending (LAD) artery **DBS :** Right groin site CD+I (central line) and Foley catheter. **OCD :** Right groin site CD+I, Ecchymotic, Pulses palpable distal, IVFs dc'd.

Table 3.5 Example of extractions of all methods (blue text is not relevant to the concept and red text is not factual to the clinical note)

Upon inspecting individual samples to evaluate cases where the proposed method performed better or sub-optimally, we found multiple key examples showing the ups and downs of our method. Table 3.5 details these results. The first observation that can be made is that our proposed approach generally generates longer responses, which is consistent with its objective of incorporating more concepts into the beams. However, these longer responses can sometimes lead to less relevant information. For instance, in the "Electrocardiogram abnormal" concept, the proposed method (OCD) included "Normal sinus rhythm, rate 75",

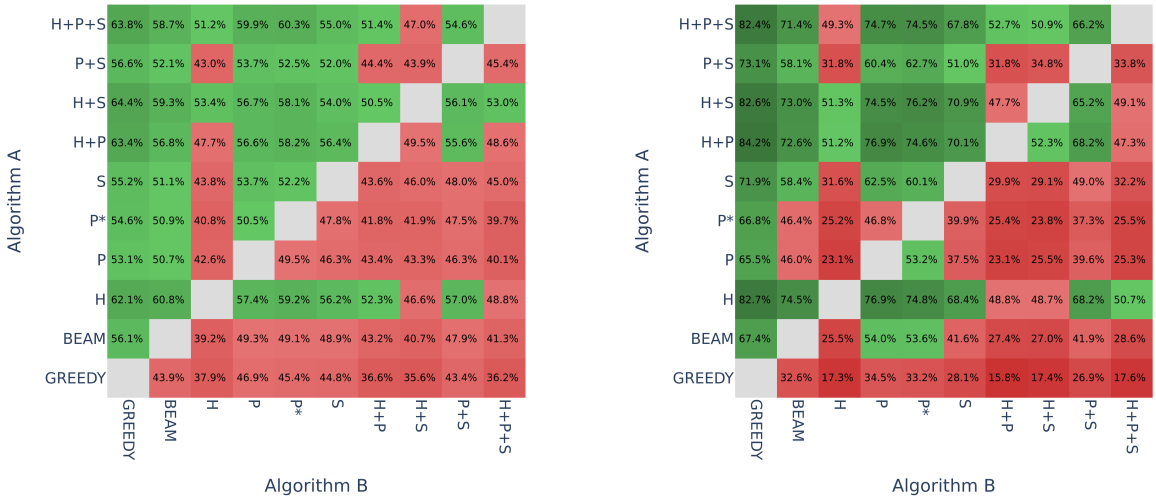
which is not indicative of an abnormality, although it is related to electrocardiograms. However, these longer responses can also lead to better extractions as with the *Clinical Finding* concept. Plus, our method improves factuality when compared to greedy search (GS) and diverse beam search (DBS) when the model chooses to not answer even though the concept is present in the clinical note. For example, the *Disease* shows an example where GS and DBS returned *N/A* while OCD provided *Low-grade fever, incision site infections (CDI) at right buttock and right breast* which is both factual and relevant to the clinical note. However, this can also lead the model to generate irrelevant answers when the concept is not present in the note, because abstention is not prioritized by our algorithm. The last example in Table 3.5 shows how our decoding process improves factuality both compared to greedy search and diverse beam search by prioritizing formulations that resemble the clinical note.

Ablation Study

We also perform an ablation study on the effect of each score (hierarchy, property and similarity) on the model's performance. To do this, we create a subset of 100 clinical notes derived from our original test set selected randomly and run the extraction process with every combination of scores on Llama-8B-Instruct and OpenBioLLM on all concepts in each domain. We then evaluate the extractions using the same method as before using Prometheus-2-8x7B and report the win rates on groundedness and relevance for all combinations. Additionally, we also perform an ablation on the impact of using ROUGE in the property score. Figure 3.16 shows the results on groundedness and Figure 3.17 shows the results on relevance.

Figures 3.16 and 3.17 show that combining all three scores (Hierarchy, Property, and Similarity) generally yields the most consistent performance improvements. The H+P+S combination achieves the highest overall win rate across all models and configurations tested, demonstrating the value of this comprehensive approach. Interestingly, according to the Prometheus evaluation, the full combination doesn't always outperform more selective approaches. For instance, with Llama-8B-Instruct, the H+P+S combination only wins against the H+S combination 47% of the time. For OpenBioLLM, the primary competition comes from using the hierarchy score alone. The hierarchy score appears to have the most significant impact on improving generations both for groundedness and relevance, which validates the core premise that ontology guidance effectively steers language model outputs in the intended direction. This aligns with our hypothesis that leveraging ontological structures can meaningfully constrain generation.

The property score shows a more modest effect on performance. This limited impact likely stems from the structural characteristics of the SNOMED ontology itself, where restriction



(a) Llama-8B-Instruct (b) Llama3-OpenBioLLM-8B

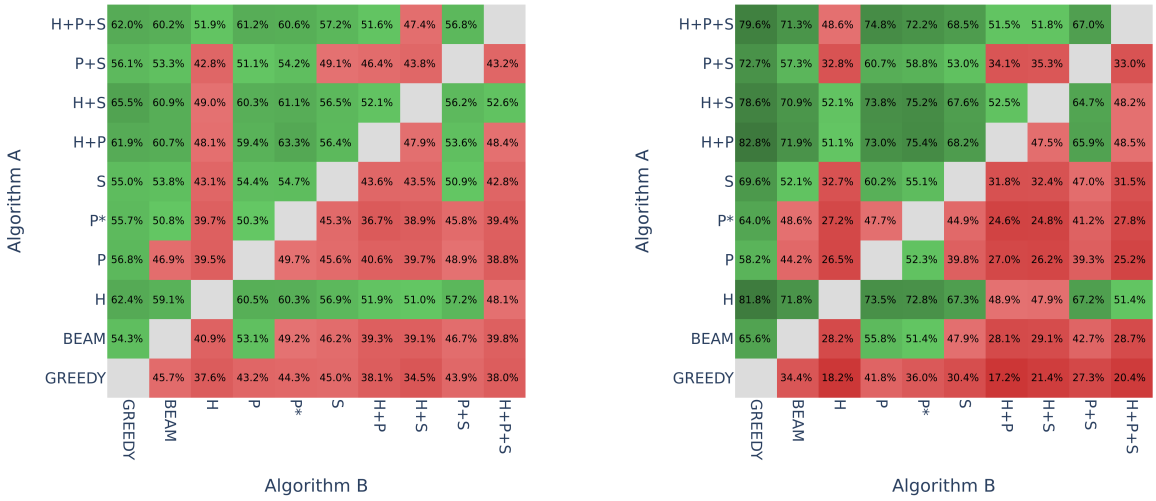
Figure 3.16 Ablation study of performance on groundedness (H=Hierarchy score, P=Property score, S=Similarity score, P*=Property score without ROUGE)

properties are relatively rare compared to hierarchical relationships. While all concepts (except the root) have ancestors, only a small portion possess restriction properties, meaning this score isn't consistently applicable during the generation process.

The similarity score demonstrates a positive influence overall, but surprisingly, it appears to have a stronger effect on relevance than on groundedness. This suggests that textual similarity with input clinical notes helps keep generations topically appropriate, even if it doesn't always guarantee factual accuracy to the same degree.

Medical vs Non-Medical Models

Our prior analysis involved internal comparison of extractions within the model, meaning we compared our ontology-constrained approach against diverse beam search and greedy search within a single model and observed this method's application across different models. To further assess the quality of extractions between medical and non-medical models, we apply the same evaluation criteria across different models. Specifically, we report the win rates of Llama-3-8B-Instruct against Llama-OpenBioLLM-8B when taking the extractions of the same decoding strategy for each model. These two models were selected, because they



(a) Llama-8B-Instruct (b) Llama3-OpenBioLLM-8B

Figure 3.17 Ablation study of performance on relevance (H=Hierarchy score, P=Property score, S=Similarity score, P*=Property score without ROUGE). We set the boost factors for all scores to 1.0 when considered and to 0 when not considered. This is performed with a beam size of 10 and a group beam size of 2.

are respectively the best performing general and medical models in our set of models. This evaluation will allow us to evaluate which types of models are better at extracting information from medical texts and assess the effect our ontology-constrained decoding strategy across models. Results are shown in Table 3.6.

Decoding Strategy	**Groundedness**	**Relevance**
Diverse Beam Search	69.50	69.00
Ontology-Constrained Decoding	52.91	54.33

Table 3.6 Win rates of Llama-3-8B-Instruct against Llama-OpenBioLLM-8B

As detailed in Table 3.6, extractions from Llama-3-8B-Instruct are more grounded and relevant than those from Llama-OpenBioLLM-8B according to Prometheus. This aligns with studies indicating that medical fine-tuning doesn't necessarily enhance a model's information extraction and summarization on medical texts [238]. However, when our constrained-decoding strategy is applied, the win rate of Llama-3-8B-Instruct drops significantly. This means that Llama-3-8B-Instruct, initialy considered superior to Llama-OpenBioLLM-8B by

Prometheus is considered equally performant when our constrained decoding method is applied to both models. This shows that the best generations of medical models are not necessarily leveraged with classical decoding strategies.

Relevance evaluation of domain-adapted summaries

We then evaluated the domain-adaptation capabilities of both general and medical instruction-tuned LLMs, both in their baseline performance and when utilized in our proposed method (see Section 3.2). In order to evaluate the relevance of generated summaries to a specific domain, we would ideally need a per sample ground truth summary for each domain. However, because no such ground truth exists, we trained a BERT-based classifier [239] to classify whether a generated summary is relevant for a domain. This model served as the evaluator for the domain-adaptation task. Denoted as the evaluator model, it predicts the domain of a given clinical note using the CATEGORY column from MIMIC-III as the target label. We focused on four main domains: Nursing, ECG, Physician and Radiology. To train the domain evaluator model, we constructed a dataset disjoint from the test set using the following strategy:

1. Initialized with all clinical notes from MIMIC-III.
2. Removed all notes belonging to our held-out test set (5005 notes).
3. Filtered the remaining notes, keeping only those associated with the Nursing, ECG, Physician and Radiology domains.
4. Subsampled each domain to a maximum of 100,000 notes.
5. Generated summaries for 50% of clinical notes using Llama-8B-Instruct and used those summaries instead of the clinical notes for training.

The process resulted in a 400,000-sample dataset, with each of the four domains contributing 100,000 clinical notes. Within each domain, half of the notes were original, and the other half were generated summaries. This final summarization step was crucial to prevent the evaluator model from relying on formatting cues to predict the domain. As illustrated in Figure 3.9, the format and length of notes vary considerably across domains (e.g., Nursing vs. ECG). By also training on summarized notes, we ensure that the evaluator focuses on the semantic content rather than superficial formatting differences, which is essential because the generated domain-adapted summaries do not follow the individual domain formats. We opted for an 8-billion parameter model for generating summaries primarily due to computational

efficiency. Models of this size are already quite capable of basic summarization tasks. We didn't need perfectly accurate, clinically validated summaries as the task was only to predict the domain from a task. Instead, our goal was to capture the main domain concepts of a note. For this purpose, an 8B model was more than sufficient, allowing us to avoid the higher computational demands of a larger model. We compute a domain score D to quantify how well each generated summary aligns with its target domain. It can be interpreted as the percentage of information in a text that can be associated to the expected domain, thus relevant to the domain. The score is calculated as the average probability score assigned by the evaluator model to the expected domain:

$$D = \frac{1}{N} \sum_{i=0}^{N} \texttt{EVALUATOR}(d(x_i))[d_i] \tag{3.5}$$

where $d(x_i)$ is the domain adapted summary of the admission x_i, d_i is the expected domain and N is the number of samples.

Results To evaluate the effectiveness of our domain adaptation approach, we compared multiple methods. First, given an LLM, we evaluated the performance of prompting the model for a domain-adapted summary using standard (greedy) generation and diverse beam search. We then applied our method on the same model to generate domain-adapted summaries. For greedy and diverse beam search, we augmented the input prompt with a prefix specifying the target domain. Our method, in contrast, directly uses the pruned CSR (see Section 3.2.4) as input to the verbalizer, without explicit domain specification.

> *Here are multiple clinical notes associated to the hospital course of a patient ordered by the time they were recorded:*
>
> *[Clinical Notes]*
>
> *Summarize the hospital course of the patient only using the information related to the "domain" medical domain in a text. Only output the summary without any additional text.*

Figure 3.18 Prompt format used for generating domain-adapted summaries of clinical notes for Greedy Search and Diverse Beam Search

Both prompts are shown in Figures 3.18 and 3.19. We then passed the summaries generated by each method through our evaluator model. This comparison allows us to assess the relative effectiveness of our method in generating domain-tailored summaries without explicit prompt engineering. Our baseline corresponds to simply prompting the model to summarize the clinical notes according to the specified domain. We evaluated these methods on the domain

Sentences were extracted from multiple clinical notes based on a medical concepts. For each clinical note, we have a dictionary where the keys are the concepts and the values are the sentences that were extracted linked to those concepts. The clinical notes are ordered by the time they were recorded. Here are the clinical notes' extractions:

[Clinical Notes]

Summarize the clinical notes of the patient based on the extractions of each clinical note in a text. Only output the summary without any additional text.

Figure 3.19 Prompt format used for generating domain-adapted summaries of clinical notes for our method

adaptation test set (see Section 3.3.1). The domain adaptation results are shown in Table 3.7.

Model	**Greedy Search**	**Diverse Beam Search**	**Ours** $\alpha = 4$	$\alpha = 2$	$\alpha = 1$
Llama-3B-Instruct	0.62	0.63	0.67	0.68	**0.70**
Llama-8B-Instruct	0.62	0.64	0.69	0.69	**0.70**
Llama-8B-medicine	0.35	0.36	0.66	0.67	**0.67**
Llama3-OpenBioLLM-8B	0.35	0.36	0.53	**0.56**	0.54

Table 3.7 Domain scores of each method on generating domain-adapted summaries. Each domain score, can be interpreted as, on average, the amount of information in a text, that can be linked to the expected domain as judged by the evaluator model.

We also report in Table 3.8 the domain score of each model and method by domain in order to evaluate whether certain methods are more efficient for certain domains.

Analysis The results presented in Table 3.7 show the effectiveness of our domain adaptation method in enhancing the ability of LLMs to generate domain-adapted summaries of clinical notes, with a peak increase of 32% observed for Llama-8B-medicine. Our approach outperforms simply prompting the model for domain-specific summaries, highlighting its practical value. These results also suggest that relevant domain concepts can be effectively derived directly from the data through our initial domain adaptation analysis. Furthermore, these findings implicitly validate our ontology-based decoding process, as the efficacy of the pruning step depends heavily on the extracted values. While all tested alpha values for the pruning phase yielded improvements, the optimal performance was generally achieved with $\alpha = 1$ (except in the case of Llama3-OpenBioLLM-8B). This finding aligns with expectations, as a lower alpha prioritizes the most frequent concepts in the domain, minimizing the

	Method	ECG	Nursing	Radiology	Physician
Llama-3B-Instruct	Greedy Search	0.68	0.70	**0.78**	0.34
	Diverse Beam Search	0.69	0.71	**0.78**	0.35
	$\alpha = 4$ (Ours)	0.88	0.74	0.68	0.39
	$\alpha = 2$ (Ours)	0.88	0.74	0.69	**0.42**
	$\alpha = 1$ (Ours)	**0.89**	**0.77**	**0.78**	0.38
Llama-8B-Instruct	Greedy Search	0.68	0.71	0.76	0.34
	Diverse Beam Search	0.69	0.72	**0.78**	0.35
	$\alpha = 4$ (Ours)	**0.90**	**0.79**	0.67	0.39
	$\alpha = 2$ (Ours)	**0.90**	0.78	0.68	**0.42**
	$\alpha = 1$ (Ours)	0.89	0.78	0.72	0.38
Llama-8B-medicine	Greedy Search	0.26	0.36	0.33	0.46
	Diverse Beam Search	0.25	0.36	0.34	**0.48**
	$\alpha = 4$ (Ours)	0.90	**0.90**	**0.64**	0.21
	$\alpha = 2$ (Ours)	**0.93**	**0.90**	0.61	0.22
	$\alpha = 1$ (Ours)	**0.93**	**0.90**	0.61	0.23
Llama3-OpenBioLLM-8B	Greedy Search	0.26	0.36	0.33	0.46
	Diverse Beam Search	0.26	0.37	0.34	**0.47**
	$\alpha = 4$ (Ours)	**0.76**	0.60	0.46	0.29
	$\alpha = 2$ (Ours)	**0.76**	0.64	**0.49**	0.33
	$\alpha = 1$ (Ours)	0.70	**0.65**	**0.49**	0.30

Table 3.8 Domain scores of each method on based on the domain

inclusion of less relevant terms. Beyond performance gains, our method significantly improves the interpretability of the summarization process by decoupling the extraction and adaptation phases. This modularity and transparency facilitate verification and validation of the generated content. Critically, the separation of these steps allows the pruning step to be applied to any domain without requiring a repeated extraction phase.

The analysis of Table 3.8 reveals that our domain adaptation method exhibits varied effectiveness across different medical domains. The approach demonstrates strong performance in the *ECG* and *Nursing* domains, consistently having higher domain scores across all models. This suggests the initial domain analysis is highly effective in extracting relevant concepts in these contexts, leading to more relevant summaries. However, our method encounters greater challenges with the *Radiology* and *Physician* domains, where improvements are generally lower. For the Radiology domain, our approach unexpectedly decreases the domain score for non-medical models (Llama-3B-Instruct and Llama-8B-Instruct) compared to their baseline Greedy and Diverse Beam Search performances. Conversely, it significantly increases the performance of medical models (Llama-8B-medicine and Llama3-OpenBioLLM-8B) in the same

domain. The opposite trend is observed for the Physician domain: our method improves the domain score for non-medical models, while it leads to a decrease in performance for medical models. This discrepancy likely stems from the fact that medical concepts within Radiology and Physician notes often encompass multiple domains, alongside less standardized documentation. This characteristic may make it more difficult for the ontology-guided pruning step and extraction step to consistently extract and prioritize the most critical domain-specific information. Therefore, while our method offers general improvements, its efficacy is most pronounced in domains with more precise and consistently structured conceptual frameworks.

Medical vs Non Medical Models Interestingly, specialized medical models (Llama-8B-medicine and Llama3-OpenBioLLM-8B) show lower baseline domain adaptation scores (0.35-0.36) compared to general-purpose models (0.62-0.64). This counterintuitive result may originate from their training approach: medical models are fine-tuned to prioritize comprehensive coverage across all medical domains simultaneously, rather than domain-specific specialization. This broad medical focus may obstruct their ability to distinguish between specific medical subdomains when explicitly prompted, while general-purpose models may rely more on surface-level formatting cues that correlate with domain categories. Despite their poor baseline performance, medical models demonstrate a better responsiveness to our ontology-based approach, with Llama-8B-medicine showing the largest improvement (32% increase). This suggests that while these models struggle with explicit domain prompting, they can still be leveraged for domain adaptation. The consistent improvement across both model types validates the robustness of our approach, though the varying optimal alpha values indicate that different models may require tailored configurations to achieve peak performance.

CHAPTER 4 AUTOMATIC HALLUCINATION EVALUATION OF MEDICAL TASKS

4.1 Introduction

As demonstrated in Chapter 3, Large Language Models (LLMs) continue to be susceptible to hallucinations [6]. Results in Section 3.3.2 have shown that the groundedness and relevance of generations are not optimal even after implementing various enhancement strategies. This underscores the critical importance of robust hallucination detection mechanisms. Current evaluation metrics exhibit accuracy limitations in certain scenarios [13, 211, 220, 231], as outlined in Section 2.5.2. The most promising approach appears to be the use of neural models as evaluators, given their stronger correlation with human judgment [226]. In Section 3.3.2, we used this approach with Prometheus-2-8x7B, a general-purpose evaluator, to show how our constrained decoding method improved groundedness on information extraction. This persistent lack of optimality in groundedness and relevance directly contributes to the challenge of mitigating hallucinations, as ungrounded or irrelevant information is a primary characteristic of such model fabrications. However, this evaluation approach still has limitations. As Prometheus wasn't fine-tuned for the medical domain, its evaluation capabilities may be suboptimal in clinical contexts. Its creators only evaluated it on general tasks rather than medical-specific ones [229]. Additionally, it performs best when comparing generated answers against a ground truth, creating challenges when no reliable ground truth exists for clinical tasks. Currently, medical model performance assessment relies heavily on expert evaluation [2, 240], which is both expensive and time-consuming for healthcare professionals. This shows the need for a dedicated medical hallucination evaluation dataset that could not only allow researchers to assess which models are better as evaluators in a clinical setting, but also allow us to train an evaluator.

In fact, currently available medical datasets often assess LLM hallucinations by focusing narrowly on single tasks such as Question Answering (QA) or Natural Language Inference (NLI), limiting their applicability across diverse medical text generation scenarios [240–245]. Furthermore, even current medical hallucination datasets typically only contain hundreds of examples, making them unsuitable for training LLMs [240]. To the best of our knowledge, the MedNLI dataset [246] appears to be the only resource that could potentially evaluate LLMs' capabilities as evaluators in clinical settings. However, MedNLI samples are not well-suited for large-scale medical hallucination evaluation due to their typically short length. Furthermore, since MedNLI is designed to determine whether a hypothesis entails a premise, it has

limited utility for training a medical judge to assess factual accuracy in LLM-generated content. This limitation arises because LLM generations commonly contain a mixture of factual and non-factual information, whereas MedNLI samples are binary in nature. Consequently, in tasks such as clinical summarization, where a single sentence or even a portion of a sentence might contain inaccuracies within an otherwise accurate summary, an LLM fine-tuned on MedNLI would likely struggle to identify these isolated errors. The model might classify the entire summary as accurate because the majority of the content is correct, overlooking small but potentially critical inaccuracies. Some researchers have attempted to address this issue by decomposing generations into atomic facts or individual sentences [13] before evaluation by a judge model. However, this decomposition approach introduces significant computational costs and requires careful hyper-parameter tuning (as we cannot predict in advance whether an entire sentence or just a portion contains the hallucination). Moreover, the decomposition process itself might rely on LLMs themselves which would introduce additional opportunities for hallucinations during the fact extraction phase.

To address these limitations, we create MedHal, a medical hallucination evaluation dataset [247]. Our work differs from traditional hallucination medical datasets in several key aspects:

1. We incorporate a diverse range of sources including clinical notes, clinical trials, and medical questions to assess hallucinations in more complex settings.

2. Our dataset is designed to train a medical evaluator capable of efficiently detecting hallucinated content.

3. We provide explanations for why statements are factual or not, creating a valuable guiding signal for LLM fine-tuning.

4.2 Methodology

This study introduces MedHal, a new dataset and benchmark created for evaluating medical hallucination detection models. MedHal contains a wide range of medical text, including clinical notes, scientific articles, and patient communication, annotated with various examples of factual errors and medical hallucinations. These hallucinations are produced through multiple strategies customized to specific task modalities, including answer replacement in question-answering (QA) and the incorporation of contradictory statements in natural language inference (NLI). MedHal's effectiveness is shown through the development of baseline models and the comparative evaluation of its performance against state-of-the-art models on the proposed benchmark as well as other benchmarks. This resource aims to support

the creation of more precise and reliable medical evaluators by offering a standardized and clinically relevant assessment framework.

The following sections outline the development of MedHal, explain the associated benchmark, and present experimental outcomes demonstrating its value in evaluating medical hallucination detection models. The approach involves converting existing medical datasets across various tasks (QA, Summarization, NLI, Information Extraction) into a unified hallucination detection task. This is accomplished by structuring the task as binary classification of a given statement as factual or non-factual, potentially based on a provided context. Additionally, we provide an explanation when a statement is not factual, a guiding indicator that can be utilized when fine-tuning LLMs. We define factuality here according to the definition given in Section 1.3.3 which states that a statement is factual only if it is backed by its context or general medical knowledge.

4.2.1 Unified Task Formulation

To create this dataset, we unified several common medical tasks, such as question-answering and summarization, into a single task. This approach allows us to leverage samples from existing datasets as a foundation for our own. The unified task begins with a statement that can be either factual or non-factual. This statement might relate to general medical knowledge or refer to a specific context. For instance, the statement "*the patient has suffered from myocardial infarction*" most likely refers to a specific context, such as the patient's clinical notes. Contrarily, the statement "*aging causes an increase in blood pressure*" refers to general medical knowledge. In our dataset, each sample includes a label indicating whether the statement is factual. For non-factual statements, an explanation is provided to detail the inconsistency. A statement is considered factual if all the information it contains can be verified either through the provided context or through general medical knowledge. Thus, each sample in the MedHal dataset is structured as follows:

- Statement: An assertion about a specific context or general medical knowledge.
- Context (Optional): Relevant contextual information pertaining to the statement.
- Label (Binary): Indicates whether the statement is factual or not.
- Explanation: One or more sentences clarifying why the statement is non-factual.

Table 4.1 presents examples illustrating how statements are generated from existing datasets. The following sections detail the transformation of various tasks (question-answering, infor-

mation extraction, natural language inference, and summarization) to align with our unified task.

Task	**Datasets**	**Sample (Input → Output)**	**Generated Statement**
Information Extraction	Augmented-Clinical Notes (ACM) [248]	A 10-year-old girl first noted a swollen left knee and underwent repeated arthrocentesis. She underwent arthroscopic surgery and was diagnosed with ... → **age: 10 years old**	The patient is 10 years old.
Summarization	SumPubMed [92]	the large genotyping studies in the last decade have revolutionize genetic studies. our current ability to ... → **genetic admixture is a common caveat for genetic association analysis. these results...**	genetic admixture is a common caveat for genetic association analysis. these results...
NLI	MedNLI [246]	Labs were notable for Cr 1.7 (baseline 0.5 per old records) and lactate 2.4. → **Patient has normal Cr**	Patient has normal Cr
QA	MedQA [249], MedMCQA [250]	Which of the following medications is most commonly used as first-line treatment for newly diagnosed type 2 diabetes mellitus in patients without contraindications? → **Metformin**	Metformin is most commonly used as first-line treatment for newly diagnosed type 2 diabetes mellitus in patients without contraindications.

Table 4.1 Example of samples are used to generate statements for each task

4.2.2 Question Answering Dataset Transformation

Question-answering (QA) datasets are structured around the presentation of a question followed by a set of potential responses, including binary (yes/no) and multiple-choice (A, B, C, ...) formats. While they are usually used to evaluate LLMs' knowledge on a certain topic or domain, we use them here to create factual and non-factual samples. To generate factual samples from QA datasets, the question and its corresponding correct answer are transformed into a declarative statement using a large language model. Conversely, non-factual samples are produced by pairing the question with incorrect answer options and subsequently con-

verting these combinations into statements via the same large language model. While a single sample from the original dataset could theoretically generate as many examples as there are possible choices, we specifically generate only a single factual and a single non-factual sample per question to ensure the dataset remains balanced. Table 4.2 illustrates an example of a multiple-choice question converted into such a statement. For this generation process, a consistent prompt template is employed for both factual and non-factual sample creation. One-shot prompting is utilized to guide the large language model in accurately converting question-answer pairs into coherent statements. The precise prompt format used is provided in Figure 4.1.

[**System prompt**] Given a medical text, a question about the text and the associated answer, your role is to transform the question into a statement by incorporating the answer with it. Do not add any details that is not mentioned in the question or the answer.

[**User**] Question: Which of the following is the best treatment for this patient?
Answer: Nitrofurantoin

[**Assistant**] Nitrofurantoin is the best treatment for this patient.

[**User**] Question: {question}
Answer: {answer}

Figure 4.1 Prompt format to generate samples for MedHal from a QA dataset

Regarding the explanations for these statements, certain QA datasets sometimes provide an explanation for why a particular answer is incorrect. In these specific instances, we directly utilize the dataset's provided explanation for our generated non-factual statements. However, in cases where the original dataset does not offer an explicit explanation for an incorrect answer, we generate an explanation for the non-factual statement by simply using its corresponding factual statement. For example, if a non-factual sample asserts "Insulin is primarily produced by the thyroid gland," its explanation would be the factual statement: "Insulin is primarily produced by the pancreas." This approach clarifies why the non-factual statement is incorrect, as the factual statement presents the true biological fact.

4.2.3 Information Extraction Dataset Transformation

Information Extraction (IE) datasets comprise a source document along with a set of text sequences, known as "extractions," that represent specific concepts or attributes within that document. For example, in a clinical setting, an extraction might represent a patient's reason for a visit, linked to the concept "visit motivation." Our work specifically utilizes clinical IE

Sample Type	Question	Answer	Generated Statement
Factual	Which of the following medications is most commonly used as first-line treatment for newly diagnosed type 2 diabetes mellitus in patients without contraindications?	Metformin	Metformin is most commonly used as first-line treatment for newly diagnosed type 2 diabetes mellitus in patients without contraindications.
Non-Factual		Insulin	Insulin is most commonly used as first-line treatment for newly diagnosed type 2 diabetes mellitus in patients without contraindications.

Table 4.2 Example of Question-Answering Dataset Transformation

datasets, where each clinical note is accompanied by a structured summary detailing various patient and admission attributes.

Sample Type	Source Document	Extraction	Statement	Explanation
Factual	A 10-year-old girl first noted a swollen left knee and underwent repeated arthrocentesis...	age: 10 years old	The patient is 10 years old	-
Non-Factual	A 10-year-old girl first noted a swollen left knee and underwent repeated arthrocentesis...	age: 16 years old	The patient is 16 years old	The patient is 10 years old

Table 4.3 Example of Information Extraction Dataset Transformation (the extraction from the non-factual statement is taken from another original sample)

To generate factual samples, we use the clinical note as the contextual basis, and its corresponding extractions are treated as declarative statements. For non-factual samples, we introduce fabricated information. This is achieved by randomly interchanging extraction val-

ues of the same concept type between different documents. For instance, medication names are swapped between different patient records. For each individual extraction within a clinical note, we generate both a factual and a non-factual sample, both of which are linked to the same original clinical note as their context. Similar to how explanations are handled for QA datasets, the explanation for a non-factual statement derived from an IE dataset is provided by its corresponding factual statement. For example, if a non-factual statement asserts "The patient's medication is Ibuprofen" but the correct medication was "Aspirin," the explanation would be the factual statement "The patient's medication is Aspirin.".

[**System prompt**] You are tasked with transforming structured medical data into natural language statements about a patient.
Each input will contain 4 elements:
- concept: The type of information being described (e.g., dosage, age, symptoms)
- value: The specific information or measurement
- category: The broad medical category this information belongs to (e.g., treatment, patient information, symptoms)
- concept_reference: The specific element that the value refers to (e.g., a specific medication, a specific symptom)

Your task is to generate a clear, grammatically correct sentence that conveys this information in a medical context. Follow these rules:
1. Use appropriate verbs based on the concept:
- For treatments: 'takes', 'receives', 'is prescribed'
- For symptoms: 'experiences', 'reports', 'presents with'
- For measurements/states: 'is', 'has', 'shows'
- For time-related concepts: 'has been', 'started', 'continues'
2. Incorporate the concept_reference when it adds clarity
3. Use present tense
4. Maintain medical terminology as provided
5. When the concept_reference is 'None' or does not add clarity, don't include it in the statement
6. The statement should be a single sentence.
Do not include any other information in the statement aside from the concept and the extraction. Only output the statement and nothing else.

[**User**] 'category': 'medical examinations',
'value': 'Severe gait disturbance secondary to hip pain',
'concept_reference': 'Physical examination'

[**Assistant**] Physical examination showed severe gait disturbance secondary to hip pain.

[**User**] 'category': {category},
'value': {value},
'concept_reference': {concept_reference}

Figure 4.2 Prompt format to generate samples for MedHal from an IE dataset

In cases where attributes have limited value diversity (e.g., "sex" with typically "Male" or

"Female"), we ensure that the swapped non-factual value is genuinely distinct from the original factual value by comparing the value swapped. If a random swap yields the same value, we re-attempt the swap until a different value is found, thereby guaranteeing the non-factual nature of the generated sample. In practice, given extractions from a clinical note that are structured as key-value pairs (e.g., "visit motivation: Lower back pain"), a large language model transforms this pair into a coherent statement (e.g., "The patient's visit motivation is lower back pain"). This is performed using the prompt template shown in the Figure 4.2 within a one-shot prompting framework. The one-shot example is adapted based on the extraction's concept to further improve the model's faithfulness on generating factual statements. An example of generated statement is shown in Table 4.3.

4.2.4 Natural Language Inference Dataset Transformation

Given the strong parallel of our task with the Natural Language Inference (NLI) task, we've found NLI datasets to be a natural fit for the MedHal benchmark. For these datasets, our factual samples are created from hypotheses entailing their premise, while non-factual samples are generated from hypotheses leading to a contradiction. The premise component of the NLI example serves as the context and the hypothesis as the statement. We filter out and ignore any samples that result in a neutral label during dataset construction, and no further preprocessing is applied to the NLI datasets beyond this filtering. It's important to note that, in this specific case, we do not provide explanations for the non-factual samples. This is because, unlike samples we generate ourselves, these NLI samples are simply transformed from existing data. Our usual method for generating explanations involves deriving them from an associated factual sample, which is not available when directly transforming NLI contradiction examples. An example of an NLI sample is shown in Table 4.4.

Premise (Context)	**Hypothesis (Statement)**	**Label**
Labs were notable for Cr 1.7 (baseline 0.5 per old records) and lactate 2.4.	Patient has elevated Cr	Entailment

Table 4.4 Example of NLI sample from MedNLI

4.2.5 Summarization Dataset Transformation

For the summarization task, we create factual samples directly from original text-summary pairs using the original text as the context and the summary as the statement. To generate

non-factual samples, we choose a sentence from an original summary and modify it using an LLM to introduce contradictory information. This altered sentence is then reinserted into the summary at its initial position. The source text that was originally summarized serves as the context, and the entire modified summary is treated as the statement for our task. The original sentence from which the non-factual version was derived serves as the explanation for the non-factual statement. When selecting sentences for modification, we ensure they are at least 100 characters long to provide sufficient context for the LLM during the hallucination generation process. Table 4.5 provides a detailed example of a sample generated through this method, while the specific prompt used for generating contradictory sentences can be found in Figure 4.3.

```
You will be given a text and a sentence that was extracted from the text. Your task
is to transform the sentence by introducing a deliberate inaccuracy. Strategies can
include:
- Changing numerical values
- Inverting the meaning
- Using antonyms
- Negating the original statement

Text: {text}
Sentence: {sentence}

Ensure the new sentence remains grammatically correct but semantically differ-
ent from the original. Only output the transformed sentence without any additional
text.
```

Figure 4.3 Prompt format to generate samples for MedHal from a Summarization dataset

The primary goal of this task is to evaluate whether models can effectively detect subtle errors embedded within longer text sequences. Unlike classical tasks such as NLI, where the entire answer is typically binary (either entirely factual or entirely non-factual), our task derived from summarization samples specifically challenges LLMs because only a small portion of the statement might be false. This localized falsity can significantly trick many models, making detection much more difficult.

4.2.6 Dataset Description

To make sure our dataset is balanced, we've made sure to include an equal number of factual and non-factual samples. We achieved this by generating one non-factual sample for every factual one. Since our dataset draws from various sources and tasks, Table 4.6 provides a complete breakdown of all datasets used for this benchmark, alongside the number of samples contributed by each. Preprocessing steps were also applied to specific datasets. For

Sample Type	Source Document	Summary	Statement	Explanation
Factual	a central feature in the maturation of hearing is a transition in the electrical signature of cochlear hair cells from spontaneous calcium...	cochlear hair cells are high-frequency sensory receptors...	cochlear hair cells are high-frequency sensory receptors...	-
Non-Factual			cochlear hair cells are **low-frequency** sensory receptors...	According to the source document, cochlear hair cells are high-frequency sensory receptors...

Table 4.5 Example of Summarization Dataset Transformation

instance, in the ACM dataset, we only kept extractions that led to a valid JSON structure, as some original samples contained malformed JSON. Similarly, for Question Answering (QA) datasets, we used regular expressions to remove original dataset options or references (e.g., "Answer is c)") from explanations. This transformation was necessary because we repurposed these samples outside their original QA context. For example, an explanation like "Answer is c), Metformin is most commonly used as a first-line treatment, because..." would be modified to "Metformin is most commonly used as a first-line treatment, because...".

Dataset	Task	Synthetic	Content Type	# Samples	# Generated
MedMCQA [250]	QA	✗	Medical Content	183,000	70,730
MedNLI [246]	NLI	✗	Clinical Notes	11,232	7,488
ACM [248]	IE	✓	Clinical Notes	22,000	73,040
MedQA [249]	QA	✗	Medical Content	12,723	18,906
PubMedSum [251]	Sum	✗	Clinical Trials	33,772	178,657

Table 4.6 Description of datasets used to generate the MedHal benchmark

Several factors account for the variation between the initial sample counts from source datasets and the final number of samples included in our benchmark. Firstly, for tasks such as information extraction, a single sample from a source dataset can yield multiple derived samples for our benchmark. This occurs because one source text might contain numerous pieces of information, each becoming a distinct extraction in our unified task. Secondly, to

optimize computational efficiency, we intentionally limited the overall number of potential statements that could have been created from certain datasets. This reduction was achieved by generating only a random subset of the total possible samples. Thirdly, after all samples were initially prepared, any resulting token sequences exceeding 8192 tokens, as determined by the Llama3 tokenizer [132], were excluded. This threshold aligns with the typical context window of current "small" LLMs. Finally, only the training set of each individual dataset was used to create MedHal. This approach allows models trained on MedHal to be fairly assessed on the test sets of those original datasets.

The MedHal dataset is also divided into training, validation, and test sets with an 80/10/10 ratio, resulting in 313,920 samples in the training set, 17458 samples in the validation set and 17443 samples in the test set.

4.2.7 Statement Generation

To generate samples of the MedHal dataset, we leveraged the Llama-3-70B model [132]. We chose this model due to its state-of-the-art performance across a wide range of benchmarks [252]. For nearly all tasks, we employed a one-shot prompting strategy to create our samples. The exception was the summarization task. Our initial tests revealed that a one-shot setup offered no improvement over the baseline for summarization. This was likely due to the increased prompt length required as document-summary pairs are pretty large. However, for other tasks, the one-shot framework greatly improved the quality of generation, as creating statements primarily involved simple rephrasing.

4.3 Experiments

4.3.1 General Evaluation

To assess how well current models detect medical hallucinations, we evaluate their performance on MedHal's test set. This evaluation helps us understand which models are more effective at identifying hallucinated content in a medical context, and, more importantly, provides insights into whether specific fine-tuning strategies enhance performance in medical hallucination detection. Our study investigates on the performance of different categories of models :

- General models : General-purpose models
- Medical models : Models fine-tuned on medical datasets

- Evaluator models : Models fine-tuned on hallucination detection datasets (regardless of domain)

We aim to evaluate whether a specific category of models demonstrate superior performance on medical hallucination detection.

```
### Task Description
- You will evaluate whether a medical statement is factually accurate.
- The statement may reference a provided context.
- Respond with "YES" if the statement is factually correct or "NO" if it contains inaccuracies.
- In order to answer YES, everything in the statement must be supported by the context.
- In order to answer NO, there must be at least one piece of information in the statement that is not supported by the context.
- You must also provide an explanation of why you think the statement is factual or not. If it is factual, put "The statement is factual" as your explanation.
- Your answer should follow the following format :
Factual: [YES/NO]
Explanation: [Your explanation]

### Context
{context}

### Statement
{statement}

### Factual
```

Figure 4.4 Prompt format used to evaluate models on MedHal

We detail two types of metrics: factuality metrics and explanation metrics. Factuality metrics gauge a model's accuracy in correctly identifying factual and non-factual content. These metrics only consider the label (factual or non-factual) of a sample, not the explanation associated to it. To compute these metrics, we simply use the ground truth label and the class the sample was categorized into by the model. We then compute precision, recall, and F1 scores. Explanation metrics are designed to evaluate the validity of explanations provided by models for non-factual statements. These metrics specifically assess whether a model, upon detecting non-factual content, accurately pinpoints the erroneous portion of the statement. To compute these, we only consider samples where both the prediction and the ground truth label indicate a non-factual statement. This ensures that both a true explanation exists and that the model generated one. Explanation metrics include the ROUGE-1 (R1), ROUGE-2 (R2) [211] and BLEU [253] score which measure the n-gram overlap between the generated explanation and the real explanation. All results are presented in Table 4.7. Given

that models, when prompted using the format in Figure 4.4, do not consistently follow the specified output structure, we restrict the computation of factuality metrics only to samples where an answer could be extracted via regular expressions. The number of such samples corresponds to N in Table 4.7.

Type	Model	Factuality			Explanation			
		P	**R**	**F1**	**BLEU**	**R1**	**R2**	**N**
General	Llama-3.2-1B	0.51	0.18	0.26	0.01	0.08	0.03	7488
	Llama-3-8B	0.53	0.50	0.52	0.01	0.12	0.04	4600
Medical	BioMistral-7B	0.56	0.43	0.49	0.03	**0.22**	**0.11**	2011
	MedLlama-8B	0.52	0.59	0.55	0.03	0.21	0.08	**19251**
	Llama3-OpenBioLLM-8B	0.52	**0.77**	**0.62**	**0.04**	0.21	0.10	910
Evaluator	Prometheus-2-8x7B	**0.62**	0.37	0.47	-	-	-	-
	HallOumi-8B	0.59	0.54	0.56	-	-	-	-

Table 4.7 Performance of models on MedHal's test set (for general-purpose and medical models, we use the prompt template detailed in Figure 4.4, while for Prometheus-2-8x7B and HallOumi-8B, we adhere to the prompt formats recommended by their original authors, as these formats are optimized for their performance given how they were fine-tuned)

4.3.2 Impact of Fine-Tuning

Next, we fine-tune several of these models on MedHal's training set to evaluate their performance after specialization. The primary goal of this experiment was to determine if fine-tuning a medical model specifically for medical hallucination detection, or specializing a hallucination detection model on medical data, yields better performance. Additionally, we investigate whether simply fine-tuning a general-purpose model could achieve comparable results, potentially skipping the need for more specialized initial training. For this purpose, we fine-tune Llama-3-8B, Llama-3-OpenBioLLM-8B, and HallOumi-8B on MedHal. Furthermore, driven by our focus on reducing the computational power needed to detect hallucinated content in medical texts, a crucial consideration for medical facilities with limited computing infrastructure, we also fine-tune a Llama-3.2-1B model to assess the impact of a reduced model size on detection performance. The prompt used during training is shown in Figure 4.5. We fine-tune all models using the same QLora configuration. We detail the same metrics shown in Table 4.7 as well as the difference in F1 score between fine-tuned and non fine-tuned versions of each model. Results are shown in Table 4.8.

```
### Task Description
- You will evaluate whether a medical statement is factually accurate.
- The statement may reference a provided context.
- Respond with "YES" if the statement is factually correct or "NO" if it contains inaccuracies.
- In order to answer YES, everything in the statement must be supported by the context.
- In order to answer NO, there must be at least one piece of information in the statement that is not supported by the context.

### Context
{context}

### Statement
{statement}

### Factual
{label}

### Explanation
{explanation}
```

Figure 4.5 Prompt format used when to fine-tune a model on MedHal

Base Model	**Factuality**			**Explanation**			**ΔF1**
	P	**R**	**F1**	**BLEU**	**R1**	**R2**	
Llama-3.2-1B	0.75	0.77	0.76	**0.45**	0.70	0.59	**+0.50**
Llama-3-8B	**0.82**	0.73	0.77	**0.45**	**0.73**	**0.61**	+0.25
Llama3-OpenBioLLM-8B	0.77	**0.80**	**0.78**	0.10	0.22	0.17	+0.16
HallOumi-8B	0.79	0.77	**0.78**	**0.45**	0.72	**0.61**	+0.22

Table 4.8 Performance of models on MedHal's test set after fine-tuning (ΔF1 is the difference in F1-score between the fine-tuned and non fine-tuned version)

4.4 Downstream Task Evaluation

In order to assess the validity of the data used in our dataset and its potential for benchmarking models on medical hallucination detection, we evaluate the downstream capabilities of models trained using MedHal.

4.4.1 Evaluation on MedNLI

We first focus on the MedNLI dataset [246] by evaluating multiple models on its test set. It's important to note that MedHal was built only using samples from the training portion

of MedNLI, ensuring fair comparisons on the MedNLI test set without data leakage. We established baseline performance for various models and then compared these baselines to their multiple versions. First, we compare the models to their counterpart fine-tuned solely on MedNLI to predict whether a hypothesis is factual or not based on a premise versus models fine-tuned on MedHal. We fine-tune each model on MedNli for 5 epochs. Second, we also compare them to their counterpart that were fine-tuned on 1 epoch of MedHal.

Fine-Tuning	**Base Model**	**F1-Score**
None	Llama-3.2-1B	0.64
	Llama-3-8B	0.65
	BioMistral-7B	0.56
	MedLlama-3-8B	0.66
	OpenBioLLM-8B	0.64
	Prometheus-2-8x7B	0.62
	HallOumi-8B	**0.89**
MedNLI	Llama-3.2-1B	0.70
	Llama-3-8B	0.95
	OpenBioLLM-8B	**0.97**
	HallOumi-8B	**0.97**
MedHal	Llama-3.2-1B	0.89
	Llama-3-8B	0.96
	OpenBioLLM-8B	0.96
	HallOumi-8B	**0.97**

Table 4.9 F1-Score on the MedNLI dataset of models that have gone through fine-tuning on different datasets

This comparison, detailed in Table 4.9, helps us assess whether the additional tasks within our MedHal benchmark, such as information extraction, summarization, and question answering, contribute to a performance boost on MedNLI. Given that our evaluation framework produces only binary labels, we restrict our analysis to the subset of MedNLI's test set containing only entailment (classified as factual) and contradiction (classified as non-factual) samples, excluding all neutral samples.

4.4.2 Evaluation on Hallucination Dataset

We also evaluate several models on a more specialized hallucination detection dataset [254]. This dataset, based on MIMIC-III, contains samples with BHC sections and corresponding summaries generated by frontier models such as GPT-4 [227]. Medical students then annotated these summaries to identify hallucinated content. Sentence portions flagged as

containing hallucinated content were categorized using more specific labels: 'Unsupported Name', 'Fact Contradicted', 'Unsupported Procedure', 'Unsupported Other', 'Unsupported Medication', 'Unsupported Number', 'Unsupported Time', 'Unsupported Location', 'Unsupported Word', 'Unsupported Condition', and 'Incorrect Fact'. As we define factuality in the MedHal dataset as information that can be supported by context or general medical knowledge, we flag summaries containing at least one of these labels as not factual, thereby creating a hallucination detection dataset. Although this dataset contains only 210 samples (152 of which contain hallucinated content), we consider this evaluation particularly valuable because the dataset is both recent and relatively unknown in the medical community. Consequently, the models we are evaluating likely were not trained on this dataset, avoiding potential bias that could occur with MedNLI, which is a mainstream dataset in the medical NLP field and may have been encountered during training. To evaluate the impact of fine-tuning on our dataset, we take a base Llama-3-8B model [132] and fine-tune it on MedHal, then compare its performance against other high-performing models. We specifically chose the base Llama-3-8B rather than models that have been fine-tuned on medical text to isolate the effect of our dataset. We report the precision, recall, and F1-score of each model for detecting hallucinated content in Table 4.10. As models not fine-tuned on a specific answer format consistently flagged summaries as factual in 0-shot settings, we report only their 1-shot performance.

Model	Precision	Recall	F1
Llama-3.2-1B (1 shot)	0.81	0.32	0.45
Llama-3.1-8B (1 shot)	**1.00**	0.13	0.22
OpenBioLLM-3-8B (1 shot)	0.85	0.60	0.70
Prometheus-2-8x7B	0.77	0.75	0.76
HallOumi-8B	0.91	0.61	0.73
MedHal-Llama-3-8B	0.73	**0.87**	**0.79**

Table 4.10 Evaluation of different models on an Hallucination Dataset

4.5 Discussion

4.5.1 General vs Medical vs Evaluator models

A clear trend emerges from Table 4.7 indicating that models specifically fine-tuned on medical text generally outperform their general-purpose counterparts. Notably, OpenBioLLM-8B demonstrates superior performance, achieving an F1-score of 0.62 for factuality detection,

which is 10 points higher than Llama-3-8B (0.52). This improved performance of medical-tuned models extends to the explanation metrics (BLEU, R1, R2) as well, suggesting a more robust understanding of factual nuances within medical contexts. Despite their lower adherence to specific formats, a tendency reflected in generally low N values (except for MedLlama), medical models still outperform general-purpose models on explanation metrics. Interestingly, models designed as dedicated evaluators do not consistently surpass medical or general-purpose models in factuality detection. For instance, Prometheus-2-8x7B underperforms Llama-3-8B in F1-score, despite having significantly more parameters. A key distinction lies in the classification strategies adopted by different model types: evaluator models, such as Prometheus-2-8x7B and HallOumi-8B, exhibit a consistent pattern of higher precision than recall. Prometheus-2-8x7B, while achieving the highest precision (0.62) for identifying factual content, demonstrates a comparatively low recall (0.37). This indicates a tendency for these models to be highly confident when classifying a statement as factual, but they are also prone to missing many truly factual statements, classifying them as non-factual. In essence, they are "skeptical" about accepting a statement as factual. This behavior aligns with a cautious approach, where it's preferable to be on the side of caution by classifying uncertain factual content as non-factual. In contrast, medical models generally display a higher recall than precision. This suggests they are more adept at identifying a greater proportion of truly factual statements, even if it comes at the cost of occasionally misclassifying some non-factual content as factual. These findings suggest that while general medical fine-tuning is crucial for achieving higher overall factuality detection performance (as evidenced by the F1-score of medical models), fine-tuning for hallucination detection (which is inherent to the evaluator models' design) appears to instill a more cautious, precision-oriented classification strategy. This skepticism in factual classification, while leading to lower recall, could be a desirable characteristic in high-stakes medical contexts where misclassifying non-factual content as factual carries greater risks.

4.5.2 Fine-tuning on MedHal

As expected, when models are fine-tuned on MedHal, we observe a substantial improvement in their performance on the test set. However, there is no significant difference in overall performance whether the base model is a medical, general-purpose, or evaluator model before being fine-tuned on MedHal. For instance, Llama-3-8B, OpenBioLLM-8B and HallOumi-8B all achieve strong F1-scores for factuality. However, a specific anomaly appears with OpenBioLLM-8B. Fine-tuning it on MedHal leads to lower performance in explanation metrics compared to other fine-tuned models. In fact, a fine-tuned Llama-3.2-1B even outperforms OpenBioLLM-8B in explanation metrics (BLEU score of 0.45 vs. 0.10). Our analysis

suggests that while specialized pre-training offers advantages, its combination with subsequent fine-tuning on the MedHal dataset for explanation generation may lead to suboptimal results. Specifically, we observed that the model tends to produce an excessive number of tokens, frequently leading to gibberish after an initial coherent output. An example of this behaviour is shown in Table 4.11. However, this behaviour could be specific to this model.

Statement	**Explanation**	**Model Output**
MRI at 32 weeks of gestation showed a left-sided cyst increased to 27 cm in diameter, and the right-sided multilocular cyst became very close to the left monocytic cyst.	The MRI at 32 weeks of gestation shows an unremarkable result.	The MRI at 32 weeks of gestation showed complete resolution. <|reserved_special_token_233|> .ci_ail treatment. <|reserved_special_token_75|> a 2-dose iv mannitol treatment and an infusion ...

Table 4.11 Example of sample where OpenBioLLM-8B generates gibberish after an initial coherent output.

4.5.3 Downstream Performance

Evaluation on MedNLI

The evaluation on the MedNLI dataset provides insights into the effectiveness of our MedHal dataset for medical hallucination detection. Results from Table 4.9 show that models fine-tuned on MedHal achieve F1-scores between 0.89 and 0.97, showing improvement over both baseline performance (0.56-0.89) and MedNLI-only fine-tuning (0.70-0.97). This comparison is valid given that MedHal was constructed using only the training portion of MedNLI, ensuring no data leakage in our evaluation.

The comparable, and often superior, performance difference between MedHal fine-tuned models and MedNLI-only fine-tuned models suggests benefits from multi-task learning in medical hallucination detection. It's important to note that while the performance of 8B models fine-tuned using MedHal is similar to their MedNLI-only fine-tuned counterparts, the MedNLI-only versions have seen the dataset more times (5 epochs). In contrast, MedHal incorporates samples from MedNLI, but the MedHal fine-tuned versions have only processed these specific MedNLI samples one time as part of a broader, multi-task dataset. Despite this difference in exposure frequency to the MedNLI data, the MedHal-tuned models achieve competitive re-

sults. A possible explanation for the similar top scores achieved by the 8B models, regardless of whether they were fine-tuned on MedNLI or MedHal, is that the MedNLI dataset might be saturated for models of this size. However, for Llama-3.2-1B, a smaller model for which the dataset might not be as saturated, the jump in performance when fine-tuned on MedHal is significant, increasing from 0.70 (MedNLI only fine-tuning) to 0.89 (MedHal fine-tuning). This 1B parameter model, when exposed to MedHal's multi-task training, achieves an F1-score that matches the baseline performance of HallOumi-8B (0.89), a model eight times its size and specifically fine-tuned with for hallucination detection.

The diverse range of tasks incorporated in MedHal, including information extraction, summarization, and question answering, appears to provide complementary learning signals that support the models' ability to distinguish between factual and non-factual medical statements. This indicates that exposure to various medical reasoning tasks during training contributes to more robust representations for factuality assessment.

Notably, our MedHal fine-tuning approach enables models to exceed the performance of HallOumi-8B, a model specifically designed for hallucination detection, improving from its baseline F1-score of 0.89 to 0.94. This result highlights the importance of domain-specific training data, as our medically-focused dataset appears more effective than general hallucination detection training for medical contexts. The improvements are observed across different model architectures, from the smaller Llama-3.2-1B to the larger 8 billion parameter models.

Evaluation on Specialized Hallucination Dataset

The evaluation on the MIMIC-III-based hallucination dataset provides validation of our approach's generalization capabilities. The MedHal-fine-tuned Llama-3-8B model achieves an F1-score of 0.79 and demonstrates recall of 0.87, indicating capability in identifying hallucinated content. The model's performance on this unseen and doctor curated dataset suggests generalization emerges from training on MedHal. It is important to mention that the dataset's class imbalance, with 152 out of 210 samples containing hallucinated content, influences the precision-recall characteristics observed across models. This imbalance naturally leads to apparently high precision scores for models that tend to classify most samples as containing hallucinations. However, our MedHal fine-tuned model is the only one achieving high recall (0.87), demonstrating superior ability to identify the minority class of factual content while maintaining competitive precision.

CHAPTER 5 CONCLUSION

In this section, we summarize the work done in this study, indicate limitations with the developed methodology and give possible research directions for future work.

5.1 Summary of Works

We introduced a novel approach for generating domain-adapted clinical summaries using Large Language Models (LLMs). Our methodology leveraged ontologies to address two key challenges in clinical summarization: adapting summaries to specific clinical domains (e.g., radiology, nursing) and hallucination reduction. Our process began with comprehensive domain analysis to identify critical ontological concepts specific to each clinical domain. We then structured clinical notes around these concepts by systematically extracting relevant information. This extraction operated in conjunction with an ontology-guided decoding process that prioritized outputs aligned with both the input notes and ontological knowledge. Using these domain-specific extractions, we selectively retrieved values pertinent to the target domain to generate customized summaries. The innovation in our approach stemmed from two key components: the ontology-constrained decoding process and the automatic domain-adaptation methodology, which together enabled the generation of both structured and unstructured domain-adapted summaries. During our evaluation, we identified significant limitations in existing methods for assessing hallucinations. To address this gap, we developed a new dataset derived from multiple clinical tasks that effectively measured model reliability in detecting medical hallucinations. Additionally, we fine-tuned evaluator models on this dataset to facilitate clinical evaluation for the broader research community and evaluate the impact of different specialized LLMs (medical, evaluator) on medical hallucination detection.

5.2 Limitations

Our research is subject to several important constraints and limitations. Regarding our constrained decoding method, the computational overhead associated with multiple inference passes presents a primary challenge. While these processes can be parallelized across clinical notes and ontology classes, the beam search requirements create substantial obstacles for scalability and deployment in practical settings. The method's sensitivity to hyperparameters (including prompt format, k value, and α) further complicates optimization efforts.

Additionally, our approach depends heavily on ontology annotators, which, while available for SNOMED-CT, may not exist for other ontologies. Finally, a notable limitation is also the absence of human expert assessment and domain-specific gold standards for the generated summaries.

Concerning our medical hallucination detection benchmark, our samples lack formal expert review despite being derived from authentic content. The modification process, which utilized a large language model, might have introduced potentially false statements due to hallucinations. Consequently, our evaluator, while demonstrating improved performance on non-synthetic data, may exhibit diminished effectiveness in real-world applications. Furthermore, since the evaluator was fine-tuned from an existing model rather than trained from scratch, it may perpetuate inherent biases from the original model.

5.3 Future Research

To enhance the applicability of our domain adaptation method, future efforts should prioritize eliminating the reliance on an ontology annotator. This would allow for seamless integration across various ontologies and domains. Furthermore, given that current evaluation metrics are human-agnostic, it's crucial to assess the impact of ontology-constrained decoding directly with medical professionals.

Regarding the development of a medical hallucination detection benchmark, we propose incorporating a validation step after statement generation to minimize dataset errors. This could be augmented with a human validation step on a subset of samples to estimate data validity. More importantly, the dataset could be expanded to include additional tasks such as text classification, token classification, and text retrieval. Ontologies could also be leveraged to generate factual and non-factual statements by randomly swapping relationships between classes. For model training, future work could involve conducting ablation studies to understand the effect of different tasks and the impact of using samples from existing synthetic datasets on model performance. Furthermore, studying the effect of balancing the dataset, both by task and by label, could be a valuable area of future work.

REFERENCES

[1] T. Tajirian, V. Stergiopoulos, G. Strudwick, L. Sequeira, M. Sanches, J. Kemp, K. Ramamoorthi, T. Zhang, and D. Jankowicz, "The influence of electronic health record use on physician burnout: Cross-sectional survey," *J Med Internet Res*, vol. 22, no. 7, p. e19274, Jul 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/32673234

[2] T. Searle, Z. Ibrahim, J. Teo, and R. J. Dobson, "Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models," *Journal of Biomedical Informatics*, vol. 141, p. 104358, May 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046423000795

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023, arXiv:1706.03762 [cs]. [Online]. Available: http://arxiv.org/abs/1706.03762

[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023. [Online]. Available: https://doi.org/10.1145/3571730

[5] N. M. Guerreiro, D. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. T. Martins, "Hallucinations in Large Multilingual Translation Models," Mar. 2023, arXiv:2303.16104 [cs]. [Online]. Available: http://arxiv.org/abs/2303.16104

[6] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, p. 1–55, Jan. 2025. [Online]. Available: http://dx.doi.org/10.1145/3703155

[7] R. Studer, V. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *Data Knowledge Engineering*, vol. 25, no. 1, pp. 161–197, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X97000566

[8] Physiopedia, "Soap notes — physiopedia,," 2025, [Online; accessed 7-April-2025]. [Online]. Available: https://www.physio-pedia.com/index.php?title=SOAP_Notes&oldid=365097

[9] M. Afzal, F. Alam, K. M. Malik, and G. M. Malik, "Clinical context–aware biomedical text summarization using deep neural network: Model development and validation," *J Med Internet Res*, vol. 22, no. 10, p. e19810, Oct 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/33095174

[10] D. O. Cajueiro, A. G. Nery, I. Tavares, M. K. D. Melo, S. A. d. Reis, L. Weigang, and V. R. R. Celestino, "A comprehensive review of automatic text summarization techniques: method, data, evaluation and coding," Oct. 2023, arXiv:2301.03403 [cs]. [Online]. Available: http://arxiv.org/abs/2301.03403

[11] P. Roit, J. Ferret, L. Shani, R. Aharoni, G. Cideron, R. Dadashi, M. Geist, S. Girgin, L. Hussenot, O. Keller, N. Momchev, S. Ramos, P. Stanczyk, N. Vieillard, O. Bachem, G. Elidan, A. Hassidim, O. Pietquin, and I. Szpektor, "Factually Consistent Summarization via Reinforcement Learning with Textual Entailment Feedback," May 2023, arXiv:2306.00186 [cs]. [Online]. Available: http://arxiv.org/abs/2306.00186

[12] P. Béchard and O. M. Ayala, "Reducing hallucination in structured outputs via Retrieval-Augmented Generation," Apr. 2024, arXiv:2404.08189 [cs]. [Online]. Available: http://arxiv.org/abs/2404.08189

[13] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," Oct. 2023, arXiv:2305.14251 [cs]. [Online]. Available: http://arxiv.org/abs/2305.14251

[14] Z. Luo, Q. Xie, and S. Ananiadou, "Factual consistency evaluation of summarization in the Era of large language models," *Expert Systems with Applications*, vol. 254, p. 124456, Nov. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417424013228

[15] G. Adams, J. Zucker, and N. Elhadad, "SPEER: Sentence-Level Planning of Long Clinical Summaries via Embedded Entity Retrieval," 2024, publisher: [object Object] Version Number: 1. [Online]. Available: https://arxiv.org/abs/2401.02369

[16] M. Afzal, F. Alam, K. M. Malik, and G. M. Malik, "Clinical Context–Aware Biomedical Text Summarization Using Deep Neural Network: Model Development and Validation," *Journal of Medical Internet Research*, vol. 22, no. 10, p. e19810, Oct. 2020. [Online]. Available: http://www.jmir.org/2020/10/e19810/

[17] K. Krishna, S. Khosla, J. P. Bigham, and Z. C. Lipton, "Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques," Jun. 2021, arXiv:2005.01795 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2005.01795

[18] S. Kwon, Z. Yang, and H. Yu, "An Automatic SOAP Classification System Using Weakly Supervision And Transfer Learning," Nov. 2022, arXiv:2211.14539 [cs]. [Online]. Available: http://arxiv.org/abs/2211.14539

[19] D. van Zandvoort, L. Wiersema, T. Huibers, S. van Dulmen, and S. Brinkkemper, "Enhancing Summarization Performance through Transformer-Based Prompt Engineering in Automated Medical Reporting," Jan. 2024, arXiv:2311.13274 [cs]. [Online]. Available: http://arxiv.org/abs/2311.13274

[20] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," Sep. 2019, arXiv:1908.08345 [cs]. [Online]. Available: http://arxiv.org/abs/1908.08345

[21] H. Zhang, P. S. Yu, and J. Zhang, "A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models," Jun. 2024, arXiv:2406.11289 [cs]. [Online]. Available: http://arxiv.org/abs/2406.11289

[22] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, "Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model," Jun. 2019, arXiv:1906.01749 [cs]. [Online]. Available: http://arxiv.org/abs/1906.01749

[23] Y. Gao, W. Zhao, and S. Eger, "SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization," 2020, version Number: 1. [Online]. Available: https://arxiv.org/abs/2005.03724

[24] S. Huang, L. Qin, and Z. Cao, "Diffusion Language Model with Query-Document Relevance for Query-Focused Summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 11 020–11 030. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.735

[25] T. Yu, Z. Ji, and P. Fung, "Improving Query-Focused Meeting Summarization with Query-Relevant Knowledge," 2023, publisher: [object Object] Version Number: 1. [Online]. Available: https://arxiv.org/abs/2309.02105

[26] Z. Xu and D. Cohen, "A Lightweight Constrained Generation Alternative for Query-focused Summarization," Apr. 2023, arXiv:2304.11721 [cs]. [Online]. Available: http://arxiv.org/abs/2304.11721

[27] J. Du and Y. Gao, "Domain Adaptation and Summary Distillation for Unsupervised Query Focused Summarization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 3, pp. 1044–1055, Mar. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10185622/

[28] A. Nenkova and L. Vanderwende, "The impact of frequency on summarization," 01 2005.

[29] B. Das and S. Chakraborty, "An improved text sentiment classification model using tf-idf and next word negation," 2018. [Online]. Available: https://arxiv.org/abs/1806.06407

[30] W. M. Darling, "Multi-document summarization from first principles," *Theory and Applications of Categories*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:17089364

[31] H. Takamura and M. Okumura, "Text summarization model based on maximum coverage problem and its variant," in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, A. Lascarides, C. Gardent, and J. Nivre, Eds. Athens, Greece: Association for Computational Linguistics, Mar. 2009, pp. 781–789. [Online]. Available: https://aclanthology.org/E09-1089/

[32] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Foundations and Trends® in Machine Learning*, vol. 5, no. 2-3, pp. 123–286, 2012, arXiv:1207.6083 [stat]. [Online]. Available: http://arxiv.org/abs/1207.6083

[33] S. Cho, L. Lebanoff, H. Foroosh, and F. Liu, "Improving the similarity measure of determinantal point processes for extractive multi-document summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1027–1038. [Online]. Available: https://aclanthology.org/P19-1098/

[34] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, p. 457–479, Dec. 2004. [Online]. Available: http://dx.doi.org/10.1613/jair.1523

[35] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, D. Lin

and D. Wu, Eds. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 404–411. [Online]. Available: https://aclanthology.org/W04-3252/

[36] H. Zheng and M. Lapata, "Sentence centrality revisited for unsupervised summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6236–6247. [Online]. Available: https://aclanthology.org/P19-1628/

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[38] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162/

[39] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: a recurrent neural network based sequence model for extractive summarization of documents," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 3075–3081.

[40] W. Xiao and G. Carenini, "Extractive summarization of long documents by combining global and local context," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3011–3021. [Online]. Available: https://aclanthology.org/D19-1298/

[41] D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang, "Heterogeneous graph neural networks for extractive document summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 6209–6219. [Online]. Available: https://aclanthology.org/2020.acl-main.553/

[42] M. Yasunaga, R. Zhang, K. Meelu, A. Pareek, K. Srinivasan, and D. Radev, "Graph-based neural multi-document summarization," in *Proceedings of the 21st*

Conference on Computational Natural Language Learning (CoNLL 2017), R. Levy and L. Specia, Eds. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 452–462. [Online]. Available: https://aclanthology.org/K17-1045/

[43] J. Xu, Z. Gan, Y. Cheng, and J. Liu, "Discourse-aware neural extractive text summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 5021–5031. [Online]. Available: https://aclanthology.org/2020.acl-main.451/

[44] X. Zhang, F. Wei, and M. Zhou, "HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5059–5069. [Online]. Available: https://aclanthology.org/P19-1499/

[45] C. Rioux, S. A. Hasan, and Y. Chali, "Fear the REAPER: A system for automatic multi-document summarization with reinforcement learning," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 681–690. [Online]. Available: https://aclanthology.org/D14-1075/

[46] S. Narayan, S. B. Cohen, and M. Lapata, "Ranking sentences for extractive summarization with reinforcement learning," 2018. [Online]. Available: https://arxiv.org/abs/1802.08636

[47] J. Su, L. Zhang, H. R. Hassanzadeh, and T. Schaaf, "Extract and Abstract with BART for Clinical Notes from Doctor-Patient Conversations," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 2488–2492. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2022/su22b_interspeech.html

[48] M. Zhong, Y. Liu, S. Ge, Y. Mao, Y. Jiao, X. Zhang, Y. Xu, C. Zhu, M. Zeng, and J. Han, "Unsupervised Multi-Granularity Summarization," Dec. 2022, arXiv:2201.12502 [cs]. [Online]. Available: http://arxiv.org/abs/2201.12502

[49] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural

Language Generation, Translation, and Comprehension," Oct. 2019, arXiv:1910.13461 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1910.13461

[50] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023. [Online]. Available: https://arxiv.org/abs/1910.10683

[51] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," 2017. [Online]. Available: https://arxiv.org/abs/1705.04304

[52] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020. [Online]. Available: https://arxiv.org/abs/2004.05150

[53] L. Huang, S. Cao, N. Parulian, H. Ji, and L. Wang, "Efficient attentions for long document summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 1419–1436. [Online]. Available: https://aclanthology.org/2021.naacl-main.112/

[54] S. Cho, K. Song, X. Wang, F. Liu, and D. Yu, "Toward unifying text segmentation and long document summarization," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 106–118. [Online]. Available: https://aclanthology.org/2022.emnlp-main.8/

[55] T. Rohde, X. Wu, and Y. Liu, "Hierarchical learning for generation with long source sequences," 2021. [Online]. Available: https://arxiv.org/abs/2104.07545

[56] Y. Liu and M. Lapata, "Hierarchical transformers for multi-document summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5070–5081. [Online]. Available: https://aclanthology.org/P19-1500/

[57] Y. Liu, J. Zhang, Y. Wan, C. Xia, L. He, and P. Yu, "HETFORMER: Heterogeneous transformer with sparse attention for long-text extractive summarization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and

Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 146–154. [Online]. Available: https://aclanthology.org/2021.emnlp-main.13/

[58] W. Xiao, I. Beltagy, G. Carenini, and A. Cohan, "PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5245–5263. [Online]. Available: https://aclanthology.org/2022.acl-long.360/

[59] C. Li, W. Xu, S. Li, and S. Gao, "Guiding generation for abstractive text summarization based on key information guide network," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 55–60. [Online]. Available: https://aclanthology.org/N18-2009/

[60] Y. Liu and P. Liu, "SimCLS: A simple framework for contrastive learning of abstractive summarization," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 1065–1072. [Online]. Available: https://aclanthology.org/2021.acl-short.135/

[61] Y. Liu, Z.-Y. Dou, and P. Liu, "Refsum: Refactoring neural summarization," 2021. [Online]. Available: https://arxiv.org/abs/2104.07210

[62] Y. Liu, P. Liu, D. Radev, and G. Neubig, "Brio: Bringing order to abstractive summarization," 2022. [Online]. Available: https://arxiv.org/abs/2203.16804

[63] J. Xie, Q. Su, S. Zhang, and X. Zhang, "Alleviating exposure bias via multi-level contrastive learning and deviation simulation in abstractive summarization," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9732–9747. [Online]. Available: https://aclanthology.org/2023.findings-acl.617/

[64] F. Schmidt, "Generalization in generation: A closer look at exposure bias," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*, A. Birch, A. Finch, H. Hayashi, I. Konstas, T. Luong, G. Neubig, Y. Oda, and K. Sudoh, Eds.

Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 157–167. [Online]. Available: https://aclanthology.org/D19-5616/

[65] C. Whitehouse, F. Huot, J. Bastings, M. Dehghani, C.-C. Lin, and M. Lapata, "Low-rank adaptation for multilingual summarization: An empirical study," 2024. [Online]. Available: https://arxiv.org/abs/2311.08572

[66] F. Nan, R. Nallapati, Z. Wang, C. Nogueira dos Santos, H. Zhu, D. Zhang, K. McKeown, and B. Xiang, "Entity-level factual consistency of abstractive text summarization," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 2727–2733. [Online]. Available: https://aclanthology.org/2021.eacl-main.235/

[67] Y. Dong, S. Wang, Z. Gan, Y. Cheng, J. C. K. Cheung, and J. Liu, "Multi-fact correction in abstractive text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 9320–9331. [Online]. Available: https://aclanthology.org/2020.emnlp-main.749/

[68] M. Cao, Y. Dong, J. Wu, and J. C. K. Cheung, "Factual error correction for abstractive summarization models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 6251–6258. [Online]. Available: https://aclanthology.org/2020.emnlp-main.506/

[69] S. Chen, F. Zhang, K. Sone, and D. Roth, "Improving faithfulness in abstractive summarization with contrast candidate generation and selection," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 5935–5941. [Online]. Available: https://aclanthology.org/2021.naacl-main.475/

[70] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee,

D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: https://arxiv.org/abs/2307.09288

[71] Y. Wang, Z. Zhang, and R. Wang, "Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method," May 2023, arXiv:2305.13412 [cs]. [Online]. Available: http://arxiv.org/abs/2305.13412

[72] G. Adams, A. Fabbri, F. Ladhak, E. Lehman, and N. Elhadad, "From sparse to dense: GPT-4 summarization with chain of density prompting," in *Proceedings of the 4th New Frontiers in Summarization Workshop*, Y. Dong, W. Xiao, L. Wang, F. Liu, and G. Carenini, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 68–74. [Online]. Available: https://aclanthology.org/2023.newsum-1.7/

[73] M. Liu, D. Chen, Y. Li, G. Fang, and Y. Shen, "Chartthinker: A contextual chain-of-thought approach to optimized chart summarization," 2024. [Online]. Available: https://arxiv.org/abs/2403.11236

[74] C. Qin and S. Joty, "LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=HCRVf71PMF

[75] M. Ravaut, H. Chen, R. Zhao, C. Qin, S. Joty, and N. Chen, "Promptsum: Parameter-efficient controllable abstractive summarization," 2023. [Online]. Available: https://arxiv.org/abs/2308.03117

[76] H. Zhang, X. Liu, and J. Zhang, "SummIt: Iterative text summarization via ChatGPT," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 644–10 657. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.714/

[77] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," Oct. 2023, arXiv:2303.08896 [cs]. [Online]. Available: http://arxiv.org/abs/2303.08896

[78] Y. Zhang, A. Ni, Z. Mao, C. H. Wu, C. Zhu, B. Deb, A. Awadallah, D. Radev, and R. Zhang, "Summ^n: A multi-stage summarization framework for long input dialogues and documents," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1592–1604. [Online]. Available: https://aclanthology.org/2022.acl-long.112/

[79] O. Ahuja, J. Xu, A. Gupta, K. Horecka, and G. Durrett, "ASPECTNEWS: Aspect-oriented summarization of news documents," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6494–6506. [Online]. Available: https://aclanthology.org/2022.acl-long.449/

[80] H. Zhang, X. Liu, and J. Zhang, "Extractive summarization via ChatGPT for faithful summary generation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3270–3278. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.214/

[81] Y.-C. Chen and M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," 2018. [Online]. Available: https://arxiv.org/abs/1805.11080

[82] Z.-Y. Dou, P. Liu, H. Hayashi, Z. Jiang, and G. Neubig, "GSum: A general framework for guided neural abstractive summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 4830–4842. [Online]. Available: https://aclanthology.org/2021.naacl-main.384/

[83] F. Wang, K. Song, H. Zhang, L. Jin, S. Cho, W. Yao, X. Wang, M. Chen, and D. Yu, "Salience allocation as guidance for abstractive summarization," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6094–6106. [Online]. Available: https://aclanthology.org/2022.emnlp-main.409/

[84] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," 2024. [Online]. Available: https://arxiv.org/abs/2404.16130

[85] S. Liu, J. Wu, J. Bao, W. Wang, N. Hovakimyan, and C. G. Healey, "Towards a robust retrieval-based summarization system," 2024. [Online]. Available: https://arxiv.org/abs/2403.19889

[86] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, "In-context retrieval-augmented language models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1316–1331, 2023. [Online]. Available: https://aclanthology.org/2023.tacl-1.75/

[87] D. Sachan, M. Lewis, M. Joshi, A. Aghajanyan, W.-t. Yih, J. Pineau, and L. Zettlemoyer, "Improving passage retrieval with zero-shot question generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3781–3797. [Online]. Available: https://aclanthology.org/2022.emnlp-main.249/

[88] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1762–1777. [Online]. Available: https://aclanthology.org/2023.acl-long.99/

[89] R. Jain, A. Jangra, S. Saha, and A. Jatowt, "A survey on medical document summarization," 2022. [Online]. Available: https://arxiv.org/abs/2212.01669

[90] J. Steinkamp, J. J. Kantrowitz, and S. Airan-Javia, "Prevalence and Sources of Duplicate Information in the Electronic Medical Record," *JAMA Network Open*, vol. 5, no. 9, pp. e2 233 348–e2 233 348, Sep. 2022, _eprint: https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2796664/steinkamp_2022_oi_2209 [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2022.33348

[91] G. Adams, E. Alsentzer, M. Ketenci, J. Zucker, and N. Elhadad, "What's in a Summary? Laying the Groundwork for Advances in Hospital-Course Summarization," Apr. 2021, arXiv:2105.00816 [cs]. [Online]. Available: http://arxiv.org/abs/2105.00816

[92] V. Gupta, P. Bharti, P. Nokhiz, and H. Karnick, "SumPubMed: Summarization dataset of PubMed scientific articles," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, J. Kabbara, H. Lin, A. Paullada, and J. Vamvas, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 292–303. [Online]. Available: https://aclanthology.org/2021.acl-srw.30/

[93] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, May 2016. [Online]. Available: https://www.nature.com/articles/sdata201635

[94] H.-C. Shing, C. P. Shivade, N. Pourdamghani, F. Nan, P. Resnik, D. W. Oard, and P. Bhatia, "Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes," *ArXiv*, vol. abs/2104.13498, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233423645

[95] K. Pal, S. A. Bahrainian, L. Mercurio, and C. Eickhoff, "Neural summarization of electronic health records," 2023. [Online]. Available: https://arxiv.org/abs/2305.15222

[96] G. Adams, H.-C. Shing, Q. Sun, C. Winestock, K. McKeown, and N. Elhadad, "Learning to revise references for faithful summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4009–4027. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.296/

[97] S. H. Lee, "Natural language generation for electronic health records," *npj Digital Medicine*, vol. 1, no. 1, p. 63, Nov. 2018. [Online]. Available: https://www.nature.com/articles/s41746-018-0070-0

[98] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, publisher: [object Object] Version Number: 5. [Online]. Available: https://arxiv.org/abs/1901.07042

[99] X. Yang, M. Ye, Q. You, and F. Ma, "Writing by Memorizing: Hierarchical Retrieval-based Medical Report Generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online:

Association for Computational Linguistics, 2021, pp. 5000–5009. [Online]. Available: https://aclanthology.org/2021.acl-long.387

[100] Y.-N. Chuang, R. Tang, X. Jiang, and X. Hu, "SPeC: A Soft Prompt-Based Calibration on Performance Variability of Large Language Model in Clinical Notes Summarization," Aug. 2023, arXiv:2303.13035 [cs]. [Online]. Available: http://arxiv.org/abs/2303.13035

[101] Z. Guan, Z. Wu, Z. Liu, D. Wu, H. Ren, Q. Li, X. Li, and N. Liu, "CohortGPT: An Enhanced GPT for Participant Recruitment in Clinical Study," Jul. 2023, arXiv:2307.11346 [cs]. [Online]. Available: http://arxiv.org/abs/2307.11346

[102] M. Liu, D. Zhang, W. Tan, and H. Zhang, "DeakinNLP at ProbSum 2023: Clinical Progress Note Summarization with Rules and Language ModelsClinical Progress Note Summarization with Rules and Languague Models," in *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 491–496. [Online]. Available: https://aclanthology.org/2023.bionlp-1.47

[103] A. Sorita, P. M. Robelia, S. B. Kattel, C. P. McCoy, A. S. Keller, J. Almasri, M. H. Murad, J. S. Newman, and D. T. Kashiwagi, "The Ideal Hospital Discharge Summary: A Survey of U.S. Physicians," *Journal of Patient Safety*, vol. 17, no. 7, 2021. [Online]. Available: https://journals.lww.com/journalpatientsafety/fulltext/2021/10000/the_ideal_hospital_discharge_summary___a_survey_of.16.aspx

[104] F. Moramarco, A. Papadopoulos Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, and A. Savkov, "Human evaluation and correlation with automatic metrics in consultation note generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5739–5754. [Online]. Available: https://aclanthology.org/2022.acl-long.394/

[105] S. Kwon, Z. Yang, and H. Yu, "An automatic soap classification system using weakly supervision and transfer learning," 2022. [Online]. Available: https://arxiv.org/abs/2211.14539

[106] M. Afzal, F. Alam, K. M. Malik, and G. M. Malik, "Clinical context–aware biomedical text summarization using deep neural network: Model development and validation,"

J Med Internet Res, vol. 22, no. 10, p. e19810, Oct 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/33095174

[107] I. of Medicine (US) Committee on Toxicology and E. H. I. R. for Health Professionals, "Understanding the information needs of health professionals," Jan 1997. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK45472/

[108] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 1906–1919. [Online]. Available: https://aclanthology.org/2020.acl-main.173/

[109] K. Filippova, "Controlled hallucinations: Learning to generate faithfully from noisy data," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 864–870. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.76/

[110] F. Ridder and M. Schilling, "The hallurag dataset: Detecting closed-domain hallucinations in rag applications using an llm's internal states," 2024. [Online]. Available: https://arxiv.org/abs/2412.17056

[111] S. Cheng, L. Pan, X. Yin, X. Wang, and W. Y. Wang, "Understanding the interplay between parametric and contextual knowledge for large language models," 2024. [Online]. Available: https://arxiv.org/abs/2410.08414

[112] A. Agrawal, M. Suzgun, L. Mackey, and A. Kalai, "Do language models know when they're hallucinating references?" in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 912–928. [Online]. Available: https://aclanthology.org/2024.findings-eacl.62/

[113] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," 2017. [Online]. Available: https://arxiv.org/abs/1704.04368

[114] M. Roberti, G. Bonetta, R. Cancelliere, and P. Gallinari, *Copy Mechanism and Tailored Training for Character-Based Data-to-Text Generation.* Springer International Publishing, 2020, p. 648–664. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-46147-8_39

[115] H.-S. Chang, Z. Yao, A. Gon, H. Yu, and A. McCallum, "Revisiting the architectures like pointer networks to efficiently improve the next word distribution, summarization factuality, and beyond," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 707–12 730. [Online]. Available: https://aclanthology.org/2023.findings-acl.805/

[116] Z. Sun, Z. Si, X. Zang, K. Zheng, Y. Song, X. Zhang, and J. Xu, "Largepig: Your large language model is secretly a pointer generator," 2024. [Online]. Available: https://arxiv.org/abs/2410.11366

[117] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen, "Breaking the softmax bottleneck: A high-rank rnn language model," 2018. [Online]. Available: https://arxiv.org/abs/1711.03953

[118] H.-S. Chang and A. McCallum, "Softmax bottleneck makes language models unable to represent multi-mode word distributions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8048–8073. [Online]. Available: https://aclanthology.org/2022.acl-long.554/

[119] D. Dai, W. Jiang, Q. Dong, Y. Lyu, Q. She, and Z. Sui, "Neural knowledge bank for pretrained transformers," 2022. [Online]. Available: https://arxiv.org/abs/2208.00399

[120] E. Mitchell, C. Lin, A. Bosselut, C. D. Manning, and C. Finn, "Memory-based model editing at scale," 2022. [Online]. Available: https://arxiv.org/abs/2206.06520

[121] Z. Huang, Y. Shen, X. Zhang, J. Zhou, W. Rong, and Z. Xiong, "Transformer-patcher: One mistake worth one neuron," 2023. [Online]. Available: https://arxiv.org/abs/2301.09785

[122] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022. [Online]. Available: https://arxiv.org/abs/2203.02155

[123] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," 2024. [Online]. Available: https://arxiv.org/abs/2305.18290

[124] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela, "Kto: Model alignment as prospect theoretic optimization," 2024. [Online]. Available: https://arxiv.org/abs/2402.01306

[125] A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, L. Campbell-Gillingham, J. Uesato, P.-S. Huang, R. Comanescu, F. Yang, A. See, S. Dathathri, R. Greig, C. Chen, D. Fritz, J. S. Elias, R. Green, S. Mokrá, N. Fernando, B. Wu, R. Foley, S. Young, I. Gabriel, W. Isaac, J. Mellor, D. Hassabis, K. Kavukcuoglu, L. A. Hendricks, and G. Irving, "Improving alignment of dialogue agents via targeted human judgements," 2022. [Online]. Available: https://arxiv.org/abs/2209.14375

[126] Z. Wu, Y. Hu, W. Shi, N. Dziri, A. Suhr, P. Ammanabrolu, N. A. Smith, M. Ostendorf, and H. Hajishirzi, "Fine-Grained Human Feedback Gives Better Rewards for Language Model Training," Oct. 2023, arXiv:2306.01693 [cs]. [Online]. Available: http://arxiv.org/abs/2306.01693

[127] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," 2023. [Online]. Available: https://arxiv.org/abs/2305.20050

[128] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.03300

[129] M. Hahn, "Theoretical limitations of self-attention in neural sequence models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 156–171, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.11/

[130] D. Chiang and P. Cholak, "Overcoming a theoretical limitation of self-attention," 2022. [Online]. Available: https://arxiv.org/abs/2202.12172

[131] Z. Li, S. Zhang, H. Zhao, Y. Yang, and D. Yang, "Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer," 2023. [Online]. Available: https://arxiv.org/abs/2307.00360

[132] A. Grattafiori, A. Dubey, and A. Jauhri, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[133] Y. Wang, J. Wei, C. Y. Liu, J. Pang, Q. Liu, A. P. Shah, Y. Bao, Y. Liu, and W. Wei, "Llm unlearning via loss adjustment with only forget data," 2024. [Online]. Available: https://arxiv.org/abs/2410.11143

[134] S. Cao and L. Wang, "CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6633–6649. [Online]. Available: https://aclanthology.org/2021.emnlp-main.532/

[135] I.-C. Chern, Z. Wang, S. Das, B. Sharma, P. Liu, and G. Neubig, "Improving Factuality of Abstractive Summarization via Contrastive Reward Learning," Jul. 2023, arXiv:2307.04507 [cs]. [Online]. Available: http://arxiv.org/abs/2307.04507

[136] L. Kühnel, A. Schulz, B. Hammer, and J. Fluck, "Bert weaver: Using weight averaging to enable lifelong learning for transformer-based models in biomedical semantic search engines," 2023. [Online]. Available: https://arxiv.org/abs/2202.10101

[137] J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin, "Linear mode connectivity and the lottery ticket hypothesis," 2020. [Online]. Available: https://arxiv.org/abs/1912.05671

[138] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," 2023. [Online]. Available: https://arxiv.org/abs/2212.04089

[139] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. Gontijo-Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt, "Robust fine-tuning of zero-shot models," 2022. [Online]. Available: https://arxiv.org/abs/2109.01903

[140] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059. [Online]. Available: https://aclanthology.org/2021.emnlp-main.243/

[141] R. Hendel, M. Geva, and A. Globerson, "In-context learning creates task vectors," in *Findings of the Association for Computational Linguistics: EMNLP*

2023, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9318–9333. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.624/

[142] S. Soo, W. Teng, and C. Balaganesh, "Steering large language models with feature guided activation additions," 2025. [Online]. Available: https://arxiv.org/abs/2501.09929

[143] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg, "Inference-time intervention: Eliciting truthful answers from a language model," 2024. [Online]. Available: https://arxiv.org/abs/2306.03341

[144] Z. Wang, S. Mao, W. Wu, T. Ge, F. Wei, and H. Ji, "Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration," 2024. [Online]. Available: https://arxiv.org/abs/2307.05300

[145] M. Zheng, J. Pei, L. Logeswaran, M. Lee, and D. Jurgens, "When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2311.10054

[146] A. Lu, H. Zhang, Y. Zhang, X. Wang, and D. Yang, "Bounding the capabilities of large language models in open text generation with prompt constraints," 2023. [Online]. Available: https://arxiv.org/abs/2302.09185

[147] C. Li, J. Wang, Y. Zhang, K. Zhu, W. Hou, J. Lian, F. Luo, Q. Yang, and X. Xie, "Large language models understand and can be enhanced by emotional stimuli," 2023. [Online]. Available: https://arxiv.org/abs/2307.11760

[148] X. Wang, X. Li, Z. Yin, Y. Wu, and L. Jia, "Emotional intelligence of large language models," 2023. [Online]. Available: https://arxiv.org/abs/2307.09042

[149] Z. R. Tam, C.-K. Wu, Y.-L. Tsai, C.-Y. Lin, H.-y. Lee, and Y.-N. Chen, "Let me speak freely? a study on the impact of format restrictions on large language model performance." in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, F. Dernoncourt, D. Preoţiuc-Pietro, and A. Shimorina, Eds. Miami, Florida, US: Association for Computational Linguistics, Nov. 2024, pp. 1218–1236. [Online]. Available: https://aclanthology.org/2024.emnlp-industry.91/

[150] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, and S. Hasan, "Does prompt formatting have any impact on llm performance?" 2024. [Online]. Available: https://arxiv.org/abs/2411.10541

[151] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[152] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for GPT-3?" in *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, E. Agirre, M. Apidianaki, and I. Vulić, Eds. Dublin, Ireland and Online: Association for Computational Linguistics, May 2022, pp. 100–114. [Online]. Available: https://aclanthology.org/2022.deelio-1.10/

[153] H. Su, J. Kasai, C. H. Wu, W. Shi, T. Wang, J. Xin, R. Zhang, M. Ostendorf, L. Zettlemoyer, N. A. Smith, and T. Yu, "Selective annotation makes language models better few-shot learners," 2022. [Online]. Available: https://arxiv.org/abs/2209.01975

[154] S. An, B. Zhou, Z. Lin, Q. Fu, B. Chen, N. Zheng, W. Chen, and J.-G. Lou, "Skill-based few-shot selection for in-context learning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 472–13 492. [Online]. Available: https://aclanthology.org/2023.emnlp-main.831/

[155] H. J. Kim, H. Cho, J. Kim, T. Kim, K. M. Yoo, and S. goo Lee, "Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator," 2022. [Online]. Available: https://arxiv.org/abs/2206.08082

[156] X. Li and X. Qiu, "Finding support examples for in-context learning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6219–6235. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.411/

[157] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[158] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot, "Complexity-based prompting for multi-step reasoning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=yf1icZHC-l9

[159] X. Li and X. Qiu, "MoT: Memory-of-thought enables ChatGPT to self-improve," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6354–6374. [Online]. Available: https://aclanthology.org/2023.emnlp-main.392/

[160] Y. K. Chia, G. Chen, L. A. Tuan, S. Poria, and L. Bing, "Contrastive chain-of-thought prompting," 2023. [Online]. Available: https://arxiv.org/abs/2311.09277

[161] B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun, "Towards understanding chain-of-thought prompting: An empirical study of what matters," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 2717–2739. [Online]. Available: https://aclanthology.org/2023.acl-long.153/

[162] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi, "Least-to-most prompting enables complex reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2205.10625

[163] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.10601

[164] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," 2023. [Online]. Available: https://arxiv.org/abs/2203.11171

[165] C. Si, W. Shi, C. Zhao, L. Zettlemoyer, and J. Boyd-Graber, "Getting MoRE out of mixture of language model reasoning experts," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8234–8249. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.552/

[166] N. Miao, Y. W. Teh, and T. Rainforth, "Selfcheck: Using llms to zero-shot check their own step-by-step reasoning," 2023. [Online]. Available: https://arxiv.org/abs/2308.00436

[167] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-Verification Reduces Hallucination in Large Language Models," 2023, publisher: arXiv Version Number: 2. [Online]. Available: https://arxiv.org/abs/2309.11495

[168] Y. Weng, M. Zhu, F. Xia, B. Li, S. He, S. Liu, B. Sun, K. Liu, and J. Zhao, "Large language models are better reasoners with self-verification," 2023. [Online]. Available: https://arxiv.org/abs/2212.09561

[169] Y. Zhang, J. Yang, Y. Yuan, and A. C.-C. Yao, "Cumulative reasoning with large language models," 2025. [Online]. Available: https://arxiv.org/abs/2308.04371

[170] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, "Self-refine: Iterative refinement with self-feedback," 2023. [Online]. Available: https://arxiv.org/abs/2303.17651

[171] C. Huang, L. Huang, J. Leng, J. Liu, and J. Huang, "Efficient test-time scaling via self-calibration," 2025. [Online]. Available: https://arxiv.org/abs/2503.00031

[172] R. Cohen, M. Hamri, M. Geva, and A. Globerson, "LM vs LM: Detecting factual errors via cross examination," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 621–12 640. [Online]. Available: https://aclanthology.org/2023.emnlp-main.778/

[173] K. Krishna, S. Ramprasad, P. Gupta, B. C. Wallace, Z. C. Lipton, and J. P. Bigham, "Genaudit: Fixing factual errors in language model outputs with evidence," 2025. [Online]. Available: https://arxiv.org/abs/2402.12566

[174] J. Luo, C. Xiao, and F. Ma, "Zero-Resource Hallucination Prevention for Large Language Models," 2023, publisher: arXiv Version Number: 3. [Online]. Available: https://arxiv.org/abs/2309.02654

[175] O. Press, M. Zhang, S. Min, L. Schmidt, N. Smith, and M. Lewis, "Measuring and narrowing the compositionality gap in language models," in *Findings of the*

Association for Computational Linguistics: EMNLP 2023, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5687–5711. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.378/

[176] G. Sriramanan, S. Bharti, V. S. Sadasivan, S. Saha, P. Kattakinda, and S. Feizi, "LLM-check: Investigating detection of hallucinations in large language models," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=LYx4w3CAgy

[177] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020, publisher: [object Object] Version Number: 4. [Online]. Available: https://arxiv.org/abs/2005.11401

[178] F. Mu, Y. Jiang, L. Zhang, L. Liuchu, W. Li, P. Xie, and F. Huang, "Query routing for homogeneous tools: An instantiation in the RAG scenario," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 10 225–10 230. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.598/

[179] W. Peng, G. Li, Y. Jiang, Z. Wang, D. Ou, X. Zeng, D. Xu, T. Xu, and E. Chen, "Large language model based long-tail query rewriting in taobao search," 2024. [Online]. Available: https://arxiv.org/abs/2311.03758

[180] I. S. Singh, R. Aggarwal, I. Allahverdiyev, M. Taha, A. Akalin, K. Zhu, and S. O'Brien, "Chunkrag: Novel llm-chunk filtering method for rag systems," 2024. [Online]. Available: https://arxiv.org/abs/2410.19572

[181] H. Jiang, Q. Wu, X. Luo, D. Li, C.-Y. Lin, Y. Yang, and L. Qiu, "LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1658–1677. [Online]. Available: https://aclanthology.org/2024.acl-long.91/

[182] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," 2024. [Online]. Available: https://arxiv.org/abs/2401.05856

[183] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, and F. Silvestri, "The power of noise: Redefining retrieval for rag systems," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR 2024. ACM, Jul. 2024, p. 719–729. [Online]. Available: http://dx.doi.org/10.1145/3626772.3657834

[184] S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, and S. Singh, "Entity-based knowledge conflicts in question answering," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7052–7063. [Online]. Available: https://aclanthology.org/2021.emnlp-main.565/

[185] Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, and W. Chen, "Critic: Large language models can self-correct with tool-interactive critiquing," 2024. [Online]. Available: https://arxiv.org/abs/2305.11738

[186] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," 2023. [Online]. Available: https://arxiv.org/abs/2303.11366

[187] Y. Peng, A. D. Gotmare, M. Lyu, C. Xiong, S. Savarese, and D. Sahoo, "Perfcodegen: Improving performance of llm generated code with execution feedback," 2024. [Online]. Available: https://arxiv.org/abs/2412.03578

[188] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, 2017. [Online]. Available: http://dx.doi.org/10.18653/v1/W17-3207

[189] M. Renze and E. Guven, "The effect of sampling temperature on problem solving in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.05201

[190] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 889–898. [Online]. Available: https://aclanthology.org/P18-1082/

[191] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rygGQyrFvH

[192] J. Hewitt, C. Manning, and P. Liang, "Truncation sampling as language model desmoothing," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3414–3427. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.249/

[193] Y. Su, T. Lan, Y. Wang, D. Yogatama, L. Kong, and N. Collier, "A contrastive framework for neural text generation," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=V88BafmH9Pj

[194] X. L. Li, A. Holtzman, D. Fried, P. Liang, J. Eisner, T. Hashimoto, L. Zettlemoyer, and M. Lewis, "Contrastive decoding: Open-ended text generation as optimization," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 286–12 312. [Online]. Available: https://aclanthology.org/2023.acl-long.687/

[195] A. See, S. Roller, D. Kiela, and J. Weston, "What makes a good conversation? how controllable attributes affect human judgments," 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1702–1723. [Online]. Available: https://aclanthology.org/N19-1170/

[196] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and Play Language Models: A Simple Approach to Controlled Text Generation," Mar. 2020, arXiv:1912.02164 [cs]. [Online]. Available: http://arxiv.org/abs/1912.02164

[197] K. Yang and D. Klein, "FUDGE: Controlled Text Generation With Future Discriminators," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, 2021, pp. 3511–3535, arXiv:2104.05218 [cs]. [Online]. Available: http://arxiv.org/abs/2104.05218

[198] D. Deutsch, S. Upadhyay, and D. Roth, "A general-purpose algorithm for constrained sequential inference," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 482–492.

[199] H. Zhang, M. Dang, N. Peng, and G. V. den Broeck, "Tractable control for autoregressive language generation," 2023. [Online]. Available: https://arxiv.org/abs/2304.07438

[200] B. T. Willard and R. Louf, "Efficient Guided Generation for Large Language Models," Aug. 2023, arXiv:2307.09702 [cs]. [Online]. Available: http://arxiv.org/abs/2307.09702

[201] L. Zheng, L. Yin, Z. Xie, C. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez, C. Barrett, and Y. Sheng, "SGLang: Efficient Execution of Structured Language Model Programs," Jun. 2024, arXiv:2312.07104 [cs]. [Online]. Available: http://arxiv.org/abs/2312.07104

[202] H. Zhang, P.-N. Kung, M. Yoshida, G. V. den Broeck, and N. Peng, "Adaptable logical control for large language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.13892

[203] E. Stengel-Eskin, K. Rawlins, and B. Van Durme, "Zero and Few-shot Semantic Parsing with Ambiguous Inputs," Jan. 2024, arXiv:2306.00824 [cs]. [Online]. Available: http://arxiv.org/abs/2306.00824

[204] S. Geng, M. Josifoski, M. Peyrard, and R. West, "Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning," Nov. 2023, arXiv:2305.13971 [cs]. [Online]. Available: http://arxiv.org/abs/2305.13971

[205] K. Murray and D. Chiang, "Correcting length bias in neural machine translation," 2018. [Online]. Available: https://arxiv.org/abs/1808.10006

[206] Y. Yang, L. Huang, and M. Ma, "Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation," 2018. [Online]. Available: https://arxiv.org/abs/1808.09582

[207] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse Beam Search: Decoding Diverse Solutions from

Neural Sequence Models," Oct. 2018, arXiv:1610.02424 [cs]. [Online]. Available: http://arxiv.org/abs/1610.02424

[208] C. Hokamp and Q. Liu, "Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1535–1546. [Online]. Available: http://aclweb.org/anthology/P17-1141

[209] C. Meister, T. Vieira, and R. Cotterell, "Best-first beam search," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 795–809, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.51/

[210] A. K. Sridhar and E. Visser, "Improved beam search for hallucination mitigation in abstractive summarization," 2023. [Online]. Available: https://arxiv.org/abs/2212.02712

[211] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[212] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909/

[213] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," 2019, publisher: [object Object] Version Number: 3. [Online]. Available: https://arxiv.org/abs/1904.09675

[214] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance," 2019. [Online]. Available: https://arxiv.org/abs/1909.02622

[215] C. Wei, B. Wang, and C. C. J. Kuo, "Synwmd: Syntax-aware word mover's distance for sentence similarity evaluation," 2022. [Online]. Available: https://arxiv.org/abs/2206.10029

[216] T. Scialom, S. Lamprier, B. Piwowarski, and J. Staiano, "Answers unite! unsupervised metrics for reinforced summarization models," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3246–3256. [Online]. Available: https://aclanthology.org/D19-1320/

[217] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, A. Wang, and P. Gallinari, "QuestEval: Summarization asks for fact-based evaluation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6594–6604. [Online]. Available: https://aclanthology.org/2021.emnlp-main.529/

[218] J. Fan, D. Aumiller, and M. Gertz, "Evaluating factual consistency of texts with semantic role labeling," in *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, A. Palmer and J. Camacho-collados, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 89–100. [Online]. Available: https://aclanthology.org/2023.starsem-1.9/

[219] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 076–12 100. [Online]. Available: https://aclanthology.org/2023.emnlp-main.741/

[220] E. Durmus, H. He, and M. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 5055–5070. [Online]. Available: https://aclanthology.org/2020.acl-main.454/

[221] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 163–177, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.10/

[222] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 9332–9346. [Online]. Available: https://aclanthology.org/2020.emnlp-main.750/

[223] T. Goyal and G. Durrett, "Evaluating factuality in generation with dependency-level entailment," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 3592–3603. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.322/

[224] M. Gao, J. Ruan, R. Sun, X. Yin, S. Yang, and X. Wan, "Human-like summarization evaluation with chatgpt," 2023. [Online]. Available: https://arxiv.org/abs/2304.02554

[225] N. Wu, M. Gong, L. Shou, S. Liang, and D. Jiang, "Large language models are diverse role-players for summarization evaluation," 2023. [Online]. Available: https://arxiv.org/abs/2303.15078

[226] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 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. 46 595–46 623. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

[227] OpenAI, J. Achiam, and S. Adler, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[228] G. Team, R. Anil, and S. Borgeaud, "Gemini: A family of highly capable multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2312.11805

[229] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo, "Prometheus 2: An open source language model specialized in evaluating other language models," 2024. [Online]. Available: https://arxiv.org/abs/2405.01535

[230] G. Mehenni and A. Zouaq, "Ontology-constrained generation of domain-specific clinical summaries," 2024. [Online]. Available: https://arxiv.org/abs/2411.15666

[231] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization," Nov. 2021, arXiv:2111.09525 [cs]. [Online]. Available: http://arxiv.org/abs/2111.09525

[232] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, and R. J. B. Dobson, "Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit," *Artif. Intell. Med.*, vol. 117, p. 102083, Jul. 2021.

[233] 2024. [Online]. Available: https://huggingface.co/instruction-pretrain/medicine-Llama3-8B

[234] 2024. [Online]. Available: https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B

[235] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut, "Meditron-70b: Scaling medical pretraining for large language models," 2023. [Online]. Available: https://arxiv.org/abs/2311.16079

[236] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang, "SNOMED clinical terms: overview of the development process and project status," *Proc AMIA Symp*, pp. 662–666, 2001.

[237] S. Mohan and D. Li, "Medmentions: A large biomedical corpus annotated with umls concepts," 2019. [Online]. Available: https://arxiv.org/abs/1902.09476

[238] F. J. Dorfner, A. Dada, F. Busch, M. R. Makowski, T. Han, D. Truhn, J. Kleesiek, M. Sushil, J. Lammert, L. C. Adams, and K. K. Bressem, "Biomedical large languages models seem not to be superior to generalist models on unseen medical data," 2024. [Online]. Available: https://arxiv.org/abs/2408.13833

[239] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly Available Clinical BERT Embeddings," Jun. 2019, arXiv:1904.03323 [cs]. [Online]. Available: http://arxiv.org/abs/1904.03323

[240] S. Hegselmann, S. Shen, F. Gierse, M. Agrawal, D. Sontag, and X. Jiang, "Medical Expert Annotations of Unsupported Facts in Doctor-Written and LLM-Generated Patient Summaries." [Online]. Available: https://physionet.org/content/ann-pt-summ/

[241] A. Romanov and C. Shivade, "Lessons from natural language inference in the clinical domain." [Online]. Available: http://arxiv.org/abs/1808.06752

[242] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-halt: Medical domain hallucination test for large language models," 2023. [Online]. Available: https://arxiv.org/abs/2307.15343

[243] J. Chen, D. Yang, T. Wu, Y. Jiang, X. Hou, M. Li, S. Wang, D. Xiao, K. Li, and L. Zhang, "Detecting and evaluating medical hallucinations in large vision language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.10185

[244] Q. Yan, X. He, and X. E. Wang, "Med-HVL: Automatic medical domain hallucination evaluation for large vision-language models," in *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024. [Online]. Available: https://openreview.net/forum?id=rxx8leoPy0

[245] V. Agarwal, Y. Jin, M. Chandra, M. D. Choudhury, S. Kumar, and N. Sastry, "Medhalu: Hallucinations in responses to healthcare queries by large language models," 2024. [Online]. Available: https://arxiv.org/abs/2409.19492

[246] C. Herlihy and R. Rudinger, "MedNLI is not immune: Natural language inference artifacts in the clinical domain," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 1020–1027. [Online]. Available: https://aclanthology.org/2021.acl-short.129/

[247] G. Mehenni and A. Zouaq, "Medhal: An evaluation dataset for medical hallucination detection," 2025. [Online]. Available: https://arxiv.org/abs/2504.08596

[248] "Mednote: Augmented clinical notes," https://huggingface.co/datasets/AGBonnet/augmented-clinical-notes/blob/main/report.pdf, accessed: 2024-01-24.

[249] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," 2020. [Online]. Available: https://arxiv.org/abs/2009.13081

[250] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering," 2022. [Online]. Available: https://arxiv.org/abs/2203.14371

[251] V. Gupta, P. Bharti, P. Nokhiz, and H. Karnick, "Sumpubmed: Summarization dataset of pubmed scientific article," in *Proceedings of the 2021 Conference of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, 2021. [Online]. Available: https://vgupta123.github.io/docs/121_paper.pdf

[252] "Llama 3.3 70b vs gpt-4o," https://www.vellum.ai/blog/llama-3-3-70b-vs-gpt-4o, accessed: 2025-03-27.

[253] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040/

[254] S. Hegselmann, S. Z. Shen, F. Gierse, M. Agrawal, D. Sontag, and X. Jiang, "A data-centric approach to generate faithful and high quality patient summaries with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.15422

APPENDIX A TIES AND PARSING ERRORS OF EVALUATIONS WITH PROMETHEUS

Model	Rubric	GS vs OCD	GS vs DBS	OCD vs DBS
Llama-3B-Instruct	Factuality	2493 / 1217	3598 / 1333	2874 / 1167
	Relevance	2534 / 1656	3124 / 1135	2840 / 1638
Llama-8B-Instruct	Factuality	2008 / 5176	4703 / 4183	2228 / 5204
	Relevance	2122 / 1511	5163 / 1564	2378 / 1609
Llama-8B-medicine	Factuality	2813 / 7300	3093 / 6828	3045 / 7282
	Relevance	3019 / 1616	3301 / 1703	3449 / 1644
Llama3-OpenBioLLM	Factuality	4286 / 1222	1711 / 1437	3613 / 1334
	Relevance	3511 / 1856	1603 / 1804	2997 / 1859

Table A.1 Ties and parsing errors of different methods when evaluated with Prometheus (format in a cell is "Ties / Parsing Errors")