Abstract

On Uncertainty Quantification and Bayesian Reasoning in Clinical Applications of Large Language Models

Vimig Socrates

2024

On Uncertainty Quantification and Bayesian Reasoning in Clinical Applications of Large Language Models

A Dissertation
Presented to the Faculty of the Graduate School
Of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

By
Vimig Socrates

Dissertation Directors: Richard Andrew Taylor

December 2024

Who is this really dedicated to?

Acknowledgment

Funding

Contents

Chapter 1

Introduction

1.1 Motivation

Uncertainty is an inherent component in clinical decision making. Despite our best efforts in biology, medicine, and pharmacology, significant gaps in our knowledge of the human body often lead to the adage that medicine is more an art than a science.

Several theories have been put forth regarding the medical reasoning process, including those of **?**] and **?**]. The majority of these theories propose clinical reasoning as a form of abductive, cyclical reasoning, establishing hypotheses given clinical evidence and then evaluating these hypotheses based on further testing. However, all wrestle with the underlying lack of complete information prior to a clinical decision. In this work, we conduct a series of decision-analytic experiments to investigate how the underlying uncertainty of a patient's state interacts with clinical cognition in an state-of-the-art *in-silico* language modeling system: the large language model.

Our experiments initially establish a theory of clinical uncertainty quantification in LLMs, before evaluating diagnostic decision making in increasingly complex scenarios. In Chapter 3, we will evaluate LLM calibration under a complex, real-world clinical decision making task requiring expertise and gestalt. Following an investigation into uncertainty quantification, in Chapter 4 we further explore how our theory of clinical uncertainty influences Bayesian diagnostic reasoning in the LLM. We also see whether racial biases influence this Bayesian reasoning. Finally, in Chapter 5, we extend the initial diagnostic reasoning task to a complete clinical evaluation scenario, framed as a sequential information seeking decision process.

1.2 Paradigms in Clinical Cognition

In order to frame the LLM as a clinical cognition experiment participant, we must first understand prevailing theories in clinical cognition in humans. There are currently two main paradigms that we will describe here.

1.2.1 The Problem-Solving Approach

As a subset of human cognition, clinical cognition has often been investigated through cognitive science methods and principles. The first major investigations into physician information processing using cognitive science methods began with the seminal work of ?] in their series of experiments, entitled the *Medical Inquiry Project* research program. In it, they conduct a series of verbalization studies in which they have physicians evaluate standardized patient actors through a history of present illness and physical exam, the latter of which is conducted through conversations with a medical student acting as a "data bank". They then pause at regular intervals ("between the history and the physical examination" or "at the conclusion of the physical examination before ordering the laboratory tests")[?]. to reflect and summarize their current problem solving processes. Similar work at McMaster University was conducted that instead used a method of "stimulated recall" to encourage reflection by playing back physician-simulated patient interactions. The goal in both programs was to identify a generalized cognitive framework of expert clinical decision making. Given this hypothesis, the authors limited the breadth and depth of their studies significantly. Namely, the number of cases and the number and type of physicians were both limited, under the assumption that a general expert physician (defined not by their relevance to the clinical case at hand, but by peer-rated excellence scores) should be able to demonstrate the general cognitive problem solving process. Perhaps as a result of these limitations, the authors of the Medical Inquiry Project found that there was no objective difference in diagnostic accuracy between experts and non-experts. They further found that expertise was context-dependent and good performance on one case did not predict good performance on another, further suggesting that a single problem solving framework was unlikely and clinical knowledge influenced in expert problem solving.

While they were unable to determine the cognitive framework that differentiated effective diagnosis, they identified that all physicians followed a *Hypothetico-deductive reasoning* process. This cognitive framework described a process by which physicians generated a limited set of hypotheses (4 ± 1) early in the exam and used them to guide further data collection. While the hypothetico-deductive reasoning model has had far-reaching impacts on clinical cognition, diagnostic AI, and medical education to this day, due to the limited number of clinical cases evaluated, it could not be concluded to be the cognitive problem solving process of expert physicians, as physicians that both arrived at the correct diagnosis and those that did not, seemed to employ it. This gap led to a flurry of subsequent work attempting to explain the difference in cognitive processes in clinical problem solving.

One hypothesis (later supported by one of the authors of MIP itself [?]) was that expertise in a given clinical situation played a key role in determining the type of cognitive process employed. For example, [?], among others, find that physicians in a familiar situation (i.e. one where they have strong expertise) do not display explicit hypothesis testing [? ? ?]. They named this process the *forward reasoning* process, compared to the *backward reasoning* of the hypothetico-deductive framework. They further identified that subexpert cardiologists that arrived at the incorrect diagnosis performed a hybrid process of forward and backward reasoning, rather than simply forward reasoning. They conclude that experts begin their reasoning process in their knowledge base, rather than by developing hypotheses. Further evaluation of the knowledge structure itself determined that expert diagnosticians have learned a diverse and abstracted network of semantic relations between clinical findings and potential diagnoses [? ?]. However, forward, intuitive, and rapid reasoning is not the only necessary reasoning strategy. In fact, later, ?] found that forward reasoning breaks down under complexity and uncertainty. The relative and joint benefits of both the hypothesis testing and data-driven, forward reasoning would lead some to advocate for a cognitive dual reasoning strategy model [?]. A decision making theory borne from psychology discussed in the next section seems to bridge the gap.

1.2.2 Dual Process Theory: A Theory of Decision-Making

The prevailing theory of general cognitive decision making today is known as *dual process theory*, originally proposed by ?] as a "two-process theory of human information processing". It has since gone through several iterations, being popularized in cognitive psychology by the work of Hammond and colleagues [? ?], and in the general public by Kahneman [?]. The two systems have since been dubbed System 1 and System 2, to be described later.

However, the origin of these cognitive systems comes from early decision making literature relying on normative models of decision making. The decision-making approach to cognition forgoes protocol methods eliciting cognitive processes from experts and assumes a normative model of decision making that expert and subexpert physicians should adhere to, often based on a Bayesian or utility-based model. The first work in this space established a two stage clinical reasoning model based on aligning symptom complexes with disease complexes. In the first stage, a set of potential diagnoses are identified based on patient symptoms. In the second, a decision-analytic approach is taken to determine the conditional likelihood of a disease, given a symptom complex and lab tests. This process is iterated upon until the final diagnosis is reached [?]. However, work in cognitive science led to the realization that more could be learned from investigating the cognitive processes of experts, rather than prescribing an optimal decision making process through some optimization function [?], as there are a number of other considerations that a rational choice theory does not take into account. Thus began the aforementioned problem-solving research paradigm.

Despite this initial reticence, there was some limited work on integrating a probabilistic approach with clinical cognition [?]. Some relied on Bayes' theorem to develop a framework of updating disease estimates following the acquisition of new information [? ? ?]. We leverage this paradigm in Sections 4 and 5. Much of the work in System 1 and 2 thinking arises from attempting to understand predictable errors in probabilistic judgement [?] which in the clinical space has been called *cognitive dispositions to respond* [?]. These led to a set of *heuristics* that simplify decision making under uncertainty [?]. The existence of such heuristics in the clinical domain was quickly theorized [?]. However, heuristics are prone to biases that may lead to the incorrect conclusion, including anchoring [?], confirmation bias [?], and framing effects [?]. Therefore, it was theorized that two different cognitive processes are employed interchangeably based on the cognitive task at hand. Namely, to use Kahneman's parlance, System 1 and System 2 Thinking, as described in further detail below:

1. **System 1:** A cognitive system described as "intuitive" and "experiential". It is largely a automatic, reflexive system used to make decisions rapidly with minimal cognitive load. It uses information immediately available, while sometimes limited, to make a holistic judgement. These decisions are often governed by habit and are difficult to control or intervene upon.

2. **System 2:** A more deliberate cognitive system, used for more methodical and "rational" or "analytical" decision making. This system utilizes additional information actively collected by the individual from their environment. Decisions are made by applying a set of learned rules consciously. Decision making is also actively monitored and metacognitive steps such as doubt are governed by System 2 [? ? ?].

This new cognitive decision making paradigm was quickly adopted by clinicians and comfortably reconciles ?]'s work on data-driven, forward reasoning and the original hypothetico-deductive framework. In the clinical domain, System 1 is often referred to as clinical gestalt or a "gut feeling" and is often dependent on the pathognomicity of the presentation [?]. Recognizing patterns of clinical features and diagnoses is a function of System 1 thinking [? ?]. Conversely, System 2 thinking in the clinical domain is characterized by a slower, more taxing, analytical decision making process. This may be triggered in a physician when the presentation is unrecognized or ambiguous. System 2 can also override an initial System 1 judgement if there are clinical features that do not indicate an initial, reflexive diagnosis.

The decision on which System to use however, is interesting and modulated by a variety of criteria. Recent work shows that System 2 thinking subjugates System 1 processing when the situation is complex or ill-defined, the stakes are high, or in the context of uncertainty [? ?]. In this thesis, we are chiefly concerned with uncertainty. We believe that in such a setting, System 2 thinking should be employed and therefore, we consider a normative, decision-making approach leaning on probabilistic reasoning. We encourage the LLM to reason deliberately and analytically given the ambiguity of the clinical contexts we present it with. In addition to this theoretical frame, there are a number of practical reasons for choosing the decision-making paradigm, rather than the problem-solving one described earlier. We will enumerate these below.

1.2.3 Benefits of the Normative Approach

Both the problem-solving and decision-making cognitive theories offer productive lenses through which to view a clinical cognitive study of large language models. As mentioned, both are rooted in rigorous cognitive research that demonstrate their ability to model physician diagnostic decision making in a variety of clinical contexts. Furthermore, both have been used as inspiration for clinical AI systems, and therefore demonstrate their usefulness in developing computational models of cognition [? ?]. Lastly, we could employ both frames to design a set of experiments to evaluate the LLM as a subject in a clinical cognitive psychology experiment [?]. For instance, we could present cases to the LLM, as has been done in the past and evaluate its error modes [?], assuming that it has some clinical expertise and describe its cognitive processing as one would in the problem-solving paradigm. Conversely, we could present these same cases to the LLM and assume a normative approach that should be taken through a decision-making frame. Here, we will argue why we take this latter approach.

As the LLM will ultimately be used as a collaborative tool in clinical decision making, it is important to validate its accuracy, as well as its cognitive processes. Using a descriptive, information-processing approach, we are limited to describing and classifying its cognitive process, but aren't able to ground its responses in an objective answer. Using our normative, decision-analytic frame, we are able to evaluate LLM reasoning against a ground truth. Several normative functions have been proposed in the decision making literature, including expected utility [?] and Bayesian models [?]. We elect to use a Bayesian frame, as we are chiefly interested in patient outcomes, not cost-effectiveness or other metrics of optimal decision making strategy, as is common to utility functions.

However, in our consideration of Bayesian decision making, we must reconcile several well founded critiques of the decision-making approach by it's descriptive opponents. The first of which is well-founded that humans are not Bayesian thinkers [?], and therefore we cannot use a Bayesian grounding for the normative state. While this may be true for humans, we argue that it isn't true for LLMs, as demonstrated in our work in Section 4 as well as by others [?]. We see that LLMs are capable of engaging in reasoning that aligns explicitly with Bayesian principles. However, whether a Bayesian function provides the *optimal* decision function in a clinical scenario is more difficult to examine. In this work, we select Bayesian information utility as a reasonable choice, but it could be easily replaced by others in our evaluation framework. We also agree with proponents of the information-processing approach who maintain that the biases or "fallacies" identified by normative approaches may in fact be a function of irrational, but otherwise meaningful, components of clinical care (e.g. patient requests, cost considerations, etc.). However, these heuristics must be developed as a function of expertise [?]. Namely, expert physicians working in their area of expertise, often employ heuristics such as confirmation bias successfully, as compared to subexperts [?]. We view the LLM as a subexpert clinical decision maker for several reasons. As shown in a variety of work [?], while state-of-the-art LLMs perform well on some tasks, they are still far from expert clinical reasoners [?]. Therefore, we believe that we can use a more decision-analytic framework to evaluate them, rather than assume that their biases are a function of expertise, and therefore valid decision making practice.

Finally, while we are largely leveraging a normative evaluation framework, we are not solely concerned about the final diagnostic decision. As ?] puts it, a key difference between the two cognitive paradigms is that "Problem-solving research emphasizes the sequential process of searching for a solution path, whereas decision research focuses more on the nature of the decision outcome and how it may deviate from an acceptable normative standard." However, in Section 5, we forgo this assumption of decision-making theory and evaluate sequential Bayesian decision making. We will leverage the LLM's unique ability to explicate its reasoning as it performs decision-making to both evaluate the decision-making process, as well as the final decision itself. Given the aforementioned advantages, we believe that leveraging a decision-making cognitive framework with a Bayesian utility function will yield unique insight into an LLM's cognitive process.

1.3 LLMs as Candidate Models of Clinical Sublanguage Processing

In the previous section (1.2), we discussed the cognitive decision making frame that we would be borrowing from in our clinical cognitive evaluation of LLMs. As our chief modality of study is language, a product of large language models, in this section, we will argue why this product is worthy of study, how it relates to the clinical sublanguage, and why LLMs potentially present a unique linguistic theory of the clinical sublanguage. To begin, we provide a brief overview of Zellig Harris' theory of sublanguages. For our purposes, this will provide background for a view of the clinical sublanguage, so far as it has been constructed. Moreover, it will allow us to view the large language model as a novel, generative theory on the clinical sublanguage. This motivates our use of it in a cognitive study of uncertainty.

1.3.1 Harris' Sublanguage Theory

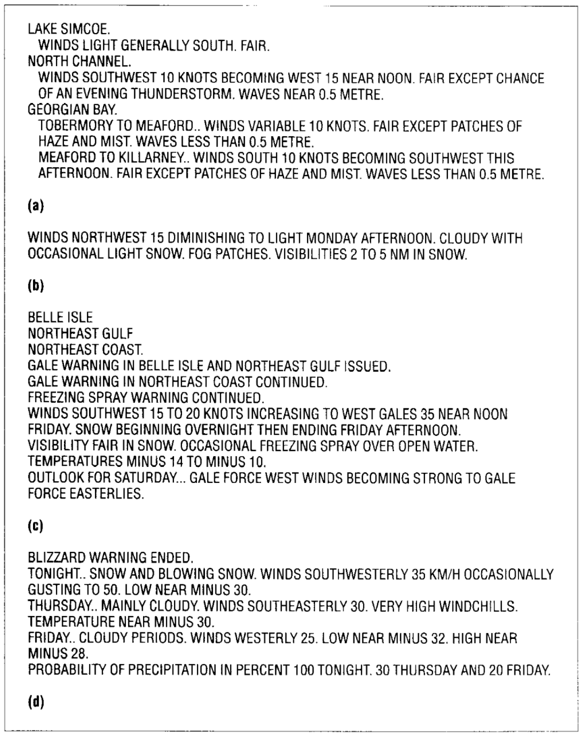
Zellig Harris (1909-1992) was an instrumental linguist, semiotician, and "methodologist" [?] that significantly influenced 20th century American linguistics. His *distributional hypothesis*, despite being misconstrued, has largely motivated much of the past few decades of research in computational linguistics. Namely, embedding models from word2vec through present-day LLMs are based on an idea of distributional semantics that was first derived from Harris' work. As we will discuss later, the original distributional hypothesis was pertaining co-occurrences of word classes and their relative positions to one another in sentence structures. It was therefore a method of formal characterization of the structure of language. Harris' explicitly states that there is no one-to-one map between the distribution of word classes and meaning, a point that has been largely ignored by current NLP researchers.

Harris made another significant contribution to formal semantic theory with his work on sublanguages. Harris defines a sublanguage as "subsets of sentences that are closed under certain operators in a language" [?]. Put simply, a grammatical operation carried out on any sentence in the sublanguage yields another sentence in the sublanguage. Sublanguages have certain properties that distinguish them from other constructs of language variation, namely, registers and dialects. Registers and dialects are both variations of language restriction that arises from different sociolinguistic contexts and use cases. Specifically, both have either specific operators or language features that are either included or excluded from the general language. Registers in particular are used for a particular function, such as newspaper headlines and recipes [?]. Despite their similarity in being a restricted subset of a language, they differ significantly from sublanguages. Mathematically, a sublanguage can be defined by the existence of a metalanguage that describes the sublanguage grammar. The sublanguage grammar can be described by a set of word-classes and the grammatical operators that exist between them. We can make this definition more tangible through an examination of Harris' seminal work on the immunology sublanguage, which just happens to relate somewhat to the clinical sublanguage central to our work.

Harris used his theory of sublanguage to investigate a particular subfield of early immunology: the search for the cell that produces antibodies conducted from 1940-65. Together with colleagues, he lays out a complete grammar describing this corpus by collecting words into classes by co-occurrence and defining a closed set of world-class sequences (sentence structures). Instead, we will just describe a few key structures that shed light on the way that sublanguages are defined. The first sentence structure Harris defines is using 3 word classes. Namely, he identifies class **A** as a subject of *found in the lymph nodes after injection of an antigen*, a verb of class **V** that consists of *is found in, is contained in, is produced by*, and an object class **T** consisting of nouns such as *lymph nodes, lymph, serum.* These create a sentence structure of **AVT** to create sentences such as *Antibodies are produced by lymph nodes* or transformed as *Lymph nodes produce antibodies* [?]. Harris further argued that precisely defining a sublanguage grammar is one way of converting language into information. Namely, he describes methods of weighing the relative evidence of claims in various papers in the corpus through the occurrence of certain sentence structures. This second-order analysis is outside the scope of this thesis, except to point out that we can derive additional information by describing a subfield through a sublanguage.

1.3.2 What makes a complete sublanguage?

Despite the considerable work put into the definition of the immunological sublanguage by Harris, he notes that further articles need to be analyzed in order to describe



LAKE SIMCOE.
WINDS LIGHT GENERALLY SOUTH. FAIR.
NORTH CHANNEL.
WINDS SOUTHWEST 10 KNOTS BECOMING WEST 15 NEAR NOON. FAIR EXCEPT CHANCE OF AN EVENING THUNDERSTORM. WAVES NEAR 0.5 METRE.
GEORGIAN BAY.
TOBERMORY TO MEAFORD.. WINDS VARIABLE 10 KNOTS. FAIR EXCEPT PATCHES OF HAZE AND MIST. WAVES LESS THAN 0.5 METRE.
MEAFORD TO KILLARNEY.. WINDS SOUTH 10 KNOTS BECOMING SOUTHWEST THIS AFTERNOON. FAIR EXCEPT PATCHES OF HAZE AND MIST. WAVES LESS THAN 0.5 METRE.

(a)

WINDS NORTHWEST 15 DIMINISHING TO LIGHT MONDAY AFTERNOON. CLOUDY WITH OCCASIONAL LIGHT SNOW. FOG PATCHES. VISIBILITIES 2 TO 5 NM IN SNOW.

(b)

BELLE ISLE
NORTHEAST GULF
NORTHEAST COAST.
GALE WARNING IN BELLE ISLE AND NORTHEAST GULF ISSUED.
GALE WARNING IN NORTHEAST COAST CONTINUED.
FREEZING SPRAY WARNING CONTINUED.
WINDS SOUTHWEST 15 TO 20 KNOTS INCREASING TO WEST GALES 35 NEAR NOON FRIDAY. SNOW BEGINNING OVERNIGHT THEN ENDING FRIDAY AFTERNOON.
VISIBILITY FAIR IN SNOW. OCCASIONAL FREEZING SPRAY OVER OPEN WATER.
TEMPERATURES MINUS 14 TO MINUS 10.
OUTLOOK FOR SATURDAY... GALE FORCE WEST WINDS BECOMING STRONG TO GALE FORCE EASTERLIES.

(c)

BLIZZARD WARNING ENDED.
TONIGHT.. SNOW AND BLOWING SNOW. WINDS SOUTHWESTERLY 35 KM/H OCCASIONALLY GUSTING TO 50. LOW NEAR MINUS 30.
THURSDAY.. MAINLY CLOUDY. WINDS SOUTHEASTERLY 30. VERY HIGH WINDCHILLS.
TEMPERATURE NEAR MINUS 30.
FRIDAY.. CLOUDY PERIODS. WINDS WESTERLY 25. LOW NEAR MINUS 32. HIGH NEAR MINUS 28.
PROBABILITY OF PRECIPITATION IN PERCENT 100 TONIGHT. 30 THURSDAY AND 20 FRIDAY.

(d)

Figure 1.1: Forecast text messages by region: (a) Great lakes marine; (b) Arctic marine; (c) Atlantic marine; (d) public *reproduced from [?]*

certain meta-science segments and grammatical structures such as quantifiers and sentences with conjunctions. We claim that he clinical sublanguage is poorly defined and fuzzy. In order to validate this claim, we first describe a complete and well-formed sublanguage used in one of the first natural language generation systems: the forecast generator (FoG), used to generate computer-readable forecasts from short-form narratives describing the results of weather simulations in both English and French [?]. The reports were disseminated as text messages and following a distinctive telegraphic style sublanguage, as shown in Figure 1.1. An initial distributional analysis of the million-word corpus led to the definition of a Backus Naur Form (BNF) grammar and distribution of a lexicon and syntactic structures found in forecasts. Combined with a state-of-the-art generative theory at the time (based on Meaning-Text Theory from Igor Mel'čuk and Aleksandr Zholkovsky [?]), they were able to deploy their NLG algorithm throughout the Canadian weather service across several forecast types. They found that adding new forecast types within the sublanguage required very little change to the sublanguage, reasonably validating their generated grammar. Further evidence came from the adaptation of the grammar to neatly accommodate language "drift" as forecaster language evolved [?]. Their sublanguage grammar was largely robust against modification despite new documents that should still exist within the sublanguage, thereby demonstrating a closed grammar.

1.3.3 The limitations of the current clinical sublanguage

Having seen the robust grammar of the FoG system, we can contrast it against that of the clinical sublanguage as defined by ?]. The authors explicitly state that this sublanguage is developed for patient reports. The work conducted on the clinical sublanguage was based on two prior NLP systems: the Linguistic String Project [?] and the MedLEE system [?], both of which used a constituent grammar formalism compared to the operator-argument formalism proposed in Harris' work. The MedLEE grammar, initially developed for the chest radiology report sublanguage, has been extended to include discharge summaries, pathology reports [?], and nursing narratives [?]. Despite its impressive versatility with limited changes to the underlying grammar, both the authors [?] and others [?] have found that its parsing was not perfect. Namely, MedLEE incorrectly categorized word classes, such as misunderstanding T1 and T2 as the 1st and 2nd thoracic vertebrae instead of T1 and T2 relaxography (MRI protocols). It also struggled with multiple modifiers of body locations (*upper inner quadrant of the left breast*), despite them being added to the vocabulary, indicating that the underlying grammar was not complete [? ?]. However, the main limitation of the MedLEE system comes from its sentence-by-sentence processing, leading to poor anaphora resolution. While sublanguage grammars conventionally consist of word classes and sentence-structures, in discourse settings, they result in discourse word classes and between-sentence structures. Such a construction would be necessary to effectively resolve the grammar proposed in MedLEE. However, the inclusion of discourse parsing would significantly complicate the grammar and could potentially make its ruleset excessively complex, though this remains to be tested. Based on this analysis of the current state of the clinical sublanguage, we conclude that further work needs to be done in establishing the grammatical rules of such a sublanguage. To that end, we propose that current distributional semantic models can act as such a bridge.

1.3.4 Introduction to Distributional Semantic Models

Before we can describe how an LLM might act as a candidate model for clinical sublanguage processing, an introduction to the transition from Harris' original sublanguage theory to distributional semantic language models is necessary. The underlying hypothesis in Harris' distributional theory was that we can describe a language as the co-occurrences of its parts in relation to one another, irrespective of history or meaning, and words can be fully specified by the sum of all the environments that they appear in [?]. As discussed, he explicitly denies the claim that there is a one-to-one relationship between these occurrences and meaning. However, in practice, this simplification has proved useful, often described succinctly as "a word is characterized by the company it keeps" [?]. This adage has led to distributional semantic models that lay the foundation for the modern day Large Language Model (LLM). Given that LLMs are neural network-based models, we will mostly focus on models following the development of dense, continuous word representations such as word2vec. Prior to these models, dimensionality reduction [? ?] or matrix decomposition [?] methods over word co-occurrence matrices were common.

Instead of first learning computationally-expensive co-occurrence matrices, word2vec models directly draw from Harris' theory by using learning word representations using a context window of surrounding words. The continuous bag-of-words (CBOW) model is trained using a "fill-in-the-blank" approach, where a fixed length representation of a word, known as its embedding, captures the likelihood of other words appearing within the context window, as shown in Figure 1.2. Words that have similar meanings are expected to impact these probabilities in similar ways, since they tend to occur in similar contexts. Similarly, the continuous skip-gram model trains word embeddings by predicting the words in the context window from a single word. Both models were trained with classic neural network loss functions and optimizers, namely cross-entropy loss and stochastic gradient descent, respectively. ?] found that the CBOW model trains faster while the skip-gram performs better with infrequent words.

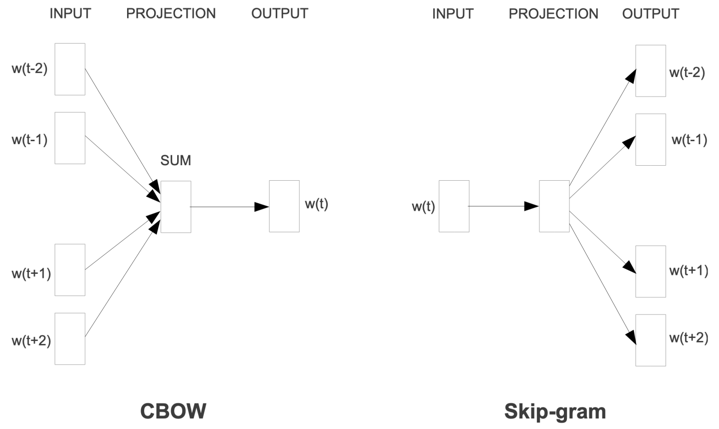


Figure 1.2: Architectures of the word2vec models *reproduced from [?]*

Given the sequential nature of human reading comprehension, several model architectures were developed that employed neural networks to model text sequences (i.e. RNNs, LSTMs) [?]. However, their slow training times led to the development of more parallelizable architectures. Hence, in 2017, the transformer was developed in the seminal paper entitled "Attention is all you Need" [?]. This model has two main features that differentiate it from all prior neural network architectures: positional embeddings and attention. As shown in Figure 1.3A, attention allows the transformer to learn based on the entire sentence, rather than just the preceding text, like in RNNs.

Eventually, the general-purpose transformer architecture was extended to explicitly learn distributional semantics through BERT [?] and GPT [?]. BERT models only used the encoder component of the transformer architecture, while GPT models only used the decoder component. Because sequence is not explicitly encoded within

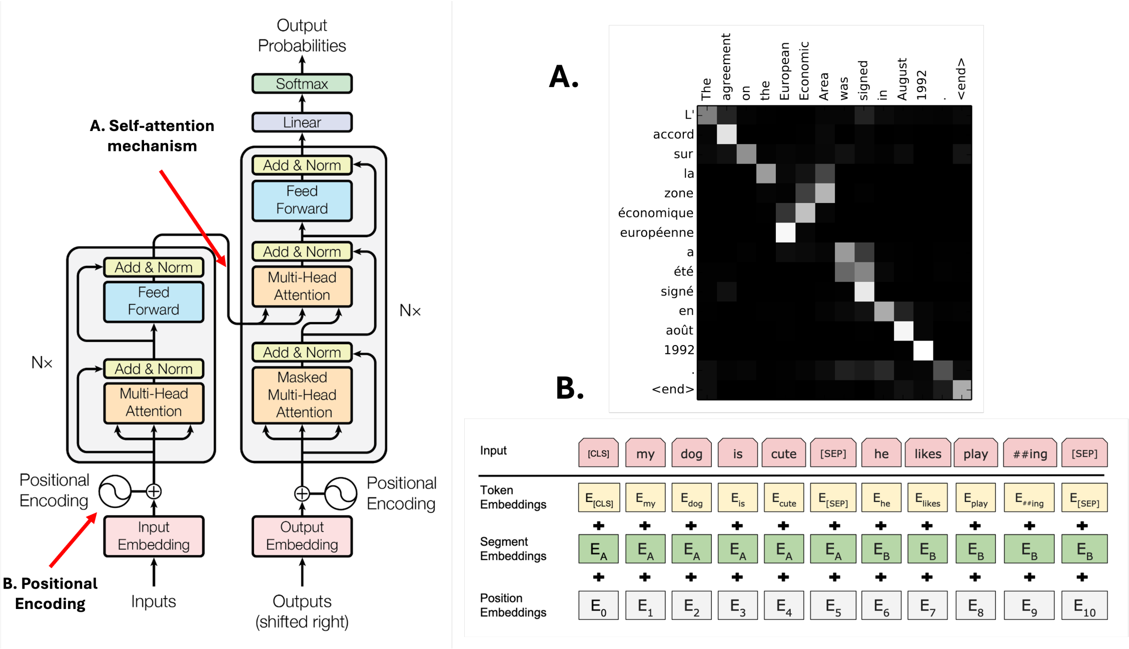


Figure 1.3: Transformer architecture with examples of (A) self-attention and (B) position embeddings *Repurposed from [? ? ?]*

transformers, it must be introduced elsewhere to ensure that the order of words is remembered. This is done with position encoding as in Figure 1.3B, in which a position embedding, along with the segment and token embeddings, define the word embedding within transformer-based models. One final improvement from the original word2vec models was developed to resolve their inability to handle out-of-vocabulary (OOV) words. As word2vec models used human-readable words, scaling them to human language would require an intractable number of words. Instead, tokenization schemes at the subword (similar to morphemes, but statistically driven) level, such as byte-pair encoding were developed that naturally compromised some expressivity for full coverage [?]. Interestingly, BERT is trained on a similar training objective as the CBOW model, known as masked language modeling, where a word is hidden and the model is trained to reproduce it. Unlike BERT, GPT models are trained on next word prediction, similar to the original RNN-based models. These models revolutionized modern-day NLP and distributional semantics models.

These initial models had an incredible number of trainable parameters ($BERT_{Base}$: 110M, $BERT_{Large}$: 340M). However, it was soon found that performance and expressiveness increase with scale, and as a function of scale, compute [? ?], leading to current sizes of 70B parameters in open models. They're huge sizes have earned then the name Large Language Models (LLMs). In addition to the increased performance, these models have also progressively increased their context windows, now able to handle up to 128K tokens, allowing them to take even larger sections of text into context in both training word embeddings and in prediction [?]. We believe that this unique set of characteristics allow the current state-of-the-art LLMs to model sublanguage grammars effectively.

1.3.5 Distributional semantic models as fuzzy sub-language grammars

The main limitation that we've seen thus far from the existing clinical sublanguage grammar is its inability to handle anaphora detection due to single-sentence structures. However, the aforementioned large context windows and self-attention mechanism of modern large language models enable them to identify between-sentence relationships. However, these models come with their own set of limitations, leading to a fuzzy sublanguage grammar. As we have discussed, a sublanguage grammar requires two main components (in addition to abiding by the rules of the larger English grammar): word classes, and relationships between word classes as sentence structures. However, distributional semantic models such as large language models do not operate on the word level to begin with. Instead, as we discussed, most use subword encoding schemes such as byte-pair encoding. Therefore, unless they are able to demonstrate co-occurrence recognition at the word level, they cannot be considered sublanguage grammars. This also has consequences for their efficacy in word-level downstream tasks. Therefore, a line of work has investigated if more knowledge-based, clinically meaningful tokenization schemes would significantly improve results. For instance, **?**] find that pretraining BERT-based models on segmented morpheme data cross-referenced with concepts in UMLS does not significantly improve either the pretraining performance of masked language modeling (MLM) nor does it improve results on word-level downstream tasks, such as entity linking and named entity recognition (NER). Additional evidence of subword tokenizers' ability to parse clinical language at the word level comes from French, where only 25% of downstream tasks including NER and part-of-speech tagging were improved by clinical morpheme-enriched tokenizers **?**]. However, some improvements have been seen **?**], leading to the conclusion that current subword tokenization schemes enable large language models to estimate semantic word-classes from subwords, although explicit morpheme-based tokenization would explicitly model the semantics of the clinical sublanguage better at the cost of less lexical coverage. The absence of the word as the most reduced semantic unit in large language models introduces the first element of fuzziness when compared to a strict sublanguage grammar.

In addition to the lack of explicit clinical morphemes in large language models, the word classes themselves, as well as the sentence structures are both also not explicit, leading to a further relaxation of Harris' sublanguage grammar requirements. While large language models do not explicitly model semantic word-classes in a sublanguage grammar, they can be considered leaky abstraction based on co-occurrence matrices. The next-word prediction and MLM pretraining tasks introduce an inductive bias for generating contextual word representations, including varying word senses across different contexts [**?**]. Notably, sublanguage grammar word classes are often defined by co-occurrences, making this a reasonable model for capturing features of the sublanguage. Another way of validating the claim of strong biases towards word classes in a sublanguage is by looking for irrelevant information. In the fact verification literature, **?**] find that instruction-tuned models, such as InstructGPT and ChatGPT, rarely ($\leq$ 1.8%) generate information irrelevant to the prompt. This is perhaps a function of their RLHF fine-tuning, in which such generation would be heavily nonhumanistic and therefore penalized. For our purposes though, while it is possible for them to generate words outside of the semantic classes in the sublanguage, it is highly unlikely.

The last element of fuzziness in LLMs that may break a formal sublanguage grammar definition would be their lack of explicit accepted sentence structures. While sublanguage grammars are rigid and comprehensive, their parsing can introduce ambiguity. For instance, garden path sentences (a classic example is *Time flies like an arrow; fruit flies like a banana*) are sentences where the expected parse does not align with the syntactically correct parse tree. Another source of ambiguity is abbreviation disambiguation. One can imagine the following sentence within the clinical sublanguage of the sentence structure $Finding + v_{show} + change$:

> *Patient's LFTs are trending downwards, which could suggest improvement; will continue monitoring to confirm stabilization.*

Without additional context, *LFT* could be confused for either lung function test or liver function test, significantly differentiating the underlying pathology. Both would fit within the aforementioned sublanguage sentence structure as a *Finding* but have different semantic meaning. Given the already ambiguous nature of parsing sublanguage grammars, large language models introduce additional ambiguity by not formally modeling these grammars. However, work in machine translation to code has shown that LLMs are also capable of translating natural language to domain-specific languages (DSLs) [**?** **?**]. Therefore, we are confident that these models can effectively capture contextual word-class relationships within sublanguage grammars, even without the grammar being explicitly encoded. Having demonstrated that word classes and syntactic sentence structures can be represented, albeit not formally, by large language models, we are reasonably confident that state-of-the-art models can generate text proficiently within the clinical sublanguage. However, in this thesis, our goal is not only to generate text in the clinical sublanguage but also to investigate how cognitive reasoning aligns between humans and LLMs under conditions of uncertainty in decision-making. Therefore, we next discuss how a normative cognitive framework, combined with the contextual biases of LLMs, can be used to investigate uncertainty-driven clinical reasoning as it relates to physicians.

1.4 An LLM as a Participant in a Cognitive Clinical Study of Uncertainty

So far, we have argued that the normative approach affords us certain benefits when using the LLM as a cognitive study participant, such as viewing the LLM as a subexpert physician and evaluating the LLM under a standardized ground truth. We have also validated that the LLM is able to generate text within the clinical sublanguage. Therefore, we must motivate the purpose of investigating its ability to align to a normative approach to diagnostic reasoning through the clinical sublanguage. We argue that there are both technical and sociotechnical reasons for doing so.

Firstly, as discussed, uncertainty is an inherent part of the clinical decision making process, particularly in diagnosis. In a real-world clinical context, it is impossible to have 100% certainty before making a diagnosis and further management. Instead, the art of medicine is to weigh the relative risks while gathering more information. As it is unlikely that LLMs will be used with no oversight, they will often function as triggers for further manual review. As such, the weighing of relative risks by the LLM must be better characterized. As we will see in Chapter 2, while LLMs have been demonstrated to quantify uncertainty effectively in medical challenge questions (i.e. USMLE question-answering), these are not representative of real-world tasks that physicians and AI will collaborate on. Therefore, our first experiment in Ch. 3 is designed around such a task in which the benefit of discontinuing medications is weighed against the potential cost of adverse withdrawal reactions or return of a medical condition [**?**].

Secondly, it has been shown that LLMs perpetuate racial and gender biases in a variety of generative language tasks, including medical education, differential diagnosis and medical plan generation [**?**], while having been instruction-tuned to limit such biases. Therefore, it is clear that while they will not explicitly respond in a sociodemographically biased manner, they may harbor implicit biases as a function of their training data and procedure. However, this underlying bias has not been evaluated in the estimation of risk. Therefore, in Chapter 4, we investigate the potential for racial bias in a Bayesian clinical diagnostic frame. This will allow us to determine if LLMs estimate risk variably, in an unexpected way. In particular, we look at estimates of sensitivity/specificity and likelihood ratios. We hope to learn contexts in which LLMs' risk evaluations can be trusted in cases where it cannot, allowing for both better understanding of current model capabilities and providing a path for future standardization.

Thirdly, although much of this risk-weighing may not be included in the clinical note (aside from more recent medical-decision making (MDM) sections and medicolegal clarifications), it is at least evidenced by the decisions that the physician makes. Similarly, while the LLM may "explain" its reasoning, it has been shown that they are surprisingly brittle following their own explanations in subsequent reasoning steps [**?**]. As LLMs are used for progressively more complicated tasks such as planning, it would be helpful to determine what influences an LLM's sequence of clinical decisions. Our decision process of choice is diagnosis of chest pain in the Emergency Department. We elect to use the normative approach from cognitive science to establish the "optimal" decision making process and compare the LLM to that. In future work, we hope to conduct a decision analysis of expert physicians to compare against, instead of the simulation study proposed in Chapter 5. Such fine-grained evaluations will enable greater confidence in LLMs and identify differences in physician and LLM behavior. Knowing where they differ will allow us to better model physician decision making in LLMs and propose more standardized decision making processes in physicians.

Across these three experiments, we hope to rigorously test LLMs' decision-making capabilities under uncertain clinical situations both as a mechanism of risk quantification as well as its impact on diagnostic narrowing. In the coming Chapter, we will describe the associated progress on uncertainty quantification and Bayesian reasoning in LLMs and how our approach contributes to this literature.

Chapter 2

Related Work

Having introduced the historical underpinnings of our work and placed our experiments within context, we turn to a survey of the more recent literature, in an effort to demonstrate the research gaps that currently exist and how our work fills them. In particular, I will review current literature on uncertainty quantification and Bayesian reasoning in Large Language Models, notably focusing on the clinical domain and closed LLMs.

2.1 Definitions of Uncertainty in Machine Learning

An adequate model of uncertainty is vital in healthcare applications. Despite advancements in explainable AI, LLMs are inherently black-box models, as their underlying decision-making process is largely nonlinear and complex. Therefore, it is important for models to have a consistent and accurate estimation of their own uncertainty. Due to the relatively recent emergence of the field of uncertainty quantification (UQ), it remains somewhat unstructured, with key definitions of uncertainty still subject to ongoing debate. Here, we adopt a definition of uncertainty from a recent review that addresses uncertainty at various stages of the neural network training/prediction pipeline. We will also distinguish the two main types of uncertainty: aleatoric and epistemic [**?**].

We first define a neural network as a non-linear function f_θ parameterized by network weights θ that maps a measureable input set $\mathbb{X}$ to a measureable set $\mathbb{Y}$ as such:

$$f_\theta : \mathbb{X} \to \mathbb{Y} \qquad f_\theta(x) = y \tag{2.1}$$

We can further define a training dataset in the supervised learning case:

$$\mathcal{D} = (\mathcal{X}, \mathcal{Y}) = \{x_n, y_n\}_{n=1}^{N} \subseteq \mathbb{D} \tag{2.2}$$

where $\mathcal{D} \subseteq \mathbb{D} = \mathcal{X} \times \mathcal{Y}$ and N is the number of samples in the training set. A new data sample $x^* \in \mathbb{X}$, a neural network on $\mathcal{D}$ can be used to predict a corresponding target $f_\theta(x^*) = y^*$. This can also be extended to zero-shot evaluation as the neural network (a black-box LLM in our case) is trained on $\mathcal{D}$ as the set of text in the English language. Given this definition, there are 4 places uncertainty can arise from: (1) the *data acquisition* process, (2) the *neural network building* process, (3) the *applied inference* model, (4) the *prediction's uncertainty* model. Uncertainties associated with (3) and (4) are differentiated by out-of-domain errors vs. errors caused by in-domain input data x^* and the model f_θ [**?**].

We are mainly interested in the *prediction's uncertainty* as we cannot control the other sources of uncertainty in a black box model and we mainly would like to evaluate uncertainty as a function of prediction to promote physician-AI collaboration. Therefore, we are mostly interested in the uncertainty of a prediction y^*. The Bayesian framework offers a solid basis for analyzing this uncertainty. We can consider the probability distribution for a prediction y^*, based on a sample input x^* as:

$$p(y^*|x^*) = \int_D p(y^*|\mathcal{D}, x^*). \tag{2.3}$$

We can also define a maximum a posteriori (MAP) estimation over the distribution of y^* by:

$$y^* = arg \ \max_y \ p(y|x^*). \tag{2.4}$$

Unfortunately, the distribution in (2.3) can only be approximated by the data given in D. We can decompose the probability of any particular y^* as the two contributing types of uncertainty:

$$p(y^*|\mathcal{D}, x^*) = \int_D \underbrace{p(y^*|x^*, \theta)}_{Data} \ \underbrace{p(\theta|D)}_{Model} \, d\theta. \tag{2.5}$$

In (2.5), $p(y^*|x^*, \theta)$ term describes the *data* or *aleatoric* uncertainty, which is caused by a loss of information when mapping from some real-world event to an input sample. For instance, in prediction of antibiotics prescription for a UTI diagnosis from urinalysis reports, this could arise from a physician that is running late to see the next patient and quickly documents their interpretation for the test, but does not describe the decision-making for prescribing antibiotics. This form input data uncertainty can generally not be reduced. Conversely, $p(\theta|D)$ describes *model* or *epistemic* uncertainty, which arises from shortcomings of the model, poor training procedures, or bad coverage over the training domain. This type of uncertainty is the one we will focus on and expect our models to estimate.

A closely related concept to UQ is *calibration*, which measures how effectively we quantify uncertainty. Informally, an estimated confidence $\hat{P}$ would be well-calibrated if it represented the true probability. For instance, for 100 predictions, each with confidence 0.7, we would expect 70 of them to be correctly classified. Mathematically, we can defined this as:

$$\forall p \in [0,1]: \quad \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{y_{i,k} \cdot \mathbb{I} f_\theta(x_i)_k = p}{\mathbb{I} f_\theta(x_i)_k} \xrightarrow{N \to \infty} p. \tag{2.6}$$

In the above equation, $\mathbb{I}\cdot$ is the indicator function that is 1 if the condition is true, and 0 otherwise, and $y_{i,k}$ is the kth entry in the one-hot encoded ground truth vector for training sample (x_i, y_i) [? ?]. Put simply, the fraction of cases where a prediction equals a class over all classes (i.e. the confidence) should be equal to the true probability. This map must be a limit as the true probability p is a continuous random variable, so cannot be estimated with measurable input set $\mathcal{X}$. Having formally defined uncertainty and calibration, we can move on to describe the recent work done in UQ and calibration of neural networks and transformers.

2.2 Uncertainty Quantification in the General Domain

Much of the work in UQ in conventional neural networks has focused on calibration, as it has an intuitive definition and can be easily quantified. Therefore, we will largely focus on calibration in this section. However, calibration can be thought of as a quantifiable prerequisite for downstream use cases of UQ, and so we will also introduce relevant downstream tasks such as selective prediction.

2.2.1 Calibration in Larger Neural Networks

The first work on calibrating neural networks began before deep learning models were widely used. ?] find that simple neural networks predicting binary classes are more well-calibrated than other common ML methods of the time, including boosted trees and SVMs. They mostly showed performance on non-clinical data, although they did include two medical datasets including the MEDIS dataset [?] used for pneumonia prediction. Despite this initial demonstration, later work found that more modern neural networks were no longer well calibrated. The authors demonstrated this primarily on computer vision tasks, but also used the state-of-the-art NLP models of the time, namely, Deep Averaging Networks (DANs) and TreeLSTMs, on classic news datasets such as *20 News*, *Reuters*, and the Stanford Sentiment Treebank [?]. The authors identify some factors about modern neural networks that contribute to poor calibration. Namely, the cross entropy loss function often used in neural networks is commonly overfit to, which leads to improved classification accuracy at the expense of increased miscalibration. They also found that miscalibration increases with model size, potentially concerning given the scaling laws that have lead to recent large models. As pretrained transformer-based architectures began to be used for transfer learning, the impact of out-of-domain (OOD) inputs on calibration was also investigated, finding that UQ degrades with dataset shift [?]. This trend also seems to hold in smaller, pretrained language models [?].

However, the direct relationship between model size and Expected Calibration Error (ECE) [?], an approximate measure of the difference between confidence and accuracy, has raised concerns with training larger models for safety-critical applications. Calibration evaluation on these larger models have demonstrated that unlike previous deep neural networks, models such as MLP-Mixer [?] and Vision Transformers [?] are well-calibrated and robust to distribution shift. Notably, accuracy and calibration are correlated under distribution shift, meaning that optimizing for accuracy may also improve calibration [?]. This suggests that foundation models such as LLMs could be well-calibrated in downstream tasks, if we consider their zero-shot application in classification tasks as OOD. However, this evaluation focused solely on image processing models and did not include language models, and so a more direct evaluation is necessary. We review UQ and calibration literature in general domain LLMs in the next section.

2.2.2 Confidence Elicitation in Large Language Models

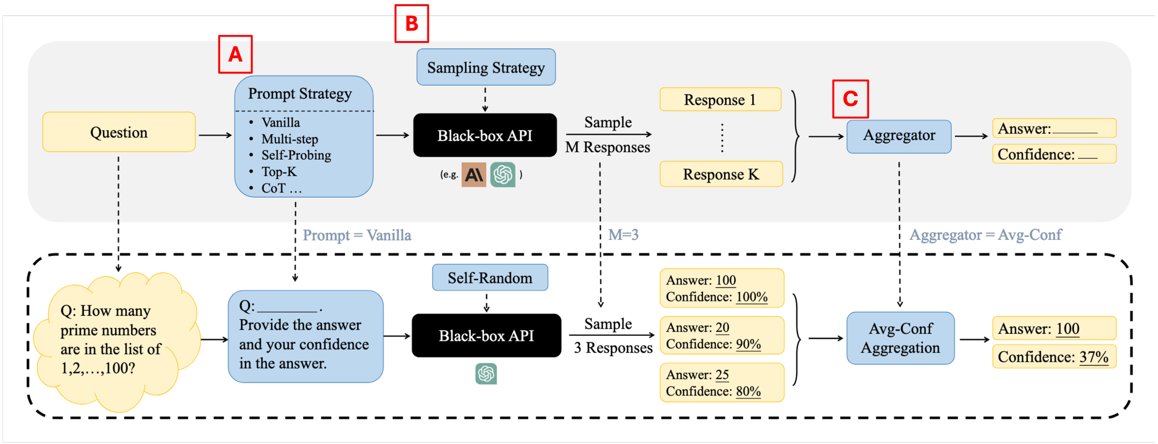


Figure 2.1: Various methods of confidence elicitation from large language models (*modified from* ? *]*).

Although calibration has been evaluated in transformer-based vision models models, the same patterns may not necessarily apply to large language models. In fact, the relationship between model scale and uncertainty quantification is complex [?]. To evaluate calibration trends in LLMs, a number of large-scale empirical studies have been conducted. Since LLMs communicate through text, the field of UQ has also been called *confidence elicitation*. In particular we will focus on *black box* models, as these are the ones that we consider in our own experiments. However, there is significant overlap in confidence elicitation methods, even when we have access to the weights and log probabilities, such as in white-box models.

All confidence elicitation in black-box models can be divided into three components, prompting strategies, sampling strategies, and aggregation methods to establish consistency [?]. White-box models can also include probes, but these require access to the model weights [?]. The three components can be seen in Figure 2.1. The upper row of the figure demonstrates the generalized confidence elicitation framework. Namely, (A) shows that some example prompt variations and (B) indicates the types of sampling strategies that can be employed. In some cases, sampling isn't performed, such as in *verbalized confidence*. Verbalized confidence explicitly asks the LLM for its confidence, as opposed to using a variety of proxies. Following a response from the LLM, we can aggregate results to arrive at a final confidence with (C). The lower section of the figure demonstrates one particular actualization of this methodology from ?]. In this work, the authors use a "top-k" prompting strategy in which they prompt the model for k answers, along with their estimated confidences. They also use chain-of-thought (CoT) and multi-step prompting, the latter of which first requests the answer and then verbalized confidence as separate questions. Other prompting strategies include those that ask the LLM to review its own answer as an estimate of confidence (i.e. *"How likely is the above answer to be correct"*) [?].

Prior work has found that *verbalized confidence* methods driven by prompting strategies can often lead to overconfidence [? ?]. As a result, sampling methods have been suggested as an alternative. In this case, the prompt may include any of the strategies above, but the confidence estimate is derived from sampling the response of the LLM multiple times and computing the confidence through an aggregation strategy. In particular, this can be done by directly sampling the answer and computing how frequently a particular answer is suggested [?] or by evaluating a surrogate token probability $P(I\ know) = P(IK)$ [?] as shown below:

> Q: Do you know the answer to the following question: $question (Yes/No/Maybe)?
> A: $answer
> *(from ?])*

This method requires that token-level probabilities are available from the LLM. In most black-box LLMs, this is the case, although usually only the top k log-probabilities are accessible. Still, this is enough to compute the confidence. Across all these methods, ?] find that no single prompting strategy consistently outperforms all other methods. In general, they found that self-probing seemed to be the most well-calibrated for GPT-4 models, like those we evaluate in our studies. Further, they find that in tasks requiring professional knowledge like clinical decision-making, LLMs still struggle to predict their incorrect predictions. In general, the authors recommend a Top-K prompting approach with self-random sampling (repeating the prompt multiple times with high temperature). For aggregating confidence, they propose an average confidence method, in which confidence is estimated as an answer-weighted average over K runs of the prompt. In our experiments, we evaluate both this method and simpler alternatives, consistent with the conclusion that the efficacy of confidence elicitation techniques is highly domain- and task-dependent.

2.2.3 How is Selective Prediction related to UQ?

As previously discussed, a well-calibrated model aligns its confidence levels with the true probability of the predicted outcome. However, why might we want to align confidence with true probabilities? One reason would be the use of these probabilities for a downstream task. Given that clinical LLMs will largely be used in conjunction with clinicians, one of the main use cases for well-calibrated models is human-in-the-loop decision-making. Namely, we want clinical AI models to abstain cases where they are unsure of the answer. We call this paradigm *Selective Prediction* [**?** **?**]. Given that LLMs are text generation engines, this process is sometimes called *selective generation* instead.

While LLMs appear to be well-calibrated, selective prediction extends well-calibrated confidence estimations for reflection and filtered decision-making. Given that well-calibrated confidence estimates are required, most of the methods from Section 2.2.2 can be used to filter results. When the $P(True)$, similar to the $P(IK)$ method is used to threshold answers above 0.5, accuracy increased across 5 question-answering datasets in mathematics, code, and general knowledge [**?**]. However, unless LLMs are forced into a classification framework like in our experiments, determining calibration in natural language generation becomes difficult due to "semantic equivalence" (i.e. different sentences can mean the same thing). Therefore, **?**] propose a novel uncertainty metric, termed semantic entropy, which estimates confidence by measuring the distribution of responses across clusters of semantically similar answers. This method can be thought of as a sampling approach over text generations. Similarly, **?**] use a self-probing prompting strategy to filter low-quality responses and return a response of "I don't know" in PALM-2 LARGE and GPT-3 models. Both studies find that selective generation leads to responses that not only improve accuracy, but also generate higher quality content according to human reviewers. More generally, selective prediction over well-calibrated results improve results across both question-answering and natural language generation tasks. In our case, we show how selective prediction methods can be used to filter out low-quality predictions in a clinical recommendation task (Section 3).

2.3 Uncertainty Quantification in the Clinical Domain

In the previous section, we discussed progress on uncertainty quantification (UQ) in the general domain. However, significantly less progress has been made on UQ and selective prediction in the clinical domain. In this section, we first discuss the progress on UQ in clinical applications across various modalities, and then focus on the text modality with large language models.

2.3.1 Prediction Uncertainty in Healthcare

The sources of uncertainty in healthcare are the same as those in the general domain. These include errors in measurement that were described previously as *aleatoric* uncertainty and errors associated with the model selection and training process known as *epistemic* uncertainty. In addition, *dataset shift*, in which the training data differs from the data being evaluated at inference time can also cause uncertainty. Aleatoric uncertainty can arise from lower inter-rater reliability (IRR) in ambiguous cases, such as in the detection of pneumonia as a radiographic diagnosis which is inherently more subjective than the detection of pulmonary opacity [**?**]. Due to the highly contextual nature of deprescribing, we also notice poor IRR in our own experiments (Ch. 3). Even with perfect information however, patient privacy requirements limit the size and coverage of clinical datasets. This leads to errors arising from dataset shift and epistemic uncertainty. For instance, an ML model that detects the location of organs on an MRI may not have seen a patient with situs inversus (mirrored organs), and so should report high predictive uncertainty so the patient can be manually reviewed by an expert physician [**?**]. In another example, **?**] found that patient-specific factors caused significant variability in uncertainty estimates when predicting in-patient mortality and differential diagnoses from ICU datasets. In such cases, the optimal action for an ML model is abstention. However, such selective prediction requires well-calibrated uncertainty estimates. A scoping review from 2022 on uncertainty quantification studies in healthcare found that the vast majority described medical imaging applications, with only 6/30 studies using other data modalities. Of those, none used clinical text at all [**?**]. Moreover, four of the most highly cited clinical ML models have no mechanism for uncertainty quantification and abstention [**?**]. Our analysis collectively highlights the need for further research into the patterns of uncertainty quantification in language models for clinical applications.

2.3.2 Uncertainty Quantification in clinical LLMs

Although there is limited research on the uncertainty quantification of LLMs, some preliminary work has been conducted using the aforementioned prompting strategies on medical challenge questions. Namely, as a part of model development, Med-PaLM [**?**] and NYUTron [**?**] both derived calibration curves on MedQA and 30-day readmission, respectively. The Med-PaLM evaluation used a self-random sampling approach while NYUTron used a verbalized confidence approach. Both models were found to be well-calibrated based on these preliminary evaluations. Previously, Codex was also found to be well-calibrated on MedQA-USMLE and MedMCQA [**?**]. Initial evaluations indicated that clinical large language models were well-calibrated across various medical decision-making tasks.

However, more thorough analysis revealed gaps in their uncertainty quantification capabilities. Two studies involving various black-box (e.g., GPT-4, GPT-4o, Claude Opus, Gemini) and white-box models (e.g., Llama2, Llama3, Mixtral) demonstrated

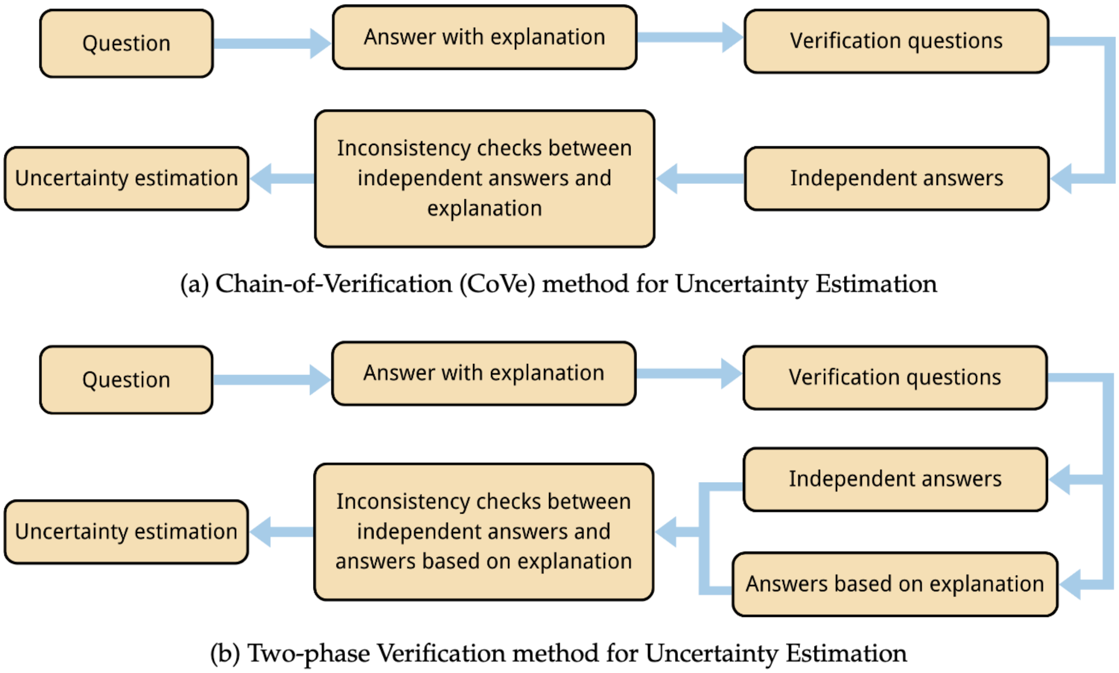

Question
Answer with explanation
Verification questions
Independent answers
Inconsistency checks between independent answers and explanation
Uncertainty estimation
(a) Chain-of-Verification (CoVe) method for Uncertainty Estimation
Question
Answer with explanation
Verification questions
Independent answers
Answers based on explanation
Inconsistency checks between independent answers and answers based on explanation
Uncertainty estimation
(b) Two-phase Verification method for Uncertainty Estimation

Figure 2.2: *reproduced from* **?** *]*

consistent overconfidence patterns in LLMs when verbalizing confidence on clinical challenge questions, similar to those observed in the general domain **? ?**]. **?**] took this further by determining that sampling strategies outperformed both verbalized confidence and token-level probability methods when answering both USMLE-style questions and diagnosing NEJM Case Reports. In addition to the standard confidence elicitation methods, **?**] employed an advanced form of the self-probing prompting strategy based on Chain-of-Verification (CoVe), known as Two-phase verification shown in Figure 2.2. When compared with the previously described semantic entropy, they found that their method improved accuracy and UQ, robust to QA datasets and model types.

While there has been growing interest and some significant foundational work in UQ and selective prediction in clinical reasoning with LLMs, the vast majority has been using challenge question-answering datasets. Although these datasets serve as valuable benchmarks for general clinical reasoning, we must be careful in extrapolating these findings to real-world clinical tasks. Firstly, the clinical vignettes used in USMLE questions and case reports are well-formed English sentences that would have been seen by the LLM, potentially even within its training data. However, it is much more likely that an LLM deployed in practice would be reasoning over notes which are far less well-formed. Moreover, a model in practice may have to integrate multiple modalities of data (e.g. vitals, lab results, notes, demographics), while all information presented in medical challenge questions has already been parsed and formed into natural language, further complicating the clinical decision making task.

Given these limitations, we build on prior work that identified effective uncertainty quantification techniques, applying them to evaluate LLMs' selective prediction capabilities in real-world clinical tasks. We aim for this to be the first study evaluating uncertainty quantification in LLMs using clinical notes for the critical task of deprescribing recommendations.

2.4 Bayesian Reasoning in Large Language Models

In addition to selective prediction in Chapter 3, we quantify uncertainty through a clinical decision-making Bayesian framework in Chapters 4 and 5. A key assumption of our evaluation framework is the LLM's capability to reason within the Bayesian framework. Therefore, in this section, we review the current literature on Bayesian reasoning in LLMs. Beyond general Bayesian reasoning, effective clinical decision-making also demands explicit numerical and causal reasoning, which we address by summarizing recent efforts to evaluate these reasoning abilities in LLMs.

2.4.1 Probabilistic Reasoning Capabilities of LLMs

In order for an LLM may reason within a probabilistic Bayesian framework, they must first be able to generally reason about probabilities and distributions. **?**] validate whether state-of-the-art LLMs including Gemini 1.0 Ultra, GPT4-Turbo, and GPT3.5-Turbo can estimate percentiles, draw samples from a distribution, and calculate probabilities from both ideal distributions and real-world distributions. In the clinical domain, they use Fitbit data from 100K users with 4 data points: step count, resting heart rate, sleep duration, and exercise minutes. They find that of the 3 models, GPT4-Turbo can effectively estimate percentiles along 3/4 of these data elements. They also find that including real-world context or providing information approximating a Normal distribution improves results. In general, this provides some evidence that LLMs have the numeracy to reason with probabilities. When integrating a Bayesian network, the LLM can approach this in two main ways: either with explicit Bayesian inference, or implicitly with verbalized Bayesian reasoning. Both methods have been evaluated. When modeling probabilities outside of natural language, the LLMs biggest tasks are information extraction from a problem statement. For instance, **?**] present a framework that uses LLMs to translate world knowledge and observations into a probabilistic language of thought (PLoT) using a language called Church, a Turing-universal probabilistic programming language. This is essentially a language-to-code translation task, after which Church can answer the query. To extend this, the LLM can be used to identify abductive factors that influence a particular outcome (i.e. *You want to charge your phone while using it* → *The charger is portable* and *The user needs to stay close to the charger* etc.). These potential factors can be used to train a simple BN that is then used to estimate outcome probabilities explicitly, improving performance on both commonsense and temporal reasoning [**?**

]. Instead of explicitly modeling the BN, the LLM can be placed within a Bayesian frame and reason with the implicit relations between various related elements. For instance, **?**] generate the BLInD dataset which provides context associated with synthetic Bayesian networks. They then ask Llama 3, GPT 3.5, and GPT4 to answer conditional probability queries (CPQ) by first extracting probability numbers and generating a text-based BN. The LLM can use this information, as well as its inherent numeracy, to reason about the CPQ and arrive at an estimate effectively. Other work has also demonstrated LLMs' abilities to perform BN structure learning, either for inference [**?**], or to reduce the human-driven effort necessary to generate a BN [**?**]. In general, LLMs seem to be able to both reason about probabilities distributions, as well as explicitly and implicitly reason about information contained within Bayesian networks for inference.

2.4.2 Decision Making under Uncertainty with LLMs

Our goal in evaluating LLMs within a Bayesian framework is to assess decision-making under uncertainty, a fundamental aspect of the clinical domain. Decision-making can be classified under the *Agent-based* framework in LLMs [**?**]. LLM agents not only parse text and generate responses, but take actions based on these responses and continue to the next conversation turn. For instance, to answer the query *"What other devices apart from Apple Remotes can control the program the Apple Remote was originally designed to interact with?"*: an LLM might draw purely from its internal, parametric knowledge, while an Agent might perform online searches to answer the sub-questions *"What program was originally designed to work with the Apple Remote"* and use the response of *"Front Row"* and conduct another search. With sequential reasoning and acting, the LLM is able to arrive at the final answer. Likewise, in Ch. 5, we ask the LLM to gather diagnostic information and use it to estimate disease probability risk until a final diagnosis is achieved. While we are the first to use agent-based simulation of diagnosis under a Bayesian framework, decision-making under uncertainty with LLM agents isn't new. Several studies have developed uncertainty-aware LLM agents, such as DeLLMa [**?**] or Uncertainty-Aware Language Agent (UALA) [**?**]. DeLLMa leverages classic *utility theory* by eliciting preferences towards a utility function for a set of m states related to a particular goal (e.g. *"I'm a farmer in California. What fruit should I grow next year?"*). By maximizing this utility function, the LLM can make an informed decision under uncertainty. UALA is simpler in that it performs CoT reasoning and estimates its uncertainty. Depending on its confidence, it either searches for more information with a search engine or defers to the user if the search results also have low confidence. Through these pipelines, both models make sequential decision steps under uncertainty, albeit not within a Bayesian framework like our work.

Perhaps the work closest to ours from a decision-making perspective comes from **?**]. The authors used data from the MIMIC-III dataset to evaluate LLM performance in a turn-based conversational setting, where the LLM sequentially requests information. The LLM could request four types of information: the history of present illness (HPI), physical examination results, laboratory tests, and imaging findings. Initially, the LLM is provided with the HPI and can subsequently request any of the other three types of information, specifying modalities for lab tests and imaging. When confident, the LLM could provide a diagnosis instead of seeking additional information. This study focused on patients with appendicitis, cholecystitis, diverticulitis, or pancreatitis, comparing LLM diagnoses against those of expert physicians, including treatment recommendations and the sequence of information requests. Their findings indicate that LLMs struggle to adhere to diagnostic and treatment guidelines, fail to interpret laboratory results accurately, and are sensitive to both the amount and order of information requested. Despite similarities to our work in Chapter 5, our study incorporates significant differences that extend beyond their approach.

Most importantly, we ground our evaluation in a validated Bayesian network. This offers several advantages. We are able to determine an information-theoretic "optimal" next decision at each stage in the conversation, allowing us to derive an optimal pathway to diagnosis, rather than simply recording the order of events, we can determine the "correct" order of events. Further, we do not have to deny the LLM any information when requested, as we can use counterfactual conditional probability queries to determine the most likely test results, rather than responding that the information wasn't gathered like in **?**]. This ensures that the LLM is not biased towards aligning with physician decision-making, especially in cases where the physicians may be incorrect. Unfortunately, the definition of a Bayesian network is a requirement to use our pipeline, but we believe the work necessary to establish a gold-standard network is worth the benefits in evaluation.

2.5 A Lack of UQ and Bayesian Reasoning in Clinical Domain

As demonstrated, there is a significant gap in studies on uncertainty quantification and Bayesian reasoning within the clinical domain. Notably, most UQ and calibration assessments have relied on medical challenge question-answering datasets. We argue that uncertainty quantification using clinical notes presents a considerably greater challenge and merits independent evaluation. We present the results of this investigation in Chapter 3. Likewise, there has been limited evaluation of Bayesian reasoning capabilities in LLMs. A few papers have discussed probabilistic diagnostic reasoning, either as a prompting strategy [**?**] or within the pretest/post-test evaluation framework [**?**]. While the second paper overlaps with our work in Ch. 4, we leverage established prompting strategies from NLP literature, such as chain-of-thought reasoning, to enhance model performance. Furthermore, we assess probability estimates within a bias mitigation framework, examining the influence of race and disease condition on post-test probability outcomes. Through these approaches, we expand the limited body of research on probabilistic diagnostic reasoning using LLMs.

These initial studies largely evaluate an LLM's ability to estimate probabilities in a single pretest/post-test probability scenario. However, diagnosis of a patient involves sequential wayfinding [?] and decision-making. In our studies (Ch. 5), we evaluate an LLM's full diagnostic process grounded in a Bayesian network. This work can be thought of as bringing together the sequential reasoning paradigm from ?] with the Bayesian reasoning framework in BLInD [?]. However, to actualize a Bayesian reasoning framework with real-world data, we must generate a realistic BN and align it with the LLM, a process we describe in the chapter.

We initially validate this process by extending some of the prior work in diagnostic probability estimation by leveraging state-of-the-art prompting strategies from the NLP literature and investigating biases in estimation (Ch. 4). As a result, we believe that our studies in UQ and Bayesian reasoning increase the overall understanding of probabilistic reasoning and selective decision making in LLMs used for clinical applications. These advancements are anticipated to facilitate more effective clinician-AI collaboration and mitigate the risk of adverse outcomes in safety-critical clinical environments.

Chapter 3

Uncertainty Quantification in a Real-World Clinical Task: Deprescribing

3.1 Introduction

The aim of this thesis is to assess uncertainty quantification (UQ) and, in doing so, evaluate the Bayesian diagnostic reasoning capabilities of large language models (LLMs). A fundamental requirement for Bayesian diagnostic reasoning is a comprehensive understanding and quantification of clinical uncertainty. Clinical judgment frequently demands both qualitative and quantitative interpretations of uncertainty [?]. For instance, the LLM must be able to quantify the uncertainty associated with a differential list of diagnoses to determine the most useful diagnostic test to run next, similar to how a physician would reason about a patient during diagnostic wayfinding. However, in a similar scenario, the relative risks associated with a particular patient situation may convince a physician to choose a less sensitive diagnostic test with lower risk. For instance, a particular patient may be afraid to get an CT for a head injury given their lack of experience with radiation imaging, despite it being the most sensitive to determine TBI. Therefore, the physician may need to perform an US instead, despite the quantitative risk suggesting an MRI. Therefore, while the quantitative interpretation of uncertainty may be relevant in determining the "ideal" decision making process, qualitative elements influence decision making. These sometimes competing priorities contribute to the art and science of medicine and require clinical gestalt. An LLM that attempts to align with current clinical practice must handle both these qualitative and quantitative interpretations of uncertainty. In this work, we tackle the former through an inherently qualitative uncertainty quantification task: the process of recommendation deprescribing in older adults (i.e. ≥ 65 years old).

Deprescribing is defined as the systematic process of identifying and discontinuing drugs, called potentially inappropriate medications (PIMs), whose present or potential harms outweigh benefits provided to the patient within the context of their individual care goals and quality of life [? ?]. This process is primarily conducted in patients that are at-risk for drug-related negative outcomes: those with polypharmacy regimes. Widely defined as the regular use of at least five medications, polypharmacy is common in older adults and at-risk populations[?]. In fact, approximately 30% of patients aged 65 years or older have polypharmacy[?], and nearly half of older emergency department (ED) patients are discharged with one or more new medications[?]. Although necessary and beneficial for some patients, polypharmacy can increase risk of negative consequences for patients, including emergency department (ED) visits, adverse drug events (ADEs), falls, disability, and inappropriate medication use[?].

Deprescribing tools, such as the Screening Tool of Older People's Prescriptions (STOPP) and Beers criteria, have been developed to help providers assess and identify PIMs based on a patient's medication list[? ? ?]. These explicit assessments are criterion-based with clear standards, but are often impractical to implement in time-constrained clinical settings, such as the emergency department[?]. Attempts to digitize these criteria into electronic clinical decision support have raised difficulties, typically requiring a labor-intensive coding process and unstructured information from patient records to contextualize certain criteria[? ?]. Large language models (LLMs) have been shown to interpret complex clinical situations and offer recommendations, from differential diagnoses to care management, leading to growing interest in their application in the medical field[? ? ? ?]. Moreover, they have been shown to extract medication-related data such as medication name, dosage, and frequency, necessary for application of deprescribing criteria[?]. Lastly, LLMs are excellent in-context learners, requiring very little labeled data to make predictions[?] reducing the annotation burden for time-constrained EM physicians while improving the use of unstructured patient records to contextualize patient medication lists. However, the majority of clinical reasoning evaluations on LLMs have been conducted using standardized exams (USMLE) or online case reports[? ?]. Both of these exam types are multiple choice and require clinical gestalt and reasoning to make decisions under qualitative clinical uncertainty. In some cases, patient considerations need to be taken into account. In most, explicit quantification of benefits and risks is unnecessary, and instead relative risks are weighed against one another, in combination with guidelines and rules of thumb. However, LLMs ability to perform this form of qualitative reasoning over physician-generated text (such as clinical notes) remains unclear.

In this chapter, we propose to evaluate the performance of an end-to-end LLM-based pipeline in recommending deprescribing options for ED patients at discharge based on explicit deprescribing criteria. In addition to helping address gaps in electronic deprescribing by using an LLM to reduce manual development in CDS tools, this work will help elucidate if LLMs have an internal model of qualitative uncertainty that aligns with that of physicians and medical students. We also investigate if we can use this model to improve their performance using selective prediction methods, allowing for a more collaborative physician-AI system.

3.2 Methods

We describe the methods used to evaluate a qualitative uncertainty quantification task, namely, deprescribing medications in older adults.

3.2.1 Patient Cohort

Our cohort consists of all older adults ($\geq$ 65 yo.) patients with polypharmacy ($\geq$ 5 active outpatient medications) presenting to the Yale New Haven-Health Emergency Department between January-March 2022 totaling 10,977 patients across 15,161 encounters. We select a random, convenience sample of 100 unique patients with 898 medications based on budget constraints and power analysis based on the unique medications across all patients (**Need to put in the power analysis here: XXX**). On average, a patient in our cohort had 9 medications and all were considered separately by both the LLM and annotators, given the same patient information for filtering and recommendation.

3.2.2 Consensus-Based High-Yield Criteria Evaluation

In an informal pilot study, one of the main causes of discrepancies between physicians and LLMs arose from ambiguous inclusion/exclusion conditions in deprescribing criteria (e.g. "Statins for primary cardiovascular prevention in persons aged $\geq$ 85 and established frailty with expected life expectancy likely less than 3 years."). Both established frailty and expected life expectancy are difficult to quantify and therefore implement. As a result, we first conducted a rigorous consensus-based evaluation of deprescribing criteria from three different recommendation lists: STOPP, Beers, and GEMS-Rx.

To identify high-yield criteria likely to be amenable to automated review, we evaluated a total of 180 recommendations across two dimensions: Clinical Applicability and EHR Computability, under the assumption that high-yield criteria must both be highly clinically applicable (e.g. pose high risk to the patient, be feasible for deprescribing in various clinical contexts) and identifiable within the EHR. The consensus panel consisted of 6 board-certified physicians (in Emergency Medicine, Internal Medicine, Cardiology, Med-Peds, Geriatrics) and 1 ED pharmacist at YNHH.

Each member of the group individually reviewed each of the criteria and rated them on a 5-point Likert scale for each of 5 questions of clinical applicability and 4 questions of EHR computability as shown in Figure A1. In our final selection, we averaged all EHR computability responses across panelists and selected only clinical risk to the patient to represent Clinical Applicability. Due to the high number of potentially high-yield criteria, we then selected the top 50% criteria in terms of EHR computability and risk to patients arriving at 81 total criteria across all three deprescribing lists. A plot of EHR computability and patient risk is shown in Figure 2. While we captured information on deprescribing feasibility in various contexts in our consensus study for further qualitative evaluation, it was not included in selection of high-yield criteria as they should be relevant irrespective of the clinical context under which they may be feasible.

3.2.3 Deprescribing Recommendations by GPT-4o

Given the cognitive burden required to evaluate the large list of criteria, we hypothesize that LLMs will be more effective than clinical experts in identifying relevant clinical criteria, but worse at applying them given the ambiguous nature of even high-yield inclusion/exclusion conditions. To determine an LLM?s efficacy on these two questions, as well as to reduce the context size provided to the model, we developed a 2-step pipeline, as shown in Figure 1. In step 1, GPT-4o is prompted to filter the full list of high-yield criteria solely based on the patient?s medication list, ignoring inclusion/exclusion conditions. This both reduces confusion due to large input context sizes and ensures that extraneous context doesn?t distract the LLM. In step 2, GPT-4o is prompted to use its previously filtered criteria list, along with structured (e.g. demographics, lab values, vitals, and PMH) and unstructured (most recent progress note and discharge summary) information, to determine if the patient satisfies any deprescribing criteria and therefore should be recommended for deprescribing. Aside from the sampling-based method, described below, all LLM calls are performed with almost no variation (*temperature*=0, set seed).

3.2.4 Selective Prediction

For both steps, we also collect GPT-4o?s confidence associated with its decisions using two confidence elicitation methods. The LLM?s confidence will be used to determine if the model should abstain due to low certainty regarding its own decision. In practice, this case would be considered too difficult for the LLM and forwarded to an expert reviewer. This human-in-the-loop decision making pipeline is known as selective prediction and has been commonly found to improve performance in non-text-based applications. In this work, we evaluate whether LLMs? seemingly well-calibrated confidence on question-answering tasks such as the USMLE transfer to a real-world clinical recommendation task.

We select validated, effective versions of the two main confidence elicitation methods: chain-of-thought-driven verbalized confidence and self-random sampling with average-confidence aggregation. For the prompt-based method, we ask the LLM to explicitly estimate its confidence for both steps following its decision. For the sampling-based method, we repeat the sample question with high temperature (T=0.8) several (N=5) times and use a verbalized confidence-weighted confidence estimate. We evaluate both selective prediction methods using risk-coverage curves, substituting coverage for LLM deferring fraction.

3.2.5 Comparison and Adjudication with Clinical Experts

We compare GPT-4o?s filtered criteria lists and recommendations against human clinical experts to determine the LLM?s performance. Due to the low IRR found in our pilot study, we elect to have two senior (M4) medical students evaluate all medications in the test cohort and then have discrepancies between med student and the LLM adjudicated by two senior, board-certified ED physicians. In other words, we do not assume that medical students are the gold standard. For each medication, a medical student would determine (1) if there exists a relevant high-yield criteria based on the medication list, and (2) whether the medication should be recommended for deprescribing. Of the 898 total medications, 75 were repeated by both medical students to validate the low inter-rater reliability (IRR) between junior clinical experts. We then compute discrepancies between medical students and the LLM and provide them to the ED physicians to determine who was correct and in cases where the LLM was incorrect, why this was the case. We leverage a prior evaluation framework to determine LLM error modes. We measure the IRR (Cohen?s k: Eligibility?0.619, Deprescribing?0.764) between the two senior ED physicians to ensure that coding practices were standardized prior to adjudication on the full set of GPT-medical student discrepancies. A discussion of the low k for eligibility is explained in the discussion.

3.3 Results

As discovered in our pilot study, the medical student inter-rater reliability (IRR) in both steps was low (Cohen?s k: step 1-0.741 , step 2-0.082) across 75 medications. This led to an exploration of their performance compared to GPT-4o. As shown in Figure 3, the majority of medications (56.1%) were not included in the high-yield deprescribing criteria. However, there were a number of discrepancies that needed to be adjudicated. Of these, 90 (10.9%) were lower priority as they determined eligibility of a medication, but either the med student or the GPT model ultimately decided its recommendation for deprescribing was not necessary. However, the majority (135 - 16.4%) could potentially lead to a change in medication management. A significant source of discrepancy comes from the LLM?s significantly higher likelihood to recommend deprescribing (14.5%) compared to the medical students (6.3%).

Upon adjudication by senior physicians, we found that across all discrepancies, **XXX** performed better than **XXX**. When looking at the error modes that led to GPT errors, when filtering criteria, adjudicators found that **XXX** was most common, followed by **XXX**. Conversely, in making recommendations to deprescribe, **XXX** was most common. Finally, we investigated whether eliciting confidence estimates from the LLM and using them to determine if the model should abstain would improve results. We compare both verbalized confidence with sampling-based confidence to find that sampling-based confidence outperformed verbalized when filtering criteria (max F1-Score: 0.811), but verbalized confidence performed better when making deprescribing recommendations (max F1-Score: 0.145) as shown in Figures 4 and 5. Additionally, regardless of method, selective prediction improved performance in eligibility as the deferring fraction increased (confidence threshold increased), while the inverse was true when making deprescribing recommendations, albeit with a low F1-score to begin with. In general, **XXX** are able to determine criteria-eligible medication better, while **XXX** are able to make deprescribing recommendations better. Moreover, selective prediction is not effective across the board, but is task-dependent.

3.4 Discussion

3.5 Relevance to Initial Hypothesis

Chapter 4

LLM Biases in Bayesian Diagnostic Reasoning

4.1 Introduction

In Chapter 3, we

The interpretation of diagnostic testing is a central component of clinical decision making. Diagnostic testing is the single most performed medical activity, with more than half of all clinical decisions made on the basis of laboratory testing1,2. A clear understanding of test characteristics and the judicious ordering of tests is vital in both reducing healthcare expenditures and improving overall quality of care for patients3,4. Classical diagnostic reasoning involves updating initial estimates of disease likelihood as diagnostic testing is ordered and results are returned. This process can be modeled through Bayes' rule (Figure 1)5. Unfortunately, significant evidence suggests that physicians are inaccurate in their estimates of disease probabilities and test characteristics, often falling for cognitive biases6?10.

Large language models (LLMs) have been shown to effectively perform numerical reasoning in a variety of settings.11,12 A recent study examined GPT-4's ability to estimate disease probabilities via Bayes' rule across four patient vignettes, demonstrating that the LLM was comparable to physicians for estimating post-test probability after a positive diagnostic test and more accurate than physicians after negative diagnostic tests across these vignettes.6 From this study, however, it was challenging to determine where the LLM misstepped in its erroneous answers, be it in its estimation of sensitivity or specificity of the diagnostic test, versus in its application of Bayes' rule. Nor did it examine how variation in the prompt provided to the LLM?be it specifically instructing the LLM to apply Bayes' rule or the use of chain-of-thought reasoning13?may impact results. Finally, each disease was only evaluated as a single condition, limiting the conclusions that can be drawn on disease-dependent variation in LLM disease probability estimation.

In addition, despite LLMs' significant potential to improve medical practice, there is concern that they will further exacerbate health inequities due to hidden sociodemographic biases arising from training data.14,15 In particular, Zack et al. tested the impact of varying demographics on the generation of clinical vignettes, finding that for conditions with similar prevalence by race and gender, such as COVID-19, GPT-4 was significantly more likely to generate vignettes with men16. Similarly, Nastasi et al. found that ChatGPT recommended a community clinic for an uninsured patient, while recommending the ED for an insured patient, despite the need for emergency care17. Despite initial work done in evaluating the impact of LLMs' biases on clinical decision making, previous studies have not evaluated how these biases may impact diagnostics, such as in the context of a clinical decision support tool.

Our work aims to expand on previous research by evaluating both disease variation and racial biases in estimating risk through a Bayesian diagnostic framework by two state-of-the-art LLMs of different reasoning capabilities: GPT 4o-mini and GPT 4o. By modifying the availability of likelihood ratio information, we seek to identify specific points in the reasoning process that influence risk estimation. To investigate potential biases in the diagnostic risk assessments of these LLMs, we evaluate several vignettes across 4 distinct conditions and systematically vary race throughout our evaluation.

4.2 Methods

4.2.1 Vignette Selection

To evaluate the impacts of disease and race on Bayesian reasoning in LLMs, we gathered clinical vignettes from four common ED presentations with clearly defined diagnostic tests and validated positive and negative likelihood ratios (LRs). Specifically, we chose previously developed vignettes from Brush et al18. The vignettes included clinical presentations, history, physical exam findings, as well as vitals and lab values when relevant. A total of 43 vignettes were used, covering 4 ED diagnoses: acute coronary syndrome (ACS), congestive heart failure (CHF), pneumonia, and pulmonary embolism (PE). Each vignette was also accompanied by a diagnostic laboratory test or radiologic investigation that was conducted following initial presentation.

Originally designed to assess Bayesian reasoning in medical students, each vignette begins by asking for an initial pre-test probability of disease on a scale from 0 to 100%. After determining the pre-test probability, participants were given the results of a relevant diagnostic test, reported as either positive or negative, along with literature-validated data on sensitivity, specificity, and both positive (LR+) and negative (LR-) likelihood ratios. Finally, participants used the initial presentation and test results to estimate a post-test probability of disease. We adapted these vignettes to require the LLM to estimate sensitivity and specificity, rather than relying on literature-derived values from Brush et al., allowing for a more fine-grained evaluation of its reasoning process. We selected vignettes that were physiologically likely, regardless of race for our racial bias evaluation. A description of the vignettes is shown in Table 1. All initial presentations included the patient's age, gender, and race.

4.2.2 Bias Dimensions

LLMs have the potential to perpetuate context biases due to both disease domain and race. To assess the impact of disease-specific bias, we average over all vignettes associated with a specific disease in our evaluation. To assess the impact of racial bias on risk estimation, we vary race in the history of present illness (HPI) section of the vignette using the U.S. Department of the Interior racial categories: American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, and White.

4.2.3 Bias-Aware Diagnostic Reasoning Evaluation

We conducted all experiments through Azure OpenAI using both GPT 4o and GPT 4o-mini employing prompts specifying chain-of-thought (CoT)19. CoT reasoning has been shown to significantly improve reasoning in LLMs, allowing us to test the full capabilities of both models. Given that LLMs are inherently stochastic generative models, we sampled the full distribution of disease probability estimates for a given vignette by setting a high temperature temperature (0.8, from [0-1]) and repeating each vignette 10 times with GPT 4o-mini and 5 times with GPT 4o for both positive and negative test results. The high temperature ensures that the LLM generates different responses for each trial, allowing us to characterize the full range of probability estimates an LLM might generate20.

We evaluate performance of both LLMs by computing the post-test probability error, defined as the difference between the true post-test probability, computed using the estimated pre-test probability and literature-derived likelihood ratios, and the estimated post-test probability generated from the LLM. The estimated pre-test probability is assumed to be valid, acknowledging that prevalence can depend on various factors such as geographic location and presenting signs or symptoms. Consequently, our focus is on evaluating the models' capacity for Bayesian reasoning rather than their ability to estimate initial clinical disease risk. We also compute the estimated likelihood ratios from estimated sensitivity and specificity, and compare these to the provided LRs in Brush et al.

To estimate post-test probability, likelihood ratio information was provided to the models in 3 different ways. First, we evaluate implicit estimation of disease risk by estimating both pre-test probabilities and post-test probabilities following test results, without any likelihood ratio information or explicit reference to Bayesian reasoning. Next, the LLMs estimate sensitivity and specificity, given the initial presentation of a patient and the related diagnostic test to encourage Bayesian reasoning while still limiting external information. Next, the LLMs were tasked with predicting post-test risk, given their estimated LRs. Finally, we substitute the estimated LRs with the literature-specified LRs from Brush et al. as a baseline. This allowed evaluation of diagnostic reasoning given implicit, partial and full Bayesian information.

Finally, to examine racial bias, for each vignette, we alter the patient's race in the HPI. Given the repeated trials described above, this yields a total of 5,160 for GPT 4o-mini and 2,580 runs for GPT 4o. In addition to estimating disease probability distributions, we hypothesize that differences in diagnostic accuracy are inherent biases baked into the models from training data, not a function of their stochasticity, so conduct LR estimation in a mostly-deterministic setting (temperature=0.0; set random seed).

4.3 Results

In Figure 1, we present the post-test probability error of GPT 4o-mini and GPT 4o with three elements of likelihood ratio information: none, estimated, and true LRs. Across all disease conditions, GPT 4o-mini estimated disease risk more accurately when given either estimated (post-test prob. error: -1.75 ± 19.69) or literature-derived LRs (0.38 ± 8.22), compared to when the model estimated risk with implicit Bayesian reasoning (4.01 ± 24.29). This trend was also evident in GPT 4o, but with even greater accuracy when provided literature-derived LRs (0.02 ± 0.52). Both models variably demonstrated under- or overestimation of disease risk by disease condition. For instance, both models significantly overestimated disease risk, regardless of Troponin results when no LRs were provided (GPT 4o-mini: -3.50 ± 32.72, GPT 4o: -14.66 ± 19.68).

We also show differences in estimated LRs by race, across two temperature values for both LLMs in Figure 2. In general, all models, regardless of temperature, underestimated the diagnostic accuracy of chest X-rays in diagnosis of CHF, while overestimating the power of Troponin I in diagnosis of ACS. However, as compared to GPT 4o-mini, GPT 4o estimated LRs for D-dimer in diagnosis of PE more accurately.

4.4 Discussion

In this study, we conducted a rigorous evaluation of potential disease and race biases in disease probability estimation by large language models, employing a Bayesian pre-test/post-test probability framework. Utilizing validated vignettes for 4 common ED diagnoses, we evaluated both GPT 4o and GPT 4o-mini LLMs using the Azure OpenAI API assessing their ability to estimate pre-test probabilities, likelihood ratios, and post-test probabilities, while systematically varying the ED diagnosis of interest and race.

In general, our results showed that LLMs struggled to accurately estimate post-test probabilities when no diagnostic test characteristics were estimated or provided. However, when using estimated or literature-derived LRs, in almost all cases, LLMs were able to employ Bayes' rule to improve their estimation of post-test disease risk.

In particular, ACS and pneumonia are estimated incorrectly more so than other conditions by both models, potentially due to the vignettes with these ED diagnoses being cases that could be easily confused, namely ACS vs. hypertensive emergency and pneumocystis pneumonia vs. influenza-like illness.

To understand why post-test probability estimates vary so significantly by condition, we hypothesized that LLMs misestimate the diagnostic accuracy of tests differently across conditions, affecting their ability to estimate post-test probabilities. To test this, we compared literature-derived likelihood ratios (LRs) with LLM-estimated LRs. For both models, it appears that conditions where the LLM has a poor understanding of the diagnostic accuracy of a particular condition (e.g. positive and negative results for CHF and pneumonia in GPT-4o) led to worse diagnostic probability error rates. In contrast, LRs for D-dimer in diagnosing PE were accurately estimated leading to lower error rates for both positive and negative lab results. This suggests that proper understanding of the predictive power of diagnostic tests is instrumental in disease probability estimates by LLMs.

When breaking down the evaluation by race, while there were differences in post-test probability error across all three types of LRs, they were not statistically significant. Upon qualitative error analysis, we find that GPT 4o-mini explicitly mentioned race in **XXX**% of responses, while GPT 4o mentioned race in **XXX**%. While ultimately not impacting probability estimations, underlying racial bias may need to be considered when using LLMs to estimate disease risk.

In this work, we conducted a comprehensive evaluation of disease probability estimates from two state-of-the-art LLMs, identifying notable differences across disease conditions when implicit and partial Bayesian information was provided. These differences appear to stem from limitations in estimating the predictive power of diagnostic tests in both models. Encouragingly, error rates did not significantly differ by race, though race was mentioned in both models' reasoning processes. Our findings underscore the importance of detailed assessment of reasoning steps when using LLMs as clinical risk predictors, as performance may vary across disease contexts, each requiring independent evaluation.

4.5 Relevance to Initial Hypothesis

Chapter 5

Decision-Analytic LLM Diagnostic Reasoning with Bayesian Networks

5.1 Introduction

5.2 Methods

5.3 Results

5.4 Discussion

5.5 Relevance to Initial Hypothesis

Chapter 6

Conclusion

6.1 Summary of accomplishments and contributions

6.2 Generalizability of the results

6.3 Future work

6.4 Conclusions

- Summary of accomplishments and contributions
- Assessment of hypothesis (from Chapter 1) in light of what has been discussed
- Generalizability of the results
- Range of applicability
- Future work
- Conclusions

Contents

Appendix A

Proof of Mathematical Theorems

This appendix contains the proofs of all claims made in main body of the dissertation.

A.1 Proof of Theorem 1

The proof is left as an exercise to the reader.