Proceedings of Machine Learning Research 252:1–51, 2024 Machine Learning for Healthcare

MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance

Jihye Choi[*1] JIHYE@CS.WISC.EDU
Nils Palumbo[*1] NPALUMBO@WISC.EDU
Prasad Chalasani[2] PCHALASANI@GMAIL.COM
Matthew M. Engelhard[3] M.ENGELHARD@DUKE.EDU
Somesh Jha[1,2] JHA@CS.WISC.EDU
Anivarya Kumar[3] ANIVARYA.KUMAR@DUKE.EDU
David Page[3] DAVID.PAGE@DUKE.EDU
[1] *Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA*
[2] *Langroid*
[3] *Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA*

Abstract

In the era of Large Language Models (LLMs), given their remarkable text understanding and generation abilities, there is an unprecedented opportunity to develop new, LLM-based methods for trustworthy medical knowledge synthesis, extraction, and summarization. This paper focuses on the problem of Pharmacovigilance (PhV), where the significance and challenges lie in identifying Adverse Drug Events (ADEs) from diverse text sources, such as medical literature, clinical notes, and drug labels. Unfortunately, this task is hindered by factors including variations in the terminologies of drugs and outcomes, and ADE descriptions often being buried in large amounts of narrative text. We present MALADE, the first effective collaborative multi-agent system powered by LLM with Retrieval Augmented Generation for ADE extraction from drug label data. This technique involves augmenting a query to an LLM with relevant information extracted from text resources, and instructing the LLM to compose a response consistent with the augmented data. MALADE is a general LLM-agnostic architecture, and its unique capabilities are: (1) leveraging a variety of external sources, such as medical literature, drug labels, and FDA tools (*e.g.*, OpenFDA drug information API), (2) extracting drug-outcome association in a structured format along with the strength of the association, and (3) providing explanations for established associations. Instantiated with GPT-4 Turbo or GPT-4o, and FDA drug label data, MALADE demonstrates its efficacy with an Area Under ROC Curve of 0.90 against the OMOP Ground Truth table of ADEs. Our implementation leverages the Langroid multi-agent LLM framework and can be found at `https://github.com/jihyechoi77/malade`.

1. Introduction

Pharmacovigilance (PhV) is the science of identification and prevention of adverse drug events (ADEs) caused by pharmaceutical products after they are introduced to the market. PhV is of enormous importance to both the pharmaceutical industry and public health, as it

---

[*] Equal contribution, listed alphabetically by last name

© 2024 J. Choi, N. Palumbo, P. Chalasani, M.M. Engelhard, S. Jha, A. Kumar & D. Page.

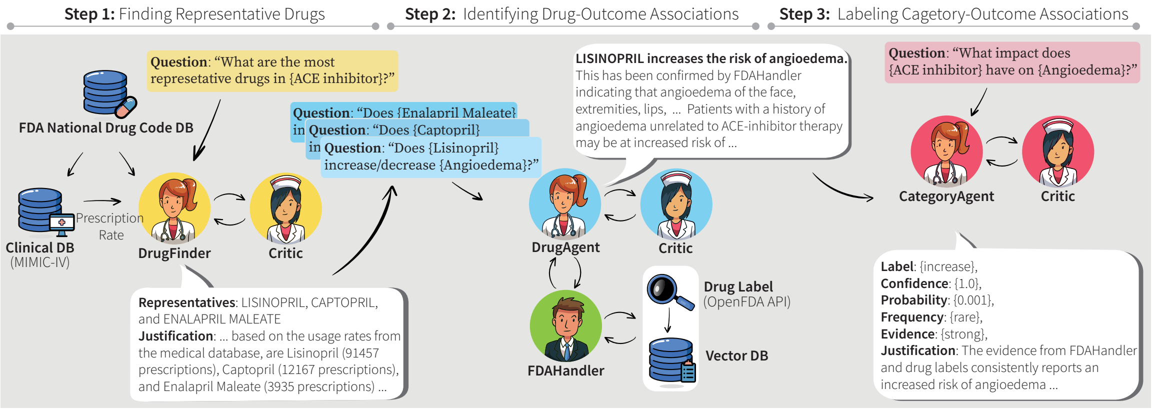

Step 1: Finding Representative Drugs
Step 2: Identifying Drug-Outcome Associations
Step 3: Labeling Cagetory-Outcome Associations
Question: "What are the most represetative drugs in {ACE inhibitor}?"
FDA National Drug Code DB
Clinical DB (MIMIC-IV)
Prescription Rate
DrugFinder
Critic
Representatives: LISINOPRIL, CAPTOPRIL, and ENALAPRIL MALEATE
Justification: ... based on the usage rates from the medical database, are Lisinopril (91457 prescriptions), Captopril (12167 prescriptions), and Enalapril Maleate (3935 prescriptions) ...
Question: "Does {Enalapril Maleate} in
Question: "Does {Captopril} in
Question: "Does {Lisinopril} increase/decrease {Angioedema}?"
LISINOPRIL increases the risk of angioedema. This has been confirmed by FDAHandler indicating that angioedema of the face, extremities, lips, ... Patients with a history of angioedema unrelated to ACE-inhibitor therapy may be at increased risk of ...
DrugAgent
Critic
FDAHandler
Drug Label (OpenFDA API)
Vector DB
Question: "What impact does {ACE inhibitor} have on {Angioedema}?"
CategoryAgent
Critic
Label: {increase},
Confidence: {1.0},
Probability: {0.001},
Frequency: {rare},
Evidence: {strong},
Justification: The evidence from FDAHandler and drug labels consistently reports an increased risk of angioedema ...

Figure 1: **Real-world demonstration of our proposed multi-agent orchestration system, MALADE.** Handling the user query, "Are ACE Inhibitors associated with Angioedema?", involves a sequence of subtasks performed by three Agents: `DrugFinder`, `DrugAgent`, `CategoryAgent` (each instantiated with GPT-4 Turbo or GPT-4o). Each Agent generates a response and justification, which are validated by a corresponding Critic agent, whose feedback is used by the Agent to revise its response.

aims to safeguard the well-being of patients by detecting new safety concerns and intervening when necessary. A central problem in PhV is **ADE Extraction**: given a *drug category* C and an *adverse event* E, determine whether (and how strongly) C is associated with E. This task demands the analysis of a vast corpus of textual data sources from a variety of sources, such as patient medical records, clinical notes, social media, spontaneous reporting systems, drug labels, medical literature, and clinical trial reports. Besides the sheer volume of text from these sources, ADE extraction is further complicated by variability in the names of drugs and outcomes, and the fact that ADE descriptions are often buried in large amounts of narrative text (Le et al., 2024).

Traditionally, various classical natural language processing (NLP) and deep learning techniques have been used to address this problem (Natarajan et al., 2017; Mower et al., 2018; Tiftikci et al., 2019; Bayer et al., 2021). Compared to classical NLP methods, today's best Large Language Models (LLMs) (and even weaker open-source/local LLMs (Touvron et al., 2023; Jiang et al., 2023)) exhibit a significant advancement in text understanding and generation capabilities, and there is a great opportunity to use these models to not only improve existing ADE extraction methods, but also consider data sources that were previously not feasible to use. Recent attempts to apply LLMs to ADE Extraction only leverage off-the-shelf ChatGPT (Wang et al., 2023a), with limited performance and inconsistent reasoning for their extraction rationales (Sun et al., 2024). These limitations stem primarily from two factors: (a) accurate ADE Extraction requires access to specific data sources which LLMs may not have "seen" during their pre-training, hence relying on an LLM's "built-in" knowledge yields inaccurate results, and (b) LLMs, being probabilistic next-token predictors, may produce incorrect or unreliable results when used naively without carefully breaking down the task into simpler sub-tasks, or without mechanisms to validate and correct their responses.

In this paper, we introduce MALADE[1](**M**ultiple **A**gents powered by **L**LMs for **ADE** Extraction), the first effective multi-agent Retrieval-Augmented Generation (RAG) system for ADE Extraction. Our approach leverages two key techniques to address the above two limitations respectively: (a) RAG, equipping an LLM with up-to-date knowledge by augmenting an input query with relevant portions of text data, and prompting the LLM to generate responses consistent with the augmented information (Lewis et al., 2020); and (b) strategic orchestration of multiple LLM-based agents, each responsible for a relatively smaller sub-task of the overall ADE Extraction task (Wu et al., 2023). Specifically, our system has agents for these sub-tasks (see Figure 1): (1) identifying representative drugs for each drug category from a medical database (*e.g.*, MIMIC-IV), (2) gathering information on side effects of those drugs from external text knowledge bases (*e.g.*, FDA drug label database), and finally, (3) composing final answers summarizing the effect of the drug category on an adverse event. Each agent is assigned a specific sub-task and collaborates with others to accomplish the the ultimate goal of ADE identification. Furthermore, we enhance the reliability of our multi-agent system even further by pairing each agent with a critic agent, whose role is to verify the behaviors and responses of its counterpart.

The system, though applied here for ADE extraction specifically, illustrates how a Multi-Agent approach can be used to generate trustworthy, evidence-based summaries and confidence scores in response to challenging medical questions requiring synthesis of evidence from multiple sources of clinical knowledge and data. As such, MALADE may be viewed as a case study illustrating an approach that could later be applied to other problems in PhV, including identification of possible drug-to-drug interactions, as well as clinical problems outside of PhV, such as identifying known symptoms of a condition of interest documented in clinical notes.

In summary, our paper makes the following contributions.

Precise Evaluation. In contrast to simpler systems that only produce a binary label indicating whether or not a drug category C is associated with an adverse event E, our method produces distinct scores, including a confidence score that indicates how confident an LLM is about its label assignment. These scores permit a rigorous quantitative evaluation against the well-established Observational Medical Outcomes Partnership (OMOP) Ground Truth table of ADEs associated with common drug classes (Madigan and Ryan, 2015). We achieve an Area Under the ROC Curve (AUC) of approximately 0.85 with GPT-4 Turbo, and 0.90 with GPT-4o (Section 5). To the best of our knowledge, this is the best performance among the baselines, even though the direct comparison may be limited [2].

Grounded generation of responses and justifications. The design of MALADE offers key features essential for high-stakes applications like ADE identification: (1) A struc-

---

1. Pronounced like the French word *malade* meaning “sick” or “ill.”
2. Because none of the original clinical data-based analyses reached this high of accuracy, followup investigations have since argued that roughly this level is the best achievable by any method based on any sources for the OMOP task. Gruber et al. (2016) argued there were reproducible errors that could be blamed on the OMOP 2010 ground truth itself that could place a ceiling on the AUC achievable. Hauben et al. (2016) more specifically argued that on the negative-labeled drug event pairs the error in the ground truth should be estimated at 17%. There may be disagreement on varying strengths of different literature evidence, but if their estimate is exactly right, it could place a ceiling as low as 0.83 on the AUC achievable.

tured format for drug-to-outcome associations, including scores indicating the strength of the association and rarity of the adverse event; this is important to ensure robust downstream processing of the extracted associations. (2) Justifications for the extracted drug-outcome associations, allowing human experts to understand and validate the associations. This is possible due to the RAG component of the MALADE architecture, which allows leveraging various external sources such as medical literature, drug labels, FDA tools (*e.g.*, OpenFDA drug information API), as well as common clinical data sources such as OMOP or PCORI, and even specific EHR systems where available. (3) Observability, *i.e.*, complete, detailed logs of inter-agent dialogs and intermediate steps; these are essential for debugging and auditing the system's behavior. See Figure 1 for a real-world demonstration of MALADE.

Generalizable Insights about Machine Learning in the Context of Healthcare. Our proposed multi-agent architecture is agnostic to LLMs and data sources and is based on design primitives intended to be universal building blocks for the orchestration of multiple LLM-based agents (Section 3). Hence, although MALADE is instantiated specifically for ADE identification, our design methodology provides a generalizable blueprint for the effective construction of multi-agent systems for trustworthy medical knowledge synthesis and summarization with wide-ranging medical applications.

2. Related Work

The advent of highly-capable Large Language Models (LLMs) has sparked significant interest in applying these models to medical tasks, including diagnostics (Singhal et al., 2023a), medical question-answering (Singhal et al., 2023b; Nori et al., 2023), and medical evidence summarization (Tang et al., 2023a). An important application area is pharmacovigilance, the science of identifying and preventing adverse drug events (ADEs) caused by pharmaceutical products after they are introduced to the market. The specific problem of ADE Extraction, namely, identifying whether a specific drug (or category) is associated with a specific adverse event, is a challenging task due to variations in drug and outcome terminologies, the presence of ADE descriptions in large amounts of narrative text, and the disparate sources of such text data, which can include patient medical records, clinical notes, drug labels, medical literature, clinical trials, message boards, social media. Prior research in this field, notably works drawing on large-scale research initiatives including Sentinel (Platt et al., 2009), OMOP (Ryan et al., 2013), and OHDSI (Stang et al., 2010), has focused on developing new methods for causal discovery from purely observational data. Huang *et al.* Huang et al. (2022) investigate the use of social forums for constructing predictive models of ADEs, focusing on the performance of different data processing techniques and BERT architectures. von Csefalvay von Csefalvay (2024) introduces a novel LLM, DAEDRA, for detecting regulatory-relevant outcomes from passive pharmacovigilance reports. Sorbello *et al.* Sorbello et al. (2023) use LLMs like GPT to improve the capture of opioid drug and adverse event mentions from electronic health records. Finally, Sun *et al.* Sun et al. (2024) investigate the performance of ChatGPT for extracting adverse events from medical text sources.

These early applications of LLMs to ADE Extraction are limited in at least one of two ways: (a) they either use only the bare LLM (such as ChatGPT, or its API) without access to any external APIs, tools, or knowledge bases (Wang et al., 2023a). ADE extraction using only the LLM's "built-in" knowledge (*i.e.*, text it was exposed to during pre-training) is likely to be inaccurate and incomplete, since adverse events may be discovered in any new studies or reports; (b) all prior works use a single LLM (even when augmented with external data/tools) without any collaboration or feedback from other LLMs. Since LLMs are after all probabilistic next-token prediction models, there is no guarantee that the generated text is accurate or complete. The only way to improve the reliability of an LLM's responses in this scenario is to either resort to elaborate prompting techniques (Wei et al., 2022; Yao et al., 2024), or have a human (or an LLM Madaan et al. (2023)) in the loop to verify the generated text and iteratively refine the prompts until a satisfactory response is obtained.

To address these limitations, three paradigms have emerged in LLM practitioners' toolboxes. The first limitation is addressed by two techniques: *Retrieval Augmented Generation* (RAG) and *tool-use*. RAG addresses the knowledge limitations of LLMs by augmenting the input prompt or query with relevant information retrieved from external knowledge bases (using similarity based on vector embeddings, keywords, or a combination of both), and instructing the LLM to respond to the original query in a way that is consistent with the augmented data, and also to provide a justification for its response by citing the relevant external data (Lewis et al., 2020). Thus the RAG approach not only alleviates the limitations of relying only on an LLM's pre-trained knowledge, but also provides evidence-citation ability, which is crucial to engender trust in the LLM's responses, especially in high-stakes applications like medical decision-making. This approach has shown promise in enhancing LLM performance in biomedicine, particularly in literature information-seeking and clinical decision-making (Frisoni et al., 2022; Jin et al., 2023; Wang et al., 2023b; Zakka et al., 2024). The second technique, tool-use, involves instructing the LLM to produce *structured* text (typically JSON) which can then be easily parsed by downstream code to perform a variety of actions, including web-search, querying APIs for information, querying databases, and performing computations (Ruan et al., 2023; Li et al., 2023).

The emergence of multi-agent systems addresses the second limitation (of using single LLMs) – this approach aims to harness the collective capabilities of multiple LLMs (Xi et al., 2023; Hong et al., 2024). Such systems introduce cooperative learning and feedback mechanisms between LLM-based agents, which simulate human-like communication, consultation and debate processes, enabling them to tackle even more complex tasks than a single-agent with RAG. In medical reasoning tasks, for instance, multi-agent collaboration can mirror hospital consultation mechanisms (Tang et al., 2023b). Our work extends this trajectory of research; to the best of our knowledge, our system MALADE is first effective multi-agent orchestration system with RAG and tool-use, tailored for a specific task in pharmacovigilance, namely ADE Extraction. In our approach, LLM-based agents collaborate, leveraging their collective expertise and the latest medical knowledge. This approach aims to improve the analysis of ADEs, offering a more robust and reliable system for pharmacovigilance.

3. Preliminaries on LLM-based Agents

While today's LLMs exhibit impressive capabilities, they remain constrained by technical and practical limitations such as brittleness, non-determinism, limited context window, inference costs, and latency (Liang et al., 2023), with the implication that one cannot simply give high-level instructions to an LLM and expect it to accomplish a complex task. Consequently, to best harness the capabilities of LLMs as components of a complex application, it is necessary to decompose the task into smaller sub-tasks and manage multiple LLM conversations, each with its own set of specifically-defined instructions, state, and data sources. This leads naturally to the notion of an *agent* as an LLM-powered entity responsible for a well-defined small sub-task. In Section 3.1, we introduce the key abstractions and components needed for agent-oriented programming, and Section 3.2 describes multi-agent orchestration. Our implementation leverages the open-source multi-agent LLM framework Langroid (Chalasani et al., 2023), which supports these abstractions and mechanisms.

3.1. Agent-oriented Programming

Agent, as an intelligent message transformer. A natural and convenient abstraction in designing a complex LLM-powered system is the notion of an *agent* that is instructed to be responsible for a specific aspect of the overall task. In terms of code, an `Agent` is essentially a class representing an intelligent entity that can respond to *messages*, *i.e.*, an agent is simply a *message transformer*. An agent typically encapsulates an (interface to an) LLM, and may also be equipped with so-called *tools* (as described below) and *external documents/data* (*e.g.*, a vector database, as described below). Much like a team of humans, agents interact by exchanging messages, in a manner reminiscent of the *actor framework* in programming languages (Hewitt, 2010). An *orchestration mechanism* is needed to manage the flow of messages between agents, to ensure that progress is made towards completion of the task, and to handle the inevitable cases where an agent deviates from instructions. In this work we adopt this *multi-agent programming* paradigm, where agents are first-class citizens, acting as message transformers, and communicate by exchanging messages.

To build useful applications with LLMs, we need to endow them with the ability to trigger actions (such as API calls, computations, database queries, etc) and access external documents. *Tools* and *Retrieval Augmented Generation* (RAG) provide these capabilities, described next.

Tools, also known as functions or plugins. An LLM is essentially a text transformer; *i.e.*, in response to some input text (known as a *prompt*), it produces a response. Free-form text responses are ideal when we want to generate a description, answer, or summary for human consumption, or even a question for another agent to answer. However, in some cases, we would like the responses to trigger external *actions*, such as an API call, code execution, or a database query. In such cases, we would instruct the LLM to produce a *structured* output, typically in JSON format, with various pre-specified fields, such as code, an SQL query, parameters of an API call, and so on. These structured responses have come to be known as *tools*, and the LLM is said to *use* a tool when it produces a structured response corresponding to a specific tool. To elicit a tool response from an LLM, it needs to be instructed on the expected tool format and the conditions under which it should use the tool. To actually use a tool emitted by an LLM, a *tool handler* method must be defined as well. The tool handler for a given tool is triggered when it is recognized in the LLM's response. See Appendix A.1 for an example describing the LLM's interaction with a database.

Starting with the view of an LLM as a text transformer, it turns out that one can express the notion of an agent, a tool, and other related concepts in terms of different function signatures, as shown in Table 4 in Appendix A.3.

Retrieval Augmented Generation (RAG). Using an LLM in isolation has two major constraints: (a) the responses are confined to the knowledge from its pre-training, hence cannot answer questions specific to private/enterprise documents, or up-to-date information past its training cutoff date; and (b) there is no way to verify the validity of the generated answers. RAG is the most popular technique to address both limitations by making LLMs generate responses based on specific documents or data and justify the answer by presenting source citations (Lewis et al., 2020). The basic idea of RAG is as follows: when a query Q is made to an LLM-agent, a set of k documents (or portions thereof) $D = \{d_1, d_2, \ldots, d_k\}$ most "relevant" to the query are *retrieved* from a document-store, and the original query Q is *augmented* with D to a new prompt of the form, "Given the passages below: [d_1, d_2, …, d_k], answer this question: Q based ONLY on these passages, and indicate which passages support your answer". See Appendix A.2 for more details on RAG.

3.2. Multi-Agent Orchestration

As mentioned above, when building an LLM-based multi-agent system, an orchestration mechanism is critical to manage the flow of messages between agents, to ensure task progress, and handle deviations from instructions. In this work, we leverage a simple yet versatile orchestration mechanism that seamlessly handles user interaction, tool handling, and sub-task delegation. The orchestration mechanism is encapsulated in a Task class that wraps an Agent, and one initiates a task by invoking its `run` method which has type signature `string` → `string`, identical to the type signature of an Agent's own native "response" methods (corresponding to the LLM, tool-handler, and human user). The Task maintains a "current pending message" (CPM) to be acted on by one of the "responders" of the Task, which include the agent's own response methods as well as `run` methods of sub-tasks. The `run` method executes a series of "steps" until a task termination condition is reached. In each step, a valid response to the CPM is sought by iterating over the responders, and the CPM is updated with the response. See Appendix A.4 for more details.

4. MALADE: Proposed Multi-Agent System for ADE Extraction

In this section, we describe our RAG-based Multi-Agent architecture, MALADE, for identifying associations between drug categories and outcomes. We first give a high-level outline of the objectives of the key sub-tasks in Section 4.1, and delve into their implementation details in Section 4.2 - 4.5. See Figure 1 for an illustrative depiction of the overall pipeline.

We emphasize that developing a multi-agent RAG system tailored for ADE extraction is a highly non-trivial undertaking, requiring careful handling of several issues: (a) the complex structure of FDA label data, which can be challenging for naively applied RAG techniques; (b) the difficulty of correctly grouping prescribed drugs (*e.g.*, assigning the appropriate National Drug Code) based on the varied text descriptions present in medical databases; (c) LLM brittleness such as deviation from instructions, hallucinations, and inaccurate or incorrect responses; (d) Infinite loops, fixed points, and deadlocks, which can arise in inter-agent interactions unless carefully managed.

4.1. Objectives of Key Sub-tasks

Our ultimate goal is to identify the risk of an adverse event associated with a drug *category*. We developed our system, MALADE, to be able to respond to questions of the form:

> "Does drug category C *increases* the risk of a specific (adverse) health outcome H, *decrease* it, or is there *no clear effect?*. And what is the evidence?" For instance, C could be "ACE inhibitors", and H could be "angioedema".

Given a query of this form, the system executes the following steps: given C and H,

STEP 1: Find the extensive list of drug names that belong to C by searching the FDA's National Drug Code (NDC) database. Among them, `DrugFinder` identifies drugs D representing each category; top-k distinct drug names that are most commonly prescribed in a clinical database (*e.g.*, MIMIC-IV).

STEP 2: For each representative drug D in C, `DrugAgent` generates a free-form (*i.e.*, unstructured) text summary about the effect of D on H. These summaries are generated by referring to up-to-date external pharmaceutical reference sources (*e.g.*, FDA drug label database), which indicate potential adverse outcomes and evidence for the risks.

STEP 3: `CategoryAgent` combines the drug-level information from STEP 2, and generates a structured report; consisting of a label (one of "increase", "decrease", or "no-effect") indicating the potential effect of C on the risk of H, a confidence score for this label, structured descriptions of levels of risk, and strength of evidence.

Our system extracts the association between C and H by establishing the associations between each drug within C and H, rather than directly linking C to H. This construction is motivated by that the reference sources for drug label data, such as the FDA drug label database in our implementation, are typically structured by individual drugs rather than broad drug categories; hence necessitating STEP 1. It is important to note that applying our system to real patient data requires a complete list of drugs, including both brand and generic names, which can be used to map the actual prescribed drugs recorded in electronic health record (EHR) data to their corresponding categories.

Each of `DrugFinder`, `DrugAgent`, and `CategoryAgent` is coupled with a Critic agent, which provides feedback on the primary agent's output. The primary agent then regenerates its output based on this feedback. This Agent-Critic interaction continues until the Critic approves the agent's response. This design pattern significantly enhances the reliability of our system, as detailed further in Section 4.2.

4.2. Agent-Critic Interaction

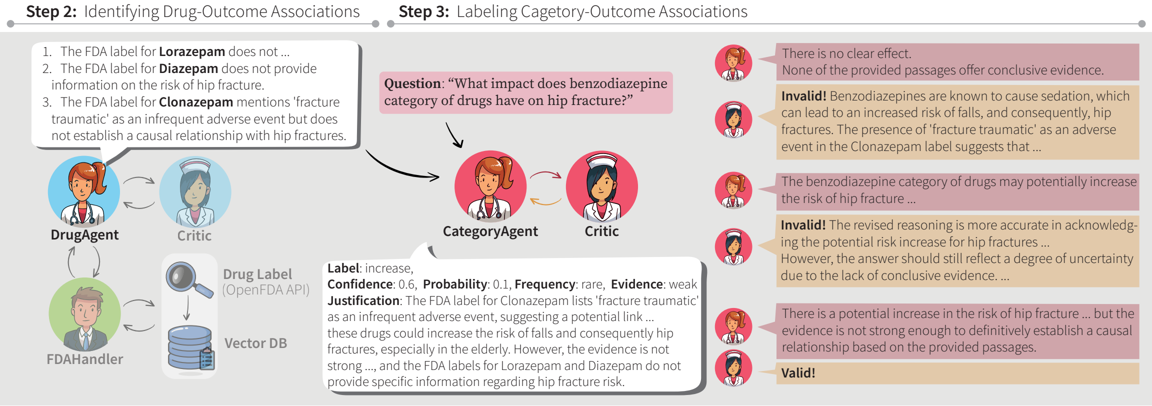
This is the core multi-agent interaction pattern that underlies our system, and is reminiscent of Actor/Critic methods in reinforcement learning (Konda and Tsitsiklis, 1999).

Agent. In an Agent-Critic pair, the Agent is the primary entity that handles external-facing input and output. It receives a specific goal, instructions on how to accomplish the goal, and access to tools and resources. In our context, the goal is generally a form of specialized question-answering; resources can be data sources, or even other agents or multi-agent systems, that the agent can draw upon when answering the question; tools are structured responses needed to trigger calls to APIs, database look-ups, or computations.

The primary function of the Agent is to construct a sequence of queries to these resources to fulfill its goal. The Agent is instructed to compose a semi-structured message consisting of its answer, its reasoning steps and a justification (citing sources where possible) of its answer in a semi-structured format, and seek feedback on these from the Critic, as below.

Critic. This is another agent, paired with the one described above. The Critic's role is to validate the Agent's reasoning steps and compliance with instructions, and provide feedback to the Agent, which has been shown to improve the quality of LLM-generated outputs Madaan et al. (2023). The Agent iterates on its response based on this feedback, until the Critic is satisfied, at which point the Agent signals completion and outputs the results (see Figure 2).

While the Agent/Critic pattern may appear simple, this interaction is extremely powerful, and can significantly improve the reliability of the task completion. This synergistic relationship mirrors a pattern in interactive proof systems used in complexity theory; a prover (*i.e.*, Agent) presents a solution, and a verifier (*i.e.*, Critic) checks the validity of this solution. The verifier cannot solve the problem on its own but is capable of checking the prover's solution efficiently, which is relatively easier (Babai, 1985). This way, even if the Agent's task is complex, the Critic can efficiently verify the correctness of the solution, thereby enhancing reliability.


Step 2: Identifying Drug-Outcome Associations
Step 3: Labeling Cagetory-Outcome Associations
DrugAgent
Critic
FDAHandler
Drug Label (OpenFDA API)
Vector DB
CategoryAgent
Critic

1. The FDA label for **Lorazepam** does not ...
2. The FDA label for **Diazepam** does not provide information on the risk of hip fracture.
3. The FDA label for **Clonazepam** mentions 'fracture traumatic' as an infrequent adverse event but does not establish a causal relationship with hip fractures.

Question: "What impact does benzodiazepine category of drugs have on hip fracture?"

Label: increase,
Confidence: 0.6, **Probability**: 0.1, **Frequency**: rare, **Evidence**: weak
Justification: The FDA label for Clonazepam lists 'fracture traumatic' as an infrequent adverse event, suggesting a potential link ... these drugs could increase the risk of falls and consequently hip fractures, especially in the elderly. However, the evidence is not strong ..., and the FDA labels for Lorazepam and Diazepam do not provide specific information regarding hip fracture risk.

There is no clear effect.
None of the provided passages offer conclusive evidence.

Invalid! Benzodiazepines are known to cause sedation, which can lead to an increased risk of falls, and consequently, hip fractures. The presence of 'fracture traumatic' as an adverse event in the Clonazepam label suggests that ...

The benzodiazepine category of drugs may potentially increase the risk of hip fracture ...

Invalid! The revised reasoning is more accurate in acknowledging the potential risk increase for hip fractures ...
However, the answer should still reflect a degree of uncertainty due to the lack of conclusive evidence. ...

There is a potential increase in the risk of hip fracture ... but the evidence is not strong enough to definitively establish a causal relationship based on the provided passages.

Valid!

Figure 2: **Real-world demonstration of Agent-Critic interactions in MALADE.** Given the question of identifying the association between *Benzodiazepines* and *Hip Fracture*, we illustrate how `CategoryAgent` corrects its answers over iterations until the paired Critic is satisfied. See Appendix 3.3 for full prompts between the two agents. Agents are instantiated using GPT-4 Turbo.

4.3. STEP 1: Finding Representative Drugs

We first construct a reasonably complete set of all drugs that can possibly belong to the category, by querying FDA's NDC database, which contains records of specific drugs, tagged with pharmacological class information of various types (including chemical classes, mechanisms of action, and established pharmacologic classes). Specifically, we extract all drugs with names or classes matching the relevant search term or terms (*e.g.*, "antibiotic" or any of the sub-categories considered by OMOP, for example, erythromycin). Since this list may contain some drugs that do not actually belong to the class (*e.g.*, a search for "typical antipsychotics" returns atypical antipsychotics as well), we rely on an additional filtering phase to construct the final, reasonably accurate list of drugs in the category. For each drug D in this "complete" list, we obtain its prescription rate via a SQL query to the MIMIC-IV prescriptions table.

Note that we chose to implement the above two SQL query steps directly without using an LLM to generate the queries. This is an instance of an important design principle we adhere to in our system, which we call the *LLM Minimization* principle: for tasks that can be expressed deterministically and explicitly in a standard programming paradigm, handle them directly without using LLMs to enhance reliability and reduce token and latency costs.

`DrugFinder`. Now that we have a reasonably complete list of candidate drug names that belong to the category of interest, along with their prescription rates, we want to identify three distinct, most commonly prescribed drugs that belong to the category. This task is complicated by several difficulties: the same drug may appear in this list with different names; some pairs of drugs may be essentially the same but only differ in formulation and delivery method, and a judgment must be made as to whether these are sufficiently different to be considered pharmacologically distinct; and some of these drugs may not actually belong to the category. This task thus requires a grouping operation, related to the task of identifying standardized drug codes from text descriptions, well known to be challenging (Le et al., 2024). Hence, this is very difficult to explicitly define in a deterministic manner that covers all cases (unlike the above database tasks), and hence is well-suited to LLMs, particularly those such as GPT-4 Turbo which are known to have been trained on vast amounts of medical texts in general (and drug-related ones in particular). We assign this task to the `DrugFinder`, which is an Agent/Critic system where the Critic agent helps improve the paired agent's output via iterative feedback; in particular, the Critic corrects the Agent when it incorrectly classifies drugs as pharmacologically distinct.

4.4. STEP 2: Identifying Drug-Outcome Associations

`DrugAgent` is an Agent/Critic system whose task is to identify whether a given drug has an established effect on the risk of a given outcome, based on FDA drug label database, and output a summary of relevant information, including the level of identified risk and the evidence for such an effect. This agent does not have direct access to the FDA Drug Label data, but can receive this information via another agent, `FDAHandler`. `FDAHandler` is equipped with tools to invoke the OpenFDA API for drug label data, and answers questions in the context of information retrieved based on the queries. Information received from this API is ingested into a vector database, so the agent first uses a tool to query this vector database, and only resorts to the OpenFDA API tool if the vector database does not contain the relevant information.

4.5. STEP 3: Labeling Drug Category-Outcome Associations

To identify the association between a drug category C and an adverse health outcome H, we concurrently run a batch of queries[3] to copies of `DrugAgent`, one for each drug D in the category, of the form: "Does drug D increase or decrease the risk of condition H?". The results are sent to `CategoryAgent`, described next.

`CategoryAgent` is an Agent/Critic system that performs the final classification step; its goal is to generate a label identifying whether a category of drugs increases or decreases the risk of a condition, or has no effect. In addition to the label, `CategoryAgent` produces a number of additional outputs, all of which are combined into a JSON-structured string, including: (a) a *confidence* score in [0,1], indicating the confidence in the assigned label, (c) *strength of evidence,* one of "none", "weak", or "strong", and (d) *frequency of the effect,* one of "none," "rare", or "common". In this sense, `DrugAgent` serves as a function of the following type: `[string]` → {`‘‘increase’’,‘‘decrease’’,‘‘no-effect’’`} × `[0,1]` × {`‘‘none’’,‘‘weak’’,‘‘strong’’`} × {`‘‘none’’,‘‘rare,’’,‘‘common’’`}. The structured output of `CategoryAgent` facilitates downstream post-processing to produce a final evaluation, with no further LLM involvement (Section 5.1).

5. Experiments

This paper presents MALADE, the first LLM-based multi-agent architecture that is capable of producing a structured report with characterizations and scores related to the risk of an adverse health outcome H from a drug category C, based on FDA drug label data. We evaluate our method against a widely used benchmark, the OMOP Evaluation Ground Truth task (Madigan and Ryan, 2015), henceforth referred to as the OMOP ADE task (Section 5.1), to answer the following three research questions:

RQ1: How effectively does MALADE identify ADEs? (Section 5.2)

RQ2: Does Agent-Critic interaction, the core design pattern underlying MALADE, effectively enhance the reliability of the system? (Section 5.3)

RQ3: What useful insights do the justifications by MALADE provide for further system improvement? (Section 5.4)

5.1. Evaluation Setup

The objective of OMOP ADE task is to assign one of three labels ("increase," "decrease," and "no-effect") to each (C, H) pair, denoting whether C increases, decreases, or has no effect on the risk of H, respectively. There are 10 drug categories, some of which consist of a single drug, and 10 health outcomes (refer to Table 6 for the complete list). Notably, while only three labels are valid outputs, not all (C, H) pairs are deemed sufficiently certain to be

---

[3] For any OMOP drug categories which contain multiple sub-categories, we execute the full process for each sub-category (identifying a set of representatives for each sub-category), merging the outputs of the classification agent, taking the highest risk indicated for any sub-category as the risk for the full category.

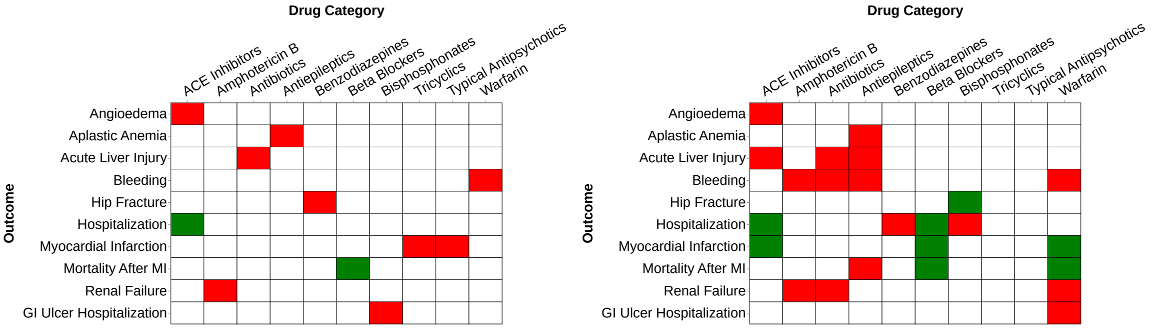
used in the evaluation. The authors of OMOP ADE task mark certain pairs as *uncertain*, to which we assign "no-effect" labels with the special restriction that it should not be used in the evaluation. See Appendix B.3 for further details.

Metrics. For quantitative evaluation, we convert the task into binary classification with two different focuses of analysis: (1) classifying effect vs no-effect, where the labels "increase" and "decrease" are considered the positive class, and "no-effect" is the negative class (namely, **effect-based** classification); and (2) classifying ADE vs. non-ADE, where only "increase" is considered the positive class, and the other two labels are the negative class (namely, **ADE-based** classification). For both choices, we report AUC and F1 scores, which are common evaluation metrics for binary classification (Madigan and Ryan, 2015; Bayer et al., 2021). Corresponding to the above two binary-classification methods, this results in "effect-based AUC, F1" and "ADE-based AUC, F1." The AUC metric captures how well the scores produced by MALADE discriminate between the classes, while the F1 score assesses the accuracy of the assigned labels in classifying both positive and negative instances.

5.2. RQ1: MALADE effectively identifies ADEs

In the evaluations of MALADE, we consider two LLMs, GPT-4 Turbo and GPT-4o. For GPT-4o, we limit the number of rounds of feedback from Critics to 5, after which it is required to accept. Figure 3 compares the ground truth labels of OMOP ADE task with ADE labels identified by MALADE (with GPT-4 Turbo). Considering the uncertainty inherent in the label of certain (drug category, outcome) pairs (Madigan and Ryan, 2015), these indicate strong performance on the task. See Figure 10 of Appendix B for results on GPT-4o. We also present the confusion matrix of the MALADE labels in Figure 4.

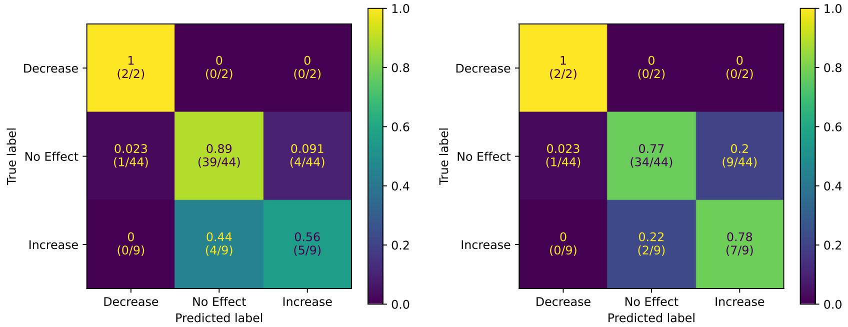
Moreover, we report the performance of MALADE in terms of AUC and F1 metrics (see Table 1). Recall that `CategoryAgent` outputs a confidence score ranging from 0 to 1 for its predicted labels, namely "increase," "no-effect," or "decrease.". This score reflects the agent's certainty regarding the accuracy of the predicted outcome. For quantitative evaluation as in Table 1, we transform these tripartite label-confidence scores into binary classification probabilities, suitable for effect-based or adverse drug event (ADE)-based analysis.



	ACE Inhibitors	Amphotericin B	Antibiotics	Antiepileptics	Benzodiazepines	Beta Blockers	Bisphosphonates	Tricyclics	Typical Antipsychotics	Warfarin
Angioedema	🟥									
Aplastic Anemia				🟥						
Acute Liver Injury			🟥							
Bleeding										🟥
Hip Fracture					🟥					
Hospitalization	🟩									
Myocardial Infarction								🟥	🟥	
Mortality After MI						🟩				
Renal Failure		🟥								
GI Ulcer Hospitalization							🟥			

	ACE Inhibitors	Amphotericin B	Antibiotics	Antiepileptics	Benzodiazepines	Beta Blockers	Bisphosphonates	Tricyclics	Typical Antipsychotics	Warfarin
Angioedema	🟥									
Aplastic Anemia				🟥						
Acute Liver Injury	🟥		🟥	🟥						
Bleeding		🟥	🟥	🟥						🟥
Hip Fracture							🟩			
Hospitalization	🟩				🟥	🟩	🟥			
Myocardial Infarction	🟩					🟩				🟩
Mortality After MI				🟥		🟩				🟩
Renal Failure		🟥	🟥							🟥
GI Ulcer Hospitalization										🟥

Figure 3: **Ground truth (left) vs. predictions by MALADE (right) for OMOP ADE task.** Red, green, and white cells represent "increase", "decrease", and "no-effect" labels, respectively.



(a) With GPT-4 Turbo [4]

(b) With GPT-4o.

Figure 4: **Confusion matrix for MALADE.**

Converting the three-class labels to a binary format requires a clear method for correlating each confidence score with a probabilistic value for the respective binary classification task.

The three labels exhibit a natural progression: “decrease”, “no-effect”, and “increase” imply an *ascending* likelihood that a drug category is associated with the adverse outcome of interest, signifying a rising probability score for the positive class in ADE-based classification. Furthermore, an *increase* in confidence of ”no-effect” or ”decrease” corresponds to a *decrease* in the ADE score, while an *increase* in confidence of the ”increase” label corresponds to an *increase* in the ADE score. These observations guide us in formulating an intuitive conversion of the label-confidence scores into ADE probability scores; taking $(1 - c_{\text{de}})/3, (2 - c_{\text{no}})/3$, and $(2 + c_{\text{in}})/3$, respectively, where $c_{\text{de}}, c_{\text{no}}$, and c_{in} are the LLM output confidence score when the assigned label is “decrease”, “no-effect”, and “increase”, respectively. This transformation preserves the semantic ordering of the classes, as well as the valence of confidence in each class. To illustrate, increasing confidence in “decrease” or “no-effect” suggests that the LLM is *less* confident that C causes H. We derive an effect-score similarly, except that both “increase” and “decrease” are now positive classes; taking $(1 - c_{\text{no}})/2$ and $(1 + c_{\text{in/de}})/2$, respectively.

The results in Table 1 indicate that the confidence scores output by the model are well-calibrated. We observe that MALADE performs well both at distinguishing ADEs from non-ADEs and at identifying the presence/absence of an effect in general. We include ROC curves and sensitivity vs. specificity curves in Figure 11 and Figure 12 of Appendix B, respectively. We conduct experiments with additional scoring functions, in particular, the model’s estimates of the *probabilities* that C will cause or prevent H; see Appendix B.1.

5.3. RQ2: Agent-Critic interaction enhances reliability

Our primary tool to analyze the effectiveness of the Agent-Critic pattern in MALADE is by ablation; in particular, we evaluate modified versions of MALADE, with and without

---

[4] We observe that GPT-4 Turbo tends to assign “increase” rather confidently even when the evidence is weak. To further enhance the reliability of the assigned labels, we take an additional postprocessing step; replacing *unreliable* predictions with “no-effect”. See Appendix B.4 for detailed discussions on label postprocessing.

Model	Metric	Effect-based	ADE-based
GPT-4o	AUC with confidence	0.883	0.903
GPT-4o	F1 score	0.600	0.560
GPT-4 Turbo	AUC with confidence	0.831	0.851
GPT-4 Turbo	F1 score [4]	0.609	0.556

Table 1: **Quantitative evaluation of MALADE.** "Effect-based" captures the classification between the presence and the absence of any ADE, while "ADE-based" represent's the ability of MALADE to distinguish drugs with increased risk from those with decreased risk or no effect.

Critics	RAG	Confidence AUC		F1 Score	
		ADE	Effect	ADE	Effect
✓	✓	0.851	0.831	0.556	0.609
×	✓	0.825	0.819	0.556	0.609
✓	×	0.924	0.929	0.526	0.609
×	×	0.920	0.926	0.556	0.636

Table 2: **Ablation results on MALADE.**

Agent	Correction
`DrugAgent`	4.24 %
`CategoryAgent`	44.52 %

Table 3: **Percentage of agent responses corrected by the Critic.**

feedback from the Critic components of `DrugAgent` and `CategoryAgent`, with and without RAG for FDAHandler. The results are shown in Table 2.

We observe that, both in the case with and without RAG, Critics improve the quality of the confidence scores, increasing both ADE-based and Effect-based AUCs. We additionally observe strong performance without RAG (in which case Critics slightly improve AUCs but decrease F1 scores), suggesting that GPT-4's internal medical knowledge is frequently sufficient for the OMOP ADE task. However, to ensure that MALADE is a realistic prototype for future pharmacovigilance systems, we consider only instances of MALADE with RAG for our main analysis; LLM-based systems without RAG are prone to hallucinations, and are limited by a static pool of information to draw upon. They lack the ability to produce citations, which is vital for trust in these systems, particularly in the medical domain. Integrating RAG enables the system to access and leverage the most current information from (for example) FDA label data, ensuring the system's responses are grounded with the up-to-date knowledge available. Refer to Appendix D for further details on ablation results and discussions.

We continue our investigation of the effectiveness of Agent-Critic interaction by analyzing the frequency of Critic interventions to rectify errors in Agent responses. We identify corrections made by the Critic as examples in which the Agent and Critic engaged in more than one round of interaction. Results are shown in Table 3.

We find that the frequency with which the Critic catches a flaw varies significantly by Agent. `CategoryAgent` in particular incurs errors, necessitating the help of the Critic and is generally corrected due to flaws in its medical reasoning, hence the Critic can directly prevent an incorrect response. In the example of an actual run of MALADE(Figure 2), when asked about the effects of benzodiazepines on hip fracture, `CategoryAgent` first answered "no-effect", which was flagged as an error, as the sedative and muscle relaxant properties of benzodiazepines can increase the risk of falls and hence hip fractures, and as `DrugAgent` had noted that traumatic fractures were listed as an ADE in the drug labels. This feedback was forwarded to `CategoryAgent` and used to revise its answer to "increase". We find that `DrugAgent` generally produces reliable responses; however note that it occasionally makes no calls to the Critic, hence the Agent fails to validate its answer. We observe that this can occur when the FDA drug label does not contain information related to the condition, and the Agent concludes that no validation is necessary.

5.4. RQ3: MALADE provides justifications that are aligned with human expert reasoning, and help understand its failure modes

We extract the justifications produced by `CategoryAgent` from a full run of MALADE for OMOP ADE task for review by a clinician. We observe that the agent exhibits valid medical reasoning in most cases, in particular, 85% of its justifications align with the reasoning of the clinician.

More importantly, examining the provided justifications helps us understand the common patterns of failures and provides guidance on the further improvement of the system. For instance, `CategoryAgent` occasionally assigns "increase" to drug categories based on weak evidence, overestimating its strength It may also overlook risks not explicitly mentioned in the drug label data, particularly when `DrugAgent` fails to provide sufficient context. In addition, `CategoryAgent` may fail to identify potential therapeutic effects not specified in the drug label data in association with a condition. We observe that it does not recognize the antihistamine properties of tricyclic antidepressants. In one case, evidence against gastric and duodenal ulcers caused by alendronate led `CategoryAgent` to dismiss results regarding esophageal ulcers.

While MALADE exhibits correct medical reasoning in general and hence achieves strong and reliable performance on ADE identification, we highlight that understanding its failures is essential for its further improvements, as discussed in Section 6. Extracts from the logs showing both correct and incorrect behavior by MALADE are in Appendix C. See Appendix B.5 for a discussion of the justifications produced by `DrugFinder`.

6. Discussion

Generalizable insights about collaborative LLM-powered agents in the context of healthcare. We have observed the strong performance of MALADE for ADE extraction, indicating the potential of multi-agent systems toward broader PhV application. Importantly, the principles guiding the design of our system, including 1) the Agent-Critic interaction, 2) the decomposition of a complex task into sub-tasks, and 3) LLM minimization, are quite general. These principles extend beyond PhV, and can be applied to many other problems in clinical medicine which require trustworthy, automated responses to challenging questions that must be answered based on multiple competing, and potentially conflicting, sources of knowledge or data. Thus, MALADE may be viewed not only as a system for ADE extraction, but also as a roadmap for development of other multi-agent systems that generate precise, evidence-based responses to such questions.

General principle 1) Agent-Critic interaction. The Agent/Critic pattern, as discussed in Section 4, is essential to the design of our system, and serves as a powerful tool to enhance accuracy of an LLM-based system. Indeed, we have observed several instances where the Critic corrected the parent Agent's initial response, as in the example mentioned in Section 4.2. However, we should note that if improperly configured, Critics can be harmful to the performance of a system, both in terms of efficiency (since the repeated rounds of interaction between the Agent and Critic can significantly increase token cost and runtime), and reliability. Since a Critic strictly enforces the provided guidelines, incorrect guidelines can significantly harm performance; in some cases, excessively strict requirements can lead to infinite loops, as the Agent and Critic will deadlock, neither able to satisfy the other's requirements. We observed this effect in early versions of MALADE; resolving the infinite loop issue required specific instructions listing acceptable behavior. For instance, the Critic for `DrugAgent` needed to be explicitly told to accept statements that the effect of a drug was uncertain due to a lack of information from the FDA labels; without this, infinite loops occurred in some drug-outcome combinations.

General principle 2) Decomposition of a complex task. The principle of decomposition, mirroring the analogous principle of general software development, is the Unix philosophy as applied to multi-agent systems. Individual agents should be minimal, in that they should "do one thing and do it well". This decomposition principle is evident in the hierarchy of specialized agents in the design of MALADE (*i.e.*, `DrugFinder`, `DrugAgent`, and `CategoryAgent` taking charge of each sub-task in Section 4.3 - 4.5). In addition to promoting modularity and maintainability, decomposition also promotes reliability, especially when combined with another key design principle, LLM only when necessary.

General principle 3) LLM only when necessary. As LLMs have surprising capabilities, one might be tempted to take an "LLM-maximalist" approach, where LLMs are responsible for all aspects of the task. Unfortunately, this can be both costly and unreliable, since using proprietary LLMs (*e.g.*, GPT-4) behind paid APIs incurs a significant "token cost" as well as "time cost" (due to the latency of the responses API calls). Instead, we carefully identified deterministic, well-defined algorithmic parts of the task and performed these in standard code. We relied on LLM-powered agents only for the specific tasks requiring language understanding, reasoning, and text generation. This principle guides key choices in MALADE: for instance, to retrieve prescription frequencies of drugs in a category, instead of having an LLM generate the needed SQL queries to the MIMIC-IV database, we observed that these queries are a simple function of the list of drugs, and directly generated the query in standard code. A similar choice was made for the FDA API queries to retrieve drug labels.

Such "LLM only when necessary" principle also illustrates the key utility of tool-use (also known as function-calling): in addition to providing the LLM the ability to perform external actions and to retrieve external data, it allows offloading execution of complex code from the LLM, hence dramatically reducing cost and increasing reliability. A multi-agent orchestration system, in this sense, can be seen as *control flow* for the LLM.

Limitations and Future Work. One key limitation of MALADE is that we rely entirely on textual FDA label data. In particular, if the information is not specifically included in the label data, MALADE cannot reliably identify the strength of any associations raised in the data. This resulted in several flawed predictions, as discussed in Section 5.4. To remedy this, we envision that extracting ADEs from EHR data is a promising direction for future work. As a first step, this would enable estimating the rarity of certain adverse events noted without further detail in the label data; in principle, a multi-agent system with access to EHR data may be able to identify ADEs directly. This would require the LLM to perform causal discovery from historical data (answering, "Is the drug causing this event?").

Another interesting avenue for future work is a detailed evaluation with local, open-source LLMs such as LlaMA (Touvron et al., 2023), Grok (xAI), and Mistral (Jiang et al., 2023), which have privacy and cost advantages over the proprietary LLMs (such as GPT-4, Claude, and variants) behind paid APIs. Unfortunately, our initial experiments with local LLMs exhibited many failure modes due to deviation from instructions and incorrect tool use. These are in principle possible to remedy by further breaking down tasks into simpler subtasks, and more sophisticated multi-agent validation and correction mechanisms.

Besides these broad limitations and avenues of future work, a few specific improvements are possible. Our system requires some minimal human input at the initial step, in particular, the names of the drug categories must be put into the form expected by the FDA's databases; in particular, acronyms are expanded and plurals and punctuation are removed. This task is quite likely amenable to LLMs, which are capable of acronym identification, and could attempt additional transformations for more robust output (for example, identifying all synonyms of a pharmacological class; the union of the drugs identified with each search would then be forwarded to `DrugFinder`). In addition, increased usage of structured input and output is a potential enhancement; for example, `DrugAgent`'s reliability might be enhanced by replacing the free-form text output, using instructions enforcing the presence of certain information, such as the reliability of information and the risk.

7. Conclusion

We consider the problem of ADE Extraction from FDA Drug Labels, a key task in Pharmacovigilance (PhV), and propose a solution using MALADE, based on collaboration among multiple LLM-powered agents equipped with Retrieval Augmented Generation (RAG). Our system goes significantly beyond simplistic techniques that only produce a binary label of presence/absence of association between a drug category and an ADE: it produces a structured report containing justification for the generated label, and scores characterizing probability of occurrence, confidence, strength of evidence, and rarity of the association between a drug category and an ADE. The scores permit rigorous quantitative evaluation of the system's performance against the widely-used OMOP Ground Truth table of ADEs, and the results are impressive. We introduce the agent/critic pattern, a powerful and general design pattern for reliable multi-agent systems. We hope that our multi-agent architecture and guiding principles will inspire future work on multi-agent approaches to broader PhV and general medical tasks.

References

László Babai. Trading group theory for randomness. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 421–429, 1985.

Samuel Bayer, Cheryl Clark, Oanh Dang, John Aberdeen, Sonja Brajovic, Kimberley Swank, Lynette Hirschman, and Robert Ball. Ade eval: an evaluation of text processing systems for adverse event extraction from drug labels for pharmacovigilance. *Drug safety*, 44:83–94, 2021.

Prasad Chalasani, Nils Palumbo, Mohannad Alhanahnah, and Somesh Jha. Langroid: Multi-agent framework for llm applications. https://github.com/langroid/langroid, 2023.

Jeff Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5770–5793, 2022.

Susan Gruber, Aloka Chakravarty, Susan R Heckbert, Mark Levenson, David Martin, Jennifer C Nelson, Bruce M Psaty, Simone Pinheiro, Christian G Reich, Sengwee Toh, et al. Design and analysis choices for safety surveillance evaluations need to be tuned to the specifics of the hypothesized drug–outcome association. *Pharmacoepidemiology and drug safety*, 25(9):973–981, 2016.

Manfred Hauben, Jeffrey K Aronson, and Robin E Ferner. Evidence of misclassification of drug–event associations classified as gold standard 'negative controls' by the observational medical outcomes partnership (omop). *Drug safety*, 39:421–432, 2016.

Carl Hewitt. Actor model. *arXiv:1008.1459 [cs.PL]*, Aug 2010. doi: 10.48550/arXiv.1008.1459. URL https://doi.org/10.48550/arXiv.1008.1459. arXiv: 1008.1459.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Jhih-Yuan Huang, Wei-Po Lee, and King-Der Lee. Predicting adverse drug reactions from social media posts: Data balance, feature selection and deep learning. *Healthcare*, 10(4), 2022. ISSN 2227-9032. doi: 10.3390/healthcare10040618. URL https://www.mdpi.com/2227-9032/10/4/618.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Qiao Jin, Robert Leaman, and Zhiyong Lu. Retrieve, summarize, and verify: how will chatgpt affect information seeking from the medical literature? *Journal of the American Society of Nephrology*, 34(8):1302–1304, 2023.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Huyen Le, Ru Chen, Stephen Harris, Hong Fang, Beverly Lyn-Cook, Huixiao Hong, Weigong Ge, Paul Rogers, Weida Tong, and Wen Zou. Rxnorm for drug name normalization: a case study of prescription opioids in the fda adverse events reporting system. *Frontiers in Bioinformatics*, 3, 2024. ISSN 2673-7647. doi: 10.3389/fbinf.2023.1328613. URL https://www.frontiersin.org/articles/10.3389/fbinf.2023.1328613.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-bank: A comprehensive benchmark for tool-augmented LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3116, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.187. URL https://aclanthology.org/2023.emnlp-main.187.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

David Madigan and Patrick Ryan. Learning from observational databases: Lessons from omop and ohdsi. 2015. URL https://www.niss.org/sites/default/files/Session3-DaveMadigan_PatrickRyanTalk_mar2015.pdf.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Justin Mower, Devika Subramanian, and Trevor Cohen. Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications. *Journal of the American Medical Informatics Association*, 25(10):1339–1350, 2018.

Sriraam Natarajan, Vishal Bangera, Tushar Khot, Jose Picado, Anurag Wazalwar, Vitor Santos Costa, David Page, and Michael Caldwell. Markov logic networks for adverse drug event extraction from text. *Knowledge and information systems*, 51:435–457, 2017.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Richard Platt, Marcus Wilson, K Arnold Chan, Joshua S Benner, Janet Marchibroda, and Mark McClellan. The new sentinel network—improving the evidence of medical-product safety. *New England Journal of Medicine*, 361(7):645–647, 2009.

Jingqing Ruan, YiHong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Hangyu Mao, Ziyue Li, Xingyu Zeng, Rui Zhao, et al. Tptu: Task planning and tool usage of large language model-based ai agents. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

Patrick B Ryan, Martijn J Schuemie, Emily Welebob, Jon Duke, Sarah Valentine, and Abraham G Hartzema. Defining a reference set to support methodological research in drug safety. *Drug safety*, 36:33–47, 2013.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023a.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023b.

Alfred Sorbello, Syed Arefinul Haque, Rashedul Hasan, Richard Jermyn, Ahmad Hussein, Alex Vega, Krzysztof Zembrzuski, Anna Ripple, and Mitra Ahadpour. Artificial intelligence–enabled software prototype to inform opioid pharmacovigilance from electronic health records: Development and usability study. *JMIR AI*, 2:e45000, Jul 2023. ISSN 2817-1705. doi: 10.2196/45000. URL https://ai.jmir.org/2023/1/e45000.

Paul E Stang, Patrick B Ryan, Judith A Racoosin, J Marc Overhage, Abraham G Hartzema, Christian Reich, Emily Welebob, Thomas Scarnecchia, and Janet Woodcock. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Annals of internal medicine*, 153(9):600–606, 2010.

Zhaoyue Sun, Gabriele Pergola, Byron Wallace, and Yulan He. Leveraging ChatGPT in pharmacovigilance event extraction: An empirical study. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–357, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-short.30.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158, 2023a.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023b.

Mert Tiftikci, Arzucan Özgür, Yongqun He, and Junguk Hur. Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels. *BMC bioinformatics*, 20: 1–9, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Chris von Csefalvay. Daedra: A language model for predicting outcomes in passive pharmacovigilance reporting. *arXiv preprint arXiv:2402.10951*, 2024.

H. Wang, Y.J. Ding, and Y. Luo. Future of chatgpt in pharmacovigilance. *Drug Safety*, 46:711–713, 2023a.

Yubo Wang, Xueguang Ma, and Wenhu Chen. Augmenting black-box llms with medical textbooks for clinical question answering. *arXiv preprint arXiv:2309.02233*, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

xAI. grok-1. URL https://github.com/xai-org/grok-1.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Rita Lee, Joanna Melia, Joanna Nelson, Karim Sallam, Stacey Tullis, Melissa Ann Vogelsong, John Patrick Cunningham, and William Hiesinger. Almanac — retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068, 2024.

Supplementary Material

Section A includes an in-depth description of the core primitives of our multi-agent framework. Section B offers the experimental details, including the system prompts for each agent, the details on our OMOP evaluation, and discussions of the postprocessing of the generated scores and justifications. In Section C, we analyze both successful and unsuccessful instances of MALADE, presenting comprehensive logs for selected examples. Section D presents an ablation study that evaluates the individual contributions of key components to the overall system efficacy; namely, the iterative refinement facilitated by Agent-Critic interactions, and the integration of external knowledge through RAG. Finally, in Section E, we assess how much the variance of numerical outputs by the random sampling of LLMs affects the variance of scores output by the entire MALADE system.

Appendix A. Agent-Oriented Programming

This section describes the core abstractions needed to implement a complex LLM-based application such as MALADE. The open-source multi-agent LLM framework langroid (Chalasani et al., 2023) has an elegant, intuitive and flexible implementation of these abstractions, and MALADE is built on top of this library.

A.1. Tool Use: Example

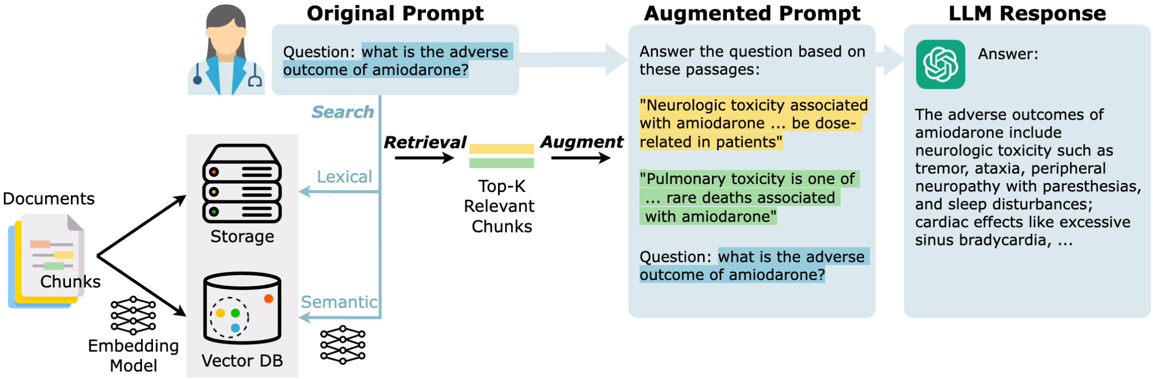
As a simple example, a SQL query tool can be specified as a JSON structure with a `sql` field (containing the SQL query) and a `db` field (containing the name of the database). The LLM may be instructed with a system prompt of the form:

```
When the user asks a question about employees,
use the SQLTool described in the below schema,
and the results of this tool will be sent back to you, and you can use these
to respond to the user's question, or correct your SQL query
if there is a syntax error.
```

The tool handler would detect this specific tool in the LLM's response, parse this JSON structure, extract the `sql` and `db` fields, run the query on the specified database, and return the result if the query ran successfully, otherwise return an error message. Depending on how the multi-agent system is organized, the query result or error message may be handled by the same agent (*i.e.*, its LLM), which may either summarize the results in narrative form, or revise the query if the error message indicates a syntax error.

A.2. Retrieval Augmented Generation

RAG involves two phases: (a) a *ingestion* phase, where documents are sharded into reasonable-size chunks and ingested into a suitable type of document-store, and (b) a query phase, where top-k document-chunks most *relevant* to the query are retrieved from the document-store, and the LLM is prompted to answer the query given these chunks (see Figure 5 for illustrative description). Not surprisingly, the performance (*i.e.*, precision and recall of answers) of a RAG system depends critically on how we define the relevance of document chunks to the query so that they will contain sufficient information for the LLM to compose a reasonable



Question: what is the adverse outcome of amiodarone?

Answer the question based on these passages:

"Neurologic toxicity associated with amiodarone … be dose-related in patients"

"Pulmonary toxicity is one of … rare deaths associated with amiodarone"

Question: what is the adverse outcome of amiodarone?

Answer:

The adverse outcomes of amiodarone include neurologic toxicity such as tremor, ataxia, peripheral neuropathy with paresthesias, and sleep disturbances; cardiac effects like excessive sinus bradycardia, …

Figure 5: **A simple agent with RAG**. During the ingestion phase, documents are sharded into document chunks. At the query phase, top-k chunks most relevant to the original query are retrieved, based on lexical relevance and semantic relevance. Now we prompt the LLM with the augmented query to ground its response to the documents.

answer. In this work, we use a combination of two standard notions of relevance: (a) lexical relevance, which is based on word overlap between the query and the document-chunk (*i.e.*, keyword search), while (b) *semantic* relevance focuses on the similarity of "meaning". The latter is based on the intuition that specially-trained embedding models can encode text as fixed-length *embedding vectors* that roughly capture the "meaning" of the text, and thus two texts are considered semantically similar if their embedding vectors are "close" as measured by a metric such as cosine similarity (Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2019). During the ingestion phase, each document chunk is mapped to an embedding vector using an embedding model and this vector is indexed into a *vector database*, along with a pointer to the chunk contents as metadata. During the query phase, the same embedding model is used to map the query into a vector, and then the top-k nearest-neighbors of this vector (based on cosine similarity) are found from the vector database, and their corresponding document chunks are retrieved.

A.3. From LLM to Agent-Oriented Programming

If we view an LLM as a function with signature `string` → `string`, it is possible to express the concept of an agent, tool, and other constructs in terms of derived function signatures, as shown in Table 4.

A.4. Detailed Description of Multi-Agent Orchestration

When building an LLM-based multi-agent system, an orchestration mechanism is critical to manage the flow of messages between agents, to ensure task progress, and handle deviations from instructions. In this work, we leverage Langroid's simple yet versatile orchestration mechanism that seamlessly handles:

- user interaction
- tool handling

Function Description	Function Signature
LLM	`string` → `string` `string` for the original query.
Chat interface	`[string]` × `string` → `string` `[string]` is for previous messages[5].
Agent	`string` × `[string]` × `string` → `string` `string` is for system prompt.
Agent with tool	`string` × (`string` → `T`) × (`T` → `string`) × `[string]` × `string` → `string`
Parser with type `T`	`string` → `T`
Callback with type `T`	`T` → `string`
General Agent with state type `S`	`S` × `string` × (`string` → `T`) × (`S` × `T` → `S` × `string`) × `[string]` × `string` → `S` × `string`

Table 4: **From LLM to agent-oriented programming.** An LLM is essentially a message transformer. Adding "tool" (or function calling) capability to LLM requires a parser and a callback that performs arbitrary computation and returns a string. The serialized instances of `T` correspond to a language L; as, by assumption, the LLM is capable of producing outputs in L, this allows the LLM to express the intention to execute Callback with arbitrary instances of `T`. Finally, we incorporate state by making Agent and Callback transducers, and have the general form in the last row.

- sub-task delegation

Recall that we view an agent as a message transformer; it may transform an incoming message using one of its three "native" responder methods, all of which have the same function signature: `string` → `string`:

- `llm_response` returns the LLM's response to the input message. Whenever this method is invoked, the agent updates its dialog history (typically consisting of alternating user and LLM messages).
- `user_response` prompts the user for input and returns their response.
- `agent_response` by default only handles a "tool message," *i.e.*, one that contains an llm-generated structured response, performs any requested actions, and returns the result as a string. An `agent_response` method can have other uses besides handling tool messages, such as handling scenarios where an LLM "forgot" to use a tool, or used a tool incorrectly, and so on.

To see why it is useful to have these responder methods, consider first a simple example of creating a basic chat loop with the user. It is trivial to create such a loop by alternating between `user_response` and `llm_response` . Now suppose we instruct the agent to either directly answer the user's question or perform a web-search. Then it is possible that sometimes the `llm_response` will produce a "tool message", say WebSearchTool, which we would handle with the `agent_response` method. This requires a slightly different, and more involved, way of iterating among the agent's responder methods. From a coding perspective, it is useful to hide the actual iteration logic by wrapping an Agent class in a separate class, which we call a Task, which encapsulates all of the orchestration logic. Users of the Task class can then define the agent, tools, and any sub-tasks, wrap the agent in a task object of class Task, and simply call task.run(), letting the Task class deal with the details of orchestrating the agent's responder methods, determining task completion, and invoking sub-tasks.

The orchestration mechanism of a `Task` object works as follows. When a task object is created from an agent, a sequence of eligible responders is created, which includes the agent's three "native" responder agents in the sequence: `agent_response` , `llm_response` , `user_response` . The type signature of the run is `string` → `string`, just like the `Agent`'s native responder methods, and this is the key to seamless delegation of tasks to sub-tasks. A list of subtasks can be added to a `task` object via `task.add_sub_tasks([t1, t2, ... ])`, where `t1, t2, ...` are other `Task` objects. The result of this is that the run method of each sub-task is appended to the sequence of eligible responders in the parent task object.

A task always maintains a current pending message (CPM), which is the latest message "awaiting" a valid response from a responder. At a high level the run method of a task attempts to repeatedly find a valid response to the CPM, until the task is done. This is achieved by repeatedly invoking the step method, which represents a "turn" in the conversation. The step method sequentially tries the eligible responders from the beginning of the eligible-responders list, until it finds a valid response, defined as a non-null or terminating message (i.e. one that signals that the task is done). In particular, this step() algorithm implies that a Task delegates to a sub-task only if the task's native responders have no valid response.

There are a few simple rules that govern how step works: (a) a responder entity (either a native entity or a sub-task) cannot respond if it just responded in the previous step (this prevents a responder from "talking to itself", (b) when a response contains "DONE" the task is ready to exit and return the CPM as the result of the task, (c) when an entity "in charge" of the task has a null response, the task is considered finished and ready to exit, (d) if the response of an entity or subtask is a structured message containing a recipient field, then the specified recipient task or entity will be the only one eligible to respond at the next step.

Once a valid response is found in a step, the CPM is updated to this response, and the next step starts the search for a valid response from the beginning of the eligible responders list. When a response signals that the task is done (e.g. contains the special string "DONE"), the run method returns the CPM as the result of the task. This is a highly simplified account of the orchestration mechanism, and the actual implementation is more involved.

The above simple design is surprising powerful and can support a wide variety of task structures, including trees and DAGs. As a simple illustrative example, tool-handling has a natural implementation. The LLM is instructed to use a certain JSON-structured message as a tool, and thus the `llm_response` method can produce a structured message. This structured message is then handled by the `agent_response` method, and the resulting message updates the CPM. The `llm_response` method then becomes eligible to respond again, and the process continues.

Figure 6 shows a schematic of the task orchestration and delegation mechanism.

Appendix B. Detailed Descriptions on MALADE Implementation

B.1. Prompts to Each Agent

STEP1: finding representative drugs under each drug category. This is the full prompt to `DrugFinder`:

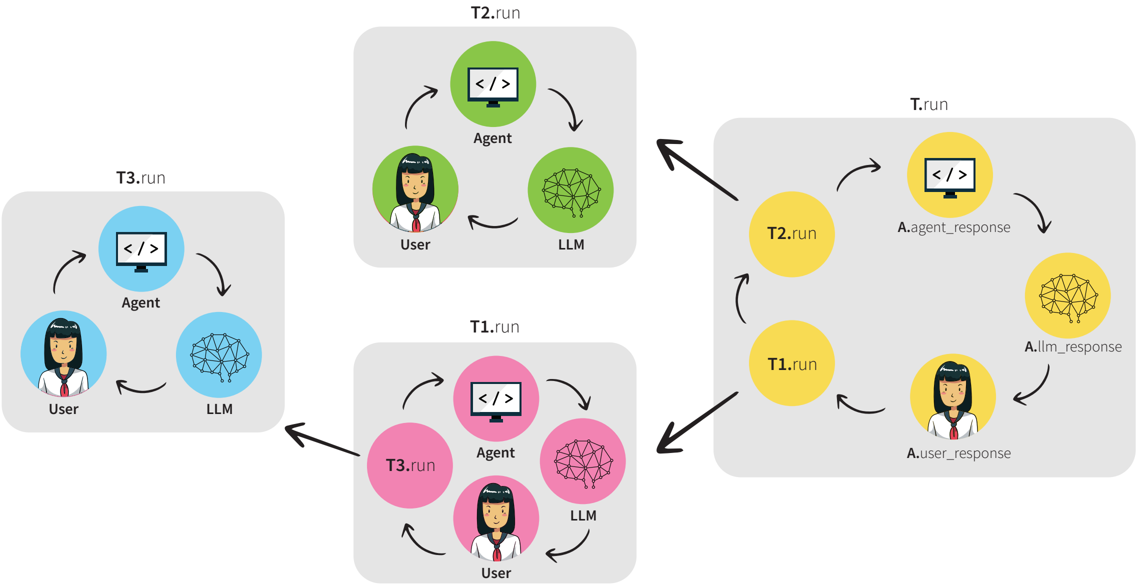


Figure 6: Example of how iteration among responder methods works when a task T has sub-tasks [T1, T2] and T1 has a sub-task T3.

```
You are a helpful assistant with general medical and pharmacological knowledge
    . I will provide you with a list of drugs, and the result of a query on a
    medical database with their usage rates; your goal is to find N
    representative drugs in category \{cat\} out of the provided drugs.

Prefer generic names if possible, and do not include both a brand and generic
    name for the same drug in your list.

If possible, prefer drugs with different active ingredients
(i.e. avoid derivatives of a drug already in the list),
keeping your choices to the most basic variant of a given drug
from the list (use the total prescription rate of variants of the same base
    drug to select the top drugs); disregard this if you cannot find N with
    this restriction. If fewer than N meet the conditions, you may include
    fewer than N (but never more).

The names of the selected representatives must EXACTLY match one of the
    provided drugs; choose the names from the original list, not the database
    query.

You must provide your final answer with the ‘final_answer‘ tool/function; make
     sure to clearly state my question, as well as the reasoning used to derive
     the answer. Include the requirements on your answer in the ‘question‘
    field.

Once the critic is satisfied with your answer, send me the answer with the ‘
    submit_answer‘ tool/function.
```

This is the full prompt to the Critic agent:

```
You are also an expert in medical and pharmacological reasoning.

Your goal is to ensure that the selected drugs are distinct members of the
    category \{cat\} of drugs. You will consider information provided directly
    to the user to be reliable (for example, this might include prescription
    rates and a complete list of drugs in category \{cat\}). Unless this
    contradicts your pharmacological knowledge, the user’s choices of
    representatives for a category are acceptable unless they do not represent
    the basic form of a given drug.
```

STEP2: identifying the interaction between each drug and each outcome. Below is the full prompt to `DrugAgent`.

```
You will receive questions involving medical data.
You are experienced in general medical reasoning, but must consult references
    for any specific medical knowledge required to answer my questions.

You have access to ‘FDAHandler‘, who will answer questions you ask about
    specific drugs using FDA data. You must use the ‘recipient_message‘ tool/
    function to ask these questions, and the ‘intended_recipient‘ MUST be ‘
    FDAHandler‘ anytime you use this tool.
Ensure that you ask FDAHandler for the specific information you need.

As some potential complications are listed in FDA labels as lacking a verified
     causal relationship, make certain that your final answer expresses the
    degree of reliability of your answer. Similarly, make sure to clearly
    express the degree of risk associated (i.e. is the condition a rare or a
    common side effect, or does a drug rarely or frequently result in reduced
    risk of a condition).

If FDAHandler cannot answer your question then your answer
should be {NO_ANSWER}, because the FDA label data does not
specify the answer. If FDAHandler answers with {NO_ANSWER}
that means that the FDA label for the drug does not
contain the information requested (and, in particular, it
means that it does not mention the condition); hence, your
answer should be {NO_ANSWER}. This indicates that there
may not be any effect on the risk of the condition, make sure to explain this
    in your justification.

IMPORTANT: if multiple attempts fail to retrieve any relevant information,
    there is no need to continue asking questions to FDAHandler; assume that
    the information is not in the FDA labels and so FDAHandler cannot answer.

You MUST specifically tell the critic why you could not
find an answer to the question; be sure to specify that
the FDAHandler answered with {NO_ANSWER} if that is the reason.

You must provide your final answer with the ‘final_answer‘
tool/function; make sure to clearly state my question, the
```

```
reasoning used to derive the answer, including the questions asked to
    FDAHandler and a summary of the results, as well as your final answer in
    the ‘answer‘ field.

Once the critic is satisfied with your answer, say {DONE},
and give me the answer and justification for it. Make sure
to provide your answer again, do not just use the answer
sent to the critic. Include any relevant details provided by FDAAgent.

If the critic is satisfied and your answer is {NO_ANSWER},
say {DONE} {NO_ANSWER} and provide a justification.
IMPORTANT: say {DONE} specifically, not DONE.
```

This is the full prompt to the Critic agent:

```
You are also experienced in medical reasoning, and have general medical
    knowledge. Unless the responses are inconsistent with your medical (or
    common-sense) knowledge, you generally trust responses from FDAHandler.

The answer should express the strength of evidence for the answer and the
    magnitude of the effect. If the user states that FDAAgent does not have
    this information, you should accept it.

If the answer given contains {NO_ANSWER}, accept it as long as the answer
    clearly expresses why it was not possible to answer the question. If it
    states that this is because FDAHandler responded with {NO_ANSWER}, you
    should accept it as sufficient justification.
Otherwise, ask the user to express whether FDAHandler responded with {
    NO_ANSWER}, and, if not, to state why it was not possible to answer the
    question. If it does so, the answer is acceptable and the other
    requirements need not be enforced.
```

In this case, the Critic agent similarly behaves as a medical expert; in general, the Critic must always behave as if proficient with any task that the orchestrator agent will do; this is specified as: “You are also experienced in medical reasoning, and have general medical knowledge. Unless the responses are inconsistent with your medical (or common-sense) knowledge, you generally trust responses from FDAHandler.”

It is told to trust the agents’ responses as any necessary validation of the responses from the two agents should happen on their side; the criticism should focus on the orchestrator itself.

Below is the full prompt to `FDAHandler`:

```
You will try your best to answer my questions, in this order of preference:

1. Ask me for some relevant text, and I will send you.
   Use the ‘relevant_extracts‘ tool/function-call for this purpose.
   Once you receive the text, you can use it to answer my question.
   If the question asks for information about a specific drug, make sure to
       begin by including that drug in the ‘filter_drugs‘ field. If I say {
       NO_ANSWER}, it means I found no relevant docs, and you can try the next
        step, using a web search.
```

```
2. If you are still unable to answer, you can use the `
    relevant_search_extracts` tool/function-call to get some text from a web
    search. Once you receive the text, you can use it to answer my question. If
     you need to identify the drugs in a category, use the `
    drug_category_search` tool/function-call instead.
3. If you are still unable to answer, and used `filter_drugs` in your initial
    attempt with `relevant_extracts`, try again without a filter.
4. If you still can't answer, simply say {DONE} {NO_ANSWER}

If given a question asking about a drug "X and Y", this is a
combination drug, so your initial searches should be for "X and Y" not "X" or
    "Y".

If asked a question about drugs in broad category, make to consider EVERY drug
     in the category, and in particular, if the question asks for which drugs
    in the category something is true, make CERTAIN that your answer correctly
    lists ALL drugs in the category where the condition holds.

IMPORTANT: some fields in the FDA label data retrieved
by `relevant_search_extracts' and `relevant_extracts` have the level of
    reliability of information specified prior to it (for example, statements
    of the level of reliability may precede each section of adverse reactions,
    the immediately preceding such statement is the one that corresponds to any
     given reported interaction). Make certain that your answer reflects the
    specified level of reliability.
Similarly, when asked about the effect of a drug on a condition, ALWAYS
    express the magitude of the effect (i.e. how frequently the drug results in
     the condition or how frequently the drug improves the condition); whenever
     possible, make sure to explicitly state whether a condition is rarely or
    commonly reported.

ANSWER FORMAT:

ALWAYS present your answer in one of the below 2 formats:

1. In case you COULD NOT find an answer:

{DONE} {NO_ANSWER}

2. In case you ARE able to find an answer:

{DONE}
ANSWER: [Your concise answer, with a brief summary of necessary context.
    ALWAYS clarify the level of reliability of the information, if specified in
     the extracts. If applicable, ALWAYS express the magnitude of any increase
    or decrease in risk and any associated information.]
SOURCE: aspirin label
EXTRACT_START_END: Aspirin can cause ... with any medicine.

For the EXTRACT_START_END, ONLY show up to the first 3 words and last 3 words.
```

STEP3: labeling the association between each drug category and each outcome. This is the full prompt to `CategoryAgent`:

```
You are experienced in general medical reasoning and have general medical knowledge.

You will be provided a list of passages answering, for each of a set of drugs X, whether drug X increases or decreases the risk of {condition}. They all belong to category {cat_name}.

You must provide your final answer with the ‘final_answer‘ tool/function; make sure to clearly state my question, the reasoning used to derive the answer,
including the evidence from the passages, as well as your final answer in the ‘answer‘ field.

Once the critic is satisfied, submit your answer with the ‘category_effect‘ tool, making sure that the answer, ‘label‘, is one of the following: "increase," "decrease," or "no-effect," and make sure to include your justification. DO NOT use this tool before you have used the ‘final_answer‘ tool and have had your answer accepted by the critic.

Your ‘justification‘ must clearly express the magnitude of risk indicated and the strength of evidence. Provide a ‘confidence‘ value between 0 and 1 indicating the confidence in your assigned ‘label‘ and a ‘probability‘ value indicating the probability that the drug will cause the condition (or prevent the condition) in a given patient.

Express the frequency that the drug has an effect as either "none," "rare," or "common" with the ‘frequency‘ field and express the strength of ‘evidence‘ as either "strong" (for example, evidence is strong when shown in a cal trial) or "weak" (for example, this applies to purely correlational evidence) or "none" if no evidence exists.
```

This is the full prompt to the Critic agent:

```
You are also experienced in medical reasoning, and have general medical knowledge. Unless the responses are inconsistent with your medical (or common-sense) knowledge, you generally trust responses from FDAHandler. Similarly, you trust that the user’s statements about passages are correct without the need to review them directly.

The answer provided should indicate an increase, decrease, or no effect on the risk, and must be no effect if no evidence linking the drug category to the risk of the condition exists.

The answer should be drawn from the specified passages, hence, the absence of information related to a condition in the FDA data for all drugs in a category should be enough to conclude that there is no effect for that drug.

The answer should express the degree of certainty and the magnitude of change in risk, ensure that the provided answer is consistent with the evidence.
```

B.2. Probability-based scoring

In addition to the confidence-based scoring discussed in Section 5, we consider probability-based scoring. In particular, we ask the model to specify the probability of an evant, specifically, the event that a drug in category C causes or prevents H.

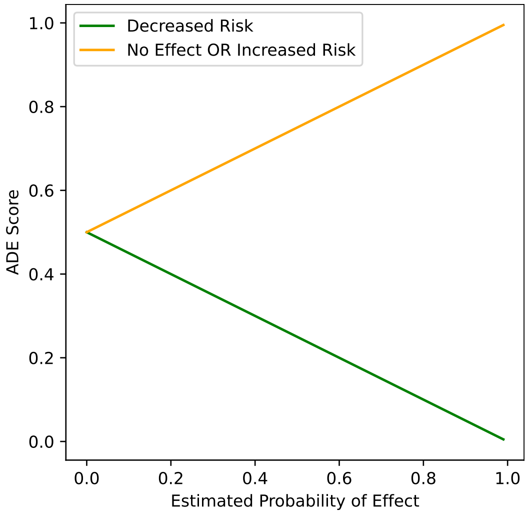


Figure 7: **Derivation of ADE Scores from Event Probability.** The x-axis represents the LLM's output probability estimate, and the color indicates the mapping for the corresponding set of labels.

As in Section 5, we must derive confidence scores in ADE and effects from the output probability estimate; as the probability is already in terms of the probability of any effect, either harmful or beneficial, we use the probability directly for Effect AUC. For ADE AUC, we use the tranformation shown in Figure 7.

We make the assumption that, if the LLM specifies a probability p with a label other than “decrease,” that probability expresses the probability of a harmful effect. Hence, the derived score decreases linearly with increasing probability when the label is “decrease,” and increases linearly when the label is anything else. With this assumption, we additionally maintain the semantic ordering of the LLM's implied confidence in ADE, and hence this is a well-defined confidence score.

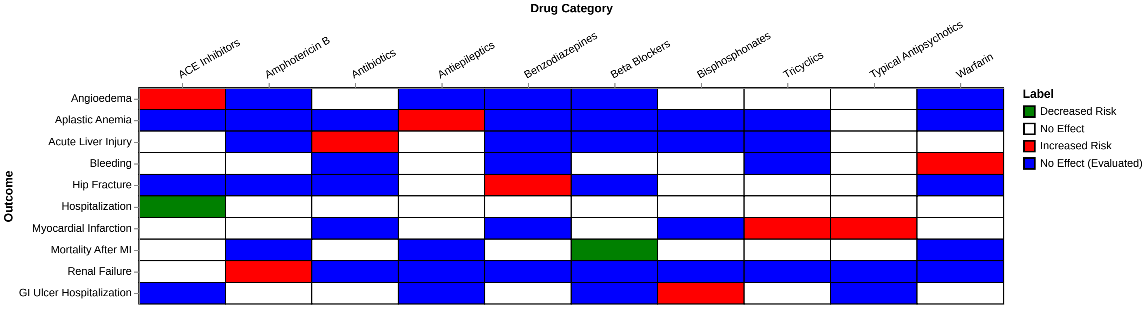
The results with probability-based scoring are shown in Table 5, we observe that the probabilities are less reliable (unsurprising as the FDA label data does not always contain the information necessary for a reliable estimate). See Appendix D for further discussion on the potential unreliability of the probability estimates.

B.3. OMOP ADE task Details

The ground truth for the OMOP ADE task is shown in Figure 8. As noted in Section 5.1, while the OMOP ADE task permits only three output labels for the effect of a drug category on an outcome, some drug category, outcome pairs are considered uncertain (which we treat as a “no-effect” label which is not used in evaluation). In Figure 8 the “No Effect” cells considered reliable are shown in blue, while the uncertain “no-effect” cells are those in white.

Model	Metric	Effect-based	ADE-based
GPT-4o	AUC with confidence	0.8833	0.9034
GPT-4o	AUC with probability	0.6715	0.6534
GPT-4 Turbo	AUC with confidence	0.8306	0.8514
GPT-4 Turbo	AUC with probability	0.8058	0.7935

Table 5: **Comparison of confidence and probability based scoring for MALADE.** “Effect-based” captures the classification between the presence and the absence of any ADE, while “ADE-based” represent’s the ability of MALADE to distinguish drugs with increased risk from those with decreased risk or no effect.



	ACE Inhibitors	Amphotericin B	Antibiotics	Antiepileptics	Benzodiazepines	Beta Blockers	Bisphosphonates	Tricyclics	Typical Antipsychotics	Warfarin
Angioedema	🟥	🟦	⬜	🟦	🟦	🟦	⬜	⬜	⬜	🟦
Aplastic Anemia	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	⬜	🟦
Acute Liver Injury	⬜	🟦	🟥	⬜	🟦	🟦	🟦	🟦	⬜	⬜
Bleeding	⬜	⬜	🟦	⬜	🟦	⬜	⬜	🟦	⬜	🟥
Hip Fracture	🟦	🟦	🟦	⬜	🟥	🟦	⬜	⬜	⬜	🟦
Hospitalization	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Myocardial Infarction	⬜	⬜	🟦	⬜	🟦	⬜	🟦	🟥	🟥	⬜
Mortality After MI	⬜	🟦	⬜	🟦	⬜	🟩	⬜	⬜	⬜	🟦
Renal Failure	⬜	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
GI Ulcer Hospitalization	🟦	⬜	⬜	🟦	⬜	🟦	🟥	⬜	🟦	⬜

Figure 8: OMOP ground truth.

In particular, the cells in white are *not* used in evaluation (*i.e.*, for AUC computation, confusion matrices, and F1 scores). The cells used for evaluation are shown in red, blue, and green.

Table 7 present the representatives selected for each category of drugs, respectively, produced by GPT-4 Turbo.

B.4. Effectiveness of Label Postprocessing with GPT-4 Turbo

In this subsection, we illustrate that the postprocessing of labels in Section 5.1 significantly improves the accuracy of ADE identification by MALADE instantiated with GPT-4 Turbo.

Drug Categories	ACE Inhibitors, Amphotericin B, Antibiotics (Erythromycin, Sulfonamide, Tetracycline), Antiepileptics (Carbamazepine, Phenytoin), Benzodiazepines, Beta blockers, Bisphosphonates (Alendronate), Tricyclic antidepressants, Typical antipsychotics, Warfarin
Outcome	Angioedema, Aplastic anemia, Acute liver injury, Bleeding, Hip fracture, Hospitalization, Myocardial infarction, Mortality after myocardial infarction, Renal failure, Gastrointestinal ulcer hospitalization

Table 6: OMOP drug categories and conditions. Parenthesized lists contain the subcategories of the broad drug category considered.

Drug or Drug Category	Representative Drug(s)
ACE Inhibitors	Lisinopril, Captopril, and Enalapril Maleate
Amphotericin B	Ambisome, Amphotericin B, and Abelcet
Erythromycin	Erythromycin, Erythromycin Ethylsuccinate, and Erythromycin
Sulfonamides	Silver Sulfadiazine, Bactrim, and Sulfadiazine
Tetracyclines	Doxycycline Hyclate, Tigecycline, and Minocycline
Carbamazepine	Carbamazepine
Phenytoin	Phenytoin Sodium, Phenytoin, and Extended Phenytoin Sodium
Benzodiazepines	Lorazepam, Diazepam, and Clonazepam
Beta Blockers	Metoprolol Tartrate, Labetalol, and Atenolol
Alendronate	Alendronate Sodium and Alendronate
Tricyclics	Doxepin HCL, Desipramine, and Amitriptyline HCL
Typical Antipsychotics	Haloperidol, Thiothixene, and Pimozide
Warfarin	Warfarin

Table 7: Drug Representatives selected for each OMOP category (or subcategory).

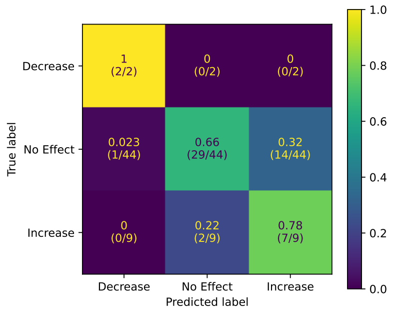
We take an additional postprocessing step to further enhance the quality of the assigned labels, replacing *unreliable* predictions with "no-effect," unless stated otherwise. Specifically, we consider outputs for which the LLM reported weak evidence and rare incidences of effects as unreliable. Additionally, we deem outputs for which the LLM selected small round numbers for the probability (*i.e.*, 0.1 and 0.01) as unreliable, as such values are often chosen in the absence of strong evidence, resembling typical human preferences for round numbers. We apply this postprocessing except in the case of AUC, as uncertainty should be reflected directly in the confidence scores.

We obtain an effect-based F1 score of 0.5294 without postprocessing, and 0.6087 with postprocessing. We obtain an ADE-based F1 score of 0.4828 without postprocessing, and 0.5556 with postprocessing. Figures ?? and ?? show the confusion matrices and predictions, respectively, of MALADE without postprocessing on GPT-4 Turbo (compared to the results in Section 5.2).

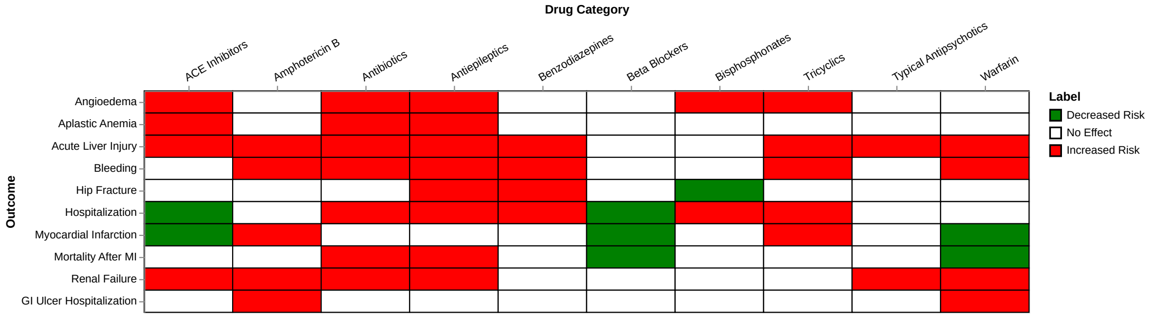
B.5. Discussion of Justifications Produced by `DrugFinder`

As in Section 5.4, we extract the justifications produced by `DrugFinder` in a run of MALADE for review by a clinician. The most common flaw in its reasoning is that `DrugFinder` frequently fails to recognize OMOP categories which consist of a single drug; beyond that, the agent's justifications are generally correct, with only one significant other error occurring.

This error isn't directly caused by `DrugFinder`, but, rather, was a result of our OpenFDA retrieval process failing to identify certain common typical antipsychotics, for example, fluphenazine is labeled as a phenothiazine in the National Drug Code DB but not as a typical antipsychotic; hence a more robust search, taking subcategories into account, for drugs in a category is needed.



(a) Confusion matrix for MALADE



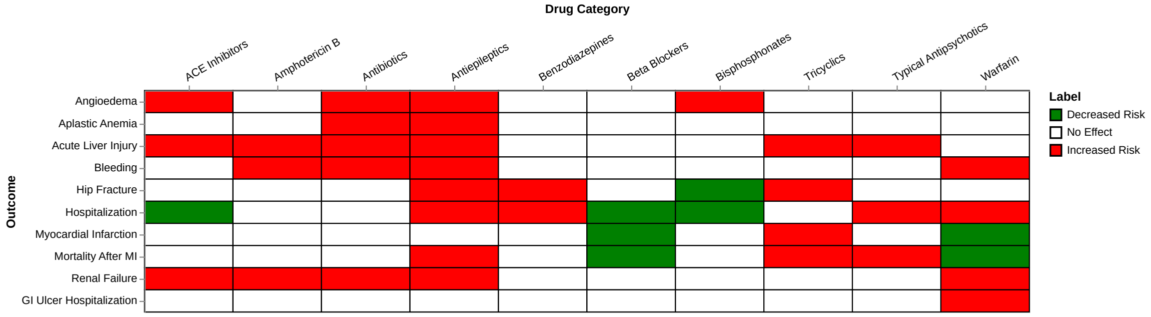
Drug Category

Outcome	ACE Inhibitors	Amphotericin B	Antibiotics	Antiepileptics	Benzodiazepines	Beta Blockers	Bisphosphonates	Tricyclics	Typical Antipsychotics	Warfarin
Angioedema	🟥	⬜	🟥	🟥	⬜	⬜	🟥	🟥	⬜	⬜
Aplastic Anemia	🟥	⬜	🟥	🟥	⬜	⬜	⬜	⬜	⬜	⬜
Acute Liver Injury	🟥	🟥	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥
Bleeding	⬜	🟥	🟥	🟥	🟥	⬜	⬜	🟥	⬜	🟥
Hip Fracture	⬜	⬜	⬜	🟥	🟥	⬜	🟩	⬜	⬜	⬜
Hospitalization	🟩	⬜	🟥	🟥	🟥	🟩	🟥	🟥	⬜	⬜
Myocardial Infarction	🟩	🟥	⬜	⬜	⬜	🟩	⬜	🟥	⬜	🟩
Mortality After MI	⬜	⬜	🟥	🟥	⬜	🟩	⬜	⬜	⬜	🟩
Renal Failure	🟥	🟥	🟥	🟥	⬜	⬜	⬜	⬜	🟥	🟥
GI Ulcer Hospitalization	⬜	🟥	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥

Label
🟩 Decreased Risk
⬜ No Effect
🟥 Increased Risk

(b) Predictions of MALADE

Figure 9: Results with GPT-4 Turbo, without postprocessing.



Drug Category

Outcome	ACE Inhibitors	Amphotericin B	Antibiotics	Antiepileptics	Benzodiazepines	Beta Blockers	Bisphosphonates	Tricyclics	Typical Antipsychotics	Warfarin
Angioedema	🟥	⬜	🟥	🟥	⬜	⬜	🟥	⬜	⬜	⬜
Aplastic Anemia	⬜	⬜	🟥	🟥	⬜	⬜	⬜	⬜	⬜	⬜
Acute Liver Injury	🟥	🟥	🟥	🟥	⬜	⬜	⬜	🟥	🟥	⬜
Bleeding	⬜	🟥	🟥	🟥	⬜	⬜	⬜	⬜	⬜	🟥
Hip Fracture	⬜	⬜	⬜	🟥	🟥	⬜	🟩	🟥	⬜	⬜
Hospitalization	🟩	⬜	⬜	🟥	🟥	🟩	🟩	⬜	🟥	🟥
Myocardial Infarction	⬜	⬜	⬜	⬜	⬜	🟩	⬜	🟥	⬜	🟩
Mortality After MI	⬜	⬜	⬜	🟥	⬜	🟩	⬜	🟥	🟥	🟩
Renal Failure	🟥	🟥	🟥	🟥	⬜	⬜	⬜	⬜	⬜	🟥
GI Ulcer Hospitalization	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥

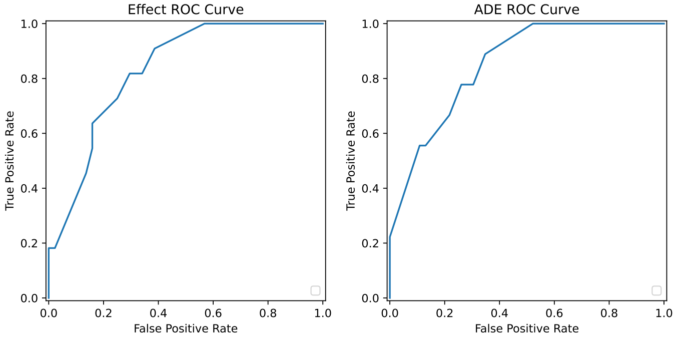
Label
🟩 Decreased Risk
⬜ No Effect
🟥 Increased Risk

Figure 10: Predictions of MALADE run on GPT-4o.

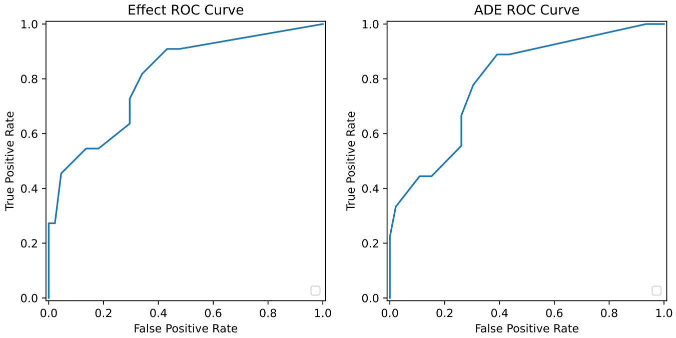
Appendix C. Detailed behavior of MALADE: Successful and Failed Behavior

C.1. Successful Behavior

For an example of a successful run of MALADE, consider the effect of ACE inhibitors on agioedema. As we will see, `DrugAgent` will retrieve the reports of ADEs from the FDA labels, and `CategoryAgent` will conclude that they are strong evidence for increased risk.



(a) Using the model's generated confidence values.



(b) Deriving confidence from the model's generated probability values.

Figure 11: **ROC curves for MALADE on OMOP**

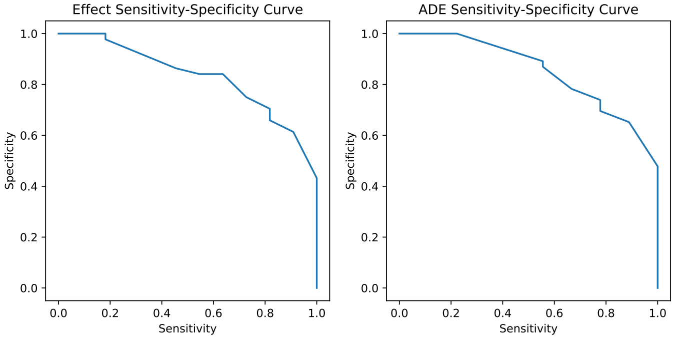
3.1. DrugAgent Step

This step will be run three times, for each of the representative drugs (Lisinopril, Captopril, and Enalapril Maleate). We'll show the results for Lisinopril.

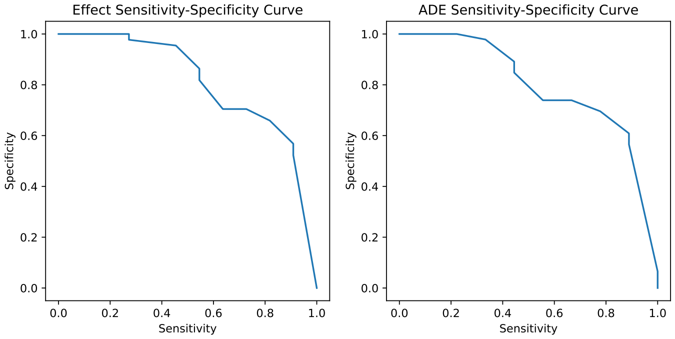
The original question to `DrugAgent` is "Does LISINOPRIL increase or decrease the risk of angioedema?"

As `DrugAgent` requires evidence from the drug labels to answer the question, the agent immediately asks `FDAHandler`, specifying it as the *recipient* of the message using a tool call, as follows:

```
FUNC: {
  "name": "relevant_extracts",
  "to": "",
  "arguments": {
    "query": "Does LISINOPRIL increase the risk of angioedema?",
    "filter_drugs": [
      "Lisinopril"
    ]
  }
}
```



(a) Using the model's generated confidence values.



(b) Deriving confidence from the model's generated probability values.

Figure 12: **Sensitivity vs. specificity curves for MALADE on OMOP**

The agent searches its vector DB for relevant passages from the drug's label, using another tool call:

```
FUNC: {
  "name": "relevant_extracts",
  "to": "",
  "arguments": {
    "query": "Does LISINOPRIL increase the risk of angioedema?",
    "filter_drugs": [
      "Lisinopril"
    ]
  }
}
```

A large number of passages are retrieved, including:

```
LISINOPRIL: warnings: WARNINGS General Lisinopril Anaphylactoid and
 Possibly Related Reactions: Presumably because angiotensin-converting
 enzyme inhibitors affect the metabolism of eicosanoids and
 polypeptides, including endogenous bradykinin, patients receiving ACE
 inhibitors (including lisinopril and hydrochlorothiazide tablets) may
 be subject to a variety of adverse reactions, some of them
 serious. Head and Neck Angioedema: Angioedema of the face,
 extremities, lips, tongue, glottis and/or larynx has been reported
 rarely in patients treated with angiotensin converting enzyme
 inhibitors, including lisinopril. ACE inhibitors have been associated
 with a higher rate of angioedema in Black than in non-Black patients.
```

with which `FDAHandler` produces its answer, stating that it is done:

```
<DONE>
ANSWER: Yes, LISINOPRIL has been associated with an increased risk of
 angioedema. Angioedema of the face, extremities, lips, tongue,
 glottis, and/or larynx has been reported rarely in patients treated
 with angiotensin-converting enzyme inhibitors, including LISINOPRIL.
  This can occur at any time during treatment. The incidence of
 angioedema is higher in Black patients compared to non-Black
 patients. In very rare cases, fatalities have been reported due to
 angioedema associated with laryngeal edema or tongue edema. Patients
 with a history of angioedema unrelated to ACE-inhibitor therapy may
 be at increased risk of angioedema while receiving an ACE inhibitor.
  Intestinal angioedema has also been reported in patients treated
 with ACE inhibitors, presenting with abdominal pain, with or without
 nausea or vomiting.

SOURCE: LISINOPRIL label
EXTRACT_START_END: Head and Neck Angioedema: Angioedema of the
 face...associated with laryngeal edema or tongue edema.
```

`DrugAgent` expresses its reasoning in structured form, which is sent to the Critic:

```
FUNC: {
  "name": "final_answer",
  "to": "",
  "arguments": {
    "question": "Does LISINOPRIL increase or decrease the risk of angioedema?",
    "steps": [
      "I asked FDAHandler if LISINOPRIL increases the risk of angioedema.",
      "FDAHandler confirmed that LISINOPRIL has been associated with
      an increased risk of angioedema. Angioedema involving the face,
      extremities, lips, tongue, glottis, and/or larynx has been
      reported rarely in patients treated with angiotensin-converting
      enzyme inhibitors, including LISINOPRIL. The incidence is higher
      in Black patients compared to non-Black patients, and very rare
      cases have resulted in fatalities. Patients with a history of
      angioedema unrelated to ACE-inhibitor therapy may be at
      increased risk of angioedema while receiving an ACE
      inhibitor. Intestinal angioedema has also been reported rarely."
    ],
    "answer": "LISINOPRIL increases the risk of angioedema."
  }
}
```

This is processed into natural language form before being forwarded to the LLM.

```
        Question: Does LISINOPRIL increase or decrease the risk of angioedema?
        -----
        Reasoning:
1. I asked FDAHandler if LISINOPRIL increases the risk of angioedema.
2. FDAHandler confirmed that LISINOPRIL has been associated with an
 increased risk of angioedema. Angioedema involving the face,
 extremities, lips, tongue, glottis, and/or larynx has been reported
 rarely in patients treated with angiotensin-converting enzyme
 inhibitors, including LISINOPRIL. The incidence is higher in Black
 patients compared to non-Black patients, and very rare cases have
 resulted in fatalities. Patients with a history of angioedema
 unrelated to ACE-inhibitor therapy may be at increased risk of
 angioedema while receiving an ACE inhibitor.
 Intestinal angioedema has also been reported rarely.
        -----
        Final answer: LISINOPRIL increases the risk of angioedema.
```

The Critic has no feedback (which indicates acceptance):

```
FUNC: {
  "name": "feedback",
  "to": "",
  "arguments": {
    "critique": ""
 }
}
```

As with the structured description of the reasoning steps, this is processed into natural language:

```
Your reasoning is valid, no feedback was provided.
```

From here on, we will show only the natural language form for these steps. The agent then outputs its answer to the original question, with the additional context from `FDAHandler`.

```
<DONE>

LISINOPRIL increases the risk of angioedema. This has been confirmed
 by FDAHandler, who referenced the LISINOPRIL label indicating that
 angioedema of the face, extremities, lips, tongue, glottis, and/or
 larynx has been reported rarely in patients treated with
 angiotensin-converting enzyme inhibitors, including LISINOPRIL. The
 incidence of angioedema is higher in Black patients compared to
 non-Black patients, and in very rare cases, fatalities have been
 reported due to angioedema associated with laryngeal edema or tongue
 edema. Patients with a history of angioedema unrelated to
 ACE-inhibitor therapy may be at increased risk of angioedema while
 receiving an ACE inhibitor. Intestinal angioedema has also been
 reported rarely in patients treated with ACE inhibitors.
```

3.2. CategoryAgent Step

Together with the query “Does the angiotensin converting enzyme inhibitor category of drugs increase the risk of angioedema, decrease it, or is there no clear effect?” the agent is provided with the information from the outputs from the last step, for each of the three drugs, as follows:

```
                Passages:
                1. Drug LISINOPRIL: LISINOPRIL increases the risk of
 angioedema. This has been confirmed by FDAHandler, who referenced the
 LISINOPRIL label indicating that angioedema of the face, extremities,
 lips, tongue, glottis, and/or larynx has been reported rarely in
 patients treated with angiotensin-converting enzyme inhibitors,
 including LISINOPRIL. The incidence of angioedema is higher in Black
 patients compared to non-Black patients, and in very rare cases,
 fatalities have been reported due to angioedema associated with
 laryngeal edema or tongue edema. Patients with a history of
 angioedema unrelated to ACE-inhibitor therapy may be at increased
 risk of angioedema while receiving an ACE inhibitor. Intestinal
 angioedema has also been reported rarely in patients treated with ACE
 inhibitors.
2. Drug CAPTOPRIL: CAPTOPRIL increases the risk of angioedema. This
 has been reported in approximately one in 1000 patients, and can
 involve various parts of the body including extremities, face, lips,
 mucous membranes, tongue, glottis, or larynx. Fatal airway
 obstruction has occurred in some cases, and the risk is higher in
 black patients compared to non-black patients.
3. Drug ENALAPRIL MALEATE: The final answer to the question "Does
 ENALAPRIL MALEATE increase or decrease the risk of angioedema?" is
 that ENALAPRIL MALEATE increases the risk of angioedema. This
 information is based on the response from FDAHandler, which
 referenced the ENALAPRIL MALEATE label indicating that angioedema has
 been reported in patients receiving the drug. The incidence is higher
 in black patients compared to non-black patients, and angioedema
 associated with laryngeal edema can be fatal. If angioedema occurs,
 treatment with ENALAPRIL MALEATE should be discontinued and
 appropriate therapy instituted immediately.
                ---------
                Does the angiotensin converting enzyme inhibitor
 category of drugs increase the risk of angioedema, decrease it, or is
 there no clear effect?
```

`CategoryAgent` immediately concludes that ACE inhibitors increase the risk of angioedema, and sends its reasoning to the Critic:

```
        Question: Does the angiotensin converting enzyme inhibitor
 category of drugs increase the risk of angioedema, decrease it, or is
 there no clear effect?
        -----
        Reasoning:
        1. All three passages provided indicate that the respective
 angiotensin converting enzyme inhibitors (LISINOPRIL, CAPTOPRIL, and
 ENALAPRIL MALEATE) increase the risk of angioedema.
2. Passage 1 states that LISINOPRIL increases the risk of angioedema,
 with rare reports of angioedema in patients treated with ACE
 inhibitors, including LISINOPRIL.
3. Passage 2 indicates that CAPTOPRIL increases the risk of
 angioedema, with an incidence reported in approximately one in 1000
 patients.
4. Passage 3 confirms that ENALAPRIL MALEATE increases the risk of
 angioedema, as reported in patients receiving the drug.
5. The increased risk of angioedema is consistent across all three
 drugs mentioned, which are part of the angiotensin converting enzyme
 inhibitor category.
6. The evidence is strong as it is based on drug labels and FDAHandler
 references, which are authoritative sources.
        -----
        Final answer: The angiotensin converting enzyme inhibitor
 category of drugs increases the risk of angioedema.
```

The critic approves:

```
        Feedback: The reasoning provided is sound and consistent with
 medical knowledge. Angiotensin converting enzyme inhibitors are known
 to increase the risk of angioedema. The user has correctly identified
 that all three passages from the FDA data indicate an increased risk
 of angioedema with the use of LISINOPRIL, CAPTOPRIL, and ENALAPRIL
 MALEATE, which are all ACE inhibitors. The conclusion that the ACE
 inhibitor category of drugs increases the risk of angioedema is
 supported by the evidence provided and is in line with established
 medical understanding.

        If any flaws in the reasoning used to produce your answer were
 identified, you must try again.
```

And the agent outputs “increase” with very high confidence:

```
FUNC: {
  "name": "category_effect_tool",
  "to": "",
  "arguments": {
    "label": "increase",
    "confidence": 1,
    "probability": 0.001,
    "frequency": "rare",
    "evidence": "strong",
    "justification": "The evidence from FDAHandler and drug labels for
 LISINOPRIL, CAPTOPRIL, and ENALAPRIL MALEATE consistently reports an
 increased risk of angioedema with the use of these ACE
 inhibitors. The incidence of angioedema is reported as rare, with
 occurrences such as one in 1000 patients for CAPTOPRIL. The evidence
 is considered strong due to the authoritative nature of the sources."
  }
}
```

3.3. CategoryAgent and Critic

Here we present the full prompt of interactions between `CategoryAgent` and Critic, given the task of identifying the association between "benzodiazepine" and "hip fracture".

At round 1, the output from `CategoryAgent`:

```
Question: Does the benzodiazepine category of drugs increase the risk of hip
    fracture, decrease it,
or is there no clear effect?
-----
Reasoning:
1. The FDA label for Lorazepam does not specify information regarding the risk of
    hip fracture.
2. The FDA label data for Diazepam does not provide information on the risk of hip
     fracture.
3. The FDA label for Clonazepam mentions 'fracture traumatic' as an infrequent
    adverse event but
does not establish a causal relationship with hip fractures or provide specific
    data on hip fracture risk.
4. Since none of the provided passages offer conclusive evidence regarding the
    effect of benzodiazepines
on the risk of hip fracture, we cannot determine if the category as a whole has a
    clear effect
on hip fracture risk.
-----
Final answer: There is no clear effect of the benzodiazepine category of drugs on
    the risk of hip fracture
based on the provided passages.
```

Feedback by Critic:

```
The reasoning provided is flawed. While it is correct that the FDA label for
    Lorazepam and Diazepam does not specify information regarding the risk of hip
    fracture, the mention of 'fracture traumatic' as an infrequent adverse event
    in the Clonazepam label should not be dismissed. Benzodiazepines are known to
    cause sedation and muscle relaxation, which can lead to an increased risk of
    falls, and consequently, hip fractures, especially in the elderly. The
    presence of 'fracture traumatic' as an adverse event in the Clonazepam label
    suggests that there may be a link between benzodiazepine use and fractures.
    Therefore, the conclusion should reflect a potential increase in the risk of
    hip fracture associated with benzodiazepines, even if the evidence is not
    conclusive for a causal relationship.

If any flaws in the reasoning used to produce your answer were identified, you
    must try again.
```

At round 2, the corrected output from `CategoryAgent`:

```
Question: Does the benzodiazepine category of drugs increase the risk of hip
    fracture, decrease it, or is there no clear effect?
-----
Reasoning:
1. The FDA label for Lorazepam does not specify information regarding the risk of
    hip fracture.
```

```
2. The FDA label data for Diazepam does not provide information on the risk of hip
     fracture.
3. The FDA label for Clonazepam mentions 'fracture traumatic' as an infrequent
    adverse event but
does not establish a causal relationship with hip fractures or provide specific
    data on hip fracture risk.
4. Considering the pharmacological effects of benzodiazepines, which include
    sedation and muscle relaxation, there is a plausible mechanism by which these
    drugs could increase the risk of falls and consequently hip fractures,
    especially in the elderly.
5. The mention of 'fracture traumatic' as an infrequent adverse event in the
    Clonazepam label, while not conclusive, suggests that there may be a link
    between benzodiazepine use and fractures.
6. Given the plausible mechanism and the mention of fractures as an adverse event
    for Clonazepam, it is reasonable to infer that benzodiazepines may potentially
     increase the risk of hip fracture, even though the evidence is not definitive
     for a causal relationship.
-----
Final answer: The benzodiazepine category of drugs may potentially increase the
    risk of hip fracture, although the evidence is not definitive for a causal
    relationship based on the provided passages.
```

Feedback by Critic:

```
The revised reasoning is more accurate in acknowledging the potential risk
    increase for hip fractures associated with benzodiazepines, given the
    pharmacological effects and the mention of 'fracture traumatic' as an adverse
    event for Clonazepam. However, the final answer should still reflect a degree
    of uncertainty due to the lack of conclusive evidence. It would be more
    appropriate to state that there is a potential increase in the risk of hip
    fracture with benzodiazepine use, but the evidence is not strong enough to
    definitively establish a causal relationship.

If any flaws in the reasoning used to produce your answer were identified, you
    must try again.
```

At round 3, the corrected output from `CategoryAgent`:

```
Question: Does the benzodiazepine category of drugs increase the risk of hip
    fracture, decrease it,
or is there no clear effect?
-----
Reasoning:
1. The FDA label for Lorazepam does not specify information regarding the risk of
    hip fracture.
2. The FDA label data for Diazepam does not provide information on the risk of hip
     fracture.
3. The FDA label for Clonazepam mentions 'fracture traumatic' as an infrequent
    adverse event but does not establish a causal relationship with hip fractures
    or provide specific data on hip fracture risk.
4. Considering the pharmacological effects of benzodiazepines, which include
    sedation and muscle relaxation, there is a plausible mechanism by which these
    drugs could increase the risk of falls and consequently hip fractures,
    especially in the elderly.
5. The mention of 'fracture traumatic' as an infrequent adverse event in the
    Clonazepam label, while not conclusive, suggests that there may be a link
    between benzodiazepine use and fractures.
6. Given the plausible mechanism and the mention of fractures as an adverse event
    for Clonazepam,
it is reasonable to infer that benzodiazepines may potentially increase the risk
    of hip fracture, even though the evidence is not definitive for a causal
    relationship.
7. However, due to the lack of conclusive evidence, the final answer should
    reflect uncertainty regarding the causal relationship between benzodiazepine
    use and hip fracture risk.
-----
Final answer: There is a potential increase in the risk of hip fracture associated
     with benzodiazepine use, but the evidence is not strong enough to
    definitively establish a causal relationship based on the provided passages.
```

Feedback by Critic:

```
Your reasoning is valid, no feedback was provided.
```

Final output by `CategoryAgent`:

```
"label": "increase",
"confidence": 0.6,
"probability": 0.1,
"frequency": "rare",
"evidence": "weak",
"justification": "The FDA label for Clonazepam lists 'fracture traumatic' as an
    infrequent adverse event, suggesting a potential link between benzodiazepine
    use and fractures.
Given the pharmacological effects of benzodiazepines, which include sedation and
    muscle relaxation, there is a plausible mechanism by which these drugs could
    increase the risk of falls and consequently hip fractures, especially in the
    elderly. However, the evidence is not strong enough to definitively establish
    a causal relationship, and the FDA labels for Lorazepam and Diazepam do not
    provide specific information regarding hip fracture risk."
```

C.2. Failed Behavior

MALADE, however, fails to identify the effect of alendronate on gastrointestinal ulcer hospitalization. In particular, `DrugAgent` misinterprets a statement in the FDA labels indicating a lack of evidence in clinical trials and forwards an overly broad statement that indicates that clinical trials failed to show evidence of an effect to `CategoryAgent`; hence `CategoryAgent` outputs the incorrect prediction of "No Effect."

3.1. DrugAgent Step

This step was run twice, for each of the representative drugs, Alendronate Sodium and Alendronate (noting that in this case only one drug should have been selected). We'll consider the results for Alendronate.

The original question to `DrugAgent` is “Does ALENDRONATE SODIUM increase or decrease the risk of gastrointestinal ulcer hospitalization?”

`DrugAgent` queries `FDAHandler`:

```
FUNC: {
  "name": "recipient_message",
  "to": "",
  "arguments": {
    "intended_recipient": "FDAHandler",
    "content": "Does ALENDRONATE SODIUM increase
    the risk of gastrointestinal ulcer hospitalization?"
  }
}
```

which queries the FDA label database with:

```
FUNC: {
  "name": "relevant_search_extracts",
  "to": "",
  "arguments": {
    "query": "Does ALENDRONATE increase the risk of gastrointestinal
 ulcer hospitalization?",
    "drug": "ALENDRONATE"
  }
}
```

which returns a number of passages from the FDA labels, including:

```
ALENDRONATE: warnings_and_cautions: 1 Upper Gastrointestinal Adverse
 Reactions BINOSTO, like other bisphosphonates administered orally,
 may cause local irritation of the upper gastrointestinal
 mucosa. Because of these possible irritant effects and a potential
 for worsening of the underlying disease, caution should be used when
 BINOSTO is given to patients with active upper gastrointestinal
 problems (such as known Barrett's esophagus, dysphagia, other
 esophageal diseases, gastritis, duodenitis, or ulcers). Esophageal
 adverse experiences, such as esophagitis, esophageal ulcers and
 esophageal erosions, occasionally with bleeding and rarely followed
 by esophageal stricture or perforation, have been reported in
 patients receiving treatment with oral bisphosphonates including
 alendronate sodium. In some cases these have been severe and required hospitalization.
```

and

```
ALENDRONATE: warnings_and_cautions: There have been post-marketing
 reports of gastric and duodenal ulcers with oral bisphosphonate use,
 some severe and with complications, although no increased risk was
 observed in controlled clinical trials.
```

Using these, `FDAHandler` responds with:

```
<DONE>
ANSWER: Alendronate, like other bisphosphonates administered orally,
 may cause local irritation of the upper gastrointestinal
 mucosa. Esophageal adverse experiences such as esophagitis,
 esophageal ulcers, and esophageal erosions, occasionally with
 bleeding and rarely followed by esophageal stricture or perforation,
 have been reported in patients receiving treatment with alendronate
 sodium. In some cases, these have been severe and required
 hospitalization. There have also been post-marketing reports of
 gastric and duodenal ulcers with oral bisphosphonate use, some severe
 and with complications, although no increased risk was observed
 in controlled clinical trials.

SOURCE: ALENDRONATE label
EXTRACT_START_END: ...esophageal erosions, occasionally with bleeding
 and rarely followed by... ...gastric and duodenal ulcers with oral
 bisphosphonate use, some severe...
```

`DrugAgent` now concludes that the effect is uncertain, and sends its reasoning to the Critic. Here is where the error occurs: the phrasing "However, no increased risk was observed in controlled clinical trials." is overly broad and suggests that alendronate was shown not to increase the risk of *all* gastrointestinal ulcers in clinical trials, while, in reality, that was shown only for gastric and duodenal ulcers specifically. This will lead to the incorrect prediction by `CategoryAgent`.

```
        Question: Does ALENDRONATE increase or decrease the risk of
        gastrointestinal ulcer hospitalization?
        -----
        Reasoning:
        1. I asked FDAHandler whether ALENDRONATE increases the risk
        of gastrointestinal ulcer hospitalization.
2. FDAHandler responded that Alendronate can cause local irritation of
the upper gastrointestinal mucosa and has been associated with
esophageal adverse experiences such as esophagitis, esophageal ulcers,
and erosions, occasionally with bleeding and rarely followed by
esophageal stricture or perforation, which have required
hospitalization in some cases. There have also been post-marketing
reports of gastric and duodenal ulcers with oral bisphosphonate use,
some severe and with complications. However, no increased risk was observed
in controlled clinical trials.
        -----
        Final answer: Alendronate may increase the risk of
        gastrointestinal ulcer hospitalization, as it has been
        associated with severe esophageal adverse experiences and
        post-marketing reports of gastric and duodenal ulcers that
        required hospitalization in some cases. However, it is
        important to note that no increased risk was observed in
        controlled clinical trials, indicating that while there is a potential risk,
        it may not be common.
```

As the critic is unaware of the failure to correctly summarize the output from `FDAHandler`, it concludes that the answer is justified, returning no feedback; `DrugAgent` recieves:

```
 Your reasoning is valid, no feedback was provided.
```

and it outputs the answer with the flawed statement:

```
<DONE>
Alendronate may increase the risk of gastrointestinal ulcer hospitalization, as it
     has been associated with severe esophageal adverse experiences and post-
    marketing reports of gastric and duodenal ulcers that required hospitalization
     in some cases.
However, it is important to note that no increased risk was observed in controlled
     clinical trials, indicating that while there is a potential risk, it may not
    be common.
```

3.2. CategoryAgent step

The question posed to `CategoryAgent` with the results from `DrugAgent` is as follows:

```
                Passages:
                1. Drug ALENDRONATE SODIUM: ALENDRONATE SODIUM has
 been associated with gastrointestinal adverse effects that can be
 severe and may lead to hospitalization, such as esophageal ulcers and
 erosions. However, it is important to note that while there have been
 post-marketing reports of severe gastric and duodenal ulcers with
 complications, no increased risk of gastrointestinal ulcer
 hospitalization was observed in controlled clinical
 trials. Therefore, while there is a potential risk, the degree of
 reliability of this association is not fully established, and the
 risk of hospitalization due to gastrointestinal ulcers appears to be rare.
2. Drug ALENDRONATE: Alendronate may increase the risk of
 gastrointestinal ulcer hospitalization, as it has been associated
 with severe esophageal adverse experiences and post-marketing reports
 of gastric and duodenal ulcers that required hospitalization in some
 cases. However, it is important to note that no increased risk was
 observed in controlled clinical trials, indicating that while there
 is a potential risk, it may not be common.
                ---------
                Does the alendronate category of drugs increase the
 risk of gastrointestinal ulcer hospitalization, decrease it, or is there no clear effect?
```

Given the overly broad statement from `DrugAgent` regarding clinical trial evidence, `CategoryAgent` immediately concludes that there is no clear effect and sends its reasoning steps to the Critic:

```
Question: Does the alendronate category of drugs increase the risk of
    gastrointestinal ulcer hospitalization, decrease it, or is there no clear
    effect?
    -----
Reasoning:
1. Both passages mention that alendronate and alendronate sodium have been
    associated with gastrointestinal adverse effects, including severe gastric and
     duodenal ulcers.
2. Both passages also note that these associations come from post-marketing
    reports, which are less controlled than clinical trials.
```

```
3. Passage 1 states that no increased risk of gastrointestinal ulcer
    hospitalization was observed in controlled clinical trials for alendronate
    sodium.
4. Passage 2 similarly notes that for alendronate, no increased risk was observed
    in controlled clinical trials.
5. Both passages conclude that while there is a potential risk, the risk of
    hospitalization due to gastrointestinal ulcers appears to be rare and may not
    be common.
6. The evidence suggests that while there are reports of severe cases, the
    controlled clinical trials did not show an increased risk, indicating that the
     effect is not consistent or well-established.
       -----
Final answer: There is no clear effect of the alendronate category of drugs on the
     risk of gastrointestinal ulcer hospitalization.
```

and the Critic approves:

```
        Feedback: The reasoning provided is consistent with the
 information from the passages. It correctly identifies that while
 there are post-marketing reports of gastrointestinal adverse effects,
 controlled clinical trials did not show an increased risk of
 gastrointestinal ulcer hospitalization for drugs in the alendronate
 category. The conclusion that there is no clear effect on the risk of
 gastrointestinal ulcer hospitalization is supported by the evidence
 provided. Therefore, there are no flaws in the reasoning process.

        If any flaws in the reasoning used to produce your answer were identified,
        you must try again.
```

Finally, `CategoryAgent` outputs the incorrect final label, with high confidence:

```
  FUNC: {
  "name": "category_effect_tool",
  "to": "",
  "arguments": {
    "label": "no-effect",
    "confidence": 0.8,
    "probability": 0.05,
    "frequency": "rare",
    "evidence": "weak",
    "justification": "The alendronate category of drugs has been
 associated with gastrointestinal adverse effects in post-marketing
 reports, but controlled clinical trials did not show an increased
 risk of gastrointestinal ulcer hospitalization. The evidence suggests
 a potential risk, but it is not consistent or well-established, and
 the risk of hospitalization due to gastrointestinal ulcers appears to be rare
 and may not be common."
  }
}
```

Appendix D. Ablations

The objective of our ablation study is to examine the individual impact of two principal elements on our system's performance:

(1) the iterative refinement process among multiple agents (*i.e.*, Agent-Critic interaction)

(2) the incorporation of external knowledge via RAG.

We do not conduct an ablation of our multi-agent design as MALADE cannot be naturally expressed as a single-agent system; while expressing all agents and control flow of MALADE in a single highly complex agent is possible in principle, it is not straightforward to achieve that with any level of reliability.

For (1), we evaluate the system's performance with and without the involvement of the Critics paired with `DrugAgent` and `CategoryAgent`, respectively, by toggling them on and off individually. Refer to the first two columns in Table 8. For compuational reasons we do not ablate the critic on `DrugFinder`.

For (2), we substitute FDAHandler with a simple agent which answers the questions from `DrugAgent` purely based on LLM's internal knowledge and generates responses in a similar output format as FDAHandler.m Refer to the third column, labeled as "RAG", in Table 8.

Results are obtained by the OMOP ADE task evaluation with the corresponding modified versions of MALADE, all of which were run with GPT-4 Turbo. To alleviate the computational burden of ablations, when an ablated system's configuration is identical to MALADE's (*i.e.*, Critics on all agents and RAG enabled) up to a given step of the pipeline, we retain the output originally produced by MALADE. We address the effects of variance due to random sampling from the LLM in Appendix E. We maintain consistency with the evaluation metrics and output label post-processing as detailed in Section 5.1, reporting ADE and effect-based AUC scores (with both the output confidence scores and probabilities) and ADE- and effect-based F1 scores.

Critics		RAG	ADE-based AUC			Effect-based AUC			F1 Score	
`DrugAgent`	`CategoryAgent`		Confidence	Probability	Probability (Modified)	Confidence	Probability	Probability (Modified)	ADE	Effect
✓	✓	✓	0.8514	0.7935	0.8043	0.8306	0.8058	0.8151	0.5556	0.6087
✓	×	✓	0.8647	0.7126	0.7693	0.8512	0.7376	0.7862	0.5714	0.6154
×	✓	✓	0.9034	0.8889	0.9143	0.8864	0.8833	0.9050	0.6316	0.6667
×	×	✓	0.8249	0.8865	0.8804	0.8192	0.8812	0.8760	0.5556	0.6087
✓	✓	×	0.9239	0.7609	0.8659	0.9287	0.7955	0.8853	0.5263	0.6087
✓	×	×	0.9239	0.7633	0.8599	0.9287	0.7975	0.8802	0.5556	0.6364
×	✓	×	0.9203	0.7403	0.8623	0.9256	0.7779	0.8822	0.5263	0.6087
×	×	×	0.9203	0.7428	0.8563	0.9256	0.7800	0.8771	0.5556	0.6364

Table 8: Ablation results on MALADE.

Results in Table 8 show that, in the case with RAG, the best results are obtained with the Critic on `CategoryAgent` active but the Critic on `DrugAgent` disabled (row 3); this suggests that the feedback from `CategoryAgent`'s Critic is important for producing the most reliable results, but that `DrugAgent`'s Critic may reduce the performance of MALADE. The case with a Critic on `DrugAgent` but not `CategoryAgent` does not confirm this hypothesis (row

2), however, showing slightly improved confidence scores and F1 scores, but much worse results in terms of probabilities compared to full MALADE (row 1). Compared to this case, removing the `DrugAgent`'s Critic worsens results with the exception of probability-based evaluations; hence, it is difficult to confidently determine from these results whether `DrugAgent`'s Critic is helpful or harmful. The extremely strong results in the case with the Critic on `CategoryAgent` active but the Critic on `DrugAgent` disabled do, however, appear to outweigh the improvements observed in the second row, suggesting that the Critic on `CategoryAgent` does in fact improve the overall reliability of MALADE.

The results without RAG show, slightly improved AUCs in the presence of `DrugAgent`'s Critic; `CategoryAgent`'s Critic, on the other hand, reduces F1 scores and slightly reduces probability-based AUCs.

Despite strong performance observed with probability-based metrics with some settings, these results suggest that direct estimates of effect probabilities may not be reliable measures in future pharmacovigilance systems; to see this, compare the columns labeled "Probability" and "Probability (Modified)." The "Probability (Modified)" column is the same as "Probability" except that output probabilities are incremented by 1 when the label is "increase" for ADE-based AUC and "increase" or "decrease" for effect-based AUC. In the ADE case, this modification enforces the separation between the derived scores from samples GPT-4 Turbo labeled as increasing risk and as having no effect; the improved results observed indicate that GPT-4's probability estimates are not consistent: substantial numbers of "no-effect" cells are assigned higher probabilities of an effect occurring as compared to cells where GPT-4 *itself* identifies increased risk.

Appendix E. Variance of MALADE's outputs

We wish to evaluate how much random sampling from LLM outputs affects MALADE's outputs, and, in particular, given the potential unreliability of numerical outputs produced by LLMs Xiong et al. (2024), how much variance there is in MALADE's output confidence scores. Moreover, we aim to understand whether key components of MALADE's design, namely Critic agents and RAG, affect these numerical outputs, and, in particular, affect their consistency, as well as whether variance in the outputs by the first two agents in MALADE's pipeline (`DrugFinder` and `DrugAgent`) is a significant contributor to the overall variance of these outputs.

We proceed by selecting three representative cells from the OMOP table, restricting ourselves to the cells used for evaluation and, to ensure a well-defined ground truth label for each representative, to drug categories without subcategories. Each representative corresponds to one of the three ground truth labels (increased risk, decreased risk, and no effect). We then run ten trials on each cell with ablated versions of MALADE (constructed as in Appendix D; however, we only consider enabling or disabling Critics on all agents, including `DrugFinder` here). The results are shown in Figure 13; these experiments were run with GPT-4 Turbo.

We observe that in all cases, MALADE maintains a clear separation between the confidences for each ground truth label, with the sole exception being the case with RAG but without Critics; that case is the only one in which we observe any samples with incorrect labels; the variance is similarly increased significantly in that case.

Next, we investigate how the variance in the outputs of `DrugFinder` and `DrugAgent` contribute to the overall variance of MALADE. We compare the variance of MALADE's outputs with the initial steps of the pipeline held fixed; They indicate that holding the initial steps of the pipeline fixed does not substantially reduce variance and that `CategoryAgent` is the primary source of variance in MALADE. However, note that the variance for "no-effect" is, somewhat surprisingly, highest with the outputs of `DrugFinder` and `DrugAgent` held constant. We observe that the variance in the representative drugs affects output confidence, in particular, it affects mean confidence in ADEs for the "no effect" representative. With the representatives held fixed, that mean confidence is higher (or, equivalently, mean confidence in "no effect" is lower) compared to the case in which we resample representatives in each trial.

	Critics and RAG	Critics only	RAG only
Critics only	$<, p = 0.044$	—	—
RAG only	$<, p = 0.098$	$\neq, p = 1.000$	—
Neither critics nor RAG	$<, p = 0.027$	$<, p = 0.224$	$<, p = 0.255$

Table 9: Relationship of mean confidence in ADE for "decrease" for ablated versions of MALADE, with p-values.

	Critics and RAG	Critics only	RAG only
Critics only	$<, p = 0.022$	—	—
RAG only	$>, p = 0.076$	$>, p = 0.017$	—
Neither critics nor RAG	$<, p = 0.022$	$\neq, p = 1.000$	$<, p = 0.017$

Table 10: Relationship of mean confidence in ADE for "no-effect" for ablated versions of MALADE, with p-values.

	Critics and RAG	Critics only	RAG only
Critics only	$<, p = 0.330$	—	—
RAG only	$>, p = 0.314$	$>, p = 0.178$	—
Neither critics nor RAG	$>, p = 0.144$	$>, p = 0.067$	$>, p = 0.278$

Table 11: Relationship of mean confidence in ADE for "increase" for ablated versions of MALADE, with p-values.

Now, to understand the significance of these effects, we will perform paired t-tests for each pair of ablated variants of MALADE, for each representative. The results for "decrease" are in Table 9, results for "no-effect" are shown in Table 10, and results for "increase" are shown in Table 11. Overall, we have that the mean confidence in ADE for the representative for "decrease" is lowest (i.e. the confidence in "decrease" is highest) in the case with neither Critics nor RAG, and we have that confidence in ADE for the representative of "no-effect" is lower (and so confidence in "no-effect" is higher) in the case that we have neither Critics nor RAG as compared to the cases with both Critics and RAG and RAG alone and, in addition, that confidence in ADE is *increased* (with RAG alone as compared with Critics alone, all with p-values below 0.05 for each pair.

With p-values below 0.1, we additionally have that mean confidence in ADE for "decrease" is decreased in the case with RAG alone as compared to the case with both Critics and RAG, confidence in ADE is increased with RAG alone as compared to critics and RAG, and, finally, that confidence in ADE for "increase" is increased with neither Critics nor RAG as compared to Critics alone.

Note that, while, as seen in Figure 13, by far the largest absolute shift in confidence occurs between RAG alone and all others for the "no-effect" representative, the large variance observed in that case is responsible for the reduced significance.

Extrapolating from these representative samples, these results suggest that MALADE without RAG or Critics performs at least as well as any other configuration (with a p-value < 0.05), but that in the presence of RAG, Critics improve reliability on instances with no strong effect, i.e. those where the ground truth is "no-effect," with a p-value less than 0.1. Notably, we do not have clear evidence that Critics are necessary when there is clear evidence, unsurprising as the FDA labels may explicitly state that a condition H is a potential ADE of a drug category C or that drugs in C are indicated for H.

As discussed in Section 5.3, we consider RAG an essential component of a generalizable pharmacovigilance system. Hence, we focus on the results in the case with RAG, in which case these results suggest that the Critic components of MALADE improve reliability.

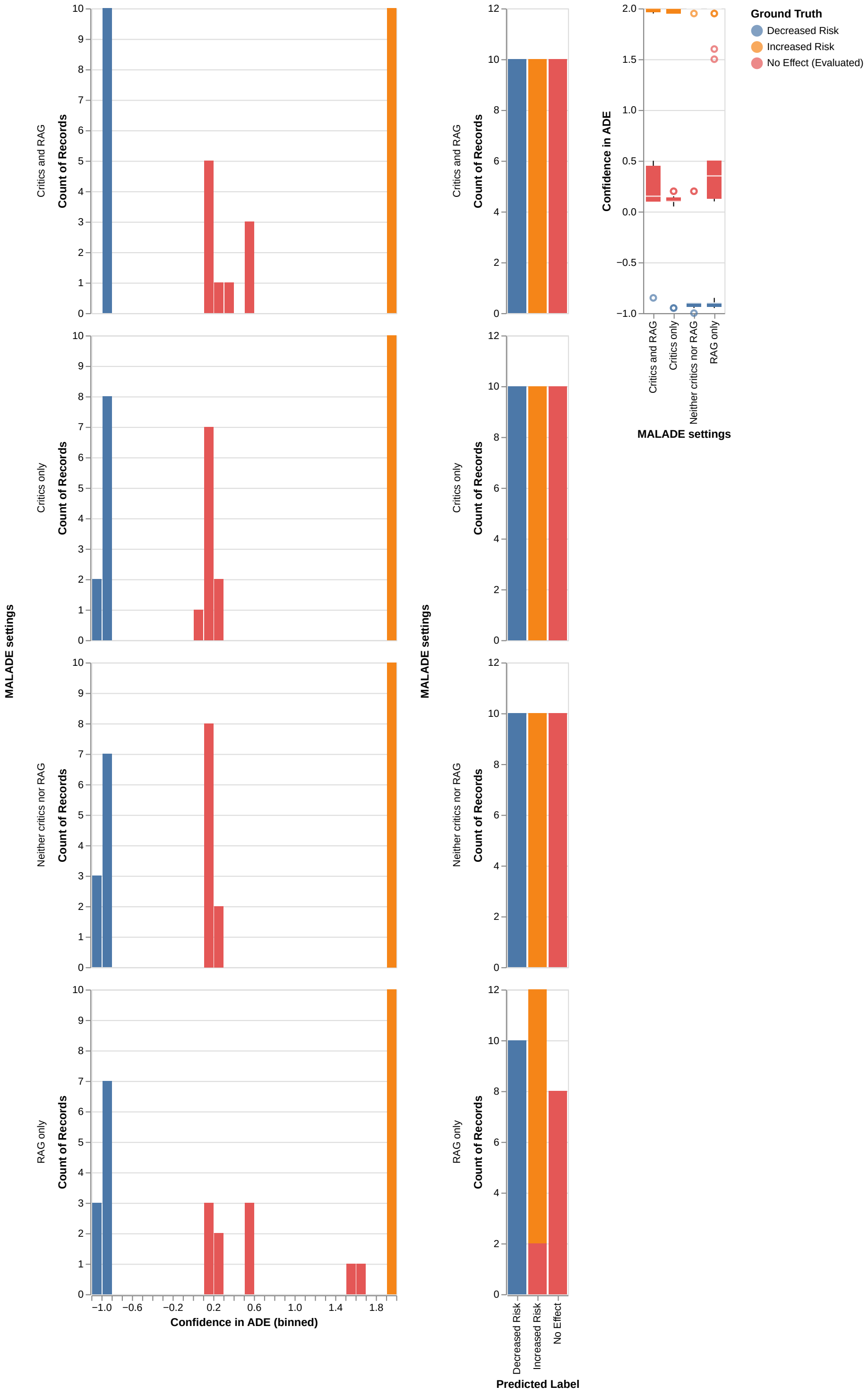


Figure 13: Histograms of confidence in ADE produced by ablations of MALADE.

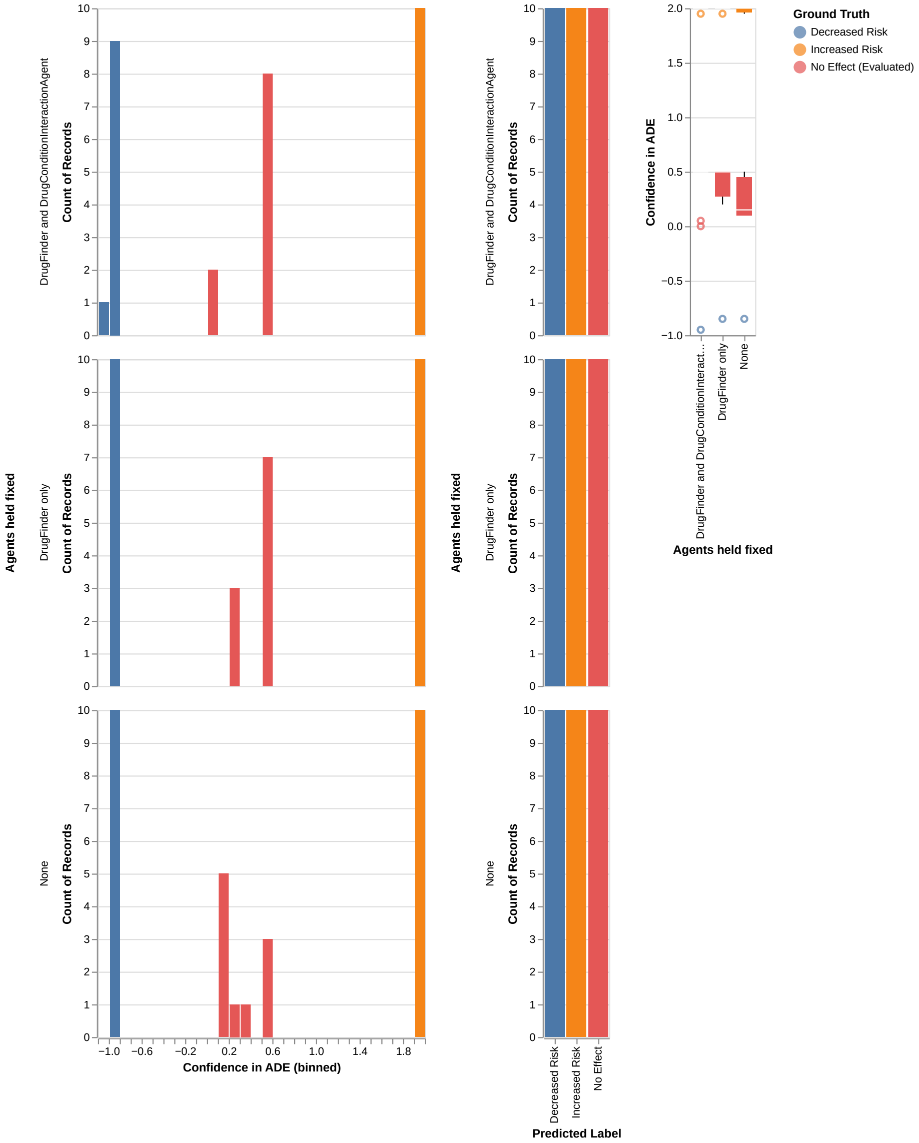


Figure 14: Histograms of confidence in ADE produced by MALADE with the outputs of each initial sequence of agents in the pipeline held fixed.