Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy

RESEARCH ARTICLE OPEN ACCESS

Enhancing Pharmacotherapy Through AI Applications

DILIConsult: A Multi-Agent Large Language Model Framework for Evaluating Drug-Induced Liver Injury in ICU Settings

Alfred Zheng Ting Ho[1] | Jeren Zheng Feng Law[1] | Aiwen Wang[1,2] | Daniel Yan Zheng Lim[3,4,5] | Jasmine Chiat Ling Ong[2,6]

[1]Department of Pharmacy and Pharmaceutical Science, Faculty of Science, National University of Singapore, Singapore City, Singapore | [2]Division of Pharmacy, Singapore General Hospital, Singapore City, Singapore | [3]Department of Gastroenterology and Hepatology, Singapore General Hospital, Singapore City, Singapore | [4]Medicine Academic Clinical Programme, Duke-NUS Medical School, Singapore City, Singapore | [5]Data Science and Artificial Intelligence Laboratory, Health Services Research Unit, Singapore General Hospital, Singapore City, Singapore | [6]Duke-NUS AI + Medical Sciences Initiative, Duke-NUS Medical School, Singapore City, Singapore

Correspondence: Jasmine Chiat Ling Ong (jasmine.ong.c.l@sgh.com.sg)

Received: 15 June 2025 | **Revised:** 5 January 2026 | **Accepted:** 6 January 2026

Keywords: chemical and drug-induced liver injury | generative artificial intelligence | large language models

ABSTRACT

Background: Large language models (LLMs) can support clinical decision-making by parsing databases and extracting relevant information. However, evaluating drug-induced liver injury (DILI) often requires processing lengthy clinical histories alongside reference materials like LiverTox, which can exceed context lengths of conventional LLMs. Challenges such as information truncation hinder standard approaches like prompt engineering and retrieval-augmented generation (RAG). To address these limitations, this study introduces DILIConsult, an agentic LLM pipeline based on GPT-4, designed to intelligently parse clinical and drug information.

Methods: To develop DILIConsult, we compared GPT-4-Turbo versus GPT-4o for extracting DILI characteristics from LiverTox descriptions. We tested two approaches to analyzing cases of suspected DILI: full-length case analysis versus sequential drug-specific evaluations. We evaluated DILIConsult on cases of suspected DILI identified from the open source Medical Information Mart for Intensive Care-IV (MIMIC-IV) ICU dataset based on American Association for the Study of Liver Diseases (AASLD) and European Association for the Study of the Liver (EASL) criteria. Outputs from DILIConsult were compared against a panel of clinicians comprising an ICU pharmacist, an ICU junior attending physician, and an ICU resident. Responses were evaluated by two senior ICU attending physicians.

Results: Using GPT-4o and a sequential approach demonstrated improved performance in the extraction of DILI characteristics and analysis of suspected DILI. DILIConsult achieved the best mean rank of 1.66 ± 0.75 in knowledge recall and ranked second for reasoning (2.00 ± 0.64) and reflection of current medical consensus (2.05 ± 0.62). DILIConsult ranked last with mean ranks of 2.07 ± 0.52 and 2.09 ± 0.72 for less omission of important information and content inaccuracy, respectively.

Conclusion: DILIConsult demonstrates the potential of LLMs to assist clinicians in evaluating DILI. The findings emphasize the importance of task division in LLM-driven workflows to minimize information loss.

Alfred Ho Zheng Ting and Jeren Law Zheng Feng contributed equally to this work.

1 | Introduction

Drug-induced liver injury (DILI) is an adverse drug reaction resulting from the use of medications, herbal supplements, or other chemical substances. Globally, its incidence is estimated to be 4.94 per 100,000 person-years, with Asia reporting the highest incidence at 17.82 per 100,000 person-years [1]. DILI is a challenging diagnosis due to heterogeneous clinical presentations, lack of a definitive diagnostic test, and rise in polypharmacy [2]. DILI can lead to severe outcomes, including acute liver injury requiring liver transplantation and death [3–5].

Diagnosis of DILI is typically one of exclusion, initiated by abnormal liver function tests (LFT), and followed by a focused clinical assessment of the patient's medical and medication history. Most guidelines [6, 7] recommend combining medication and laboratory data with evidence-based resources such as LiverTox [8], which summarizes evidence on the likelihood of DILI. However, this process can be time-consuming and highly subjective. In time-sensitive settings like intensive care units (ICUs), rapid consolidation of patient history can support faster, more appropriate diagnosis to prevent clinical deterioration.

This underscores the need for clinical decision support (CDS) tools (see Table 1 for a glossary of definitions) capable of leveraging patient-specific information with evidence-based resources to streamline clinical assessment [9]. Prior evaluations of CDS tools in medication management have shown that excessive nonspecific alerts with unclear clinical significance remain key limitations, contributing to clinician fatigue and reduced engagement. These findings highlight the importance of developing patient-specific, context-aware systems to enhance relevance and clinical adoption [12, 13]. In parallel, current rules-based clinical assessment tools such as the Roussel Uclaf Causality Assessment Method (RUCAM) offer limited utility because they fail to incorporate various drug-specific variables and DILI phenotype diversity [14, 15].

Large language models (LLMs), known for their ability to process and generate natural language, are computer algorithms which can adapt to an assortment of tasks in medicine [16]. When integrated with search engine functionality and the ability to process unstructured data, LLMs may offer more contextualized assistance. However, several challenges—such as ensuring consistently accurate outputs, maintaining data security, and upholding medical ethics—have hindered their widespread adoption in medicine [17]. A key limitation is LLMs' tendency to 'hallucinate', or generate inaccurate information [18]. These inaccuracies may be mitigated when task-specific information is provided to the LLMs, through methods known as grounding and Retrieval Augmented Generation (RAG). However, LLM performance has been noted to decline when passed longer information [19], and excessive curation of data can lead to the loss of important information.

Inspired by the ReAct framework (Reason and Act) [20] and advances in multi-agent systems (Table 1) [21], we propose a LLM workflow that decomposes DILI evaluation into a set of simpler subtasks. This method limits the breadth of external knowledge to reduce the cognitive load on the LLMs. This cooperative use of LLM agents (Table 1) in complex problem-solving scenarios has shown promise in mitigating the limitations described above [22].

In this study, our objectives were to: (i) evaluate the ability of OpenAI's GPT-4 to extract relevant information from raw patient data and LiverTox texts, (ii) develop LLM agents and a multi-agent pipeline for DILI evaluation, and (iii) assess the accuracy and clinical utility of the recommendations generated by this pipeline. We introduce DILIConsult, an LLM-driven

TABLE 1 | Glossary of key terms used in this article.

Key terms	Definitions
Clinical decision support (CDS) tools	CDS tools aim to augment clinical decision making with context-specific clinical knowledge, patient data and other relevant health information [9].
Likelihood categories	A standardized set of semiquantitative probability classifications used by the Drug-Induced Liver Injury Network (DILIN) to convey structured expert opinion regarding the causal relationship between a drug and liver injury. The categories—Definite, Highly Likely, Probable, Possible, Unlikely, and Insufficient Data—correspond to estimated probability ranges spanning from $<25\%$ to $>95\%$. For example, a drug described as highly likely to be the cause would have a probability of 75%–95%, and having "clear and convincing" evidence yet not definite [7].
Large Language Model (LLM) agents	LLMs configured as intelligent systems with internal planning and reasoning capabilities that enable them to perform multi-step tasks and interact with external tools or databases [10].
Multi-agent systems	A system of multiple specialized agents, collaborating to perform tasks [11].
R ratio	The value of alanine aminotransferase (ALT) relative to alkaline phosphatase (ALP), each normalized to their respective upper normal limits. Patients with liver injury may be categorized by phenotype based on the R ratio as such: • R ratio >5: hepatocellular • R ratio 2–5: mixed • R ratio <2: cholestatic

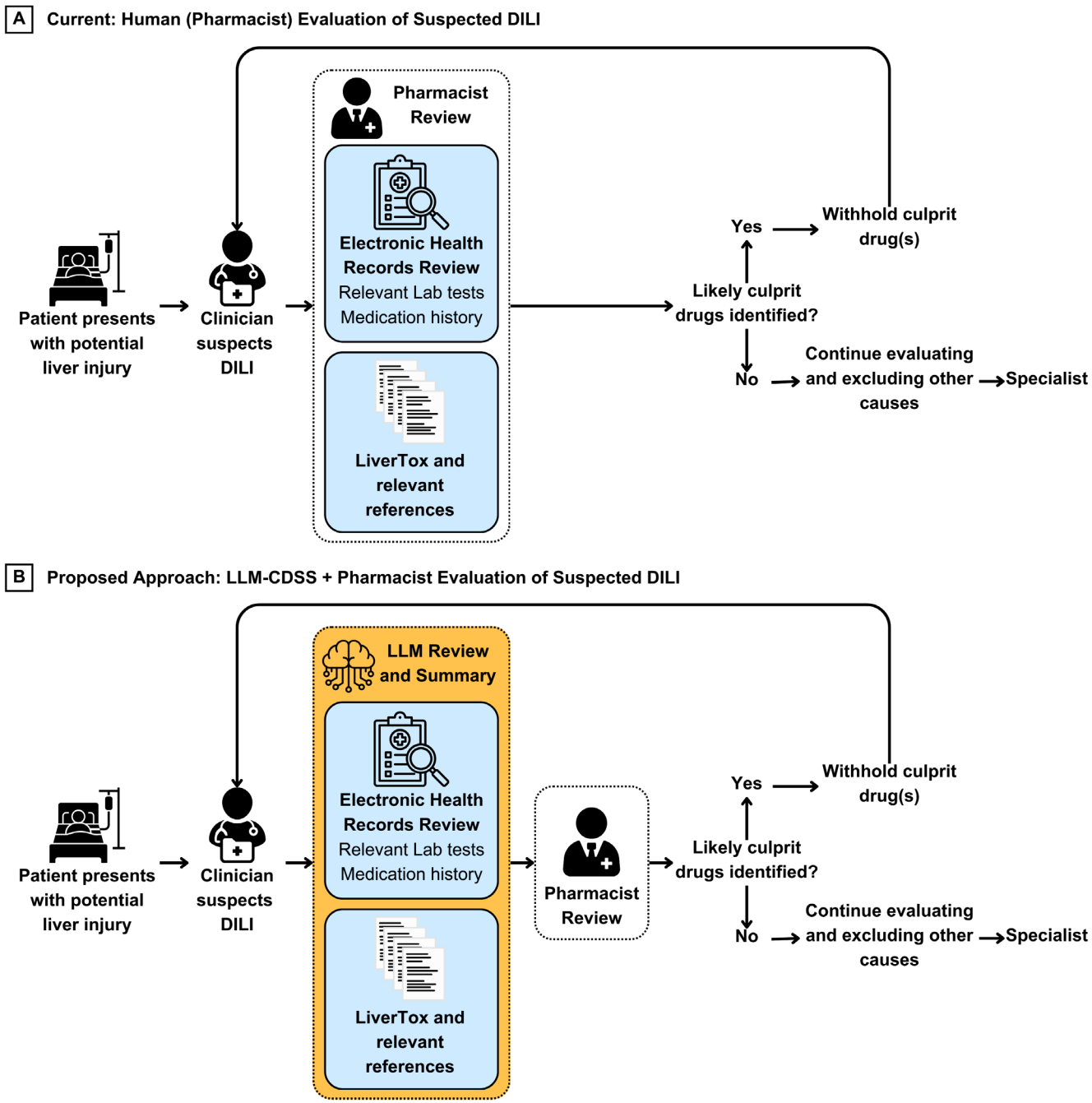

FIGURE 1 | Proposed clinical workflow. CDS, clinical decision support; DILI, drug-induced liver injury; LLM, large language model. DILIConsult is an LLM-driven pipeline designed to provide CDS to clinicians. It is intended to help pharmacists review electronic health records while cross-referencing LiverTox.

pipeline designed to offload the lengthy information-sourcing process yet still allowing clinicians to retain control of the outcome (Figure 1).

2 | Methods

2.1 | Study Design

This retrospective comparative diagnostic study comprised three phases. First, we identified and sampled suspected DILI cases. Second, we conceptualized DILIConsult by evaluating GPT-4's ability to identify DILI characteristics from patient cases and LiverTox descriptions. Finally, we assessed DILIConsult's performance by comparing its recommendations to outputs from a baseline GPT-4 and a panel of clinicians.

2.2 | Dataset Preparation

We retrieved deidentified cases of suspected DILI from the publicly available Medical Information Mart for Intensive Care IV (MIMIC-IV) database, which contains information

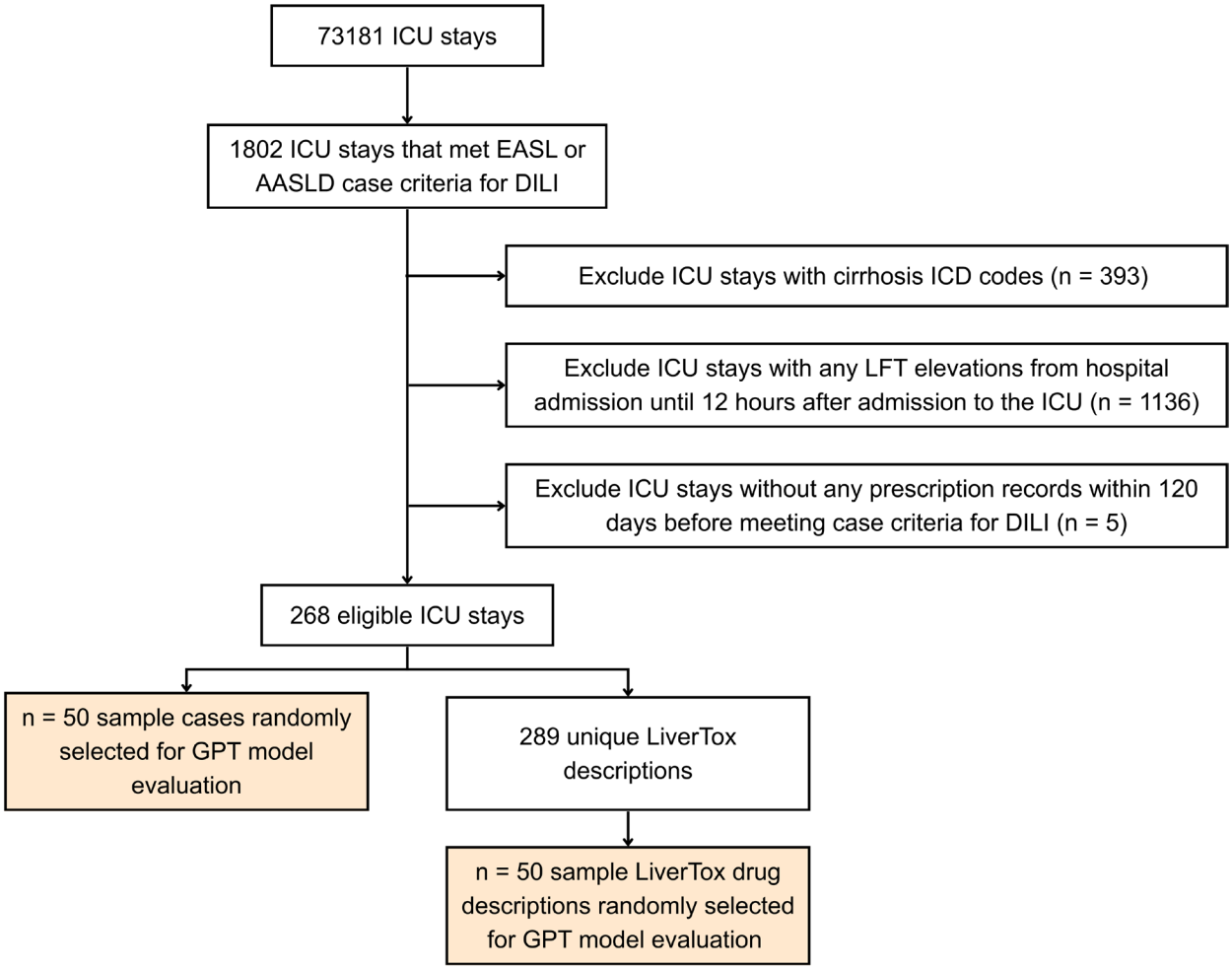

FIGURE 2 | Flow diagram showing selection of patient case and LiverTox descriptions. AASLD, American Association for the Study of Liver Diseases; DILI, drug-induced liver injury; EASL, European Society for the Study of the Liver; GPT, Generative Pre-trained Transformer; ICD, International Classification of Diseases; ICU, Intensive Care Unit; LFT, liver function test. ICU admission meeting DILI case criteria from EASL and AASLD practice guidelines were included. Admissions with competing causes of liver injury, for example liver cirrhosis diagnoses & prior LFT derangements were excluded.

from patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) emergency department or ICU between 2008 and 2019. The creation and distribution of MIMIC-IV, including waiver of informed consent, were approved by the BIDMC Institutional Review Board [23]. Prerequisite training in human-subjects research was completed before the dataset was accessed.

ICU stays were included based on European Association for the Study of the Liver (EASL) [6] or American Association for the Study of Liver Diseases (AASLD) [7] criteria for clinically significant DILI. We excluded ICU stays with any of the following: (i) assigned liver cirrhosis International Classification of Diseases (ICD) codes, (ii) serum liver enzyme elevations above upper limit of normal from hospital admission through 12 h after ICU admission, and (iii) not prescribed with any drugs in the 120 days before meeting DILI criteria.

For each stay, we extracted prescription records and laboratory results spanning 120 days before meeting DILI case criteria until discharge. Prescribed drugs were mapped to a LiverTox description, and each ICU stay was treated as a unique suspected DILI case.

Finally, to generate a representative testing dataset, we randomly sampled 50 LiverTox descriptions and 50 suspected DILI cases. An overview of the dataset preparation process is provided in Figure 2.

2.3 | Development of DILIConsult

2.3.1 | Conception of DILIConsult

DILIConsult consists of specialized GPT-4-driven 'agents' collaborating to perform DILI evaluation and provide substantiated recommendations. An overview of each component is detailed in Figure 3.

Much like multidisciplinary health care teams [24], each agent within this multi-agent framework serves a specialized role. However, consolidating lengthy contexts of patient case details

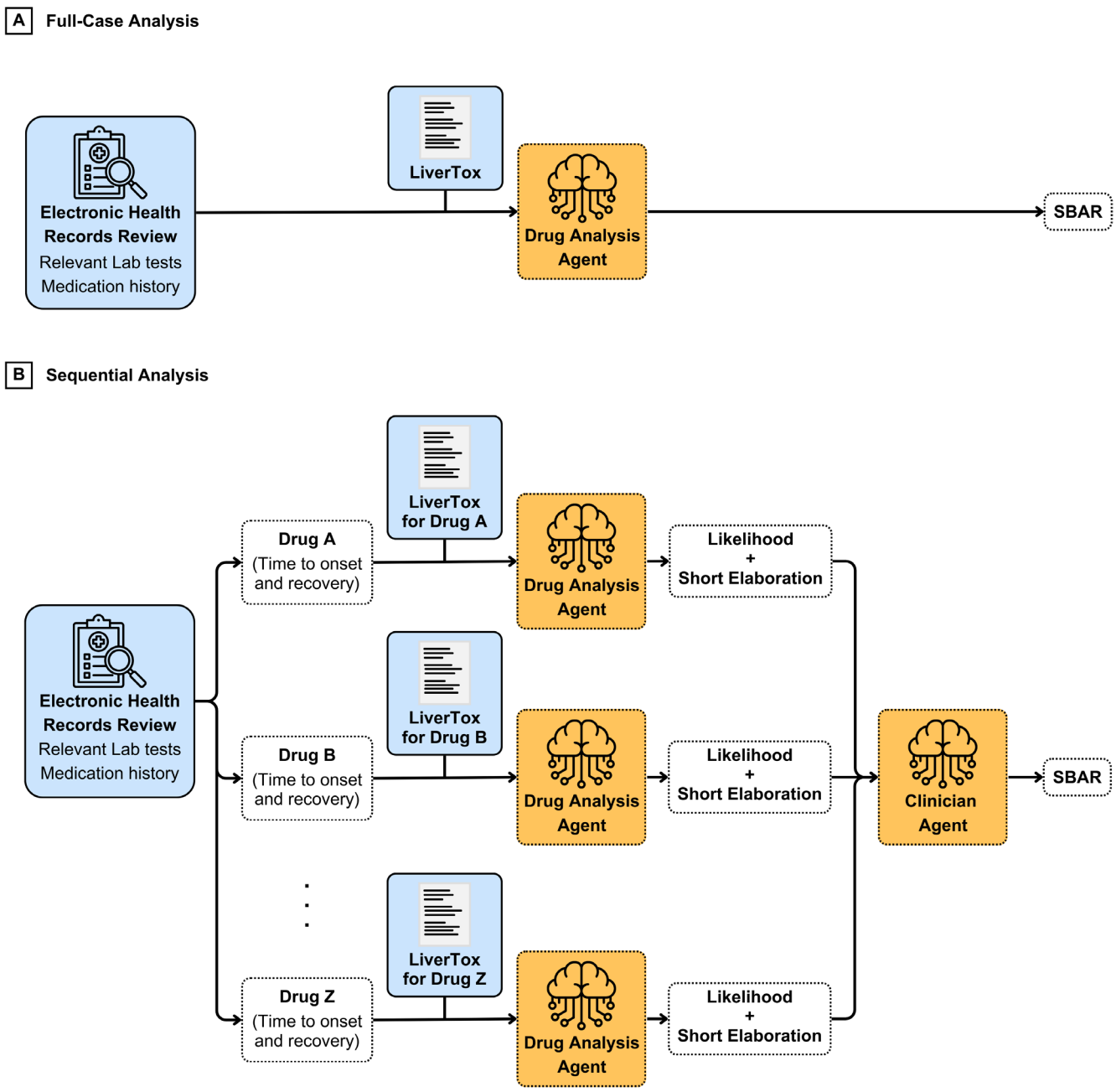

FIGURE 3 | Potential DILIConsult Architecture. GPT, Generative Pre-trained Transformer; SBAR, situation, background, assessment, recommendation. The Drug Analysis Agent and Clinician Agent hold different responsibilities. Drug Analysis Agents analyze individual drugs, and a Clinician Agent provides a SBAR answer. (A) Full-Case Analysis—A single GPT model processes all patient details and reference text (LiverTox). (B) Sequential Analysis—Drugs taken by a patient are processed one-by-one. Sequential Analysis (B) was selected to form DILIConsult architecture due to the better performance (Appendix C).

and medical literature can be a challenging task. We hypothesized that decomposing this task into subtasks of reasoning and fact-checking would improve GPT-4's performance. In this setup, a single LLM performs inference iteratively, focusing on one drug per iteration. The model's findings then flow to a separate summarization agent to form a comprehensive overview. Hence, two types of agents are utilized: the Drug Analysis Agent, which generates focused drug appraisals after each review; and the Clinician Agent, which aggregates the findings without having processed the original reference text.

To enhance the utility of DILIConsult outputs for clinical practice, LLM responses were formatted according to the SBAR (Situation, Background, Assessment, Recommendation) framework, a widely recognized tool that promotes clear and consistent communication [25, 26].

The Drug Analysis Agent was developed through stepwise evaluations of GPT-4's ability to perform three core tasks: (i) identify DILI characteristics from a patient's raw data, (ii) identify DILI characteristics from LiverTox descriptions, and (iii)

compare these characteristics to rationalize a likelihood of injury causality.

In each experiment, prompts were iteratively refined to optimize performance on a small group of test data, after which we proceeded with full-scale testing. To account for variability in responses, each test was repeated over multiple days and mean accuracy was calculated [27, 28]. The full prompt templates are provided in Appendix A.

2.3.2 | Identifying DILI Characteristics From Patient Cases

First, we evaluated GPT-4's ability to extract DILI characteristics using the set of test cases sampled above. GPT-4 was provided with patient laboratory values, R ratios, and the start and end dates of prescribed drugs, and asked to extract three key features in DILI evaluation: pattern of liver injury, onset time, and recovery time. We used the *gpt-4-0125-preview* model of GPT-4, henceforth referred to as GPT-4-Turbo.

The evaluation was divided into two arms. In the first "full-length" arm (Figure 3A), GPT-4-Turbo was given the entire patient case, including all suspect drugs, in a single prompt. We then graded the accuracy of extracted DILI features for all suspect drugs. In the second "sequential" arm (Figure 3B), we provided GPT-4-Turbo with shorter contexts, by including only one suspect drug at a time in each prompt. In both arms, each of the 50 cases was attempted three times.

Responses such as "information not available" or "not recovered" were accepted if they reflected an accurate understanding of the patient's condition. R ratio interpretations from AASLD and definitions of recovery were incorporated into the prompts to guide the determination of liver injury patterns [7, 29–30].

2.3.3 | Identifying DILI Characteristics From LiverTox Descriptions

Next, we assessed GPT-4's ability to extract DILI characteristics from LiverTox, using the "Hepatotoxicity" sections of the 50 sampled LiverTox descriptions. GPT-4 was tasked with identifying the phenotypes of liver injury associated with each drug and corresponding DILI features. Definitions of DILI phenotypes from LiverTox were incorporated into the prompts [31].

We then evaluated GPT-4's responses based on its accuracy in identifying all associated phenotypes and their characteristics. A response was considered accurate only if every injury phenotype along with the corresponding DILI features was correctly identified. Minor variations in the time to onset and recovery were permitted as long as they maintained the same semantic meaning as the original text. In cases where the LiverTox description did not specify a certain feature, an answer signifying an absence like "NIL" was accepted.

This evaluation was also conducted across two model variants: GPT-4-Turbo and *gpt-4o-2024-05-13* (otherwise known as GPT-4o). Each model was run 30 times per LiverTox description.

2.3.4 | Consolidation and Testing of Drug Analysis Agents

Following the initial testing phases, we then designed our Drug Analysis Agents to provide a DILI likelihood rating and explanation for the suspect drug. This model was tested on 10 suspect drugs across 31 patient cases, evaluating the accuracy of its explanations and the prompt was improved iteratively. Missing details that could make the drug less likely to be the cause were also expected to be highlighted. An example of this interaction is provided in Figure 4A,B.

2.4 | Evaluating the Utility of DILIConsult

2.4.1 | Experimental Setup

To complete the DILIConsult pipeline, we integrated the Drug Analysis Agent with a Clinician Agent responsible for generating a consolidated recommendation. (Figure 3) We then tasked DILIConsult with providing recommendations for our sampled cases of DILI. Similar to the previous evaluation phases, prompts were refined until consistent satisfactory SBAR output (Figure 4C) was shown on a small group of test data.

As comparators, we included a base GPT-4 model without modifications and a panel of clinicians consisting of a pharmacist (more than 10 years of experience), a medical officer (more than 5 years), and an associate consultant in critical care (more than 5 years). These comparators were given the same patient cases and reference texts and asked to provide their recommendations in the SBAR format. The model used for both the base GPT-4 and DILIConsult will be the more accurate model in the LiverTox evaluation.

2.4.2 | Expert Panel Grading

To compare the recommendations generated by the three arms, we formed an expert grading panel consisting of two critical care consultants, each with more than 10 years of experience. The panel was provided with recommendations from DILIConsult and comparator arms in a three-way ranking study and asked to evaluate each recommendation across 22 patient cases using the following questions:

- Which answer demonstrates better recall of knowledge? (i.e., mention of an accurate and/or relevant fact for answering the question)
- Which answer contains more inaccurate content?
- Which answer better reflects the current consensus of the scientific and clinical community?
- Which answer demonstrates better reasoning steps?
- Which answer omits more important information?

These questions were adapted from the benchmark of Med-PaLM-2 [33] to assess DILIConsult's ability to reduce hallucinations, utilize retrieved information, and provide well-reasoned recommendations.

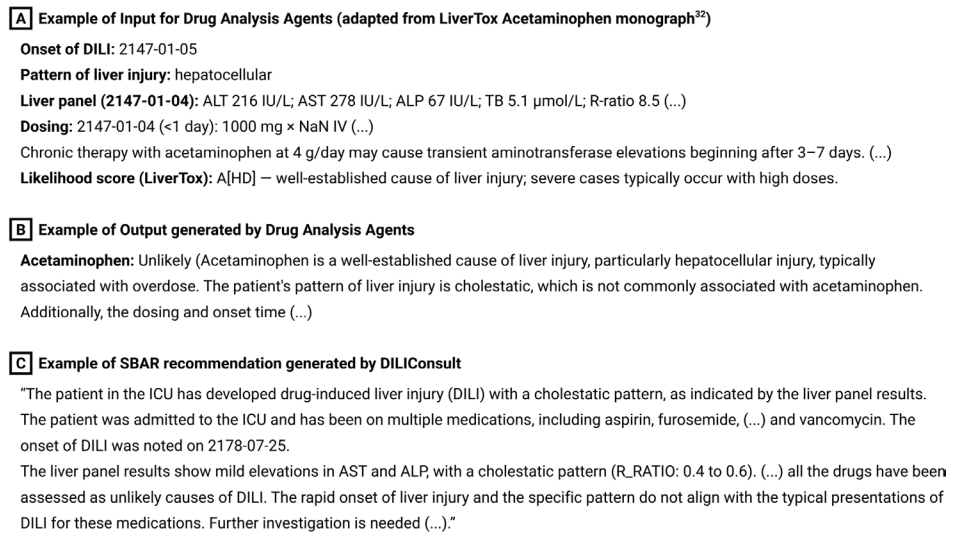

FIGURE 4 | Example input, output and SBAR recommendation. ALT, Alanine aminotransferase; AST, Aspartate aminotransferase; ALP, alkaline phosphatase; ICU, Intensive Care Unit; SBAR, situation, background, assessment, recommendation; TB, total bilirubin. (A) DILI characteristics of the suspected case, followed by text from the "Hepatotoxicity" sections of associated LiverTox monographs, are presented to the Drug Analysis Agent as input. (B) The Drug Analysis Agent provides an output summarizing the likelihood of DILI causality and its rationale. (C) An example of the SBAR recommendation generated by DILIConsult. Text has been truncated for readability.

"The patient in the ICU has developed drug-induced liver injury (DILI) with a cholestatic pattern, as indicated by the liver panel results.

The patient was admitted to the ICU and has been on multiple medications, including aspirin, furosemide, (...) and vancomycin. The onset of DILI was noted on 2178-07-25.

The liver panel results show mild elevations in AST and ALP, with a cholestatic pattern (R_RATIO: 0.4 to 0.6). (...) all the drugs have been assessed as unlikely causes of DILI. The rapid onset of liver injury and the specific pattern do not align with the typical presentations of DILI for these medications.

Further investigation is needed (...)."

FIGURE 4C | Example of SBAR recommendation generated by DILIConsult. AST, Aspartate aminotransferase; ALP, alkaline phosphatase; ICU, Intensive Care Unit; SBAR, situation, background, assessment, recommendation. Text has been truncated for readability.

To ensure blinding, all recommendations were anonymized and presented to the expert panel in randomized order.

2.5 | Statistical Analysis

To evaluate differences in the mean accuracy ± standard deviation (SD) of GPT-4 when analyzing patient cases and LiverTox descriptions, we conducted a paired *t*-test at a one-tailed significance level of 0.05. Friedman's test was applied to identify statistically significant differences in the expert panel's ranking of recommendations, followed by Nemenyi's post hoc test for pairwise comparison at a two-tailed significance level of 0.05.

3 | Results

3.1 | Patient Case Analysis

When assessing the overall DILI evaluation characteristics, the mean accuracy for the full-length approach was 40.9% ± 39.1%, compared to a higher mean accuracy of 52.0% ± 40.4% that used a sequential approach. The mean difference in accuracy was significant at 11.1% (95% confidence interval [CI]: 2.3%–19.9%, $p = 0.007$).

To identify specific areas of strengths and weaknesses, we analyzed GPT-4's performance on individual characteristics. The mean difference in accuracy for onset and recovery was statistically significant at 9.4% (95% CI: 2.7%–16.0%, $p = 0.003$) and 11.7% (95% CI: 2.8%–20.5%, $p = 0.005$), respectively. However, the difference in identification of the pattern of liver injury was not statistically significant at −0.4% (95% CI: −5.3% to 4.4%, $p = 0.43$) in this analysis.

The overall percentage of error-free attempts was 26.7% and 16.0% for the full-length approach and sequential approach, respectively.

Common mistakes made in the full-length approach were incomplete responses (22/71, 31.0%) and incorrect timing calculations (43/71, 60.6%) for onset. Comparatively, all errors for

TABLE 2 | Mean ranks for various models across different axes.

	Mean ranks (SD)[a]			
	Clinician	**Base GPT-4**	**DILIConsult**	***p***
Knowledge Recall (Q1)	2.30 (0.72)	2.05 (0.62)	1.66 (0.75)	0.06
Less inaccurate content (Q2)	1.98 (0.61)	1.93 (0.66)	2.09 (0.72)	0.79
Reflects current consensus (Q3)	1.64 (0.82)	2.32 (0.75)	2.05 (0.62)	0.05
Reasoning steps (Q4)	1.93 (0.94)	2.06 (0.74)	2.00 (0.64)	0.89
Less omission of important information (Q5)	1.91 (0.75)	2.02 (0.63)	2.07 (0.52)	0.78

Abbreviation: SD, standard deviation.
[a]Mean rank values are reported, with lower value ranks indicating better performance (i.e., 1 = best, 3 = worst).

onset in the sequential approach came from wrongly extracted durations (115/115, 100%). For the recovery characteristic, the highest proportion of errors for both approaches stemmed from incorrectly assuming that certain information was present, such as inferring that the patient has recovered when they have not. (41/89, 46.1% for full-length approach; 73/97, 75.3% for sequential approach).

Strong performance in identifying the pattern of liver injury indicates a solid understanding of the definitions provided to the model. The sequential approach was selected for the DILIConsult architecture given the higher accuracy. Even though mistakes were still present, these errors were consistent and could be mitigated by pre-processing the features involving durations.

The above statistical analysis and errors are reported in Appendix C.

3.2 | Analysis of LiverTox Descriptions

Overall, GPT-4 identified DILI phenotypes from LiverTox descriptions with considerable accuracy. After repeated testing, GPT-4-Turbo identified associated DILI phenotypes and all of their characteristics, if they were present, with an accuracy of 81.5% ± 35.6%. In comparison, GPT-4o had a higher mean accuracy of 96.9% ± 15.0%. A significant difference in mean accuracy was found between the two models (mean difference = 15.4%, 95% CI: 4.5%–26.4%, $p = 0.003$).

The most common mistake (182/334, 54.5%) made by GPT-4-Turbo was failing to identify a liver injury phenotype mentioned in the LiverTox descriptions. Notably, such mistakes were often consistent across multiple attempts of the same case, suggesting a systemic misinterpretation. Conversely, if all phenotypes were correctly identified in an initial attempt, the model consistently maintained this accuracy across all subsequent attempts.

Furthermore, for certain LiverTox descriptions, GPT-4-Turbo could recognize that a drug was associated with liver injury presenting with multiple patterns of liver enzyme elevations (e.g., cholestatic and mixed patterns), but it struggled to specify the exact phenotypes (cholestatic and mixed hepatitis, respectively). Instead, GPT-4-Turbo tended to conflate these into a single phenotype with two possible patterns of liver enzyme elevations. This suggested that certain texts may have included subtle distinctions between liver injury phenotypes and their patterns of enzyme elevations that were challenging for GPT-4-Turbo to accurately discern. Although GPT-4o was not free of such mistakes, an overall error reduction of 86.2% was seen. A detailed breakdown of the types of incorrect responses given by GPT-4-Turbo and GPT-4o can be found in Appendix C. As GPT-4o was the more accurate model in all parameters, it was used in DILIConsult.

3.3 | Comparison Between DILIConsult, Base GPT, and Human Analysis

Across our evaluation criteria, we observed mixed performance in the mean rankings of DILIConsult's recommendations compared to the Clinician and base GPT-4 arms (Table 2). However, Friedman's test indicated that these differences were not statistically significant.

Despite this, DILIConsult demonstrated stronger performance in the Knowledge Recall aspect where it achieved a mean rank of 1.66 ± 0.75 ($p = 0.06$), surpassing both the Clinician and base GPT-4. In terms of medical consensus (2.05 ± 0.62 vs. 1.64 ± 0.82, $p = 0.05$) and reasoning (2.00 ± 0.64 vs. 1.93 ± 0.94, $p = 0.89$), DILIConsult ranked second behind the Clinician arm, respectively. However, in the areas of omitting important information ($p = 0.78$) and inaccuracy ($p = 0.79$) in content, DILIConsult ranked slightly behind the other two arms, with mean ranks of 2.07 ± 0.52 and 2.09 ± 0.72, respectively.

The Nemenyi's post hoc test was not conducted as there was no statistical significance found.

An aggregate performance score and individual case performance are presented in a graphical manner in Figure 5. Although no statistical significance was reached for each individual criterion, overall preference appears to trend towards the clinician answers.

4 | Discussion

Our study is the first in describing the use of context-aware LLMs to facilitate clinical evaluation of DILI. We found that GPT-4o was able to achieve acceptable levels of accuracy in identifying DILI phenotypes and characteristics from LiverTox descriptions as

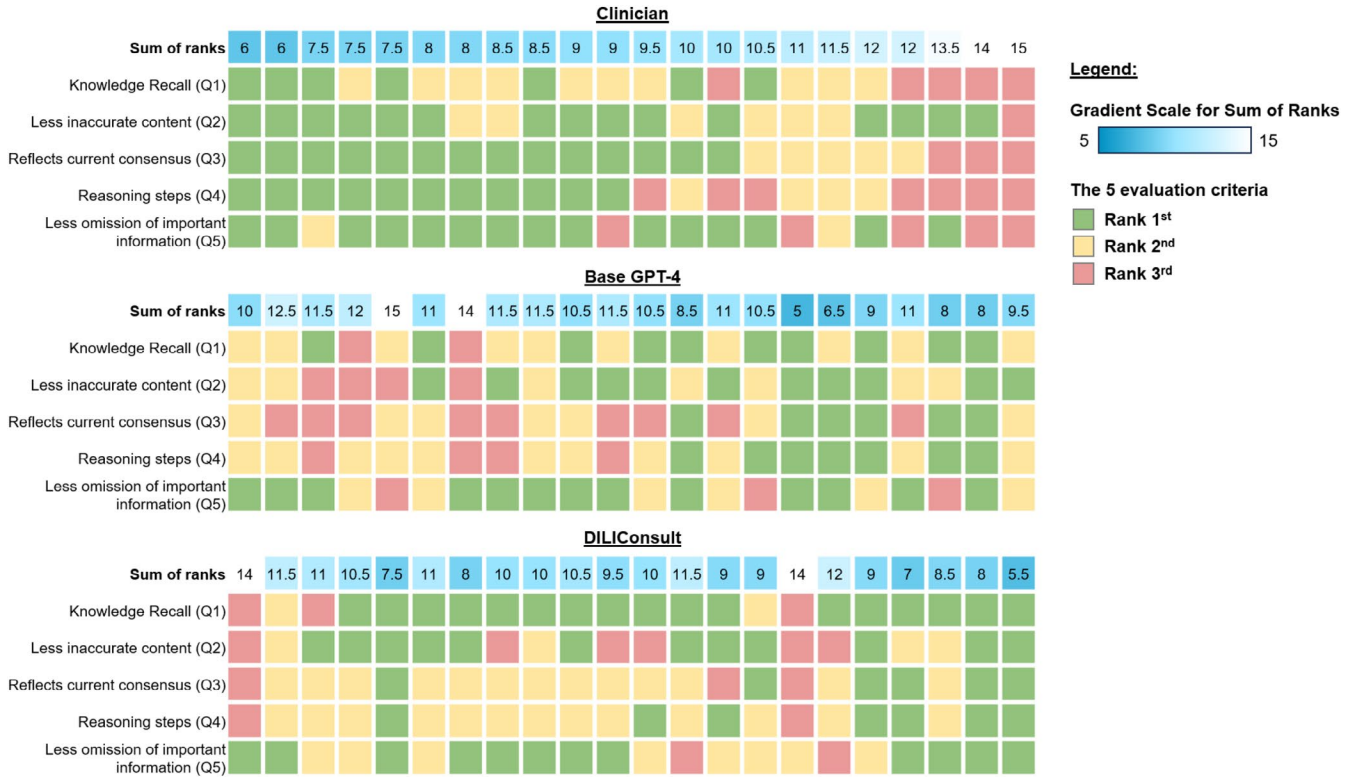

FIGURE 5 | Heatmap showing sum of ranks and individual case preference. Heatmap displaying performance of the 22 cases across the three experimental arms (Clinician, Base GPT-4, and DILIConsult). Each column corresponds to a case, with the sum of rank presented as the column header and the breakdown of preferences for each criterion shown in the subsequent rows. Green denotes 1st rank, yellow denotes 2nd rank, and red denotes 3rd rank. For Questions 2 and 5, the ranking scale was reversed so that the higher-performing responses are consistently shown as green. Overall, a lower sum of ranks and more green cases reflect a preferred experimental arm.

opposed to GPT-4-Turbo. In the patient case analysis, several drugs were missed by GPT in its report, but this issue was mitigated in the sequential approach which narrowed the task and eliminated the need for the GPT to segment the task on its own. When determining the time to onset and recovery, statistically significant improvements of using the sequential approach were observed. However, the overall accuracy was still insufficient. To address this, the DILI characteristics were pre-processed and included in the Drug Analysis Agent prompt. (see Appendix A for example).

In the three-way comparison, no statistically significant preference was found among expert graders for all five criteria. DILIConsult generally supported its argument with more information from the case and reference text, whereas the base GPT-4 provided more condensed explanations. We also noted that clinicians sometimes proposed that none of the drugs were potential contributors to DILI, whereas both DILIConsult and base GPT-4 always recommended discontinuing at least one drug. This may reflect a tendency for sycophancy in the LLM, where the generated content aims to align with the prompt (in this case, that there is a drug in the set examined that contributed to DILI), rather than necessarily challenging the assumption. This can be tackled in future work, potentially by refining the prompting strategy.

Nonetheless, both LLM approaches occasionally overlooked drugs which are highly likely causes, such as doxorubicin. In another case, only the clinicians connected an elevated INR to the use of warfarin instead of liver injury, suggesting that the careful prompting and exclusive use of DILI references may have reduced the LLM's ability to consider alternative explanations.

Our study aligns with prior work which has shown comparable performance of generative artificial intelligence (AI) to nonexpert physicians and physicians in general, but worse than expert physicians [34]. However, our approach provides greater transparency of the LLM's reasoning processes and addresses a key concern of clinicians and a barrier to overcome for implementation into health care settings [35, 36]. Although the individual outputs of the Drug Analysis Agents were not evaluated, they are able to provide interpretable insights into the eventual SBAR communication output. Some studies which evaluate the potential of Multi-Agents have shown favorable outcomes [10, 37], but our study was the first to assess the performance of a Multi-Agent framework against clinicians on clinical cases, to our knowledge. With the use of the well-established SBAR framework, our study facilitates the translation of our findings into real-world clinical practice. The intentional use of SBAR format to promote safe effective communications addresses a key gap seen often in the busy clinical space and is structured to facilitate clearer presentation of poor reasoning or hallucination by the LLM as a safety net when utilizing such technology at current form.

However, using multiple agents in a single case adds complexity. Like large teams, more agents may increase the risk of

miscommunication or inconsistent output. This underlines the importance of striking an optimal balance in the number of agents and careful design to ensure coherent communication. In future iterations of the multi-agent system, more safeguards should be implemented to ensure consistency and alignment across agents.

A key limitation of our study was the small sample size of only 22 cases of suspected DILI included in our final evaluation of DILIConsult. The limited number of cases reduced the power of our analysis and made it challenging to detect meaningful differences between the study arms. Another limitation is that additional patient data, including radiography reports, were not utilized. This limits the scope of DILIConsult to drug-related differentials, which may be resolved by adding agents to exclude other diagnoses.

Furthermore, the use of multiple agents or agentic workflows involves higher compute power, results in longer latency between input and output, as well as incurs higher costs due to more calls to the application programming interface (API). These factors pose challenges to real world implementation and limit scalability.

Lastly, the standardized prompting used for LLM responses may have introduced distinguishable language patterns in their output, potentially compromising blinding. Prior work shows that LLM-generated scientific manuscripts differ from human writing in measurable areas such as readability and lexical diversity, making them identifiable to human expert reviewers [38, 39]. Future studies should consider prompt variation, stylistic masking, or human-in-the-loop editing to better balance linguistic features and preserve blinding integrity.

5 | Conclusion

We introduced DILIConsult, a novel LLM-based tool for DILI evaluation which demonstrated some promise in knowledge recall and generating interpretable outputs. Previous works have explored the use of conventional natural language processing (e.g., Word2Vec, BERT) or machine learning methods (e.g., random forest) to summarize literature and classify potential DILI [40–42]. With the advent of LLMs, clinician-AI interaction is becoming more natural and intuitive by tapping into the natural language processing capabilities. Additionally, conversational interfaces allow dynamic, context-aware exchanges, aiding clinicians in understanding and integrating AI recommendations. We propose that DILIConsult can bridge the gap between AI and user adoption, towards seamless integration into clinical practice.

Author Contributions

Alfred Zheng Ting Ho: conceptualization, formal analysis, investigation, methodology, software, visualization, writing – original draft. **Jeren Zheng Feng Law:** conceptualization, formal analysis, investigation, methodology, software, visualization, writing – original draft. **Aiwen Wang:** methodology, project administration, supervision, writing – review and editing. **Daniel Yan Zheng Lim:** methodology, supervision, writing – review and editing. **Jasmine Chiat Ling Ong:** conceptualization, data curation, investigation, methodology, project administration, writing – review and editing, supervision.

Acknowledgments

This work was supported by the National University of Singapore, Department of Pharmacy and Pharmaceutical Science.

Funding

This research was funded by the National University of Singapore, Department of Pharmacy and Pharmaceutical Science.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data used in this study were obtained from the Medical Information Mart for Intensive Care IV (MIMIC-IV), which is publicly available but requires completion of ethics training and signing a data use agreement. Researchers can request access via PhysioNet (https://physionet.org/). All study-specific prompts and sample model outputs generated during this study are provided in the Supporting Information.

References

1. M. Li, Y. Wang, T. T. Lv, et al., "Mapping the Incidence of Drug-Induced Liver Injury: A Systematic Review and Meta-Analysis," *Journal of Digestive Diseases* 24, no. 5 (2023): 332–339, https://doi.org/10.1111/1751-2980.13205.

2. S. Weber and A. L. Gerbes, "Challenges and Future of Drug-Induced Liver Injury Research—Laboratory Tests," *International Journal of Molecular Sciences* 23, no. 11 (2022): 6049, https://doi.org/10.3390/ijms23116049.

3. S. David and J. P. Hamilton, "Drug-Induced Liver Injury," *US Gastroenterology and Hepatology Review* 6 (2010): 73–80.

4. P. H. Hayashi, D. Rockey, R. J. Fontana, et al., "Death and Liver Transplantation Within Two Years of Onset of Drug-Induced Liver Injury," *Hepatology* 66, no. 4 (2017): 1275–1285, https://doi.org/10.1002/hep.29283.

5. K. T. Suk and D. J. Kim, "Drug-Induced Liver Injury: Present and Future," *Clinical and Molecular Hepatology* 18, no. 3 (2012): 249–257, https://doi.org/10.3350/cmh.2012.18.3.249.

6. R. J. Andrade, G. P. Aithal, E. S. Björnsson, et al., "EASL Clinical Practice Guidelines: Drug-Induced Liver Injury," *Journal of Hepatology* 70, no. 6 (2019): 1222–1261, https://doi.org/10.1016/j.jhep.2019.02.014.

7. R. J. Fontana, I. Liou, A. Reuben, et al., "AASLD Practice Guidance on Drug, Herbal, and Dietary Supplement-Induced Liver Injury," *Hepatology* 77, no. 3 (2023): 1036–1065, https://doi.org/10.1002/hep.32689.

8. LiverTox: Clinical and Research Information on Drug-Induced Liver Injury, "National Institute of Diabetes and Digestive and Kidney Diseases" (2012), http://www.ncbi.nlm.nih.gov/books/NBK547852/.

9. R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success," *npj Digital Medicine* 3 (2020): 17, https://doi.org/10.1038/s41746-020-0221-y.

10. N. Mehandru, B. Y. Miao, E. R. Almaraz, M. Sushil, A. J. Butte, and A. Alaa, "Evaluating Large Language Models as Agents in the Clinic," *npj Digital Medicine* 7, no. 1 (2024): 84, https://doi.org/10.1038/s41746-024-01083-y.

11. M. Moritz, E. Topol, and P. Rajpurkar, "Coordinated AI Agents for Advancing Healthcare," *Nature Biomedical Engineering* 9, no. 4 (2025): 432–438, https://doi.org/10.1038/s41551-025-01363-2.

12. K. B. Nguyen, S. Jacobs, N. Tasnim, and J. P. Knorr, "Evaluation of a Clinical Decision Support Alert to Identify Hepatic Dysfunction and Need for Medication Therapy Adjustment in Hospitalized Patients," *American Journal of Health-System Pharmacy* 82, no. 2 (2025): S2885–S2893, https://doi.org/10.1093/ajhp/zxae327.

13. P. L. Smithburger, M. S. Buckley, S. Bejian, K. Burenheide, and S. L. Kane-Gill, "A Critical Evaluation of Clinical Decision Support for the Detection of Drug–Drug Interactions," *Expert Opinion on Drug Safety* 10, no. 6 (2011): 871–882, https://doi.org/10.1517/14740338.2011.583916.

14. Causality, *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury* (National Institute of Diabetes and Digestive and Kidney Diseases, 2012), http://www.ncbi.nlm.nih.gov/books/NBK548049/.

15. D. C. Rockey, L. B. Seeff, J. Rochon, et al., "Causality Assessment in Drug-Induced Liver Injury Using a Structured Expert Opinion Process: Comparison to the Roussel-Uclaf Causality Assessment Method," *Hepatology* 51, no. 6 (2010): 2117–2126, https://doi.org/10.1002/hep.23577.

16. T. H. Kung, M. Cheatham, A. Medenilla, et al., "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models," *PLOS Digital Health* 2, no. 2 (2023): e0000198, https://doi.org/10.1371/journal.pdig.0000198.

17. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large Language Models in Medicine," *Nature Medicine* 29, no. 8 (2023): 1930–1940, https://doi.org/10.1038/s41591-023-02448-8.

18. B. R. Beaulieu-Jones, M. T. Berrigan, S. Shah, J. S. Marwaha, S. L. Lai, and G. A. Brat, "Evaluating Capabilities of Large Language Models: Performance of GPT-4 on Surgical Knowledge Assessments," *Surgery* 175, no. 4 (2024): 936–942, https://doi.org/10.1016/j.surg.2023.12.014.

19. 17. T. Li, G. Zhang, Q. D. Do, X. Yue, and W. Chen, "Long-Context LLMs Struggle With Long In-Context Learning," *Transactions on Machine Learning Research* (2024), https://openreview.net/forum?id=Cw2xlg0e46.

20. S. Yao, J. Zhao, D. Yu, et al., "ReAct: Synergizing Reasoning and Acting in Language Models" (2023), https://doi.org/10.48550/arXiv.2210.03629.

21. Z. Fan, L. Wei, J. Tang, et al., "AI Hospital: Benchmarking Large Language Models in a Multi-Agent Medical Interaction Simulator," in *Proceedings of the 31st International Conference on Computational Linguistics*, ed. O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert (Association for Computational Linguistics, 2025), 10183–10213.

22. H. Chen, S. Zhang, L. Zhang, et al., "Multi Role ChatGPT Framework for Transforming Medical Data Analysis," *Scientific Reports* 14, no. 1 (2024): 13930, https://doi.org/10.1038/s41598-024-64585-5.

23. A. E. W. Johnson, L. Bulgarelli, L. Shen, et al., "MIMIC-IV, a Freely Accessible Electronic Health Record Dataset," *Scientific Data* 10, no. 1 (2023): 1, https://doi.org/10.1038/s41597-022-01899-x.

24. C. Harrington, "Composition of an Ideal Medical Care Team," *Delaware Journal of Public Health* 8, no. 5 (2022): 150–153, https://doi.org/10.32481/djph.2022.12.033.

25. M. Müller, J. Jürgens, M. Redaèlli, K. Klingberg, W. E. Hautz, and S. Stock, "Impact of the Communication and Patient Hand-Off Tool SBAR on Patient Safety: A Systematic Review," *BMJ Open* 8, no. 8 (2018): e022202, https://doi.org/10.1136/bmjopen-2018-022202.

26. M. Leonard, S. Graham, and D. Bonacum, "The Human Factor: The Critical Importance of Effective Teamwork and Communication in Providing Safe Care," *BMJ Quality and Safety* 13, no. 1 (2004): i85–i90, https://doi.org/10.1136/qshc.2004.010033.

27. H. Arora, N. Kaplan-Damary, and H. S. Stern, "Combining Reproducibility and Repeatability Studies With Applications in Forensic Science," *Law, Probability, and Risk* 22, no. 1 (2023): mgad007, https://doi.org/10.1093/lpr/mgad007.

28. J. M. Franc, L. Cheng, A. Hart, R. Hata, and A. Hertelendy, "Repeatability, Reproducibility, and Diagnostic Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Emergency Department Triage Using the Canadian Triage and Acuity Scale," *Canadian Journal of Emergency Medicine* 26, no. 1 (2024): 40–46, https://doi.org/10.1007/s43678-023-00616-w.

29. G. Danan and R. Teschke, "RUCAM in Drug and Herb Induced Liver Injury: The Update," *International Journal of Molecular Sciences* 17, no. 1 (2015): 14, https://doi.org/10.3390/ijms17010014.

30. P. H. Hayashi, M. I. Lucena, R. J. Fontana, et al., "A Revised Electronic Version of RUCAM for the Diagnosis of Drug Induced Liver Injury," *Hepatology* 76, no. 1 (2022): 18–31, https://doi.org/10.1002/hep.32327.

31. Phenotypes of Drug Induced Liver Injury," in *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury* (National Institute of Diabetes and Digestive and Kidney Diseases, 2012), http://www.ncbi.nlm.nih.gov/books/NBK548473/.

32. Acetaminophen," in *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury* (National Institute of Diabetes and Digestive and Kidney Diseases, 2012), http://www.ncbi.nlm.nih.gov/books/NBK548162/.

33. K. Singhal, S. Azizi, T. Tu, et al., "Large Language Models Encode Clinical Knowledge," *Nature* 620, no. 7972 (2023): 172–180, https://doi.org/10.1038/s41586-023-06291-2.

34. H. Takita, D. Kabata, S. L. Walston, et al., "A Systematic Review and Meta-Analysis of Diagnostic Performance Comparison Between Generative AI and Physicians," *npj Digital Medicine* 8, no. 1 (2025): 175, https://doi.org/10.1038/s41746-025-01543-z.

35. E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and Barriers of Using Large Language Models (LLM) Such as ChatGPT for Diagnostic Medicine With a Focus on Digital Pathology – A Recent Scoping Review," *Diagnostic Pathology* 19 (2024): 43, https://doi.org/10.1186/s13000-024-01464-7.

36. S. Reddy, W. Rogers, V. P. Makinen, et al., "Evaluation Framework to Guide Implementation of AI Systems Into Healthcare Settings," *BMJ Health & Care Informatics* 28, no. 1 (2021): e100444, https://doi.org/10.1136/bmjhci-2021-100444.

37. L. Yue, S. Xing, J. Chen, and T. Fu, "ClinicalAgent: Clinical Trial Multi-Agent System With Large Language Model-Based Reasoning," in *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. BCB '24* (Association for Computing Machinery, 2024), 1–10, https://doi.org/10.1145/3698587.3701359.

38. R. Khera, A. F. Pedroso, V. K. Keloth, H. Xu, G. S. Silva, and L. H. Schwamm, "Scientific Writing in the Era of Large Language Models: A Computational Analysis of AI- Versus Human-Created Content," *Stroke* 56, no. 10 (2025): 3078–3083, https://doi.org/10.1161/STROKEAHA.125.051913.

39. K. R. Nathani, A. M. Nathani, M. Delawan, A. Safdar, and M. Bydon, "Can Artificial Intelligence Write Science? A Comparative Analysis of Human-Written and Artificial Intelligence-Generated Scientific Writings," *Journal of Neurosurgery. Spine* 43, no. 6 (2025): 767–772, https://doi.org/10.3171/2025.4.SPINE25519.

40. H. Hong, S. Thakkar, M. Chen, and W. Tong, "Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using A Large Set of FDA-Approved Drugs," *Scientific Reports* 7, no. 1 (2017): 17311, https://doi.org/10.1038/s41598-017-17701-7.

41. Y. Wu, Z. Liu, L. Wu, M. Chen, and W. Tong, "BERT-Based Natural Language Processing of Drug Labeling Documents: A Case Study

for Classifying Drug-Induced Liver Injury Risk," *Frontiers in Artificial Intelligence* 4 (2021): 729834, https://doi.org/10.3389/frai.2021.729834.

42. J. H. Oh, A. Tannenbaum, and J. O. Deasy, "Improved Prediction of Drug-Induced Liver Injury Literature Using Natural Language Processing and Machine Learning Methods," *Frontiers in Genetics* 14 (2023): 1161047, https://doi.org/10.3389/fgene.2023.1161047.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** phar70131-sup-0001-DataS1.docx.

Appendix A

Development Process of Drug Analysis Agent and Clinician Agent

For each experiment, we used the LangChain library in Python to interface with OpenAI's API (temperature = 0 for the least random result; all other settings at default values).

The prompt for the Drug Analysis Agent was developed through a stepwise evaluation. Each prompt comprises a System Message and a Human Message. The System Message is shared between both Drug Analysis Agents and Clinician Agents as an overarching study background. The Human Message is customized so that each agent has a different task and context. An example of the prompt is provided below. In this appendix, the prompt is divided into sections for easier reading but is presented as one continuous text to the LLM.

System Message

The System Message is used by LangChain schema to provide initial instructions to the LLM.

Sections of the prompt	Remarks
The aim of this study is to gather expert assessments on a series of patient cases where drug-induced liver injury (DILI) is a concern. Your analysis as a healthcare professional is vital in categorizing the likelihood of certain drugs causing liver injury in these cases. You are required to consider whether the information from LiverTox applies to the specific patient case.	To provide background on the task.

Human Message (Drug Analysis Agent)

The main bulk of the prompt, the context, is provided as a Human Message.

Sections of the prompt	Remarks
You are a doctor. Your task is to assess the likelihood category for a single drug in being the cause of liver injury. Explain your thoughts clearly and succinctly, as if you are presenting to your consultant. You are just looking at one drug. Your evaluation will be used by another LLM to evaluate which of the drugs are most likely to be the cause of DILI in the patient case. Hence, keep it succinct but include all important information.	To define the role of the LLM.
## Before you start: – Even though you should take the likelihood SCORE from LiverTox into account (A, B, C, D, E, or X), you should also consider other factors of clinical phenotype, onset time and recovery time. Do not merely translate the SCORE into the CATEGORY. – The only information given regarding the patient is: Liver panel, onset time, recovery time and dosing. All other missing information (e.g., symptoms) does not mean an absence of those findings, so you should encourage investigation into those areas.	Task guidelines are provided. Common misinterpretations identified when evaluating GPT's performance on LiverTox text (Section 2.3.3) are addressed in this segment.
## For the likelihood score, please refer to the definitions below: – Definite (> 95% chance): Liver injury is typical for the drug or herbal product ("signature" or pattern of injury, timing of onset, recovery). The evidence for causality is "beyond a reasonable doubt". – Highly Likely (75%–95%): The evidence for causality is 'clear and convincing' but not definite. Probable (50%–74%): The causality is supported by 'the preponderance of evidence' as implicating the drug but the evidence cannot be considered definite or highly likely. – Possible (25%–49%): The causality is not supported by 'the preponderance of evidence'; however, one cannot definitively exclude the possibility. – Unlikely (< 25%): The evidence for causality is 'highly unlikely' based upon the available information. – Insufficient data: Key elements of the drug exposure history, initial presentation, alternative diagnoses and/or diagnostic evaluation prevent one from determining a causality score.	To guide the LLM in consistent word usage, definitions on subjective terms were provided. This helps to guide shared use of words where semantic meaning may change in other contexts.
## Follow these steps in your analysis. Step 1: Identify whether a drug is associated with liver injury. Step 2: Identify the phenotypes of liver injury associated with the drug. A drug can be associated with one phenotype or multiple phenotypes. Note that you should only identify phenotypes from the list provided. Step 3: For every identified phenotype, identify the associated pattern(s) of liver injury, onset time, and recovery time. Tips for identifying phenotypes and their associated pattern(s) of liver injury: – "Enzyme elevations without jaundice" (liver test abnormalities) phenotype: predominant ALT and AST (aminotransferases) elevations are associated with a hepatocellular pattern of liver injury, while predominant ALP elevations are associated with a cholestatic pattern. If the text does not mention which specific liver enzymes are elevated, there is no associated pattern of liver injury for this phenotype, so answer FALSE for "mixed," "hepatocellular" and "cholestatic". – "Mixed hepatitis" phenotype is ALWAYS associated with a mixed pattern of liver injury only. "Acute hepatitis" phenotype is ALWAYS associated with a hepatocellular pattern of liver injury only. "Cholestatic hepatitis" phenotype is ALWAYS associated with a cholestatic pattern of liver injury only. If the text describes multiple injury patterns, make sure to capture that these are distinct injury phenotypes (e.g., (e.g., "pattern of liver injury varies from cholestatic to mixed to hepatocellular" should be captured as three separate phenotypes: acute hepatitis, mixed hepatitis and cholestatic hepatitis instead of one phenotype with three injury patterns) – "Chronic hepatitis": typically hepatocellular pattern of injury with persistent enzyme elevations Step 4: Compare the patient case and LiverTox text by clearly stating the observations for each side. Step 5: Identify specific missing information (e.g., symptoms, other lab tests) which can be investigated further.	To guide the LLM in reasoning, steps were included after development from prior evaluation (Section 2.3.3). Phenotype definitions were also provided where deficits in understanding were identified.

Sections of the prompt	Remarks
Patient Case Onset of DILI: 2184-07-04 Clinical Phenotype: mixed The recovery period starts from the peak of liver injury to the point where either of the following are fulfilled: 1. *R*-value ≥ 5 AND ALT levels fall to below 50% of their peak; OR 2. *R*-value < 5, AND ALP or Bilirubin levels decline to below 50% of their peak. If the above criteria are not met, the patient has not yet recovered. # Medication Name: Linezolid - Onset: 7 days - Recovery: 5 days - Dosing: - 2184-06-27 to 2184-07-03 (6 days): 600 mg x2.0 IV - 2184-07-04 to 2184-07-08 (5 days): 600 mg x2.0 PO/NG # Liver Panel Result: 2184-06-25 22:50:00: ALT: 16.0 IU/L, AST: 33.0 IU/L, ALP: 64.0 IU/L, TB: 10.3 umol/L, R_RATIO: 0.8 2184-06-28 02:18:00: ALT: 6.0 IU/L, AST: 18.0 IU/L, ALP: 57.0 IU/L, TB: 82.1 umol/L, R_RATIO: 0.3 (...) Upper Limits of Normal (ULN): ALT: 40.0 IU/L, AST: 40.0 IU/L, ALP: 130.0 IU/L, TB: 25.6 umol/L	Patient case details are provided in a standardized format after rounds of testing. Onset, recovery and clinical phenotype were provided, as the LLM in Section 2.3.2 was found to face challenges in translating a numerical scale to a categorical phenotype. This prompt structure was developed in Section 2.3.2.
# LiverTox Text Chapter: Linezolid Therapy with linezolid has been associated with mild and transient elevations in serum aminotransferase and alkaline phosphatase levels in 1% to 10% of patients (...)	Context from reference sources is provided in the same prompt.
Pydantic Schema: Drug	Refer to Appendix B for Pydantic schema.

Human Message (Clinician Agent)

Sections of the Prompt	Purpose
You are part of a group of clinicians. Your task is to present your analysis of a patient case in the ICU. Objective: Determine which of the drugs are most likely to be the cause of DILI. Present your answer in SBAR format.	To define the role of the LLM.
You should provide specific recommendations. Explain your thoughts clearly and succinctly, as if you are presenting to your consultant. ## This is what makes a good SBAR: Situation—What is going on with the patient? What is the situation you are communicating about? Background—What is the background or context on this patient? (e.g., patient's diagnosis, or reason for admission, medical status, relevant history and patient chart review) Do NOT list out everything. Assessment—What is the problem? Give specific information on recent laboratories. This section can include a clinical impression. Recommendation—What is the next step in the management of the patient? An informed suggestion for the continued care of the patient. The immediate need is explained clearly and specifically, including what is necessary to address the problem.	Guiding questions on what constitutes a good SBAR (Situation, Background, Assessment, Recommendation) is provided.
## Important guidelines: – Support your assessment of the patient's lab results and medication history with LiverTox. If specific presentations or numbers help you arrive at your conclusion, you should cite them. – Knowledge gaps (e.g., symptoms of fever which will aid in confirmation) should also be highlighted and investigated. – Although you need to adhere to a fixed format, your SBAR should still be coherent and logical as a whole. – The consultant already has access to the drugs prescribed and liver panel. You do not need to restate everything.	Some claims were not supported with data and clinical findings. This section encourages the LLM to substantiate any claim.
## Likelihood categories are determined by LLM Drug Analyzers. Please refer to the definitions below: – Definite (> 95% chance): Liver injury is typical for the drug or herbal product ("signature" or pattern of injury, timing of onset, recovery). The evidence for causality is "beyond a reasonable doubt". – Highly Likely (75%–95%): The evidence for causality is 'clear and convincing' but not definite. – Probable (50%–74%): The causality is supported by 'the preponderance of evidence' as implicating the drug but the evidence cannot be considered definite or highly likely. – Possible (25%–49%): The causality is not supported by 'the preponderance of evidence'; however, one cannot definitively exclude the possibility. – Unlikely (< 25%): The evidence for causality is 'highly unlikely' based upon the available information. – Insufficient data: Key elements of the drug exposure history, initial presentation, alternative diagnoses and/or diagnostic evaluation prevent one from determining a causality score.	Definitions on subjective likelihood terms were provided. This ensures a common interpretation of the likelihoods stated in summaries generated by the Drug Analysis Agents.

Please return nothing but a JSON in the following format:
{
"situation": "situation of SBAR,"
"background": "background of SBAR,"
"assessment": "assessment of SBAR,"
"recommendation": 'recommendation of SBAR
}
	Sections of the SBAR were defined to ensure that all information was provided.

Patient Case Onset of DILI: 2122-01-09 Pattern of Liver Injury: cholestatic # Liver Panel Result: 2121-12-25 12:47:00: ALT: Not available, AST: 9.0 IU/L, ALP: 259.0 IU/L, TB: 8.6 umol/L, R_RATIO: Not available (...)	The format of the patient case was refined through rounds of testing.
The following are analyses from other clinicians with reference to LiverTox. They should be interpreted as such: (Drug Name): (Likelihood Category) (Evaluation by other agents) **Likelihood Assessments:** 1. Acetaminophen: Unlikely (Acetaminophen is a well-established cause of liver injury, particularly hepatocellular injury, (...) 2. Calcium Carbonate: Unlikely (Calcium Carbonate does not have a LiverTox entry, indicating a lack of documented association with liver injury. (...)	This section comprises the summaries generated from the Drug Analysis Agents.

Appendix B

Pydantic Schema

Grading in the initial phases was limited to fixed DILI parameters like the phenotype and onset. However, allowing the LLM to answer in free-text provides room for vague answers. To reduce this variability, Pydantic schema were used to constrain the outputs.

Pydantic schema	Remarks
class Drug(BaseModel): "Likelihood categories for a single drug" drugname: str = Field(..., description = "Medication Name of the drug. This is NOT the Chapter Name. This should NOT have the likelihood score. Always ensure the full name of the medication (e.g., % and brackets) is included.") category: LIKELIHOOD_CATEGORES = Field(..., description = "The assigned likelihood category you would like to assign.") elaboration: str = Field(..., description = "Short elaboration on your choice of likelihood category.")	In the Drug schema, 3 fields are defined which the LLM must provide in its answer: the name of the drug, the likelihood category and a short elaboration. Likelihood categories (e.g., Definite, Probable, Unlikely) are defined separately in the main prompt. The short elaboration by the Drug Analysis Agent functions as a condensed analysis provided to the next LLM, the Clinician Agent. These intermediate evaluations and explanations are meant only for inter-LLM communication and not evaluated by the expert panel.

Appendix C

Results From the Initial Testing Phases

In the main manuscript, the evaluation of LLM in the extraction of information from patient cases are described in Section 2.3.2, and results reported in Section 3.1. The results are presented in tabular form in Table C1.

A breakdown of errors made during extraction are reported in Table C2.

The approach used to evaluate extraction from LiverTox descriptions is described in Section 2.3.3 of the main manuscript, and the associated results presented in Section 3.2 are tabulated in Tables C3 and C4.

Multiple error types which occur within the same case and attempt are counted separately.

TABLE C1 | Statistical analysis for accuracy of full-length and sequential approaches by characteristic.

DILI evaluation characteristic	Mean of full-length approach, % (SD)	Mean of sequential approach, % (SD)	Mean difference, % (95% CI)	*p*
Overall	40.87 (39.09)	51.96 (40.44)	11.09 [2.31, 19.86]	0.007
Time to onset	76.97 (29.67)	86.36 (16.01)	9.38 [2.74, 16.03]	0.003
Time to recovery	55.36 (40.75)	67.02 (42.29)	11.66 [2.84, 20.48]	0.005
Pattern of liver injury	89.33 (30.45)	88.90 (28.52)	−0.44 [−5.29, 4.41]	0.428

TABLE C2 | Errors made in the analysis of patient cases.

	No. (%) of errors for time to onset		No. (%) of errors for time to recovery	
	Full-length (*n* = 71)	Sequential (*n* = 115)	Full-length (*n* = 89)	Sequential (*n* = 97)
Drug omission	22 (30.99)	0 (0)	24 (26.97)	0 (0)
Timing discrepancy	43 (60.56)	115 (100)	24 (26.97)	24 (24.74)
Misjudgement of information availability	6 (8.45)	0 (0)	41 (46.07)	73 (75.26)

TABLE C3 | Statistical analysis for accuracy of analysis of LiverTox descriptions.

	Mean of GPT-4 turbo, % (SD)	Mean of GPT-4o, % (SD)	Mean difference, % (95% CI)	*p*
Overall	81.50 (35.53)	96.93 (14.96)	15.47 [4.52, 26.42]	0.003
Missing phenotype	87.87 (28.50)	99.47 (0.04)	11.6 [3.92, 19.27]	0.003
Missing pattern of liver injury	95.67 (17.72)	99.47 (3.77)	3.80 [−1.39, 8.99]	0.147
Replied with incorrect phenotype	98.00 (12.74)	98 (10.96)	0.00 [−4.85, 4.85]	1.000
Replied with incorrect pattern of liver injury	96.20 (15.79)	100.00 (0.00)	3.80 [−0.69, 8.29]	0.095

TABLE C4 | Errors made in the analysis of LiverTox descriptions.

	No. (%) of errors	
Category	GPT-4 Turbo (*n* = 334)	GPT-4o (*n* = 46)
Overall	334	46
Missing phenotype not identified	182 (54.5)	8 (17.4)
Missing pattern of liver injury not identified	65 (19.5)	8 (17.4)
Replied with incorrect phenotype identified	30 (8.9)	30 (65.2)
Replied with incorrect pattern identified	57 (17.0)	0 (0.0)