Knowledge-Based Systems 328 (2025) 114184

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

HELIOT: LLM-Based CDSS for adverse drug reaction management

Gabriele De Vito [a,*], Filomena Ferrucci [a], Athanasios Angelakis [b,c,d]

[a] Department of Computer Science, University of Salerno, Via Giovanni Paolo II 132, 84084 Fisciano, Salerno, Italy
[b] Department of Epidemiology and Data Science, Amsterdam University Medical Center, Amsterdam, The Netherlands
[c] Digital Health; Methodology, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands
[d] Data Science Center, University of Amsterdam, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Clinical decision support systems
Large language models
Drug administration
Adverse drug reaction management

ABSTRACT

Medication errors significantly threaten patient safety, leading to adverse drug events and substantial economic burdens on healthcare systems. Clinical Decision Support Systems (CDSSs) aimed at mitigating these errors often face limitations when processing unstructured clinical data, including reliance on static databases and rule-based algorithms, frequently generating excessive alerts that lead to alert fatigue among healthcare providers. This paper introduces HELIOT, an innovative CDSS for adverse drug reaction management that processes free-text clinical information using Large Language Models (LLMs) integrated with a comprehensive pharmaceutical data repository. HELIOT leverages advanced natural language processing capabilities to interpret medical narratives, extract relevant drug reaction information from unstructured clinical notes, and learn from past patient-specific medication tolerances to reduce false alerts, enabling more nuanced and contextual adverse drug event warnings across primary care, specialist consultations, and hospital settings. Evaluation using three state-of-the-art LLMs on synthetic and real-world datasets demonstrates classification accuracy ranging from 98.77 % to 99.80 % with zero false negatives for life-threatening reactions. This high accuracy enabled HELIOT to achieve a 50–53 % reduction in interruptive alerts compared to traditional CDSSs while maintaining perfect safety profiles. To support clinical deployment, the system incorporates a confidence-based risk stratification framework that enables automated decisions for high-certainty cases while ensuring appropriate clinical oversight for uncertain classifications. Clinical usability evaluation with healthcare professionals validated these achievements, revealing strong acceptance and unanimous preference for HELIOT's contextual approach over traditional systems. These findings show promise; however, broader clinical trials remain essential to confirm effectiveness across diverse healthcare environments.

1. Introduction

Medication errors pose significant risks to patient safety and lead to adverse drug events (ADEs) [1]. In England, an estimated 237 million such errors occur annually, with around 66 million being potentially clinically significant [1,2]. These incidents cost the National Health System (NHS) approximately £98.5 million annually, consume 181,626 bed days, and contribute to 1708 deaths [1]. Similarly, in the United States, the economic impact is substantial, with prescribing and administration mistakes costing an estimated $20 billion annually [2].

CDSSs have emerged as tools to mitigate medication errors and enhancing patient safety [3]. These systems provide healthcare professionals with evidence-based recommendations and alerts, helping to prevent potential ADEs. However, traditional CDSSs face several limitations [4–6]. They typically rely on static databases and rule-based algorithms, which may not capture the nuances of individual patient cases or the latest medical knowledge [5–7]. For instance, when a drug with potential cross-reactions is prescribed, these systems generate alerts without considering complex clinical scenarios documented in notes, such as previous patient-specific tolerance, distinctions between true allergies and minor side effects, or situation-specific risk-benefit analyses [8]. Evidence shows that incorporating this contextual information could significantly reduce unnecessary alerts [8]. Without such capabilities, the rigid alerting mechanisms lead to excessive warnings, contributing to alert fatigue among healthcare providers and potentially causing critical warnings to be overlooked [6–13].

The advent of Large Language Models (LLMs), such as GPT-4 [14], offers a promising avenue to address these limitations. LLMs possess advanced natural language processing (NLP) capabilities, and have already succeeded in various healthcare applications, such as predicting drug interactions and patient outcomes, assisting in diagnostic processes, and generating clinical notes [15–17]. We conjecture that their ability to

[*] Corresponding author.
E-mail addresses: gadevito@unisa.it (G. De Vito), fferrucci@unisa.it (F. Ferrucci), a.angelakis@amsterdamumc.nl (A. Angelakis).

https://doi.org/10.1016/j.knosys.2025.114184
Received 13 April 2025; Received in revised form 5 July 2025; Accepted 24 July 2025
Available online 31 July 2025

process and synthesize large volumes of unstructured data makes them an innovative substitute for current rule-based CDSSs, which often lack the flexibility and depth of understanding required for intricate patient-specific situations.

This paper presents HELIOT, a novel CDSS for adverse drug reaction management that processes unstructured clinical narratives, building upon our previous exploration of LLM integration in healthcare technologies [18]. The system overcomes the shortcomings of traditional CDSSs by leveraging LLMs to interpret medical narratives from clinical notes, text-based electronic health records, and transcribed patient conversations. By integrating this capability with a comprehensive pharmaceutical data repository, HELIOT improves the accuracy and reliability of adverse reaction alerts in healthcare. Our evaluation across three state-of-the-art LLMs (GPT-4o, Gemma 3, and Claude Sonnet) demonstrates consistent high performance, with classification accuracy ranging from 98.77 % to 99.80 % on both synthetic and real-world datasets. The system maintains perfect safety profiles with zero false negatives for life-threatening reactions while achieving a potential 50–53 % reduction in unnecessary interruptive alerts compared to traditional CDSSs. HELIOT also incorporates confidence analysis to manage uncertain classifications and evaluates clinical usability, with healthcare professionals confirming strong acceptance of the system's contextual approach.

The primary contributions of this paper are as follows.

- We present a novel approach to process unstructured clinical narratives for adverse drug reaction management, demonstrating how LLMs can extract and interpret patient-specific medication reaction information, including medication tolerances, adverse events, and cross-sensitivities.
- We describe the HELIOT CDSS's modular architecture, highlighting its core components and the approach employed to provide decision support services.
- We present an empirical evaluation of the HELIOT CDSS across multiple LLM architectures, including synthetic and real-world dataset validation, as well as a clinical usability assessment.
- We introduce a risk threshold framework based on confidence analysis for safe clinical deployment, providing structured guidelines for automated decision-making and clinical review requirements.
- We provide datasets and code used to develop the HELIOT prototype, including both the synthetic patient dataset and the real-world clinical dataset employed for empirical evaluation as contributions to the research community.

Structure of the paper. The remainder of this paper is organized as follows: Section 2 describes the background on CDSSs and LLMs and the related work. Section 3 describes our method, including the data pipeline, the HELIOT approach, and the empirical study design used to evaluate the proposed CDSS. Section 5 shows and discusses the empirical results. Section 6 provides the practical implications, future research lines, and limitations of our study. Section 7 concludes the paper.

2. State of art and motivation

This section reviews current CDSS challenges and limitations, discusses recent advances in LLMs for processing medical text, and presents the motivation behind our work.

2.1. CDSSs challenges and limitations

CDSSs are pivotal in improving patient safety and clinical efficiency through integration with Electronic Medical Records [3]. In the pharmacological domain, these systems must navigate complex territory, managing intricate drug interactions and patient-specific factors that require sophisticated algorithms to detect potential adverse reactions [19].

Research demonstrates the value of CDSSs in reducing medication errors, with knowledge-based systems showing a 4.4 % improvement in prescribing behavior [4] and AI-based approaches achieving enhanced accuracy in prescription verification [20]. However, a critical challenge emerges with drug reaction alert systems, where override rates remain problematically high-ranging from 43.7 % to 97 % [21], with particularly concerning rates for opioids [13]. A comprehensive national evaluation of 1599 hospitals revealed that while overall CDSS performance improved over time, the fundamental alert fatigue problem persisted, with hospitals achieving higher scores by overalerting-including inappropriate nuisance alerts that contribute to clinician burnout [22].

The root cause of this widespread alert dismissal [6–13] lies in current systems' inability to effectively interpret unstructured clinical narratives containing crucial patient-specific information about medication tolerances and reactions. Evidence indicates that incorporating previous drug tolerance data could significantly reduce unnecessary alerts [8]. While promising solutions have emerged, including ontology-based systems [23] and machine learning approaches [24], these methods fall short when representing these complex clinical scenarios, highlighting how advanced natural language processing techniques could potentially transform alert relevance and clinical acceptance.

2.2. LLMs opportunities

LLMs have revolutionized the field of NLP by leveraging the Transformer architecture with self-attention mechanisms, as introduced by Vaswani et al. [25]. Notable examples of these models include commercial offerings such as GPT-3 [26] and GPT-4 [14], as well as open-source models like BERT [27], FLAN-T5 [28], LLama [29], BLOOM [30], and GLM [31]. LLMs are trained on extensive text datasets and often contain hundreds of billions of parameters [32–34]. The initial "pre-training" phase is computationally intensive but essential for enabling these models to perform a wide range of NLP tasks, such as translation and summarization, with high proficiency [32–34].

Following pre-training, LLMs can be specialized through a fine-tuning process, which involves using smaller datasets to tailor the models for specific NLP tasks, such as question-answering or tasks within different domains. Several emergent abilities have been discovered in the context of LLMs. The key abilities include "In-context learning," "Instruction following," and "Step-by-step reasoning." "In-context learning," introduced by GPT-3, allows models to perform tasks based on examples without additional training. "Instruction following" enables the models to execute tasks based solely on given instructions, while "Step-by-step reasoning" facilitates solving complex problems through chain-of-thought prompting [35]. These sophisticated capabilities make LLMs particularly promising for healthcare applications, with recent research demonstrating their effectiveness in generating differential diagnoses [16] and enhancing clinical documentation [15]. Their unique ability to process context-rich information and perform complex reasoning aligns with the challenge of interpreting medication narratives in clinical notes-a critical area where current CDSSs fail. Recent work has explored specialized LLMs for medication guidance, with systems like ShennongMGS demonstrating the potential of fine-tuned models for adverse drug reaction prediction and medication decision support [36]. However, these approaches typically focus on general medication guidance rather than specifically addressing the challenge of interpreting unstructured clinical narratives for patient-specific adverse reaction assessment, which remains a critical gap in current CDSS implementations. Despite this natural fit between LLMs' capabilities and the challenges of medication narrative interpretation, applying LLMs to extract and interpret medication information from unstructured clinical text remains underexplored.

2.3. Motivation of our work

The potential of LLMs for healthcare applications is clear, but translating this potential into practical systems for medication safety requires addressing several domain-specific challenges. Even in healthcare settings with sophisticated Electronic Health Record systems (EHR), clinical notes documenting patient reactions to drugs, referral letters detailing treatment histories, and records of medication tolerances frequently remain as free-text narratives. Traditional CDSSs can only process structured data and cannot interpret these narrative texts [8]. This means that valuable information about a patient's drug history may be overlooked when prescribing medications. For instance, a clinical note stating "Patient experienced mild rash with amoxicillin but has since tolerated cephalexin" contains important information about drug allergies and tolerances, but traditional CDSSs cannot extract and use this information because it is in free-text format. Moreover, the quantitative impact of alert fatigue on clinical decision-making has been well-documented through large-scale evaluations. Co et al.'s analysis of national hospital data demonstrated that hospitals achieving high CDSS performance scores often did so by implementing excessive alerting mechanisms, with those alerting on low-risk prescriptions scoring 3 % higher overall but potentially compromising patient safety through alert fatigue [22]. This creates a critical paradox: systems designed to improve safety may inadvertently reduce it by overwhelming clinicians with inappropriate alerts.

In addition, the research community lacks comprehensive evaluation datasets for assessing CDSSs in this field [11]. LLMs may help address these challenges by providing the contextual understanding needed to interpret narrative texts and generate more targeted, contextually appropriate alerts, thereby reducing alert fatigue while improving medication safety. Our work aims to fill these gaps by providing open-source tools and evaluation resources to support future research in enhancing clinical decision support through natural language understanding.

3. HELIOT framework

This section presents the HELIOT CDSS, illustrating design principles and implementation details.

3.1. CDSS design and approach

The proposed framework comprises three integrated components: the decision support process, system architecture, and data adaptability mechanisms, described in detail below.

3.1.1. Decision support process

HELIOT employs a Retrieval Augmentation Generation (RAG) approach to support physicians in medication decisions based on patients' adverse reaction histories. RAG is a technique that enhances LLMs by retrieving relevant information from external knowledge sources before generating responses, making the output more informed, contextualized, and accurate. The decision process follows several integrated steps to produce clinical assessments.

The foundation of our approach begins with the parallel retrieval of drug information from specialized databases containing pharmaceutical data (active ingredients, excipients, contraindications, and side effects) while simultaneously analyzing patient clinical notes to identify potentially problematic ingredients.

A critical feature of our system is its ability to maintain continuity of care by integrating current and historical patient data. The patient database stores and updates clinical notes from previous encounters, creating a longitudinal record of adverse reactions and tolerances. When a healthcare professional consults the system about a new medication, these historical records are automatically retrieved and combined with current clinical notes. This integration provides a complete picture of the patient's reaction history, even in facilities without integrated EHR systems, ensuring that past adverse events are not overlooked in current decision-making. Once all relevant information, including drug composition, current clinical notes, and the patient's historical data, is gathered, the system executes the core decision support logic. This logic employs carefully crafted prompts (Figs. 1 and 2) that guide the LLM to analyze the clinical situation. Following the persona pattern [37], our system instructs the LLM to embody an expert physician who evaluates potential adverse reactions by examining relationships between the drug's composition and the patient's documented reaction history, including current and historical notes.

Decision Support Prompt: System Prompt

Act as an expert physician.
Your task is to check if the drug I want to prescribe is safe for the patient, focusing only on the potential drug reactions the patient has.
Drug To Prescribe: {*drug*}
Drug Active Ingredients: {*active_ingredients*}
Drug Excipients: {*excipients*}
Known Cross-reactivity: {*cross_reactivity*}
Known Excipients With Chemical Cross-reactivity: {*excipients_cross_reacts*}
Contraindications: {*contraindications*}

INSTRUCTIONS
...
CONFLICT HANDLING
...
OUTPUT FORMAT
{"a":"brief description of your analysis", "r":"final response: NO DOCUMENTED REACTIONS OR INTOLERANCES—DIRECT ACTIVE INGREDIENT REACTIVITY—DIRECT EXCIPIENT REACTIVITY—NO REACTIVITY TO PRESCRIBED DRUG'S INGREDIENTS OR EXCIPIENTS—CHEMICAL-BASED CROSS-REACTIVITY TO EXCIPIENTS—DRUG CLASS CROSS-REACTIVITY WITHOUT DOCUMENTED TOLERANCE—DRUG CLASS CROSS-REACTIVITY WITH DOCUMENTED TOLERANCE", "rt":"reaction type: None—Life-threatening—Non life-threatening immune-mediated—Non life-threatening non immune-mediated"}

Fig. 1. Decision support prompt: system prompt.

Decision Support Prompt: User Prompt

PATIENT INFORMATION: {*clinical_notes*}

Fig. 2. Decision support prompt: user prompt.

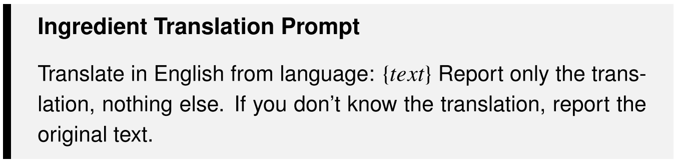
The system's final output provides an assessment structured in three parts: a clinical classification of the case (e.g., "Direct Active Ingredient Reactivity"), a categorization of the reaction severity (e.g., "Life-threatening"), and a detailed analysis explaining the rationale behind these classifications. The alert type is automatically determined based on the combination of case classification and reaction severity, as shown in Table 1.

When conflicting information emerges during the decision support process, HELIOT employs a hierarchical resolution strategy that prioritizes patient safety while incorporating clinical context. The system applies the following conflict resolution principles:

1. Patient-Specific Evidence Priority - documented patient-specific tolerance or adverse reactions in clinical notes override general pharmaceutical contraindications
2. Temporal Precedence - more recent clinical observations take precedence over older general warnings while maintaining awareness of historical patterns

Table 1
Case classification, reaction type, and alert type.

Case Classification	Reaction Type	Alert Type
No documented reactions or intolerances	None	None
No reactivity to prescribed drug's ingredients or excipients	None	None
Direct active ingredient reactivity	Life-threatening	Interruptive
	Non life-threatening immune-mediated	Interruptive
	Non life-threatening non immune-mediated	Non-interruptive
Direct excipient reactivity	Life-threatening	Interruptive
	Non life-threatening immune-mediated	Interruptive
	Non life-threatening non immune-mediated	Non-interruptive
Chemical-based cross-reactivity to excipients	Life-threatening	Interruptive
	Non life-threatening immune-mediated	Interruptive
	Non life-threatening non immune-mediated	Non-interruptive
Drug class cross-reactivity without documented tolerance	Life-threatening	Interruptive
	Non life-threatening immune-mediated	Interruptive
	Non life-threatening non immune-mediated	Non-interruptive
Drug class cross-reactivity with documented tolerance	None	None



Ingredient Translation Prompt

Translate in English from language: $\{text\}$ Report only the translation, nothing else. If you don't know the translation, report the original text.

Fig. 3. Ingredient translation prompt.

3. Severity-Based Escalation - life-threatening reactions documented in patient history always trigger alerts regardless of conflicting tolerance data
4. Clinical Context Integration - the LLM synthesizes official pharmaceutical data with direct clinical observations, reasoning through apparent contradictions (e.g., "patient has documented penicillin allergy but tolerated amoxicillin")
5. Uncertainty Acknowledgment - when conflicts cannot be definitively resolved through clinical reasoning, the LLM is instructed to explicitly state uncertainty and recommend clinical review rather than making unilateral decisions.

This strategy leverages the LLM's natural language reasoning capabilities while ensuring that clinical judgment remains paramount.

It is worth noting that HELIOT faces a significant linguistic challenge when processing clinical notes from different regions and healthcare systems. Current LLMs, as demonstrated by Wendler et al [38], exhibit an intrinsic bias toward English in their conceptual processing, potentially compromising reasoning accuracy when working with non-English medical terminology. HELIOT addresses this limitation by standardizing all medical and pharmaceutical terminology into English using a dedicated translation prompt (Fig. 3). This approach not only overcomes the language bias of underlying models but also ensures optimal correspondence with international medical ontologies that predominantly use English nomenclature.

3.1.2. System architecture

Building upon the RAG approach described above, HELIOT's architecture implements a modular, model-agnostic framework designed for deployment flexibility across different healthcare environments. The system separates data management, retrieval coordination, and decision support logic into distinct components that can be independently updated or scaled.

As shown in Fig. 4, the architecture comprises three main components: a web application for standalone operation, an API application providing RESTful services for EHR integration, and the HELIOT Controller implementing the core decision-making logic. This service-oriented design enables HELIOT to function both as an independent clinical tool and as an external microservice that existing healthcare systems can invoke without replacing their native infrastructure.

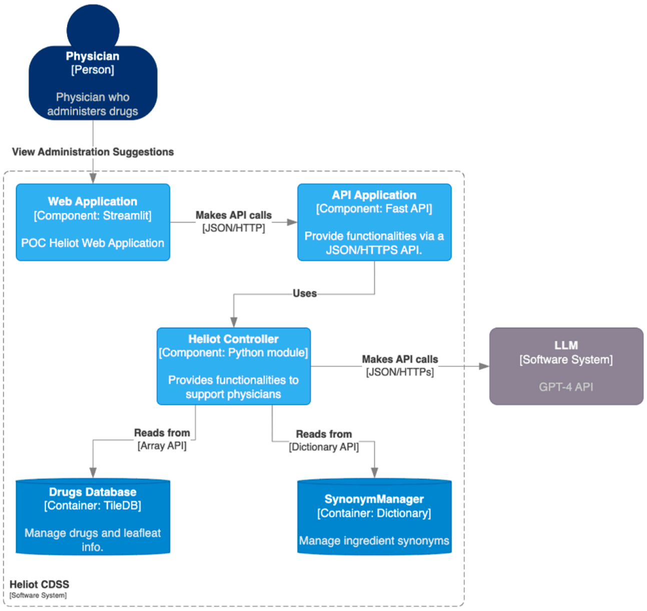


Fig. 4. HELIOT CDSS architecture.

The HELIOT Controller serves as the central orchestration engine, coordinating several specialized sub-components. The TileDB Drug and Patient Databases store comprehensive pharmaceutical information (drug identifiers, active ingredients, excipients, contraindications, and side effects) and patient clinical notes respectively. We selected TileDB [39] for its efficient data retrieval, storage compression capabilities, and versatility in supporting both local and cloud deployments. The Synonym Manager maintains an in-memory structure that maps canonical ingredient names to their corresponding synonyms, ensuring standardized ingredient recognition. The LLM component handles advanced natural language processing tasks, interpreting medical texts and generating contextually relevant recommendations.

Finally, the API application provides RESTful services consumed by either the web interface or external EHR systems through standard FHIR and HTTP protocols. To enhance user experience, the API streams responses in real-time using server-sent events, providing immediate feedback as results become available and reducing perceived latency.

The data retrieval mechanism follows a multi-layered architecture (see Fig. 5) that begins when a physician or clinician submits a prescription request through the web application or EHR system.

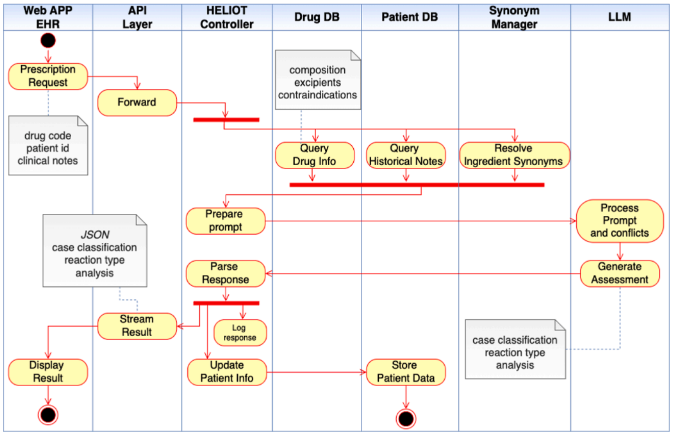


Fig. 5. HELIOT CDSS data flow.

This initial request, containing drug codes, patient IDs, and clinical notes, is forwarded through the API Layer to the HELIOT Controller. Upon receiving the request, the HELIOT Controller simultaneously initiates three parallel data retrieval operations: querying drug information from the Drug Database (local DB or through FHIR integration), retrieving historical patient notes from the Patient Database, and resolving ingredient synonyms through the Synonym Manager. Once these queries are complete, the controller prepares a prompt incorporating all retrieved data elements and leverages the LLM to process this information and generate a clinical assessment. The generated assessment undergoes parsing by the controller, which logs the response and updates patient information in the database. This architecture ensures optimal performance through parallel retrieval strategies, maintains data integrity through secure access protocols, and provides seamless integration capabilities that allow HELIOT to enhance existing healthcare workflows without disrupting established infrastructure.

3.1.3. LLM-pharmaceutical data integration framework

The coordination between pharmaceutical databases and clinical narratives operates through four distinct layers:

- **Data abstraction layer**: This layer provides uniform access to heterogeneous pharmaceutical data sources through a configurable interface that abstracts implementation-specific details. The system employs a generic loading mechanism that dynamically selects appropriate data management modules based on configuration parameters, enabling seamless integration with different pharmaceutical repositories (local databases, FHIR-compliant servers, national drug registries) without requiring architectural changes. This abstraction ensures that upper layers receive pharmaceutical data in a consistent format regardless of the underlying data source, allowing HELIOT to adapt to diverse healthcare IT environments while maintaining identical clinical decision support capabilities.
- **Retrieval coordination layer**: The system orchestrates parallel data retrieval operations to minimize response latency. When a clinician queries about a specific medication, this layer simultaneously: (1) retrieves comprehensive pharmaceutical data using drug identifiers, (2) accesses current patient clinical notes, (3) queries historical adverse reaction records, and (4) resolves ingredient synonyms through the synonym management system. This parallel processing ensures that all relevant context is available before LLM processing begins.
- **Context assembly layer**: Retrieved information is structured into coherent prompts that guide LLM reasoning. This layer combines pharmaceutical data (active ingredients, excipients, known contraindications) with patient-specific information (documented reactions, tolerances, clinical notes) into the specialized prompts shown in Figs. 1 and 2. The assembly process includes linguistic standardization (see Section 3.1.1) and maintains semantic relationships between different data elements.
- **LLM processing layer**: The assembled context is processed by the LLM backend to generate clinical recommendations. This layer is designed to be model-agnostic, supporting different LLM architectures through standardized input/output interfaces. The processing follows the decision support logic outlined in Section 3.1.1, producing structured outputs that include clinical classifications, severity assessments, and reasoning explanations.

This layered approach ensures that LLM integration remains flexible while maintaining consistent decision support quality across different technological environments.

3.1.4. Data flexibility and adaptability

The architectural design and LLM integration framework described above are built upon a foundation of data flexibility that enables HELIOT to adapt to diverse healthcare environments and evolving medical knowledge. This adaptability manifests across multiple dimensions of the system's operation. As detailed in Section 3.1.2, while drug identifiers need standardization, most critical information-including active ingredients, excipients, contraindications, and side effects-can be stored as unstructured free text, eliminating the need for complex data normalization procedures. Similarly, the patient clinical database-while maintained locally within HELIOT-demonstrates adaptability in processing clinical information. The system does not impose rigid requirements on structuring clinical notes, relying instead on the LLM's natural language processing capabilities to extract relevant adverse reaction information from various documentation styles and formats. Regarding linguistic adaptation, the standardization process employs generalized language conversion mechanisms to normalize terminology into a standard language. This approach allows healthcare facilities in regions with different primary languages to utilize the system while maintaining consistent clinical reasoning.

Beyond static data handling, the architectural flexibility extends to dynamic system maintenance and evolution. HELIOT incorporates a multi-layered updating framework to ensure the system remains current with evolving medical knowledge and regulatory requirements. The updating strategy varies depending on the deployment scenario. For standalone deployments, the local drug database receives routine monthly updates for standard pharmaceutical information, while critical safety data is incorporated within 24 h of regulatory announcements. For EHR-integrated deployments, HELIOT leverages FHIR R4 interfaces to access real-time pharmaceutical data directly from institutional repositories, ensuring automatic synchronization as institutional systems receive updates. This data-level flexibility is complemented by adaptable decision logic through the use of a prompt-based framework. As medical understanding advances, clinical protocols can be updated by modifying the structured prompts while maintaining backward compatibility through REST API versioning mechanisms. Finally, the model-agnostic architecture and microservices design support seamless component evolution. Language models can be upgraded through standardized interfaces, while individual services can be updated independently using rolling deployment strategies to ensure zero downtime. This updating framework ensures HELIOT maintains clinical accuracy and operational stability across diverse healthcare environments as medical knowledge and technology evolve.

3.2. HELIOT prototype

This section describes our implementation of the HELIOT CDSS design as a functional prototype.

3.2.1. Pharmaceutical data pipeline

The first step in developing our prototype was creating a comprehensive pharmaceutical knowledge base. To this aim, we developed a

Table 2
Drug dataset structure.

Column	Description
Drug_code	It is the ministerial code of the drug, also known as the AIC code (Marketing Authorization issued by AIFA)
Drug_name	It is the drug name
Drug_form_full_descr	It is the pharmaceutical form of the drug
Atc_code	It is the drug's ATC (Anatomical Therapeutic Chemical) code, according to the Word Health Organization classification system.
Leaflet	It is the leaflet file name.

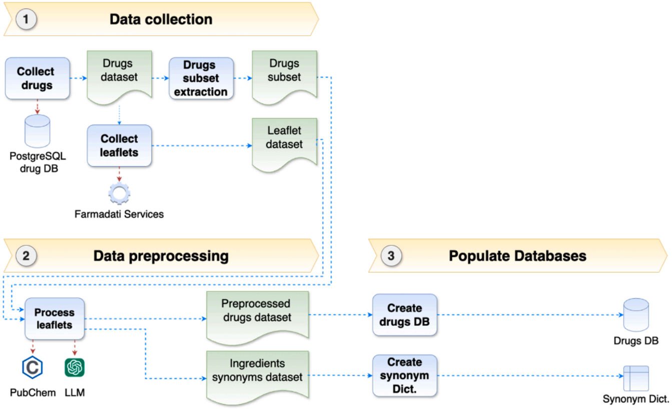


Fig. 6. Overview of the pharmaceutical data pipeline process.

Table 3
Drug distribution in dataset.

Drug Class	Perc.	No. of drugs
Opioids	65 %	653
Antibiotics	15 %	152
NSAID	5 %	47
Diuretics	2 %	24
Antiplatelet agents	2 %	16
Other	11 %	108
TOTAL		**1000**

specialized data pipeline to transform raw pharmaceutical data into HELIOT's pharmaceutical knowledge base. The pipeline (see Fig. 6) consists of three main phases: 1) Data Collection, 2) Data Preprocessing, and 3) Database Population.

The subsequent subsections offer an explanation of each phase.

3.2.1.1. Data collection. In the data collection phase, we gathered the necessary data for our experiments, specifically the datasets for drugs and leaflets.

3.2.1.1.1. Drug dataset. The first step was creating the drug dataset. To this end, an Italian company provided us with the PostgreSQL dump of the database of Italian medicines approved by the Italian Medicines Agency (AIFA).[1] The drug dataset has 106,962 drugs. Table 2 reports the dataset columns.

3.2.1.1.2. Leaflets dataset. The next step was to collect the leaflets associated with the medications. To make the pipeline production-ready, we automated this step by referring to a pharmaceutical data provider, the Farmadati company [2] that is one of Italy's major pharmaceutical data providers. It offers paid services to download the up-to-date drug database, including leaflets. Consequently, we utilized the web services provided by Farmadati to acquire leaflets for each medication in the dataset. The final leaflets dataset contains 19,188 files. The leaflets' structure includes the medication's name, qualitative and quantitative composition, pharmaceutical form, clinical information, pharmacological properties, pharmaceutical information, radiation dosimetry data for radiopharmaceuticals, and instructions on extended preparation and quality control for radiopharmaceuticals.

3.2.1.1.3. Drugs subset extraction. The final step of the data collection phase was creating a representative subset of drugs for the subsequent "data preprocessing" step. The drug dataset was randomly sampled from the drugs dataset using ATC codes. We followed the distribution reported in literature studies [8,12,13] to ensure the dataset reflected real-world drug prescription and adverse reaction patterns. The sampling process maintained the proportions of different drug classes as observed in clinical settings, with narcotic analgesics representing the largest group (65 %), followed by antibiotics (15 %), NSAID (5 %), diuretics (2 %), antiplatelet agents (2 %), and other medications (11 %). We also included common excipients associated with both immediate and delayed hypersensitivity reactions, with particular attention to preservatives and common allergens (e.g., polyethylene glycol, polysorbates, benzalkonium chloride).

Table 3 reports the final dataset distribution.

3.2.1.2. Data preprocessing. The data preprocessing phase addressed two key challenges in the leaflet data for our drug subset.

First, single leaflets often contain information for multiple pharmaceutical forms of the same medication (e.g., the ORAMORPH leaflet includes details for both syrup and oral solution forms). Since healthcare professionals prescribe specific forms, we must extract form-specific information while excluding irrelevant details. Second, the original ingredients and excipients were in Italian, requiring translation to English (consistently with the linguistic standardization approach discussed in Section 3.1) to enable synonym matching through international services like PubChem.[3] [4] This standardization is paramount as ingredients may appear under different names in medical records (e.g., "acetylsalicylic acid" vs "aspirin").

To address these challenges, we leveraged GPT-4o with two specialized prompts for processing ingredients and leaflet sections. For the sake of readability, we did not report the full prompts that can be found in our online repository [40].

During this phase, we processed 1035 unique ingredients and created an ingredient dictionary by querying PubChem REST services for comprehensive synonym lists. The preprocessing resulted in two structured datasets: "leaflet_info.csv" containing form-specific drug details and "ingredients_synonyms.csv" with standardized ingredient names and variants. Tables 4 and 5 detail the structure of these datasets.

To ensure the accuracy of the preprocessed data, two healthcare professionals (a clinical pharmacist and a physician) independently validated a random sample of 20 % of the drugs subset. This sample size was determined following established methodology for validation studies [41] and provides an appropriate statistical power for this application. The healthcare professionals focused on verifying that GPT-4o correctly extracted information specific to each pharmaceutical form from the official leaflets. They also verified the accuracy of ingredient and

[1] AIFA. https://www.aifa.gov.it/en/trova-farmaco

[2] Farmadati. https://www.farmadati.it/default.aspx

[3] PubChem substances. https://pubchem.ncbi.nlm.nih.gov/rest/pug/substance/name/{encoded_ingredient_name}/synonyms/JSON

[4] PubChem compounds. https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/name/{encoded_ingredient_name}/synonyms/JSON

Table 4
Processed leaflet dataset structure.

Column	Description
Drug_code	It is the ministerial code of the drug.
Drug_name	It is the drug name.
Drug_form	It is the pharmaceutical form of the drug.
ATC	It is the ATC code of the drug.
Composition	It contains the list of the drug's active ingredients.
Excipients	It contains the list of the drug's inactive ingredients.
Contraindications	It contains the contraindications for the drug.
Drug_interactions	It contains the drug interactions.
Side effects	It reports the side effects of the drug.
Incompatibilities	It reports the incompatibilities for the drug in the case of concurrent pharmaceutical therapies.

Table 5
Ingredient synonyms dataset structure.

Column	Description
Ingredient	It is the ingredient name extracted from the processed leaflets;
English_name	It is the English translation of the ingredient.
Synonyms	List of synonyms separated by '#' returned by the PubChem REST service.
Type	It represents the type of the ingredient, namely active or inactive.

excipient translations from Italian to English, ensuring correct matching with PubChem synonyms. The validation established that our preprocessing approach, particularly the designed GPT-4o prompts, effectively isolated and extracted form-specific information from the official leaflets while maintaining data integrity. The high inter-rater agreement (Cohen's kappa [42] = 0.95) further supported the reliability of our approach, demonstrating "almost perfect" consensus between clinical experts.

3.2.1.3. Populate HELIOT databases. We then populated two databases using the preprocessed datasets: the drug database using "leaflet_info.csv" and the synonyms database using "ingredients_synonyms.csv" (see Section 3.2.1.2). For the drug database schema, we defined drug_code, atc_code, composition, and excipients as dimensions with ASCII data type, while other columns became array attributes. We applied "ZstdFilter" compression [43] at level 3 for text-heavy attributes to optimize storage and query performance. Given the manageable number of ingredients (1,035) for the synonyms database, we implemented an in-memory data structure to ensure fast retrieval of ingredient synonyms during CDSS processing.

3.2.2. Prototype implementation

Building upon the data pipeline, we developed a fully functional proof-of-concept (POC) prototype of the HELIOT CDSS. This implementation demonstrates how the conceptual architecture described in Section 3.1.2 translates into practice with specific technology choices and integration patterns. While the architecture is designed to be model-agnostic and compatible with various LLMs (e.g., LLaMA), for this implementation we utilized GPT-4o through the OpenAI API. The prototype consists of the web application developed using the Streamlit framework, the API application developed using the FAST API and Uvicorn frameworks, and the HELIOT Controller developed using TileDB and the OpenAI API. The web application provides functionalities for processing single prescriptions with streamed results, uploading entire datasets for batch processing, and downloading results.

All the scripts, results, and source code are provided in our online repository [40].

4. Evaluation methodology

This section describes the methodology for evaluating HELIOT's effectiveness as a clinical decision-support system, including our research objectives, the creation of a specialized evaluation datasets, and the experimental design.

4.1. Research goals

The primary goal of the empirical assessment was to analyze the effectiveness of the proposed CDSS in two key dimensions: accuracy of clinical decision support and reduction of alert fatigue. The purpose was to provide empirical evidence highlighting the benefits and limitations of the HELIOT CDSS, enabling healthcare professionals to be aware of the strengths and weaknesses they would encounter through its use.

Specifically, the empirical assessment aimed to address the following research question:

RQ: How effective is HELIOT CDSS in identifying potential drug reactions and reducing alert fatigue?

This research question encompasses the system's ability to correctly identify potential adverse drug reactions and provide contextually appropriate alerts that do not overwhelm clinicians with unnecessary warnings.

4.2. Synthetic patient dataset creation

To evaluate HELIOT, we developed a synthetic patient dataset that represents the variety and complexity of adverse drug reaction scenarios encountered in clinical practice. Recent studies [44,45] support this synthetic data approach, demonstrating its potential to replicate real-world analysis results while maintaining privacy and supporting robust evaluation. Our synthetic dataset creation followed a systematic, literature-informed approach to ensure clinical accuracy and real-world representativeness. We began by analyzing distribution patterns from established literature [8,12,13] that documented comprehensive data on drug allergy alert systems, including override rates and reaction patterns across different drug categories. These studies provided empirical evidence showing override rates ranging from 43.7 % to 97 % across different therapeutic classes, with narcotic analgesics exhibiting the highest override rates and most common reactions, including itching (23.3 %), nausea (13 %), and hives (10.7 %). Based on this literature analysis, we created structured JSON configuration files that defined clinical case distributions across seven distinct classification categories (see Table 6), designed to capture the full spectrum of adverse drug reaction scenarios encountered in clinical practice. These JSON configuration files served as input specifications for automated dataset generation scripts, defining the expected number of cases for each scenario type, reaction severity levels, alert requirements, and clinical narrative templates.

The JSON configuration files specified parameters including case distributions, drug selections from our pharmaceutical database, reaction severity mappings, and clinical narrative templates for each classification type. The classifications range from patients with no documented reactions to complex cross-reactivity patterns involving chemical similarities between excipients or therapeutic drug classes. Each classification was further subdivided based on reaction severity (life-threatening, non-life-threatening immune-mediated, or non-life-threatening non-immune-mediated) and documented tolerance patterns. For cross-reactivity scenarios, we incorporated established chemical relationships between excipients based on documented cross-

Table 6
Clinical cases classifications.

Classification	Description
No documented reactions or intolerances	Patients with no history of adverse drug reactions or intolerances in their clinical records
Direct active ingredient reactivity	Patients with documented adverse reactions to the active ingredient of the prescribed drug
Direct excipient reactivity	Patients with documented reactions to an excipient present in the prescribed drug's formulation
No reactivity to prescribed drug's ingredients or excipients	Patients with documented adverse drug reactions, but not related to any component of the prescribed drug
Chemical-based cross-reactivity to excipients	Patients with adverse reactions to drugs in the same therapeutic class as the prescribed medication, without documented tolerance
Drug class cross-reactivity without documented tolerance	Patients with adverse reactions to drugs in the same therapeutic class as the prescribed medication, without documented tolerance
Drug class cross-reactivity with documented tolerance	Patients with adverse reactions to drugs in the same therapeutic class but with documented tolerance to the prescribed medication

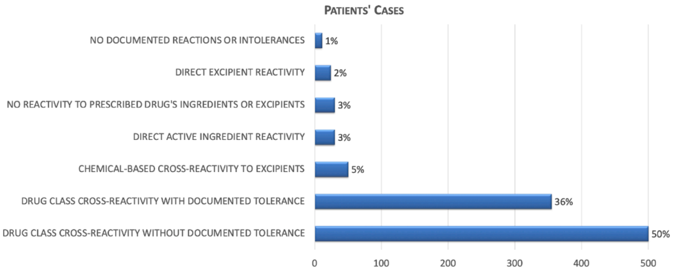


Fig. 7. Class distribution in the patients dataset.

Table 7
Synthetic dataset structure.

Field	Description
Patient ID	Unique identifier for each case
Drug Code	Ministerial code of the prescribed drug
Drug Name	Name of the prescribed drug
Clinical Note	Detailed patient history including adverse reactions
Classification	Case classification category
Alert Type	Interruptive/Non-interruptive/No alert
Reaction Type	Severity and nature of reaction
Prescribed ATC	ATC code of the prescribed drug

sensitivity patterns [46,47]. Cross-reactivity cases were designed around known chemical similarities, such as polyethylene glycol reactions with potential cross-reactivities to polysorbates, poloxamers, and cremophor, or drug class cross-reactivity scenarios based on established therapeutic relationships within antibiotic families or analgesic classes. The distribution of cases across different classifications reflects real-world clinical patterns, with drug class cross-reactivity scenarios representing the largest proportion of cases, as shown in Fig. 7. This distribution ensures testing of HELIOT's ability to handle various clinical scenarios while maintaining statistical validity for evaluation purposes. Clinical note generation was automated through specialized Python scripts that read the JSON configuration files and created patient-specific clinical narratives. These automated generation scripts processed the configuration data to create realistic clinical notes incorporating appropriate medical terminology, reaction descriptions, temporal relationships spanning from 2000 to 2025, and tolerance documentation where clinically appropriate. The generation process utilized predefined clinical templates that were refined through iterative consultation with healthcare professionals to ensure linguistic authenticity and clinical accuracy. The scripts randomly selected appropriate drugs from our pharmaceutical database based on the therapeutic classes specified in the JSON configurations (i.e., ATC codes), generated unique patient identifiers, and created detailed clinical narratives that included specific reaction histories, symptom descriptions, and documented medication tolerances where clinically appropriate. The final dataset structure is detailed in Table 7, which shows the key fields maintained for each synthetic case to ensure consistency with real-world electronic health records. The dataset validation process involved two healthcare professionals (a clinician and a physician) who complemented our earlier pharmaceutical validation team (see Section 3.2.1) with specialized knowledge in patient reactions and medical documentation. We implemented a human-in-the-loop validation approach where each synthetic case underwent independent expert review to verify clinical plausibility, appropriate reaction-drug relationships, and accurate alert classification. As demonstrated in similar healthcare Artificial Intelligence implementations where human expertise guides model decisions and validates outputs [48–50], this approach enhanced both the interpretability and reliability of the AI systems. The validation process examined multiple dimensions of each case, including the appropriateness of documented reactions for specific drugs, the clinical logic of cross-reactivity patterns, the accuracy of severity classifications, and the appropriateness of recommended alert types. For example, experts verified that documented reactions such as rash or swelling were clinically consistent with the prescribed medication's known adverse effect profile, ensuring that our synthetic cases reflected realistic clinical scenarios. Disagreements between reviewers were resolved through consensus meetings, with particular attention to complex cross-reactivity cases where alert type assignments required nuanced clinicaljudgment. The final validation achieved an inter-rater agreement of 0.91 (Cohen's kappa), indicating "almost perfect" consensus between clinical experts and confirming the reliability of our synthetic dataset creation approach. The resulting dataset contains 1000 synthetic cases with a comprehensive distribution across clinical scenarios and alert requirements. Table 8 provides the complete breakdown of case distribution by classification, reaction type, and alert type, demonstrating the dataset's coverage of diverse clinical scenarios designed to test HELIOT's ability to distinguish between different types of adverse drug reaction situations and generate contextually appropriate clinical recommendations.

The source code and configuration files for the dataset creation can be found in our online repository [40].

4.3. Real-world validation case study

Validation of HELIOT's performance beyond controlled synthetic scenarios involved obtaining real-world Electronic Health Record data from an Italian healthcare IT company that provides EHR systems to healthcare facilities. This dataset contained medication administrations for over 2000 patients with 10,000 randomly extracted administrations, all representing safe prescriptions that were actually administered to patients. Clinical notes were extracted from the allergy and intolerance sections of patient medical records, documented in Italian by healthcare professionals during routine clinical practice. After joining with our pharmaceutical subset of 1000 drugs, we obtained 162 patients with corresponding prescriptions, distributed as 114 patients without

Table 8
Case distribution by classification, reaction type, and alert type.

Case Classification	Reaction Type	Alert Type	Cases	Perc.
No documented reactions or intolerances	None	None	11	1.1 %
No reactivity to prescribed drug's ingredients or excipients	None	None	30	3.0 %
Direct active ingredient reactivity	Life-threatening	Interruptive	9	0.9 %
	Non life-threatening immune-mediated	Interruptive	12	1.2 %
	Non life-threatening non immune-mediated	Non-interruptive	9	0.9 %
Direct excipient reactivity	Life-threatening	Interruptive	6	0.6 %
	Non life-threatening immune-mediated	Interruptive	4	0.4 %
	Non life-threatening non immune-mediated	Non-interruptive	14	1.4 %
Chemical-based cross-reactivity to excipients	Life-threatening	Interruptive	15	1.5 %
	Non life-threatening immune-mediated	Interruptive	9	0.9 %
	Non life-threatening non immune-mediated	Non-interruptive	26	2.6 %
Drug class cross-reactivity without documented tolerance	Life-threatening	Interruptive	103	10.3 %
	Non life-threatening immune-mediated	Interruptive	271	27.1 %
	Non life-threatening non immune-mediated	Non-interruptive	126	12.6 %
Drug class cross-reactivity with documented tolerance	None	None	355	35.5 %
TOTAL			**1000**	**100 %**

Table 9
Classification results per category (averaged over five runs).

Case Category	Precision	Recall	F1	Cases
Chemical-based cross-reactivity to excipients	0.9804	1.0000	0.9901	50
Direct active ingredient reactivity	1.0000	1.0000	1.0000	30
Direct excipient reactivity	1.0000	0.9583	0.9787	24
Drug class cross-reactivity with documented tolerance	1.0000	1.0000	1.0000	355
Drug class cross-reactivity without documented tolerance	1.0000	1.0000	1.0000	500
No documented reactions or intolerances	0.9167	1.0000	0.9565	11
No reactivity to prescribed drug's ingredients or excipients	1.0000	0.9667	0.9831	30
Macro Average	**0.9853**	**0.9893**	**0.9869**	**1000**

Table 10
Classification results per reaction type (averaged over five runs).

Reaction Type	Precision	Recall	F1	Cases
None	1.0000	1.0000	1.0000	396
Life-threatening	1.0000	1.0000	1.0000	133
Non life-threatening immune-mediated	1.0000	1.0000	1.0000	296
Non life-threatening non immune-mediated	1.0000	1.0000	1.0000	175
Macro Average	**1.0000**	**1.0000**	**1.0000**	**1000**

Table 11
Classification results per alert type (averaged over five runs).

Alert Type	Ground Truth (%)	Heliot (%)	Traditional Systems (%)
No Alert Needed	396 (39.6 %)	396 (39.6 %)	41 (4.1 %)
Interruptive Alert	429 (42.9 %)	429 (42.9 %)	959 (95.9 %)
Non-Interruptive Alert	175 (17.5 %)	175 (17.5 %)	0 (0 %)

known allergies and 48 patients with clinical notes documenting drug allergies, adverse reactions, or side effects. Among the 48 patients with documented clinical notes, 2 cases specifically documented tolerance to the prescribed drug. While this distribution differs from literature patterns used in our synthetic dataset, this was expected given the random extraction across multiple departments and the limited scope of our pharmaceutical database.

4.4. Model selection and setup

To evaluate the HELIOT's robustness, we selected three representative LLMs with distinct characteristics and deployment approaches. The selection encompasses cloud-based and local deployment scenarios, different parameter scales, and diverse training approaches. GPT-4o was the primary evaluation model to maintain consistency throughout the research pipeline. It had previously been utilized to create synthetic datasets and develop pharmaceutical databases. This approach ensures methodological coherence across all experimental phases. The model was accessed through OpenAI's API. Google's Gemma 3 (12B parameters) [51] was selected to represent open-weight model architectures and validate performance scalability with smaller parameter counts. The model was deployed locally using LM Studio, a comprehensive platform for LLM experimentation that provides a unified OpenAI-compatible REST API for model deployment and inference across various architectures. Anthropic's Claude 4 Sonnet was included as a third validation point, accessed through Anthropic's API [52]. We set the temperature to 0.0 for all models to reduce non-deterministic outputs and maintain consistency across evaluations.

4.5. Experimental design

We employed five complementary evaluation approaches to investigate the research question

4.5.1. Evaluation using the synthetic dataset

This first evaluation approach compared responses from HELIOT CDSS against ground truth data provided by healthcare professionals in the data pipeline (see Section 4.2). This comparison was made using the HELIOT CDSS prototype web application (see Section 3.2) for interaction and validation. Fig. 8 illustrates the design of the experiment.

We first uploaded the synthetic patient dataset (described in Section 4.2) to the HELIOT DSS POC, where the data is analyzed for potential reactions. The system's results were then downloaded and compared to the ground truth, evaluating separately the "Classification" and "Reaction Type" assignments provided by the healthcare professionals. The alert type is automatically determined based on case classification and reaction type, as follows: when there is no documented tolerance, life-threatening reactions and non life-threatening immune-mediated reactions trigger interruptive alerts, while non life-threatening non immune-mediated reactions produce non-interruptive alerts. No alerts are needed

Table 12
Gemma 3: classification results per category.

Case Category	Precision	Recall	F1	Cases
Chemical-based cross-reactivity to excipients	1.0000	0.9800	0.9899	50
Direct active ingredient reactivity	1.0000	1.0000	1.0000	30
Direct excipient reactivity	0.9231	1.0000	0.9600	24
Drug class cross-reactivity with documented tolerance	1.0000	1.0000	1.0000	355
Drug class cross-reactivity without documented tolerance	1.0000	1.0000	1.0000	500
No documented reactions or intolerances	1.0000	1.0000	1.0000	11
No reactivity to prescribed drug's ingredients or excipients	1.0000	0.9667	0.9831	30
Macro Average	**0.9890**	**0.9924**	**0.9904**	**1000**

Table 13
Gemma 3: classification results per reaction type.

Reaction Type	Precision	Recall	F1	Cases
None	1.0000	0.9242	0.9606	396
Life-threatening	1.0000	1.0000	1.0000	133
Non life-threatening immune-mediated	0.9966	1.0000	0.9983	296
Non life-threatening non immune-mediated	0.8578	1.0000	0.9235	175
Macro Average	**0.9636**	**0.9811**	**0.9706**	**1000**

Table 14
Gemma 3: classification results per alert type.

Alert Type	Ground Truth (%)	Heliot (%)	Traditional Systems (%)
No Alert Needed	396 (39.6 %)	366 (36.6 %)	41 (4.1 %)
Interruptive Alert	429 (42.9 %)	430 (43.0 %)	959 (95.9 %)
Non-Interruptive Alert	175 (17.5 %)	204 (20.4 %)	0 (0 %)

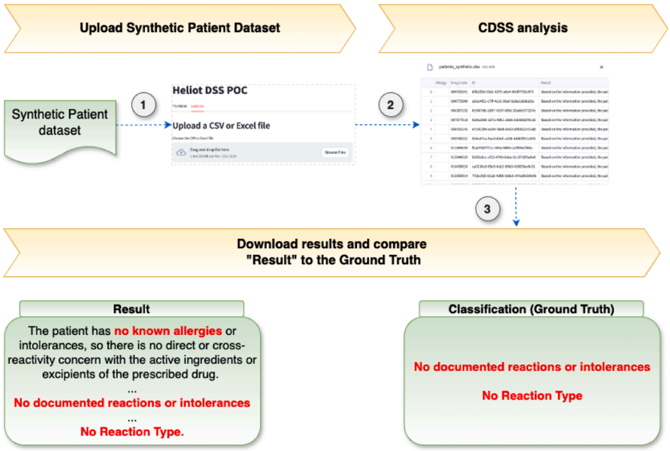

Upload Synthetic Patient Dataset
CDSS analysis
Synthetic Patient dataset
1
Heliot DSS POC
Upload a CSV or Excel file
2
3
Download results and compare "Result" to the Ground Truth
Result
The patient has no known allergies or intolerances, so there is no direct or cross-reactivity concern with the active ingredients or excipients of the prescribed drug.
...
No documented reactions or intolerances
...
No Reaction Type.
Classification (Ground Truth)
No documented reactions or intolerances
No Reaction Type

Fig. 8. Experiment design and setup.

when there are no documented reactions or when tolerance has been previously established (see Table 8 in Section 4.2).

To assess the consistency of the LLM's responses and evaluate any potential non-deterministic behavior, we repeated the experiment five times under identical conditions for all three language models (GPT-4o, Gemma 3, and Claude Sonnet), averaged the results, and calculated the Fleiss' Kappa, which is a statistical measure that evaluates the agreement between multiple raters (or iterations, in this case), across iterations [53]. This approach allowed us to verify whether the system produces consistent results across multiple runs, which is crucial for ensuring reliability in clinical applications.

To quantitatively measure the effectiveness of the HELIOT CDSS, we employed three key metrics, which are standard in evaluating classification systems and provide a comprehensive view of performance, specifically: Precision, Recall, and F1-score.

Precision is the ratio of true positives (correctly identified instances of a class) to all instances predicted as that class. Recall measures the proportion of true positive results among all actual positive cases, and indicates how well the system can identify positive instances. Finally, F1-score is the harmonic mean of precision and recall, providing a metric that balances both concerns. It is particularly useful when the class distribution is imbalanced.

We also measured the execution time of the analysis along the five runs and averaged the results to evaluate the proposed CDSS's performance comprehensively. We executed the experiment using a MacBook M3 Max with 96 GB of RAM and a unified Metal GPU.

4.5.2. *Evaluation using the real-world dataset*

For the real-world validation, we processed the 162 patient cases from the EHR dataset described in Section 4.3 through HELIOT. We maintained the same experimental design used for synthetic dataset evaluation, employing identical quantitative metrics (Precision, Recall, and F1-score) to ensure consistent performance assessment across synthetic and real-world scenarios. The number of runs for real-world validation was determined based on the consistency patterns observed in the synthetic evaluation, with deterministic models requiring single-run validation and non-deterministic models requiring multiple runs to ensure robust assessment. This real-world case study provides a crucial assessment of HELIOT's ability to process actual clinical documentation and handle the linguistic and terminological variations found in routine healthcare practice.

4.5.3. *Clinical usability evaluation*

To assess the clinical usability of HELIOT's decision support approach, we involved nine healthcare professionals. The participant demographics included healthcare professionals from Internal Medicine (78 %) and Primary Care/Family Medicine (22 %) with diverse experience levels: < 5 years (11 %), 5–10 years (33 %), 11–20 years (33 %), and > 20 years (22 %). Practice settings varied across hospitals (33 %), private practices (33 %), outpatient clinics (11 %), emergency departments (11 %), and academic medical centers (11 %). A structured questionnaire was developed to evaluate clinical accuracy, alert fatigue reduction, workflow integration, and overall system value across seven representative clinical cases spanning cross-reactivity with documented tolerance, cross-reactivity without documented tolerance, and non-immune side effects. Each case presented patient history, prescribed medication, HELIOT's output with explanation, and traditional CDSS output for comparison. The evaluation employed 5-point Likert scales to assess key dimensions, including clinical appropriateness and risk assessment accuracy, information clarity, clinical reasoning transparency, confidence in recommendations, the impact of alert fatigue, time efficiency, and the likelihood of adoption. The evaluation aimed to assess whether HELIOT's contextual approach provides clinically meaningful improvements over traditional CDSS systems in reducing alert fatigue while maintaining appropriate clinical safety measures. The complete questionnaire can be found in Appendix A.

Table 15
Claude Sonnet 4.0: classification results per category.

Case Category	Precision	Recall	F1	Cases
Chemical-based cross-reactivity to excipients	1.0000	1.0000	1.0000	50
Direct active ingredient reactivity	1.0000	1.0000	1.0000	30
Direct excipient reactivity	1.0000	1.0000	1.0000	24
Drug class cross-reactivity with documented tolerance	1.0000	0.9944	0.9972	355
Drug class cross-reactivity without documented tolerance	1.0000	1.0000	1.0000	500
No documented reactions or intolerances	1.0000	1.0000	1.0000	11
No reactivity to prescribed drug's ingredients or excipients	0.9375	1.0000	0.9677	30
Macro Average	**0.9911**	**0.9992**	**0.9950**	**1000**

4.5.4. *Safety analysis and risk assessment*

To ensure clinical safety in alert reduction scenarios, we conducted a safety analysis across all datasets and language models. This evaluation focused on quantifying clinical risk associated with "no alert needed" classifications and identifying potential missed critical events. We analyzed false negative rates for life-threatening reactions, calculated sensitivity and specificity for critical event detection, and assessed false positive rates to establish acceptable safety margins. The analysis examined consistency across different LLM architectures to validate the robustness of HELIOT's safety profile. Key metrics included negative predictive value for cases classified as requiring no intervention, false alarm rates for conservative classifications, and verification of zero missed critical events across both synthetic and real-world scenarios.

4.5.5. *Confidence analysis framework*

To establish risk thresholds for clinical deployment and assess system certainty, we conceived confidence scoring based on token log-probabilities returned by the LLM for decision support outputs. This approach enables quantitative assessment of system confidence and supports risk-stratified clinical deployment. Four confidence metrics were calculated from the LLM's JSON response:

- **Overall confidence**: Mean log-probability across the complete JSON response
- **Analysis confidence**: Log-probability of the clinical analysis text
- **Case classification confidence**: Log-probability of the case classification category
- **Reaction classification confidence**: Log-probability of the reaction type classification

These metrics provide multi-dimensional assessment of system certainty, enabling identification of cases where automated decisions may require clinical review. The confidence scoring framework supports the development of risk thresholds that balance automated efficiency with appropriate clinical oversight, particularly important for cases classified as requiring no alerts where missed critical reactions could pose safety risks. For both synthetic and real-world evaluations, confidence scores were calculated for all cases and analyzed to establish correlation patterns between confidence levels and classification accuracy. This analysis informs the development of clinical decision thresholds that can guide appropriate levels of human oversight in deployment scenarios.

5. Results

In this section, we present findings from our evaluation approach, examining system performance, consistency, usability, risk thresholds, and confidence mechanisms.

5.1. *Evaluation using the synthetic dataset*

Performance results using the synthetic dataset are reported below.

5.1.1. *GPT-4o performance analysis*

Our evaluation with GPT-4o, averaged over five runs to account for potential LLM non-deterministic behavior, demonstrates high system performance. The Fleiss kappa score across all iterations showed perfect agreement (100 %) [53,54] with a standard deviation of 0.000 for both category and reaction types classifications. This perfect agreement confirms that our system exhibited deterministic behavior, producing identical classification outputs across all five runs despite the potential for LLM-based variability.

Regarding the category classification, the system achieved excellent performance, with an overall macro-averaged precision of 0.9853, recall of 0.9893, and F1 score of 0.9869 (see Table 9). Notably, perfect classification (precision = 1.0000, recall = 1.0000, F1 = 1.0000) was achieved for three critical categories: "Direct Active Ingredient Reactivity," "Drug Class Cross-Reactivity with Documented Tolerance," and "Drug Class Cross-Reactivity Without Documented Tolerance." This perfect performance is particularly significant for the cross-reactivity categories, which comprised the majority of cases in the dataset (855 out of 1000 cases).

The "Chemical-Based Cross-Reactivity to Excipients" category showed excellent performance with a precision of 0.9804, perfect recall of 1.0000, and F1 score of 0.9901. Similar high performance was observed for "Direct Excipient Reactivity" (precision = 1.0000, recall = 0.9583, F1 = 0.9787), with only one case misclassified as "Chemical-Based Cross-Reactivity to Excipients." Importantly, this misclassification occurred between categories with similar non-life-threatening, non-immune-mediated reaction profiles.

For the "No Documented Reactions or Intolerances" category, the system achieved a precision of 0.9167 and perfect recall (1.0000), resulting in an F1 score of 0.9565. Similarly, "No Reactivity to Prescribed Drug's Ingredients or Excipients" showed perfect precision (1.0000) with slightly lower recall (0.9667) and a high F1 score of 0.9831, with one case misclassified as "No Documented Reactions or Intolerances" - both categories representing scenarios with no adverse reactions.

This pattern of errors suggests that while the system occasionally struggles to differentiate between certain similar non-serious reaction types, it maintains robust performance for scenarios with more serious clinical implications.

The evaluation of reaction type identification (see Table 10) showed perfect accuracy in distinguishing between different severity levels of adverse drug reactions, from life-threatening cases to non-immune-mediated reactions.

Regarding alert generation (see Table 11), our approach suggests improvements over traditional CDSSs, which tend to favor interruptive alerts [8,12,13]. By distinguishing between cases requiring interruptive alerts (39.6 %), those where non-interruptive alerts might suffice (17.5 %), and situations where no alert is needed (39.6 %), HELIOT shows promise in addressing alert fatigue by potentially reducing interruptive alerts by 50.3 % (395 - 41 No Alert Needed + 175 Non-Interruptive Alert) compared to traditional systems. However, these results need validation in real clinical settings, where alert override rates typically exceed 90 %.

Another key aspect derived from analyzing the results is the system's efficiency. The average execution time was 2.775 s per patient to see the overall response on the web application. Importantly, the response is streamed in real-time, similar to how ChatGPT functions, so healthcare professionals can start reading the suggestions immediately as they are generated without significant delays.

5.1.2. Gemma 3 performance analysis

Gemma 3 achieved an overall accuracy of 99.80 % on the synthetic dataset, averaged over five runs, with macro-averaged precision of 0.9890, recall of 0.9924, and F1 score of 0.9904 (see Table 12). The Fleiss kappa score across all iterations confirmed perfect agreement (100 %) and that our system exhibited deterministic behavior.

Regarding category classification, Gemma 3 reached perfect classification (precision = 1.0000, recall = 1.0000, F1 = 1.0000) for four critical categories: "Direct Active Ingredient Reactivity," "Drug Class Cross-Reactivity with Documented Tolerance," "Drug Class Cross-Reactivity without Documented Tolerance," and "No Documented Reactions or Intolerances." The "Chemical-Based Cross-Reactivity to Excipients" category showed strong performance with perfect precision (1.0000), recall of 0.9800, and F1 score of 0.9899, with one case misclassified. For "Direct Excipient Reactivity," the system achieved perfect recall (1.0000) with a precision of 0.9231 and an F1 score of 0.9600, showing 2 cases overclassified from other categories-indicating a conservative approach that prioritizes safety. The "No Reactivity to Prescribed Drug's Ingredients or Excipients" category demonstrated perfect precision (1.0000) with a recall of 0.9667 and an F1 score of 0.9831, with one case misclassified as a different non-reaction category.

The evaluation of reaction type identification (see Table 13) showed strong performance in distinguishing between severity levels, with perfect classification for life-threatening cases (precision = 1.0000, recall = 1.0000, F1 = 1.0000). Non-life-threatening immune-mediated reactions achieved near-perfect performance (precision = 0.9966, recall = 1.0000, F1 = 0.9983). For non-life-threatening, immune-mediated reactions, the system showed perfect recall (1.0000) with a precision of 0.8578 and an F1 score of 0.9235, indicating a tendency toward conservative classification. The "None" category achieved perfect precision (1.0000) with a recall of 0.9242 and an F1 score of 0.9606.

For alert generation (see Table 14), Gemma 3 demonstrated effective alert management with 36.6 % cases requiring no alerts, 20.4 % non-interruptive alerts, and 43.0 % interruptive alerts. This distribution suggests a 50.0 % reduction in unnecessary interruptive alerts compared to traditional CDSSs while maintaining perfect identification of critical cases requiring immediate attention. The slightly more conservative approach compared to GPT-4o (36.6 % vs 39.6 % no-alert cases) reflects Gemma 3's tendency to err on caution, which may be advantageous in clinical settings where patient safety is paramount. Gemma 3 evaluation was conducted using local deployment, achieving an average response time of 6.73 s per patient case. These timing results are provided for reference purposes to demonstrate the feasibility of local deployment scenarios where data privacy and infrastructure constraints require on-premises solutions.

5.1.3. Claude sonnet performance analysis

Claude Sonnet 4 achieved an overall accuracy (averaged over 5 runs) of 0.9980 with macro-averaged precision of 0.9911, recall of 0.9992, and F1 score of 0.9950, showing consistent performance compared to GPT-4o (see Table 15). Like GPT-4o and Gemma3, the Fleiss kappa score was 100 %, confirming that HELIOT provides deterministic responses.

In category classification, the LLM reached perfect performance (precision = 1.0000, recall = 1.0000, F1 = 1.0000) for five out of seven categories, including all direct reactivity categories and most cross-reactivity scenarios. The "Drug Class Cross-Reactivity with Documented Tolerance" category showed excellent performance with perfect precision (1.0000) and high recall (0.9944), resulting in an F1 score of 0.9972. Only 2 out of 355 cases were misclassified in this category. The "No Reactivity to Prescribed Drug's Ingredients or Excipients" category achieved perfect recall (1.0000) with a precision of 0.9375 and an F1 score of 0.9677. The system correctly identified all true cases but generated two false positives, representing a conservative approach that favors patient safety over precision.

Table 16
Claude Sonnet 4.0: classification results per reaction type.

Reaction Type	Precision	Recall	F1	Cases
None	1.0000	1.0000	1.0000	396
Life-threatening	1.0000	1.0000	1.0000	133
Non life-threatening immune-mediated	1.0000	1.0000	1.0000	296
Non life-threatening non immune-mediated	1.0000	1.0000	1.0000	175
Macro Average	**1.0000**	**1.0000**	**1.0000**	**1000**

Table 17
Claude Sonnet 4.0: classification results per alert type.

Alert Type	Ground Truth (%)	Heliot (%)	Traditional Systems (%)
No Alert Needed	396 (39.6 %)	396 (39.6 %)	41 (4.1 %)
Interruptive Alert	429 (42.9 %)	429 (42.9 %)	959 (95.9 %)
Non-Interruptive Alert	175 (17.5 %)	175 (17.5 %)	0 (0 %)

The evaluation of reaction type classification showed perfect accuracy across all severity levels, achieving 100 % precision, recall, and F1 scores for all categories (see Table 16), which is particularly significant for clinical safety.

Regarding alert generation, Claude Sonnet correctly identified all 396 cases requiring no alerts, 175 cases suitable for non-interruptive alerts, and 429 cases requiring interruptive alerts (see Table 17). This result represents a potential 53.0 % reduction in alert burden compared to traditional CDSSs, identical to the one observed with GPT-4o.The average execution time was 4.37 s per patient, representing a 57 % increase compared to GPT-4o (2.775 s) but still maintaining clinically acceptable response times for decision support applications.

5.2. Evaluation using the real-world dataset

Validation results employing the real-world dataset are presented in the following sections.

Nonetheless, the model obtained the same GPT-4o results in terms of alert distribution: 161 cases (99.4 %) requiring no alerts and 1 case (0.6 %) generating an interruptive alert (see Table 26).

5.2.1. GPT-4o performance analysis

Using GPT-4o as the underlying language model, HELIOT demonstrated robust performance on the real-world dataset, achieving an overall accuracy of 99.38 % with macro precision of 100.00 %, recall of 99.28 %, and F1-score of 99.63 % (see Table 18). The system correctly classified all 114 patients without documented reactions and achieved perfect performance for drug class cross-reactivity with documented tolerance cases. For patients with no reactivity to prescribed drug ingredients or excipients, GPT-4o achieved 97.83 % recall with one misclassification out of 46 cases. This single error resulted in a conservative over-classification where the system predicted drug class cross-reactivity without documented tolerance, representing a safety-first approach that avoids missed reactions.

For reaction type classification (see Table 19), HELIOT achieved similar high performance, with an overall accuracy of 99.38 %. The system correctly identified that the vast majority of cases (161 out of 162) required no reaction alerts, with only one case being conservatively overclassified as requiring a non life-threatening immune-mediated reaction alert. The alert type classification results (see Table 20) mirror the reaction type performance, with 99.38 % accuracy in determining appropriate alert levels. HELIOT correctly identified that 161 out of 162 cases

Table 18
GPT-4o classification results per category on real-world dataset.

Case Category	Precision	Recall	F1	Cases
No documented reactions or intolerances	1.0000	1.0000	1.0000	114
Drug class cross-reactivity with documented tolerance	1.0000	1.0000	1.0000	2
Drug class cross-reactivity without documented tolerance	0.0000	0.0000	0.0000	1
No reactivity to prescribed drug's ingredients or excipients	1.0000	0.9783	0.9890	45
Macro Average	**1.0000**	**0.9928**	**0.9963**	**162**

Table 19
Classification results per reaction type for GPT-4o on real-world dataset.

Reaction Type	Precision	Recall	F1	Cases
None	1.0000	0.9938	0.9969	161
Non life-threatening immune-mediated	0.0000	0.0000	0.0000	1
Macro Average	**1.0000**	**0.9938**	**0.9969**	**162**

Table 20
Classification results per alert type for GPT-4o on real-world dataset.

Alert Type	Ground Truth (%)	Heliot (%)	Traditional Systems (%)
No Alert Needed	162 (100 %)	161 (99.4 %)	160 (98.76 %)
Interruptive Alert	0 (0 %)	1 (0.6 %)	2 (1.2 %)
Non-Interruptive Alert	0 (0 %)	0 (0 %)	0 (0 %)

required no alert, with one case being conservatively flagged for an interruptive alert. This conservative approach aligns with clinical safety requirements, where false positives are preferable to missed potential adverse reactions.

In terms of response time, the evaluation corroborates the initial results, with 2.3 s on average per clinical case classification.

This real-world evaluation with GPT-4o confirms several critical aspects of the proposed CDSS clinical deployment readiness. GPT-4o successfully processed Italian clinical notes through HELIOT's linguistic standardization approach, demonstrating effectiveness across different languages and medical terminology systems. The model's natural language processing capabilities handled the documentation styles, abbreviations, and clinical terminology in actual patient records, showing strong adaptability to real-world documentation practices.

5.2.2. Gemma 3 performance analysis

The deployment of Google's Gemma 3 (12B parameters) within the HELIOT framework yielded strong performance metrics on the real-world dataset (see Table 21), with an overall accuracy of 98.77 % and macro-averaged metrics of precision 100.00 %, recall 98.58 %, and F1-score 99.28 %. While maintaining high overall performance, Gemma 3 exhibited distinct behavioral patterns compared to GPT-4o in handling edge cases and classification boundaries. Gemma 3 demonstrated perfect precision across all populated categories, correctly identifying all 113 patients with no documented reactions and both cases of drug class cross-reactivity with documented tolerance. However, the model showed a more conservative classification approach, with 2 cases from the "no reactivity to prescribed ingredients or excipients" category being over-classified as "drug class cross-reactivity without documented tolerance." This behavior resulted in a recall of 95.74 % for the former category, representing a more cautious stance that prioritizes potential risk identification over classification precision.

For reaction type classification, Gemma 3 achieved 98.77 % accuracy with perfect precision (100.00 %) and recall of 98.77 % (see Table 22). The model correctly identified most cases with no adverse reactions (160 out of 162 cases) but conservatively classified 2 cases as life-threatening reactions. This represents a more alert-prone behavior compared to GPT-4o, reflecting Gemma 3's tendency toward risk-averse classification in ambiguous scenarios.

From a clinical safety perspective, Gemma 3's conservative approach generated two interruptive alerts (1.2 % of cases) compared to GPT-4o's single alert, indicating a slightly more alert-prone behavior (see Table 23). The alert type classification mirrors the reaction classification performance, with 98.77 % accuracy and perfect precision for correctly identified cases requiring no alert (160 out of 162). While this result led to no reduction in alerts compared to traditional CDSSs for this specific dataset, the model maintained zero false negatives for critical reactions, thereby preserving the essential safety profile required for clinical deployment.

Processing efficiency with Gemma 3 was comparable to GPT-4o, averaging 2.51 s per response. The model benefited from the dataset's characteristics, where many patients reported no allergies, enabling more rapid processing of straightforward cases. Gemma 3's performance profile suggests a model architecture that errs on the side of caution, potentially valuable in clinical contexts where false positives are preferable to missed reactions, albeit at the cost of slightly increased alert frequency compared to GPT-4o's more balanced approach.

5.2.3. Claude Sonnet performance analysis

Claude Sonnet's on the real-world dataset achieved excellent results, with 99.38 % accuracy, perfect macro precision (100.00 %), and robust recall (99.29 %) and F1-score (99.64 %) (see Table 24).

The model exhibited a single but consistent error pattern while maintaining perfect accuracy for the majority classes-correctly identifying all 113 patients with no documented reactions and both cases of drug class cross-reactivity with documented tolerance. One case that should have been classified as "No Reactivity to Prescribed Drug's Ingredients or Excipients" was instead categorized as "Drug Class Cross-Reactivity Without Documented Tolerance," triggering an unnecessary interruptive alert.

Claude Sonnet attained 99.38 % accuracy for reaction type classification, with perfect precision (100.00 %) and recall of 99.38 % (see Table 25). The model correctly identified most cases with no adverse reactions (161 out of 162 cases) but classified one case as non life-threatening immune-mediated reaction. This conservative approach mirrors the pattern observed in the category classification, demonstrating consistent risk-averse behavior.

From a computational efficiency perspective, Claude Sonnet averaged 1.94 s per clinical case compared to GPT-4o's 2.3-s average on this dataset. However, this contrasts with synthetic dataset evaluation where Claude was approximately 57 % slower than GPT-4o, suggesting variable performance depending on case complexity.

The model's effective handling of Italian clinical documentation, with performance nearly identical to GPT-4o, establishes it as a reliable alternative for clinical decision support while maintaining the same conservative safety profile.

5.3. Clinical usability evaluation results

The clinical usability evaluation, conducted with nine healthcare professionals, demonstrated a positive reception of HELIOT's contextual decision support approach. Clinical appropriateness received a mean score of 4.97 out of 5 across all seven clinical cases, with risk assessment accuracy achieving unanimous perfect ratings of 5.0, indicating complete agreement with the system's clinical reasoning and safety classifi-

Table 21
Classification results per category - Gemma 3 real-world evaluation.

Case Category	Precision	Recall	F1	Cases
Drug class cross-reactivity with documented tolerance	1.0000	1.0000	1.0000	2
No documented reactions or intolerances	1.0000	1.0000	1.0000	113
No reactivity to prescribed drug's ingredients or excipients	1.0000	0.9574	0.9783	45
Drug class cross-reactivity without documented tolerance	0	0	0	2
Macro Average	**1.0000**	**0.9858**	**0.9928**	**162**

Table 22
Classification results per reaction type - Gemma 3 real-world evaluation.

Reaction Type	Precision	Recall	F1	Cases
None	1.0000	0.9877	0.9938	160
Life-threatening	0	0	0	2
Macro Average	**1.0000**	**0.9877**	**0.9938**	**162**

Table 23
Classification results per alert type - Gemma 3 real-world evaluation.

Alert Type	Ground Truth (%)	Heliot (%)	Traditional Systems (%)
No Alert Needed	162 (100 %)	160 (98.8 %)	160 (98.8 %)
Interruptive Alert	0 (0 %)	2 (1.2 %)	2 (1.2 %)
Non-Interruptive Alert	0 (0 %)	0 (0 %)	0 (0 %)

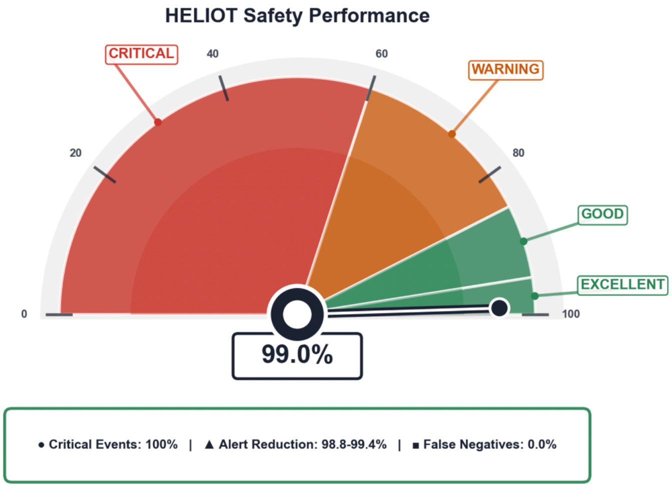


Fig. 9. HELIOT safety performance dashboard.

cations. This high clinical acceptance translated into a unanimous preference for HELIOT's alert approach over traditional CDSS across all scenarios. The system's ability to reduce alert fatigue was particularly well-received, with 67 % of participants rating the impact as "significantly reduces alert fatigue" and 33 % as "reduces alert fatigue." Supporting this effectiveness, participants considered the system's contextual understanding as "excellent" (78 %) or "good" (22 %), while information clarity received "very clear and easy to understand" ratings from 89 % of participants. The system's transparency and trustworthiness further reinforced clinical acceptance. Clinical reasoning transparency received consistently high ratings, with 89 % rating it as "excellent explanation of reasoning" and 11 % rating it as "good explanation of reasoning." This transparency contributed to high confidence levels, with 44 % of participants rating themselves as "very confident" and 56 % as "confident" in HELIOT's recommendations. These assessments directly influenced adoption intentions and perceived workflow benefits. Participants judged HELIOT's clinical decision support value highly, with 44 % considering it "extremely valuable" and 56 % considering it "very valuable." Correspondingly, the adoption likelihood was positive, with 78 % indicating they would be "very likely" to adopt the system and 22 % rating it as "likely." Time efficiency expectations aligned with these intentions, as 78 % indicated that HELIOT would "significantly reduce time spent," and 22 % noted it would "moderately reduce time spent."

Table 27 summarizes the results across all evaluated dimensions, demonstrating that HELIOT's contextual approach addresses key limitations of traditional CDSS systems while maintaining appropriate clinical safety standards.

5.4. Safety analysis and risk assessment

Safety analysis revealed remarkable profiles across both datasets. For the synthetic dataset, GPT-4o correctly identified all 133 life-threatening cases, maintaining optimal sensitivity and specificity parameters (1.0000) for critical events and detection scenarios. Gemma 3 achieved comparable results with zero undetected potential critical reactions. Despite a modest increase in false alarm rate (3.10 %) compared to GPT-4o's flawless performance, it preserved the essential safety threshold with perfect sensitivity (1.0000) for dangerous reactions and high specificity (0.9982). Claude Sonnet reached 100 % across all metrics, eliminating false negatives and positives.

The analysis using the real-world dataset confirmed HELIOT robustness in authentic clinical scenarios while maintaining zero unidentified critical events. GPT-4o correctly classified 99.4 % of cases as not requiring intervention, with a single false positive generating a conservative false alarm rate of 1.23 % and a perfect negative predictive value (1.0000). Gemma 3 showed analogous characteristics with 98.8 % correct classifications. It presented a slightly higher false alarm rate (2.47 %) with two false positives while preserving the fundamental profile of zero critical omissions. Claude Sonnet matched GPT-4o's performance with 99.4 % of cases appropriately classified and a single false positive, maintaining perfect negative predictive value and zero missed critical events. The consistency of results across different architectures confirms the robustness of the HELIOT framework (see Fig. 9). In both evaluations, all LLMs met the indispensable requirement of zero omissions for critical events. Minimal variations in false positive rates (0.00 % – 3.10 %) reflect different degrees of caution without compromising patient safety. Alert distribution showed uniformity, with all systems identifying most real cases as not requiring intervention (98.8 % –99.4 %), demonstrating that HELIOT's approach to alert fatigue reduction preserves safety standards regardless of the underlying architecture. This multi-model validation establishes reference parameters for clinical implementation. The consistent rate of zero false negatives supports the adoption of automated decisions for "no intervention required" classifications. Acceptable rates for false positives provide a prudential margin that ensures appropriate clinical attention to potentially problematic cases while significantly reducing the burden of superfluous notifications.

5.5. Confidence analysis

The confidence scoring analysis (see Table 28) across all three LLMs revealed distinct patterns that provide valuable insights for risk-

Table 24
Classification results per category - Claude Sonnet real-world evaluation.

Case Category	Precision	Recall	F1	Cases
Drug class cross-reactivity with documented tolerance	1.0000	1.0000	1.0000	2
Drug class cross-reactivity without documented tolerance	0.0000	0.0000	0.0000	1
No documented reactions or intolerances	1.0000	1.0000	1.0000	113
No reactivity to prescribed drug's ingredients or excipients	1.0000	0.9787	0.9892	46
Macro Average	**1.0000**	**0.9929**	**0.9964**	**162**

Table 25
Classification results per reaction type - Claude Sonnet real-world evaluation.

Reaction Type	Precision	Recall	F1	Cases
None	1.0000	0.9938	0.9969	161
Non Life-threatening Immune-Mediated	0	0	0	1
Macro Average	**1.0000**	**0.9938**	**0.9969**	**162**

Table 26
Classification results per alert type - Claude Sonnet real-world evaluation.

Alert Type	Ground Truth (%)	Heliot (%)	Traditional Systems (%)
No Alert Needed	162 (100 %)	161 (99.4 %)	160 (98.76 %)
Interruptive Alert	0 (0 %)	1 (0.6 %)	2 (1.2 %)
Non-Interruptive Alert	0 (0 %)	0 (0 %)	0 (0 %)

stratified clinical deployment. Each model exhibits unique confidence characteristics while maintaining consistent safety profiles.

GPT-4o demonstrated robust confidence levels across both evaluation scenarios. In the synthetic dataset, the model exhibited high mean confidence scores: overall confidence (0.855), analysis confidence (0.791), case confidence (0.999), and reaction confidence (0.994). The two observed misclassifications maintained similar confidence patterns (overall: 0.853, analysis: 0.788, case: 0.999, reaction: 0.994), occurring between clinically similar categories with non-life-threatening profiles. For real-world validation, GPT-4o showed comparable patterns with correctly classified cases achieving mean confidence scores of: overall confidence (0.870), analysis confidence (0.745), case confidence (0.999), and reaction confidence (0.999). The single misclassification revealed a notable confidence drop: overall confidence (0.819), analysis confidence (0.747), case confidence (0.975), and reaction confidence (0.966). Critically, the analysis confidence of 0.747 falls within the medium-risk threshold range, which would appropriately trigger clinical review protocols.

Gemma 3 exhibited the highest overall confidence levels in synthetic evaluation with mean scores of overall confidence (0.910), analysis confidence (0.732), case confidence (0.999), and reaction confidence (0.997). Misclassifications showed confidence degradation patterns, where overall confidence (0.897–0.905), analysis confidence (0.713–0.736), case confidence (0.979), and reaction confidence maintained near-perfect levels (0.999). In real-world scenarios, Gemma 3 demonstrated consistent confidence patterns with mean scores of overall confidence (0.899), analysis confidence (0.721), case confidence (0.986), and reaction confidence (0.997). The model's two misclassifications revealed more pronounced confidence drops: overall confidence (0.772–0.806), analysis confidence (0.604-0.626), and notably lower case confidence (0.715–0.816), indicating the model's uncertainty recognition capabilities.

Claude Sonnet displayed unique confidence patterns with moderate overall confidence (0.815) but strong analysis confidence (0.786) in synthetic evaluation. Case confidence (0.990) and reaction confidence (0.950) remained high. Notably, the model's rare misclassifications maintained high case confidence (0.999) and reaction confidence (0.999), suggesting different uncertainty expression mechanisms compared to other models. Real-world evaluation showed improved confidence levels: overall confidence (0.873), analysis confidence (0.767), case confidence (0.988), and reaction confidence (0.999). The single misclassification maintained high confidence across most metrics (overall: 0.796, analysis: 0.786, case: 0.999, reaction: 0.999), indicating the model's consistent decision-making approach even in uncertain scenarios. The multi-model analysis validates a risk threshold framework that accounts for varying confidence expression patterns across different architectures. While GPT-4o shows apparent confidence degradation in uncertain cases, Gemma 3 demonstrates more pronounced uncertainty signals, and Claude Sonnet maintains consistent high confidence even during misclassifications.

Based on the confidence analysis across all models and datasets, we propose implementing the risk threshold framework detailed in Table 29. This framework establishes four risk levels based on confidence thresholds: cases with case confidence above 0.95 and analysis confidence above 0.75 qualify for automated implementation, case confidence between 0.80–0.95 or analysis confidence between 0.60–0.75 trigger non-interruptive notifications for clinical awareness, case confidence below 0.80 or analysis confidence below 0.60 requires mandatory clinical review, and any confidence metric below 0.50 demands immediate clinical evaluation with override prevention.

This multi-dimensional confidence assessment approach accommodates the distinct uncertainty expression patterns observed across models while maintaining the demonstrated alert reduction capabilities. The framework ensures appropriate clinical oversight for uncertain decisions regardless of the underlying language model architecture, supporting robust clinical deployment across different institutional preferences for model selection.

5.6. Research question assessment

The evaluation highlights that HELIOT effectively addresses the core challenges of clinical decision support for drug administration. Across all tested language models, the system obtained high classification accuracy (98.77 % – 99.80 %) while maintaining a perfect safety profile with zero false negatives for life-threatening reactions. Most significantly, HELIOT reduced unnecessary interruptive alerts by 50–53 % compared to traditional CDSSs, directly addressing the alert fatigue problem that compromises patient safety. The system's contextual approach enables precise risk stratification, supported by a confidence-based framework that ensures appropriate clinical oversight for uncertain cases. A clinical usability evaluation confirmed strong acceptance among healthcare professionals, with unanimous preference for HELIOT's alerts over traditional systems. Response times (1.94–6.73 s) remain clinically acceptable for real-time decision support. These results align with Co et al. [22], who demonstrated that traditional CDSSs achieve performance improvements through overalerting, creating alert fatigue that undermines clinical effectiveness. HELIOT's contextual understanding successfully distinguishes between necessary and unnecessary alerts, addressing this fundamental trade-off.

Based on the evaluation results presented above, we can address our research question:

RQ: How effective is HELIOT CDSS in identifying potential drug reactions and reducing alert fatigue? HELIOT can provide effective support for drug prescription through accurate identification of potential

Table 27
Clinical usability evaluation results summary.

Evaluation Dimension	Scale	Mean Rating
Clinical Appropriateness	1–5 (Inappropriate-Appropriate)	4.97
Risk Assessment Accuracy	1–5 (Inaccurate-Accurate)	5.0
Alert Preference	HELIOT vs Traditional	100 % HELIOT
Alert Fatigue Impact	Increase-Significantly Reduce	Significantly Reduce (67 %)
Alert Contextuality	No-Excellent Understanding	Excellent (78 %)
Information Clarity	1–5 (Unclear-Clear)	5.0
Clinical Reasoning Transparency	1–5 (No-Excellent Explanation)	4.89
Time Efficiency Impact	Increase-Significantly Reduce	Significantly Reduce (78 %)
Trust and Confidence	1–5 (Not-Very Confident)	4.44
Clinical Decision Support Value	1–5 (Not-Extremely Valuable)	4.44
Likelihood of Adoption	1–5 (Unlikely-Very Likely)	4.78

Table 28
Confidence analysis results across LLMs and datasets.

Model	Dataset	Classification	Overall Conf.	Analysis Conf.	Case Conf.	Reaction Conf.	Cases (n)
GPT-4o	Synthetic	Correct	0.855	0.791	0.999	0.994	998
		Incorrect	0.853	0.788	0.999	0.994	2
	Real-world	Correct	0.870	0.745	0.999	0.999	161
		Incorrect	0.819	0.747	0.975	0.966	1
Gemma 3	Synthetic	Correct	0.910	0.732	0.999	0.997	998
		Incorrect	0.901	0.725	0.979	0.999	2
	Real-world	Correct	0.899	0.721	0.986	0.997	160
		Incorrect	0.789	0.615	0.766	0.997	2
Claude Sonnet	Synthetic	Correct	0.815	0.786	0.990	0.950	998
		Incorrect	0.797	0.786	0.999	0.999	2
	Real-world	Correct	0.873	0.767	0.988	0.999	161
		Incorrect	0.796	0.786	0.999	0.999	1

Table 29
Risk threshold framework for clinical decision support.

Confidence Range	Risk Level	Clinical Action	System Response
Case Confidence > 0.95	Low Risk	Automated decision implementation	Direct alert delivery or no alert as determined
Analysis Confidence > 0.75		Proceed without interruption	Normal workflow continuation
Case Confidence 0.80–0.95	Medium Risk	Non-interruptive notification	Background alert with rationale
Analysis Confidence 0.60–0.75		Clinical awareness recommended	Passive information display
Case Confidence < 0.80	High Risk	Mandatory clinical review	Interruptive alert requiring acknowledgment
Analysis Confidence < 0.60		Human verification required	System recommendation with uncertainty flag
Any confidence metric < 0.50	Critical Risk	Immediate clinical evaluation	Override prevention until review
Multiple low confidence scores		Expert consultation recommended	Escalation to senior clinician

adverse drug reactions while reducing alert fatigue. Clinical validation in operational environments remains necessary to confirm practical deployment effectiveness.

6. Implications, limitations and future work

This section discusses the practical implications of our findings and outlines limitations and future research directions.

6.1. Practical implications and future work

The initial results of HELIOT CDSS suggest advancements in addressing key challenges in adverse drug reaction management, particularly in healthcare settings where clinical information exists primarily in unstructured formats. By leveraging LLMs to interpret clinical narratives and generate contextual alerts, HELIOT demonstrates an approach to reduce alert fatigue while maintaining safety, a significant improvement over traditional rule-based systems that typically generate non-specific alerts for all potentially cross-reactive medications.

From an implementation perspective, HELIOT's ability to process unstructured clinical notes and generate appropriate alerts makes it particularly valuable across diverse healthcare settings. While some environments have advanced EHR systems that could feed structured data into HELIOT's database, many healthcare facilities worldwide still operate with basic or no EHR infrastructure, relying primarily on unstructured clinical notes. HELIOT's flexible architecture accommodates both scenarios: it can process data from existing EHR systems where available while functioning independently using clinical narratives in settings with limited technological infrastructure. This versatility and its microservices architecture and streaming response mechanism facilitate adaptation to different clinical workflows, from specialized hospitals to primary care practices. Moreover, the system's capability to maintain patient clinical histories could enhance continuity of care across different healthcare settings.

The impact on patient safety and healthcare costs may be significant. By improving the accuracy of adverse drug reaction alerts and reducing alert fatigue, HELIOT could help prevent ADEs while ensuring that critical warnings are not overlooked. All this leads to fewer medication errors, reduced hospital readmissions, and shorter hospital stays. The economic implications of these improvements could be substantial, as medication errors and ADEs impose significant costs on healthcare systems. The resources saved could be redirected to other critical aspects of patient care.

Looking ahead, several crucial directions for future work emerge. Large-scale clinical validation represents our immediate priority. While our pilot study with 162 real-world cases provides encouraging initial evidence, comprehensive validation requires larger datasets across diverse healthcare settings, different languages, and varied clinical documentation practices. To fully assess its clinical utility, future studies must evaluate HELIOT's performance with higher rates of contraindicated prescriptions and more complex adverse reaction scenarios. Additionally, conducting prospective clinical trials will be essential to measure the system's impact on clinical workflow, alert override rates, and patient outcomes in live healthcare environments.

Technical enhancement and optimization efforts will focus on improving system performance and deployment flexibility. We plan to systematically evaluate alternative LLM architectures, including open-source models such as LLaMA and FLAN-T5 and specialized medical LLMs like ClinicalBERT and BioBERT. This evaluation will assess accuracy metrics and factors critical for clinical deployment, including response consistency, computational requirements, and cost-effectiveness. Parallel development will implement intelligent caching mechanisms for frequently queried drug-reaction combinations and explore model distillation techniques to reduce latency while maintaining clinical decision support quality.

Extended clinical applications will leverage HELIOT's LLM-based architecture to address broader healthcare challenges beyond adverse drug reactions. Future research will explore expanding into diagnostic assistance, treatment protocol recommendations, and medication dosing optimization. We will investigate integrating patient-reported outcomes to enhance personalized recommendations by processing patient narratives about medication effectiveness and impact on quality of life. Additionally, implementing continuous learning mechanisms through real-world clinical feedback could enable HELIOT to adapt to evolving clinical practices and institution-specific prescribing patterns, creating a self-improving system that learns from clinical outcomes and prescriber experiences.

6.2. Limitations

Despite the promising results, HELIOT presents a few limitations that we have addressed through various mitigation strategies. The primary limitation of our evaluation concerns the scale of real-world validation. While we expanded beyond synthetic data to include 162 real-world patient cases, this represents a limited sample size for comprehensive clinical validation. To address this, we conducted rigorous validation and demonstrated consistent performance across three different LLM architectures (GPT-4o, Gemma 3, and Claude Sonnet), showing the robustness of our approach. Additionally, we achieved perfect safety profiles with zero false negatives for critical reactions across all models, establishing a foundation for larger-scale studies. We plan to conduct extensive multi-center validation studies with larger patient cohorts to validate these findings further.

Another problem relates to the geographic and linguistic scope of our evaluation, which focused primarily on Italian clinical documentation and European pharmaceutical standards. We managed this limitation by demonstrating that our framework successfully processes different documentation styles and clinical terminologies, and by designing the system architecture to support multiple languages and drug databases. The modular design enables easy adaptation to various healthcare systems and regulatory frameworks.

The dependency on external LLM services for cloud-based models (GPT-4o and Claude Sonnet) presents potential concerns regarding data privacy and service availability. We addressed this aspect by implementing and validating a local deployment option using Gemma 3, which achieved comparable performance (99.80 % accuracy) while maintaining data sovereignty. This dual deployment approach provides healthcare institutions with flexibility to choose between cloud-based efficiency and local privacy controls based on their specific requirements.

The management of drug leaflets containing information for multiple pharmaceutical forms remains a consideration; however, our evaluation showed that this had minimal impact on performance. We tackled this issue by providing specific prompts to extract information relevant only to the pharmaceutical form of interest, with healthcare professionals validating the extraction process. The high precision scores ($>98\,\%$) across all models demonstrate the effectiveness of this approach.

Regarding the non-deterministic nature of LLMs, our comprehensive evaluation across five runs for each model revealed perfect consistency (Fleiss kappa = 100 %) across all tested architectures, effectively eliminating this concern. We employed best practices in prompt engineering, structured templates, and clear instructions, resulting in deterministic behavior despite the theoretical potential for variability. The confidence scoring framework we developed provides additional safeguards by flagging uncertain decisions for clinical review.

6.3. Deployment challenges and mitigation strategies

The transition from research prototype to clinical implementation presents several challenges that must be addressed to ensure the successful deployment of HELIOT in real-world healthcare settings. HELIOT's modular architecture supports two primary deployment scenarios, each with distinct challenges and requirements. In integrated deployment, HELIOT operates as a background service within existing EHR systems, requiring seamless technical integration without disrupting established clinical workflows. Healthcare institutions operate diverse EHR platforms including Epic, Cerner, and Allscripts, each with unique data formats and APIs. To address this challenge, we implement FHIR-compliant APIs providing standardized data exchange formats and HL7 messaging standards for real-time data communication. The key advantage of integrated deployment is that healthcare professionals continue using their familiar EHR interfaces with no additional training requirements, as HELIOT's recommendations appear within existing alert systems and decision support modules.

Standalone deployment presents different challenges, primarily centered on user adoption and training. Healthcare professionals must learn a new interface and integrate it into their existing workflows, requiring training programs tailored for physicians, pharmacists, and nurses. Our mitigation strategy focuses on developing intuitive interfaces that mirror familiar clinical workflows and implementing a clinical champion program to provide peer support during adoption. A gradual rollout strategy beginning with specific departments allows progressive adaptation while demonstrating measurable value.

Both deployment scenarios face common technical infrastructure challenges. Cloud-native deployment options accommodate institutions with limited IT capabilities, while auto-scaling handles varying workloads without manual management. As an open-source solution, HELIOT eliminates software licensing costs, with infrastructure expenses being the primary financial consideration. This significantly reduces deployment barriers, allowing institutions to focus resources on server infrastructure rather than proprietary software licenses. Technical support including monitoring and maintenance ensures the reliability required in clinical environments, while pilot implementations allow institutions to evaluate benefits and infrastructure requirements before full-scale deployment.

6.4. Ethical considerations and regulatory compliance

The integration of AI into clinical workflows raises fundamental questions about accountability, privacy, and patient safety. HELIOT addresses these concerns through a human-centered design philosophy that positions healthcare professionals at the center of all clinical decisions. By analyzing clinical notes and drug codes, the system generates evidence-based recommendations with transparent explanations (see Section 3.1.1)-but these remain advisory insights rather than prescriptive directives. This approach creates a collaborative dynamic between human expertise and artificial intelligence, where professionals retain full authority to evaluate and accept or reject recommendations based on their clinical judgment. Such oversight preserves clinical autonomy while simultaneously acting as a natural barrier against potential algorithmic bias.

Building on this foundation, HELIOT's technical architecture prioritizes data protection through deliberate design choices. The system processes only anonymized clinical notes and standardized pharmaceutical codes, creating an inherent barrier against privacy breaches while maintaining analytical effectiveness. The platform's modular architecture offers flexible deployment options: standalone mode operates entirely within institutional boundaries, while EHR integration utilizes secure APIs to maintain consistent privacy standards without compromising functionality.

These technical safeguards represent only part of the compliance equation. HELIOT's current position in early validation phases reflects a methodical approach to regulatory approval, with comprehensive clinical validation studies planned under rigorous ethical supervision. Each phase must meet established compliance standards before it can be advanced. Nevertheless, technology alone cannot guarantee ethical implementation. Healthcare institutions bear responsibility for establishing governance frameworks-including staff training, outcome monitoring, and workflow integration. This shared responsibility model ensures that AI enhancements strengthen clinical judgment rather than replace the human expertise essential to quality patient care.

7. Conclusion

This paper presented HELIOT CDSS, a novel approach to adverse drug reaction management that leverages LLMs to interpret unstructured clinical narratives. While traditional CDSSs often struggle with free-text clinical notes and generate excessive alerts due to their rule-based nature, HELIOT demonstrated promise in processing natural language descriptions of adverse reactions and generating more contextual alerts. Our evaluation across multiple LLMs and datasets validated HELIOT's effectiveness in clinical decision support. The system successfully balances alert reduction with patient safety requirements. The implemented risk management framework provides structured approaches to handle classification uncertainty, supporting clinical deployment.

The potential impact of this approach could be considerable, particularly in healthcare environments with varying levels of EHR integration. By improving the interpretation of clinical narratives and providing more contextual alerts, HELIOT could help reduce alert fatigue while maintaining patient safety. All this could lead to better clinical outcomes and cost savings by preventing adverse drug events and reducing unnecessary alert overrides.

Future work will focus on validating the system in operational clinical environments to confirm practical deployment effectiveness and workflow integration. We will also enhance the system's capabilities for complex clinical scenarios and evaluate its workflow impact across diverse healthcare settings.

Author Agreement

The authors certify that all co-authors have seen and approved the final version of the manuscript being submitted. We warrant that this article represents our original work, has not been previously published, and is not currently under consideration for publication elsewhere.

CRediT authorship contribution statement

Gabriele De Vito: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Filomena Ferrucci:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis; **Athanasios Angelakis:** Writing – review & editing, Validation, Supervision.

Data availability

We have shared the link to our data and code in the manuscript

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Usability questionnaire

You will be presented with 7 clinical cases, grouped in scenarios, where the CDSS HELIOT provides decision support for medication prescribing. For each case, please review the patient information, clinical notes, and both HELIOT's output and a traditional CDSS output, then answer the questions below.

In these scenarios, traditional CDSSs raise **Interruptive Alert**. Interruptive Alert stops the prescribing workflow; you cannot proceed with the prescription unless you actively override the alert by providing a justification, or you choose to cancel the prescription altogether. Alert fatigue arises when such interruptive alerts are triggered frequently, even for known patient tolerances or expected side effects [6–13]. HELIOT addresses this by using **Non-Interruptive Alert** in these cases, allowing clinicians to continue their workflow without unnecessary interruptions. Therefore, it is not necessary to block the prescription workflow.

** Indicates mandatory questions*

Scenario: Cross-reactivity with documented tolerance

The following cases involve patients who have had reactions to drugs or drug classes but report tolerance for the medication to be prescribed. In addition, there are cases where patients have had side effects.

Case 1

Patient history: Patient shows Anaphylaxis reaction within Opioids class but has documented tolerance to fentanyl since 2016
Prescribed drug: Fentanil
HELIOT output: Patient has anaphylaxis to opioids but documented tolerance to fentanyl since 2016. Fentanyl is the active ingredient in the prescribed drug. Drug class cross-reactivity exists but with documented tolerance to fentanyl. Therefore the drug can be prescribed. Non-Interruptive Alert
Traditional CDSS output: Interruptive Alert

Case 2

Patient history: Urticarial reaction after taking pyrazolones (Zepelin, Zerinol, Novalgina) and sulfonamides (Bactrim); itching after taking Danzen and opioids; tremor after taking Plasil. Food intolerances to pepper, onion, salami, sausage. Avoids shellfish and mollusks and consumes foods rich in histamine (spinach, tuna, eggplant, pineapple, strawberries, kiwi) with caution.
Prescribed drug: Morphine
HELIOT output: The patient has documented reactions to opioids, such as itching. There is no direct reactivity to the active ingredient morphine hydrochloride, nor to its excipients. There is no documented tolerance to morphine in patient's medical history. Despite the side effect, the medication can still be prescribed with caution. Non-Interruptive Alert
Traditional CDSS output: Interruptive Alert

Case 3

Patient history: Reports an allergy to Amoxicillin and beta-lactamase inhibitors, tolerates Meropenem.
Prescribed drug: Meropenem
HELIOT output: The patient reports an allergy to amoxicillin and beta-lactamase inhibitors but tolerates meropenem. Although the patient is allergic to beta-lactam antibiotics, they tolerate meropenem, which is

also a beta-lactam. This case is classified as drug class cross-reactivity with documented tolerance, meaning that the drug can be prescribed. Non-Interruptive Alert
Traditional CDSS output: Interruptive Alert

Scenario: Cross-reactivity without documented tolerance

The following case involves patients who have had reactions to drugs or drug classes without reporting tolerance to the medication to be prescribed. In this scenario, CDSSs should raise interruptive alerts, because prescribing the specific drug may cause severe reaction to patients.

Case 4

Patient history: Reports anaphylaxis after taking opioids in 2021
Prescribed drug: Morphine
HELIOT output: The patient reports anaphylaxis to opioids without a specific tolerance to morphine. This case is classified as drug class cross-reactivity without documented tolerance, meaning that the drug should not be prescribed because it may cause life-threatening reactions. Interruptive Alert
Traditional CDSS output: Interruptive Alert

Scenario: Side effects, non-immune and non-life-threatening

The following cases involve patients who have experienced side effects from medications or classes of medications, for which the prescribing physician should be informed, but it is not necessary to block the prescription workflow and override alerts to move on prescribing the drug.

Case 5

Patient history: Patient showed Mild shortness of breath side effect, not indicative of hypersensitivity, with possible cross-reactivity within Opioids class, specifically to tramadol and dexketoprofen since 2014.
Prescribed drug: oxycodone
HELIOT output: The patient has mild shortness of breath with opioids, specifically tramadol, which indicates a non-immune-mediated side effect. While oxycodone also belongs to the opioid class, suggesting potential cross-reactivity, it can be prescribed with caution as there is no immune-mediated reaction or documented intolerance. Non-Interruptive Alert
Traditional CDSS output: Interruptive Alert

Case 6

Patient history: Patient showed Vomiting side effect, not indicative of hypersensitivity, to INDOMETACINA (NSAID) in 2000.
Prescribed drug: indometacin
HELIOT output: The patient experienced vomiting as a documented side effect to indometacin, which is the same active ingredient as that in the prescribed drug Indomethacin, indicating direct active ingredient reactivity. This reaction is non-life-threatening, as it is a side effect and not due to hypersensitivity, and should be considered non-immune and non-life-threatening. Non-Interruptive Alert
Traditional CDSS output: Interruptive Alert

Case 7

Patient history: The patient reports itching after taking codeine and ibuprofen in 2012.
Prescribed drug: codeine
HELIOT output: The patient reports itching after taking Codeine, an active ingredient in the prescribed drug, indicating a direct active ingredient reactivity. Given that this is a non-immune, non-life-threatening side effect, the medication can be prescribed but with caution. Non-Interruptive Alert
Traditional CDSS output: Interruptive Alert The following cases involve patients who have had reactions to drugs or drug classes but report tolerance for the medication to be prescribed. In addition, there are cases where patients have had side effects.

Section A: Clinical accuracy assessment

For each scenario (1–7), please rate:

A1. Clinical appropriateness of HELIOT's response:
Rate from 1 (completely inappropriate) to 5 (completely appropriate) *

	1	2	3	4	5
Case 1	○	○	○	○	○
Case 2	○	○	○	○	○
Case 3	○	○	○	○	○
Case 4	○	○	○	○	○
Case 5	○	○	○	○	○
Case 6	○	○	○	○	○
Case 7	○	○	○	○	○

A2. Risk assessment accuracy
How accurately does HELIOT identify and categorize clinical risks (i.e. Life-threatening situations, where drugs may cause severe reactions)? Rate from 1 (Very inaccurate) to 5 (Very accurate) *

	1	2	3	4	5
Case 1	○	○	○	○	○
Case 2	○	○	○	○	○
Case 3	○	○	○	○	○
Case 4	○	○	○	○	○
Case 5	○	○	○	○	○
Case 6	○	○	○	○	○
Case 7	○	○	○	○	○

Section B: Alert fatigue and appropriateness

B1. Alert preference *

	HELIOT's Alert	Traditional CDSS Alert	No Preference
Case 1	○	○	○
Case 2	○	○	○
Case 3	○	○	○
Case 4	○	○	○
Case 5	○	○	○
Case 6	○	○	○
Case 7	○	○	○

B2. Alert fatigue impact *
Rate how much HELIOT's approach would reduce alert fatigue compared to traditional CDSS *

○ Increase alert fatigue
○ No difference
○ Slightly reduces alert fatigue
○ Reduces alert fatigue
○ Significantly reduces alert fatigue

B3. Alert contextuality *
How well does HELIOT consider clinical context when generating alerts? *

- ○ No contextual understanding
- ○ Poor contextual understanding
- ○ Adequate contextual understanding
- ○ Good contextual understanding
- ○ Excellent contextual understanding

Section C: Usability and clinical integration

C1. Information clarity

How clear and understandable are HELIOT's recommendations? *

- ○ Very unclear
- ○ Unclear
- ○ Somewhat clear
- ○ Clear and understandable
- ○ Very clear and easy to understand

C2. Clinical reasoning transparency

How well does HELIOT explain its clinical reasoning? *

- ○ No explanation provided
- ○ Poor explanation of reasoning
- ○ Adequate explanation of reasoning
- ○ Good explanation of reasoning
- ○ Excellent explanation of reasoning

C3. Time efficiency

How would HELIOT impact the time you spend reviewing medication alerts? *

- ○ Increases time spent
- ○ No change in time spent
- ○ Slightly reduces time spent
- ○ Moderately reduces time spent
- ○ Significantly reduces time spent

C4. Trust and confidence

How confident would you feel considering HELIOT's recommendations in your decisions? *

- ○ Not confident at all
- ○ Not very confident
- ○ Somewhat confident
- ○ Confident
- ○ Very confident

Section D: Clinical workflow integration

D1. Clinical decision support value

Overall, how valuable would HELIOT be for your clinical decision-making support? *

- ○ Not valuable
- ○ Slightly valuable
- ○ Somewhat valuable
- ○ Very valuable
- ○ Extremely valuable

D2. Likelihood of adoption

How likely would you be to use HELIOT in your practice if available? *

- ○ Very unlikely
- ○ Unlikely
- ○ Somewhat likely
- ○ Likely
- ○ Very likely

Section E: Demographic information

E1. Medical speciality *

- ○ Internal Medicine
- ○ Emergency Medicine
- ○ Primary Care/Family Medicine
- ○ Clinical Pharmacology
- ○ Anesthesiology
- ○ Other: ____________________

E2. Years of clinical experience *

- ○ < 5 years
- ○ 5–10 years
- ○ 11–20 years
- ○ > 20 years

E3. Current practice setting *

- ○ Hospital (inpatient)
- ○ Outpatient clinic
- ○ Emergency department
- ○ Academic medical center
- ○ Private practice
- ○ Other: ____________________

References

[1] R.A. Elliott, E. Camacho, D. Jankovic, M.J. Sculpher, R. Faria, Economic analysis of the prevalence and clinical and economic burden of medication error in England, BMJ Qual. Saf. 30(2) (2021) 96–105.

[2] E. Ahsani-Estahbanati, V. Sergeevich Gordeev, L. Doshmangir, Interventions to reduce the incidence of medical error and its financial burden in health care systems: a systematic review of systematic reviews, Front. Med. 9 (2022) 875426.

[3] R.T. Sutton, D. Pincock, D.C. Baumgart, D.C. Sadowski, R.N. Fedorak, K.I. Kroeker, An overview of clinical decision support systems: benefits, risks, and strategies for success, NPJ Digital Med. 3 (1) (2020) 17.

[4] J.L. Kwan, L. Lo, J. Ferguson, H. Goldberg, J.P. Diaz-Martinez, G. Tomlinson, J.M. Grimshaw, K.G. Shojania, Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials, BMJ 370 (2020).

[5] U. Sarkar, L. Samal, How effective are clinical decision support systems?, 2020.

[6] P.-Y. Meunier, C. Raynaud, E. Guimaraes, F. Gueyffier, L. Letrilliart, Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review, Ann. Fam. Med. 21 (1) (2023) 57–69.

[7] P.L. Quan, S. Sánchez-Fernández, L. Parrado Gil, A. Calvo Alonso, J.M. Bodero Sánchez, A. Ortega Eslava, M. Luri, G. Gastaminza Lasarte, Usefulness of drug allergy alert systems: present and future, Curr. Treat. Options Allergy 10 (4) (2023) 413–427.

[8] T.K. Colicchio, J.J. Cimino, Beyond the override: using evidence of previous drug tolerance to suppress drug allergy alerts; a retrospective study of opioid alerts, J. Biomed. Inf. 147 (2023) 104508.

[9] B.A. Van Dort, W.Y. Zheng, V. Sundar, M.T. Baysari, Optimizing clinical decision support alerts in electronic medical records: a systematic review of reported strategies adopted by hospitals, J. Am. Med. Inf. Assoc. 28 (1) (2021) 177–183.

[10] L. Westerbeek, K.J. Ploegmakers, G.-J. De Bruijn, A.J. Linn, J.C.M. van Weert, J.G. Daams, N. van der Velde, H.C. van Weert, A. Abu-Hanna, S. Medlock, Barriers and facilitators influencing medication-related CDSS acceptance according to clinicians: a systematic review, Int. J. Med. Inf. 152 (2021) 104506.

[11] G. Van De Sijpe, C. Quintens, K. Walgraeve, E. Van Laer, J. Penny, G. De Vlieger, R. Schrijvers, P. De Munter, V. Foulon, M. Casteels, et al., Overall performance of a drug–drug interaction clinical decision support system: quantitative evaluation and end-user survey, BMC Med. Inf. Decis. Making 22 (1) (2022) 48.

[12] M. Topaz, D.L. Seger, S.P. Slight, F. Goss, K. Lai, P.G. Wickner, K. Blumenthal, N. Dhopeshwarkar, F. Chang, D.W. Bates, et al., Rising drug allergy alert overrides in electronic health records: an observational retrospective study of a decade of experience, J. Am. Med. Inf. Assoc. 23 (3) (2016) 601–608.

[13] M. Topaz, D.L. Seger, K. Lai, P.G. Wickner, F. Goss, N. Dhopeshwarkar, F. Chang, D.W. Bates, L. Zhou, High override rate for opioid drug-allergy interaction alerts: current trends and recommendations for future, in: MEDINFO 2015: eHealth-enabled Health, IOS Press, 2015, pp. 242–246.

[14] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., GPT-4 technical report, (2023) arXiv preprint arXiv:2303.08774.

[15] S. Tripathi, R. Sukumaran, T.S. Cook, Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care, J. Am. Med. Inf. Assoc. 31 (6) (2024) 1436–1440.

[16] A. Ríos-Hoyo, N.L. Shan, A. Li, A.T. Pearson, L. Pusztai, F.M. Howard, Evaluation of large language models as a diagnostic aid for complex medical cases, Front. Med. 11 (2024) 1380148.

[17] G. De Vito, F. Ferrucci, A. Angelakis, LLMs for drug-drug interaction prediction: a comprehensive comparison, (2025) arXiv preprint arXiv:2502.06890.

[18] G. De Vito, Assessing healthcare software built using IoT and LLM technologies, in: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, 2024, pp. 476–481.

[19] R. Roy, S. Marakkar, M.P. Vayalil, A. Shahanaz, A.P. Anil, S. Kunnathpeedikayil, I. Rawal, K. Shetty, Z. Shameer, S. Sathees, et al., Drug-food interactions in the era of molecular big data, machine intelligence, and personalized health, Recent Adv. Food Nutr. Agric. 13 (1) (2022) 27–50.

[20] J. Corny, A. Rajkumar, O. Martin, X. Dode, J.-P. Lajonchère, O. Billuart, Y. Bézie, A. Buronfosse, A machine learning–based clinical decision support system to identify prescriptions with a high risk of medication error, J. Am. Med. Inf. Assoc. 27 (11) (2020) 1688–1694.

[21] M. Luri, L. Leache, G. Gastaminza, A. Idoate, A. Ortega, A systematic review of drug allergy alert systems, Int. J. Med. Inf. 159 (2022) 104673.

[22] Z. Co, A.J. Holmgren, D.C. Classen, L. Newmark, D.L. Seger, M. Danforth, D.W. Bates, The tradeoffs between safety and alert fatigue: data from a national evaluation of hospital medication-related clinical decision support, J. Am. Med. Inf. Assoc. 27 (8) (2020) 1252–1258.

[23] E. Calvo-Cidoncha, C. Camacho-Hernando, F. Feu, X. Pastor-Duran, C. Codina-Jané, R. Lozano-Rubí, OntoPharma: ontology based clinical decision support system to reduce medication prescribing errors, BMC Med. Inf. Decis. Making 22 (1) (2022) 238.

[24] R. Rozenblum, R. Rodriguez-Monguio, L.A. Volk, K.J. Forsythe, S. Myers, M. McGurrin, D.H. Williams, D.W. Bates, G. Schiff, E. Seoane-Vazquez, Using a machine learning system to identify and prevent medication prescribing errors: a clinical and cost analysis evaluation, Joint Commission J. Qual. Patient Saf. 46 (1) (2020) 3 –10.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, volume 30, 2017.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language Models Are Few-Shot Learners, volume 33, 2020, pp. 1877–1901.

[27] J. Devlin, M.-W. Chang, K. Lee, K. 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, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[28] H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, J. Mach. Learn. Res. 25 (70) (2024) 1–53.

[29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., LLaMa: open and efficient foundation language models, 2023.

[30] B. Workshop, T.L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A.S. Luccioni, F. Yvon, et al., BLOOM: a 176b-parameter open-access multilingual language model, 2022.

[31] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al., GLM-130b: an open bilingual pre-trained model, 2022.

[32] M. Chen, J. Tworek, H. Jun, Q. Yuan, P.H. P.D. Oliveira, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., Evaluating large language models trained on code, 2021.

[33] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., ChatGPT for good? on opportunities and challenges of large language models for education, Learn. Individ. Differ. 103 (2023) 102274.

[34] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, 2023.

[35] M. Shanahan, Talking about large language models, 2023. 2212.03551

[36] Y. Dou, Y. Huang, X. Zhao, H. Zou, J. Shang, Y. Lu, X. Yang, J. Xiao, S. Peng, ShennongMGS: an LLM-based chinese medication guidance system, ACM Trans. Manage. Inf. Syst. 16 (2) (2025) 1–14.

[37] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D.C. Schmidt, A prompt pattern catalog to enhance prompt engineering with ChatGPT, (2023). arXiv preprint arXiv:2302.11382

[38] C. Wendler, V. Veselovsky, G. Monea, R. West, Do Lamas work in English? on the latent language of multilingual transformers, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 15366–15394.

[39] S. Papadopoulos, K. Datta, S. Madden, T. Mattson, The TileDB array data storage manager, Proc. VLDB Endow. 10 (4) (2016) 349–360.

[40] G. De Vito, A. Angelakis, F. Ferrucci, Online repository, 2024. https://github.com/gadevito/heliot.

[41] A. Flahault, M. Cadilhac, G. Thomas, Sample size calculation should be performed for design accuracy in diagnostic test studies, J. Clin. Epidemiol. 58 (8) (2005) 859–862.

[42] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46.

[43] Y. Collet, M. Kucherawy, Zstandard Compression and the Application/zstd Media Type, Technical Report, 2018.

[44] L. Kühnel, J. Schneider, I. Perrar, T. Adams, S. Moazemi, F. Prasser, U. Nöthlings, H. Fröhlich, J. Fluck, Synthetic data generation for a longitudinal cohort study–evaluation, method extension and reproduction of published data analysis results, Sci. Rep. 14 (1) (2024) 14412.

[45] M. Giuffrè, D.L. Shung, Harnessing the power of synthetic data in healthcare: innovation, application, and privacy, NPJ Digital Med. 6 (1) (2023) 186.

[46] M.A. Bruusgaard-Mouritsen, S. Nasser, L.H. Garvey, M.S. Krantz, C.A. Stone, Anaphylaxis to excipients in current clinical practice: evaluation and management, Immunol. Allergy Clin. 42 (2) (2022) 239–267.

[47] C.A. Stone, Jr, Y. Liu, M.V. Relling, M.S. Krantz, A.L. Pratt, A. Abreo, J.A. Hemler, E.J. Phillips, Immediate hypersensitivity to polyethylene glycols and polysorbates: more common than we have recognized, J. Allergy Clin. Immunol.: Pract. 7 (5) (2019) 1533–1540.

[48] M.H. Lee, D.P. Siewiorek, A. Smailagic, A. Bernardino, S. Bermúdez i Badia, Towards efficient annotations for a human-AI collaborative, clinical decision support system: a case study on physical stroke rehabilitation assessment, in: Proceedings of the 27th International Conference on Intelligent User Interfaces, 2022, pp. 4–14.

[49] C. Pais, J. Liu, R. Voigt, V. Gupta, E. Wade, M. Bayati, Large language models for preventing medication direction errors in online pharmacies, Nat. Med. 30 (2024) 1–9.

[50] F. Valente, S. Paredes, J. Henriques, T. Rocha, P. de Carvalho, J. Morais, Interpretability, personalization and reliability of a machine learning based clinical decision support system, Data Min. Knowl. Discov. 36 (3) (2022) 1140–1173.

[51] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al., Gemma 3 technical report, (2025) arXiv preprint arXiv:2503.19786.

[52] Anthropic, Claude 3.5 sonnet (version 3.5), 2024. [Online; accessed 15 November 2024], https://claude.ai/.

[53] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions, John Wiley & Sons, 2013.

[54] F. Zampetti, R. Kapur, M. Di Penta, S. Panichella, An empirical characterization of software bugs in open-source cyber–physical systems, J. Syst. Software 192 (2022) 111425.