Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 2 (36) *ISSN 2524-2296 (online)*

UDC 004.934:141.1 DOI: https://doi.org/10.30837/2522-9818.2026.2.173

Miriam Fandakova, Simon Ochotnicky, Andriy Kovalenko

INTERACTIVE PLATFORM FOR SUPPORTING CLINICAL DECISION-MAKING USING LARGE LANGUAGE MODELS (LLMs)

Current healthcare systems face increasing workload, fragmented communication, and documentation burden, which contributes to delays and diagnostic errors in early triage. The proposed solution is intended to improve consistency, speed, and standardization in early patient assessments overall. This study presents an interactive clinical decision support platform that operationalizes a workflow-constrained, two-step history-taking process to support symptom-based differential diagnosis in the pre-ambulatory phase of care. The system addresses three tasks: (T1) automated patient history elicitation via constrained dialogue (exactly two multiple-choice follow-up questions), (T2) formalization of symptom narratives into structured medical (Latinate) terminology for clinician-facing documentation, and (T3) generation of differential diagnosis recommendations under a fixed and clinically interpretable output schema. We compare a generic large language model baseline (GPT-4, zero-shot) with a domain-adapted model (Llama-3 fine-tuned using LoRA) under identical interaction, prompting, and formatting constraints. Experiments on 903 symptom–diagnosis records and a held-out set of 200 controlled vignettes show that domain adaptation yields a 10–12% macro-F1 improvement and approximately 15% higher Recall on complex cases, while producing more clinically discriminative and diagnostically relevant follow-up questions as assessed by two medical raters $(\kappa = 0.88)$. In a workflow-level evaluation, the platform reduced documentation time by 28% compared to standard intake and lowered the administrative effort required to prepare an initial clinician-facing summary for physician review. These results suggest that specialized, locally deployable fine-tuned models, combined with bounded interaction design, standardized output constraints, and workflow-oriented system integration, provide a secure, practical, and effective pathway for incorporating LLM-based decision support into routine pre-ambulatory clinical workflows while preserving safety, usability, and auditability in practice.

Keywords: clinical decision support; large language models; fine-tuning; LoRA; symptom-based diagnosis; medical AI.

Introduction

Currently, global healthcare systems are facing critical pressure driven by the rising prevalence of chronic diseases and a persistent shortage of specialized medical personnel [1, 2]. This adverse condition results in extreme clinician burnout, which directly correlates with an increased risk of diagnostic errors and prolonged waiting times [3, 4]. These systemic constraints are particularly visible in high-throughput outpatient settings, where clinicians must make rapid early judgments based on incomplete or unstructured symptom descriptions. Such early-stage clinical decision-making is inherently affected by both aleatory uncertainty arising from patient variability and epistemic uncertainty caused by incomplete or imprecise information. Prior work has shown that explicitly accounting for these uncertainty sources is critical for improving the reliability of decision-support systems, particularly when decisions must be made under constrained information conditions [5]. As a result, improving the efficiency and consistency of pre-visit information gathering has become an important target for both patient safety and operational resilience.

Analysis of modern publications

Within this context, Artificial Intelligence-based Clinical Decision Support Systems (CDSS) are increasingly explored as tools for streamlining initial triage and symptom analysis [6]. Traditional digital triage solutions typically rely on static questionnaires or rule-based branching logic. While such approaches can standardize data collection, systematic reviews report that the diagnostic and triage accuracy of online symptom checkers is variable and often below clinician performance, with heterogeneous evidence on safety and real-world effectiveness [7, 8]. These findings support the need for approaches that better handle natural-language variability and support clinically actionable structuring of pre-visit information [6, 7].

The advent of Large Language Models (LLMs) has initiated a new paradigm in natural language processing within clinical practice by enabling flexible interpretation and generation of free-text clinical narratives.

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 2 (36)

Generic models, such as GPT-4, demonstrate strong performance in medical benchmarks; however, their real-world deployment remains limited by a propensity for clinically irrelevant hallucinations and by insufficient sensitivity to local documentation practices and context [9, 10]. Recent evaluation work in healthcare emphasizes that beyond benchmark scores, LLM assessment must include safety- and workflow-relevant criteria, because these directly affect trust and deployment risk in safety-critical settings [10, 11].

Current research suggests that domain-specific fine-tuning on curated medical datasets can mitigate these deficiencies and enhance clinical usefulness [12–14]. Fine-tuning can improve task alignment and reinforce clinically appropriate output formats, which is particularly important when the system must produce concise diagnostic hypotheses and structured summaries rather than open-ended narratives [13, 14]. Parameter-efficient approaches such as LoRA additionally enable adaptation with lower computational requirements, supporting locally deployable clinical systems [15]. At the same time, practical deployment requires evaluation beyond headline benchmark scores, including robustness across heterogeneous symptom descriptions and consistency across disease categories [10, 11]. Taken together, existing work highlights a gap between benchmark-oriented model evaluations and workflow-constrained pre-ambulatory deployment, where reliability, bounded inquiry, and documentation-aligned outputs are essential.

In this paper, we present an interactive platform designed to assist clinicians in differential diagnosis during the pre-ambulatory phase. Unlike static questionnaires, our platform utilizes a multi-stage inquiry process: upon the input of initial symptoms at intake, the agent generates two targeted follow-up questions. This approach aims to simulate expert-led anamnesis, reduce ambiguity in the initial complaint, and prepare a structured clinical record prior to the clinician encounter. The primary contribution of this article is the introduction of the platform's architecture, followed by a comparative performance analysis of generic models versus models optimized through fine-tuning. In addition, we discuss practical considerations for integrating LLM-driven assistance into clinical workflows, with emphasis on reliability, output constraints, and deployment-relevant evaluation.

This study makes three primary contributions to the design of large language model–based clinical decision support systems. First, we propose a constrained, interaction-centered triage paradigm that reframes the role of LLMs from unconstrained conversational agents to bounded clinical reasoning components embedded in a realistic pre-ambulatory workflow. The proposed protocol enforces a fixed two-step anamnesis consisting of exactly two multiple-choice follow-up questions, followed by a structured clinician-facing synthesis. This design explicitly reflects time, cognitive, and safety constraints of real-world triage and differs fundamentally from prior approaches that rely on open-ended dialogue or single-pass diagnosis prediction.

Second, we demonstrate that domain-specific fine-tuning yields benefits beyond improved classification metrics by systematically improving the discriminative value of generated follow-up questions. Through supervised instruction tuning that jointly encodes question generation and diagnostic objectives, the fine-tuned model learns to prioritize high-information clinical discriminators under strict interaction limits. This finding provides empirical evidence that adaptation of LLMs can enhance the quality of clinical information elicitation itself, a dimension that is often overlooked in benchmark-driven evaluations.

Third, we introduce a modular, governance-oriented system architecture that supports secure data ingestion, privacy-preserving harmonization, LoRA-based fine-tuning, versioned model release, and controlled deployment of locally hosted models. By coupling workflow-level constraints with model adaptation and deployment governance, the proposed platform illustrates a practical pathway for integrating LLM-driven decision support into clinical environments while maintaining auditability, safety, and operational feasibility.

System architecture and design

The architecture of the platform was designed with a primary emphasis on modularity, data security, and scalability. The system is implemented as a suite of five collaborative microservices that collectively form a secure end-to-end pipeline for data collection, model training, and real-time inference. By separating ingestion, harmonisation, training, governance, and serving into dedicated components, the platform supports independent updates and reduces the risk that a failure in one module propagates across the entire system. This section describes the technical infrastructure (Figure 1) and the main data flow that enables continuous model improvement and controlled deployment.

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 2 (36) *ISSN 2524-2296 (online)*

Backend infrastructure and model training pipeline

At the core of the system is the infrastructure shown in Figure 1. The process starts at the Clinician UI, which is embedded in the hospital's Electronic Health Record (EHR) environment and enables authorized personnel to stream or upload structured clinical data using the FHIR (Fast Healthcare Interoperability Resources) standard [16]. The ingestion layer enforces access control and logging to ensure traceability of submissions and to support auditing in clinical environments. All incoming resources are persisted in a Secure Data Lake as de-identified JSON objects. Data at rest is protected using AES-256 encryption, while indexing and access are managed through a dedicated key management service to enable controlled key rotation and separation of privileges [17].

To ensure data quality and interoperability, a nightly ETL (Extract, Transform, Load) and harmonisation service is executed. This service validates schemas, normalizes laboratory units, maps clinical terminologies to a consistent representation, and tokenizes narrative notes to reduce variation caused by heterogeneous documentation practices. As an explicit privacy safeguard, when fewer than 250 patients are included in a cohort, the task applies calibrated Laplacian noise to enforce differential privacy and reduce re-identification risk in small-sample settings [18]. The harmonized data is then transferred to the Model Trainer module, where supervised fine-tuning is performed using Low-Rank Adaptation (LoRA) [19]. LoRA attaches trainable adapters to a frozen base checkpoint (Llama-3 in our implementation), enabling efficient domain adaptation with reduced computational and storage requirements compared to full fine-tuning.

The resulting adapter weights are packaged together with provenance metadata, including the dataset snapshot used for training, the harmonisation configuration, and evaluation artifacts required for release decisions. These outputs are versioned in the Model Registry, which provides a controlled mechanism for selecting approved model versions and supports rollback in case of regressions. Finally, the Conversation API loads the latest approved weights from the registry to serve real-time predictions for the chat widget, ensuring that inference is always tied to a specific governed model version. This closes the loop between data acquisition, harmonisation, training, and deployment, while maintaining clear separation between offline training workflows and online clinical use.

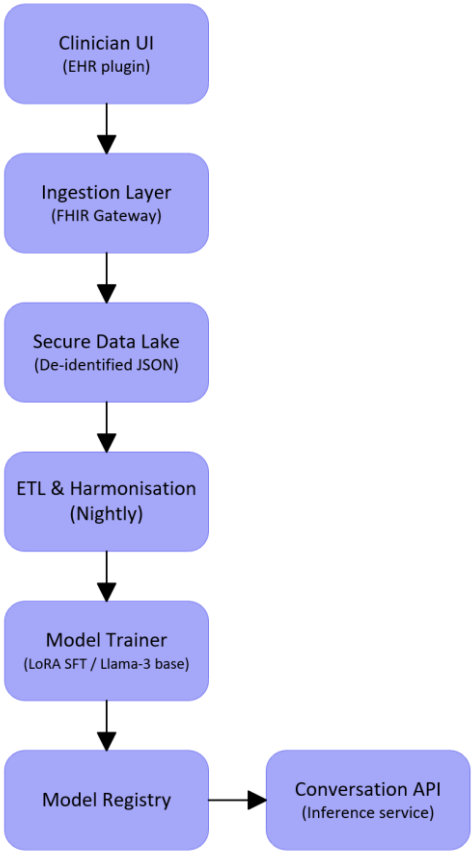

Fig. 1. Data-to-Model pipeline architecture. The diagram shows the data flow from clinician UI to deployment via the Conversation API

Figure 1 provides a detailed visualization of this end-to-end data-to-model pipeline. The process is initiated at the secure Clinician UI boundary, which employs the FHIR standard [16] to stream de-identified clinical data into the system. The Ingestion Layer serves as the first governance checkpoint, enforcing strict access controls and creating an immutable audit log before persisting data as encrypted JSON objects in the Secure Data Lake.

This raw data is then processed by the nightly ETL & Harmonisation Service, which performs three critical functions:

1) schema validation and normalization of units;

2) mapping of disparate clinical terminologies to a unified representation;

3) the application of differential privacy mechanisms (calibrated Laplacian noise) for small cohort sizes (<250 patients) to mathematically guarantee patient privacy [18].

The resulting harmonized dataset is consumed by the Model Trainer, where supervised fine-tuning of the base LLM (Llama-3) is performed using Low-Rank

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 2 (36)

Adaptation (LoRA) [19]. Unlike full fine-tuning, LoRA attaches trainable adapters to the frozen base model, which drastically reduces the number of trainable parameters and enables efficient, version-controlled storage of the resulting domain-adapted weights. These compact adapter weights, along with their associated training and evaluation metadata, are versioned and stored in the Model Registry. This registry acts as the system's "source of truth" for approved models, enabling governance, auditability, and the ability to perform instant rollbacks if a regression is detected. Finally, the Conversation API, which serves all live clinical interactions, is configured to load only the latest, governance-approved model version from this registry. This closed-loop design ensures that every patient interaction is handled by a vetted and traceable model instance, directly addressing key safety and deployment concerns in clinical AI [10, 11].

User interaction workflow

The infrastructure described above (Figure 1) provides the secure and governed foundation for online inference via the Conversation API, which exposes only approved model versions from the Model Registry. Building on this deployment layer, the platform implements a lightweight waiting-room interaction designed to standardize pre-ambulatory data collection while remaining efficient for patients. The workflow is summarized in Figure 2 and is implemented as a bounded, two-stage dialogue coordinated by two agents that invoke underlying LLMs (either generic or fine-tuned) rather than introducing additional independently trained models.

The interaction begins with Patient input, where the patient enters a free-text description of their main complaints. This input is sent to the serving layer through the Conversation API, which routes the request to the currently approved model version. The first stage is handled by Agent 1 (Questions generation). Agent 1 acts as an orchestration component responsible for selecting and structuring follow-up queries: given the initial symptom description, it calls the underlying LLM to produce exactly two targeted follow-up questions, each formulated as a four-option multiple-choice item. This design serves two purposes. First, the limited number of turns ensures predictable duration and reduces the risk of conversational drift. Second, the multiple-choice format lowers patient cognitive burden and yields standardized responses that can be reliably consumed by downstream components. In practice, the four-option responses also reduce ambiguity compared to open-ended replies and simplify downstream synthesis.

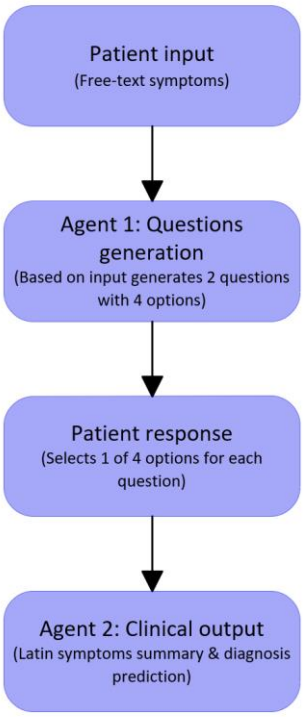

Fig. 2. User interaction workflow. The diagram shows a bounded two-agent dialogue from free-text symptoms to clinical output

After the patient answers the first question, the agent updates the dialogue state and triggers generation of the second question, conditioned on both the initial narrative and the first response. The patient then selects one of the four options for the second question. At this point, the workflow transitions to Agent 2 (Clinical output), which is responsible for synthesis rather than further elicitation.

Agent 2 aggregates the available information:

1) free-text symptoms;

2) the structured responses to both multiple-choice questions and invokes a (generic or fine-tuned) LLM to generate the final clinician-facing output.

Importantly, the agent logic enforces strict task boundaries (two questions only) and separates the elicitation objective from the synthesis objective, which improves controllability of the output.

The resulting Clinical output consists of two elements intended to support the physician during the pre-ambulatory phase:

1) a concise symptom summary expressed in professional medical (Latinate) terminology, formatted as a preliminary anamnesis/record stub;

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 2 (36) *ISSN 2524-2296 (online)*

2) a diagnosis recommendation to guide differential diagnosis.

Importantly, the agents in this workflow function as control logic that constrains the interaction (fixed number of questions, fixed answer format) and routes requests to the appropriate deployed model through the Conversation API. This preserves the governance and versioning guarantees of the backend pipeline while enabling an adaptive, patient-friendly front-end experience.

User interface implementation

The User Interface (UI) was designed with an emphasis on simplicity and accessibility, as the target audience includes patients of varying ages, health literacy levels, and acute health conditions. As shown in Figure 3, the interface evokes a conversational experience (chatbot style), which is generally perceived as more natural for describing symptoms than rigid form-based questionnaires. The minimalist layout reduces visual clutter and guides the patient toward a single primary task: providing a concise free-text description of their main complaint. The landing screen prominently communicates the application's purpose ("Patient just walked in…"), which helps establish trust and sets expectations before the automated triage begins.

In addition to free-text input, the UI includes a lightweight agent selection control that allows the patient (or assisting staff) to choose the most appropriate interaction pathway prior to submitting symptoms. This selection corresponds to a specialty-focused agent profile (e.g., dermatology for skin-related complaints), enabling the system to tailor the behavior of Agent 1 toward clinically relevant follow-up questions for the selected domain. For example, when a patient reports a rash or other cutaneous symptoms, selecting the dermatology-oriented agent biases the subsequent question generation toward discriminators such as lesion morphology, distribution, pruritus, or common triggers, improving the informativeness of the two-step dialogue while keeping the interaction bounded and easy to complete. Importantly, the agent selector is presented as a simple, clearly labeled option to avoid increasing cognitive load; if no specialty is chosen, the system defaults to a general triage configuration.

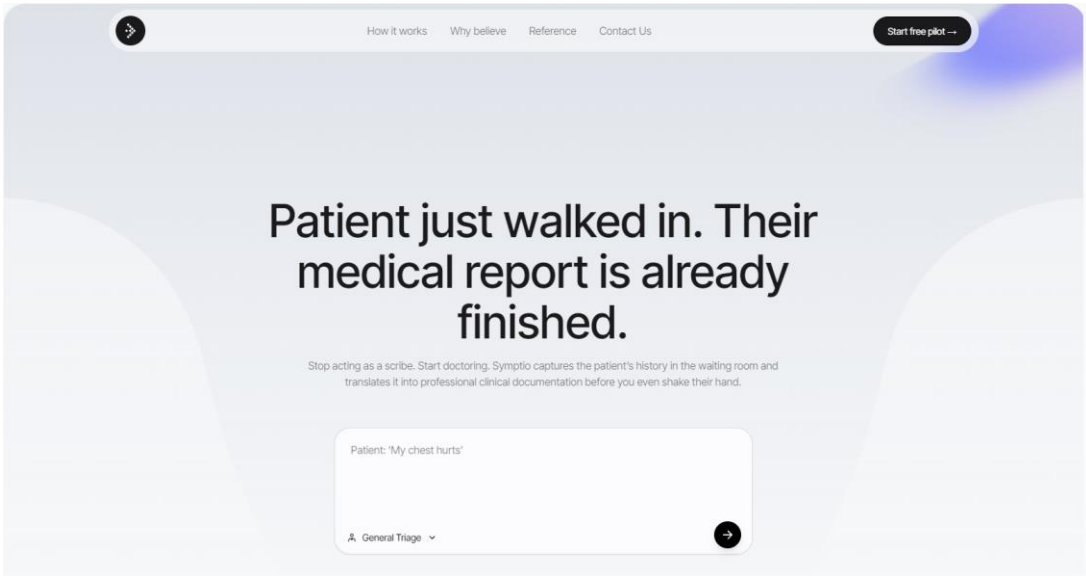

Fig. 3. Patients enter initial symptoms and may select a specialty-focused agent

From a usability perspective, the multiple-choice questions are displayed as clear, single-selection options, which supports rapid interaction in the waiting room and reduces the likelihood of incomplete or inconsistent answers. After completion, the patient-facing UI displays a confirmation message, thanking the patient for their input and informing them that their information is ready for the physician's review. Simultaneously, the platform transmits the processed data to the clinician-facing summary view within the clinical interface. In this secure environment, the physician can review the generated Latinate note stub and diagnosis suggestion alongside other relevant patient records.

Overall, the UI design supports a smooth waiting-room workflow by combining natural language symptom entry with structured, low-effort interactions, while the

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 2 (36)

optional specialty agent selection provides a mechanism for more context-aware questioning without exposing technical complexity to the patient.

Methodology and experimental organization

The objective of the experiment was to quantify the added value of domain-specific fine-tuning compared to high-performance generic models used in a zero-shot setting.

The methodology focuses on three key areas:

1) curated dataset preparation from clinical records;

2) a reproducible fine-tuning protocol;

3) a multi-criteria evaluation capturing both diagnostic performance and the quality of the interactive questioning component.

Formalization of the workflow-constrained diagnostic task

Let x denote the initial free-text symptom narrative provided by the patient. The platform implements a bounded interaction I consisting of exactly two multiple-choice questions q_1, q_2, each with four answer options, and the corresponding patient selections a_1, a_2. The diagnostic task is defined as a multi-class prediction over a fixed label set $Y = \{y_1, \ldots, y_K\}$ (here $K = 3$). Under a shared interaction specification, a model f maps the collected information to a clinician-facing output o and a diagnosis recommendation $\hat{y}$:

$$f : (x, q_1, a_1, q_2, a_2) \to (o, \hat{y}), \quad \hat{y} \in Y.$$

The interaction policy that generates questions can be expressed as:

$$q_n = g(x, q_{<n}, a_{<n}), \quad n \in \{1, 2\},$$

where g is constrained to produce exactly two questions and to follow a fixed four-option multiple-choice format.

This formulation explicitly separates:

1) elicitation (question generation) from

2) synthesis (clinical summary and diagnosis), enabling a controlled comparison between a generic zero-shot model and a domain-adapted model under identical constraints.

We test the hypotheses: H1: domain-specific fine-tuning improves diagnostic performance (macro-F1, Recall) under the bounded interaction protocol; H2: fine-tuning improves the discriminative relevance of generated follow-up questions as assessed by clinician raters; H3: the bounded interaction design supports workflow efficiency without requiring long free-form dialogue.

Dataset curation and preprocessing

For training and validation, we used the publicly available disease-diagnosis-dataset hosted on Hugging Face, which provides free-text symptom descriptions paired with diagnosis labels [20]. From the full corpus, we filtered samples belonging to three target conditions frequently encountered in acute and pre-ambulatory settings: diabetes $(n = 312)$, Chronic Obstructive Pulmonary Disease (COPD, $n = 289$), and persistent asthma $(n = 302)$. The raw dataset was subsequently cleaned and normalized (e.g., removal of duplicates and incomplete entries, whitespace and punctuation normalization, and label harmonization for synonymous disease names). After preprocessing, the dataset was downsampled to 903 samples to match the intended experimental scale while preserving the class distribution across the three conditions.

The preprocessing pipeline transformed the symptom descriptions into a task-oriented representation reflecting the intended platform use case: an initial symptom narrative followed by a short, bounded interactive step and a final clinician-facing output. Concretely, the dataset was converted into "Instruction–Response" pairs following instruction-tuning methodology [21].

The instruction/input corresponded to the initial patient complaint in free-text form, while the response/target contained:

1) two follow-up questions formatted for a constrained interaction (consistent with the platform design);

2) the final diagnosis label. The follow-up questions were generated using a controlled template-based procedure to ensure consistent formatting and to reduce spurious variability across samples.

To support reproducibility, records were normalized into a consistent schema prior to conversion (e.g., standardized handling of missing fields and consistent formatting of symptom mentions). The dataset was then partitioned into an 80% training set and a 20% validation set using a stratified split to preserve the class proportions and prevent distributional shift between splits

Model selection and fine-tuning protocol

We compared two distinct modeling approaches:

Generic baseline (zero-shot): We utilized GPT-4 (OpenAI) [21] as the reference model, queried via

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 2 (36) *ISSN 2524-2296 (online)*

a system prompt without additional training. The baseline therefore reflects a "plug-and-play" deployment scenario in which the model is expected to generalize to the triage workflow without domain adaptation.

Fine-tuned model: We selected Llama-3 [22] as the open-source base model and optimized it specifically for the platform's triage interaction and output requirements.

Fair comparison and prompting control

To ensure a fair comparison between the zero-shot baseline and the fine-tuned approach, both models were evaluated under the same interaction specification and output constraints.

Concretely, the same system-level instruction template was used to enforce:

1) generation of exactly two follow-up questions;

2) each question being presented in a four-option multiple-choice format;

3) a final clinician-facing output consisting of a concise Latinate symptom summary and a single diagnosis recommendation.

This experimental design minimizes the influence of prompt engineering and response formatting differences, such that observed performance changes can be attributed primarily to domain adaptation through fine-tuning.

The training protocol employed LoRA [19], which enables parameter-efficient fine-tuning by freezing the base model weights and inserting trainable low-rank matrices into the transformer layers. This approach is particularly suitable for iterative experimentation and controlled deployment, as only compact adapter weights need to be versioned and served. Adapters were trained using a context window of 256 tokens, a learning rate of 2×10^{-4}, and the AdamW optimizer [23]. The optimization included 2,000 warm-up steps and early stopping based on the macro-F1 score measured on the validation set. Training was performed on four NVIDIA A100-80GB accelerators, with convergence reached in approximately three hours.

Evaluation metrics

The platform's performance was evaluated across two complementary dimensions – diagnostic accuracy and interaction quality – to reflect the dual nature of the system (prediction + targeted elicitation).

Diagnostic accuracy

The model's ability to correctly predict the diagnosis was measured using Precision, Recall, and macro-averaged F1-score [24]. Evaluation was conducted on 200 synthetic patient vignettes supplemented with real laboratory values. Synthetic vignettes were used to systematically control symptom phrasing, coverage of key discriminators, and variability in presentation, while laboratory values increased clinical realism and helped test whether models can integrate structured signals with narrative text. All evaluated cases were held out from training and validation to avoid leakage.

Question relevance

Since a core feature of the platform is the generation of exactly two follow-up questions, their quality was assessed by two independent medical specialists. From a decision-theoretic perspective, the effectiveness of such a constrained interaction depends not on the number of questions, but on their discriminative importance with respect to the underlying diagnostic hypotheses. This aligns with prior research on importance analysis in decision-making systems, which demonstrates that a small number of high-impact factors can dominate the quality of the final decision outcome [25]. For each vignette, raters evaluated whether the generated questions were clinically relevant and whether they would plausibly contribute to narrowing the differential diagnosis in a pre-ambulatory setting. Inter-rater reliability was quantified using Cohen's kappa $(\kappa = 0.88)$ [26], indicating strong agreement and supporting the validity of the human evaluation protocol. In addition to relevance, the fixed multiple-choice format was considered in the assessment, as the platform explicitly constrains questions to four answer options to standardize patient responses and reduce completion burden.

Results and discussion

Experimental results confirm that domain-specific fine-tuning yields significant improvements in both diagnostic accuracy and clinical documentation efficiency compared to generic models. Importantly, the observed gains were achieved without a trade-off in output safety, as the factual error rate remained low and we did not observe an increase in hallucinations.

Diagnostic performance comparison

Comparative analysis demonstrated a consistent performance increase when using our domain-adapted model. As shown in Table 1, the Fine-Tuned model achieved higher scores across all key metrics (Recall,

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 2 (36)

F1-score) compared to the GPT-4 reference (Generic Baseline). These performance differences were statistically significant ($p < 0.05$), paired t-test) [27], indicating that the observed improvements were systematic rather than driven by isolated cases.

Table 1. *Performance comparison between generic and fine-tuned models across three diagnostic groups*

Diagnostic Group	Metric	Generic Baseline	Fine-Tuned	Δ (Improvement)
Diabetes	Recall	0.71	0.83	+0.12
	F1-score	0.73	0.83	+0.10
COPD	Recall	0.67	0.79	+0.12
	F1-score	0.69	0.80	+0.11
Asthma	Recall	0.68	0.80	+0.12
	F1-score	0.70	0.80	+0.10

The most substantial improvement (+19 percentage points in Recall) was observed in the detection of diabetic nephropathy. We attribute this gain to improved alignment with domain- and site-specific patterns, particularly in the interpretation of laboratory and biomarker cues (e.g., local conventions for reporting serum cystatin-C), which may be inconsistently expressed in generic model usage without domain adaptation. In COPD diagnosis, the fine-tuned model demonstrated an enhanced ability to recognize early signs of right heart strain in a manner consistent with GOLD-based severity cues [28]. For asthma, the detection of nocturnal symptoms improved significantly, suggesting better sensitivity to symptom variability patterns frequently found in patient narratives.

These findings align with a broader trend reported in the recent literature: constraining clinical LLM workflows and adapting models to task-specific data can improve reliability and reduce clinically unhelpful outputs relative to generic, unconstrained use [29]. In addition, studies benchmarking LLM-supported triage/diagnosis pipelines emphasize that workflow design (not only the raw model) materially affects clinical utility, particularly when outputs must be concise, structured, and decision-oriented [30].

Qualitative analysis of generated questions

A qualitative analysis revealed a fundamental difference in the nature of inquiry between the models. The Generic Baseline tended to ask questions that were semantically correct and empathetic but often clinically non-discriminatory (e.g., broad prompts such as "How are you feeling today?"). While such questions may improve conversational tone, they frequently failed to narrow the differential diagnosis efficiently under a bounded two-question interaction, and therefore contribute less under a Bayesian decision-making perspective [31].

In contrast, the Fine-Tuned model generated questions that directly tested differential diagnostic hypotheses. For example, for the input "chest pain", the model explicitly investigated discriminative features such as pain character ("Is the pain sharp or dull?") and radiation ("Does the pain radiate to the left shoulder or jaw?"). Across evaluated cases, the fine-tuned model more consistently prioritised high-yield discriminators over generic well-being probes. This behaviour can be interpreted as an implicit prioritisation of diagnostically important factors under uncertainty, rather than a purely conversational strategy. Similar prioritisation mechanisms have been formalised in importance-based decision models, where emphasising high-impact attributes leads to more reliable outcomes even when the overall information budget is limited [5, 25]. This observation is consistent with prior work indicating that off-the-shelf LLMs may produce conversationally plausible but diagnostically low-informative follow-up questions, motivating domain adaptation or workflow-level constraints for clinically useful inquiry [32].

Impact on clinical workflow

Integration of the platform into the triage process led to a 28% reduction in the time required to create an initial clinical record (from an average of 7.6 minutes to 5.5 minutes). A similar decrease was observed in keystroke count, directly reducing clinician cognitive load. In subjective evaluations using a Likert-type scale and standard analytical interpretation for ordinal response data [33], physicians reported a 2.1-point reduction in perceived administrative burden, suggesting that the benefit extends beyond objective efficiency measures.

These workflow-level improvements are directionally consistent with results reported for LLM-assisted documentation and ambient/digital scribe systems, where

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 2 (36) *ISSN 2524-2296 (online)*

reductions in documentation burden and improved clinical note efficiency have been observed in clinical or near-clinical evaluations [34]. While our system focuses specifically on the pre-ambulatory phase and uses a constrained two-question interaction rather than full-encounter transcription, the similarity in outcomes supports the broader conclusion that appropriately designed LLM integration can meaningfully reduce clerical workload and improve clinical throughput.

Limitations

Several limitations should be considered when interpreting these results. First, the evaluation was conducted on a restricted diagnostic scope (three disease groups), which may limit generalizability to broader triage settings. Second, diagnostic performance was assessed on 200 synthetic vignettes supplemented with real laboratory values; although this design enables controlled testing, it may not fully capture the variability of real-world patient narratives and comorbid presentations. Third, the interaction is intentionally constrained to two multiple-choice follow-up questions, which improves usability but may reduce performance for complex cases requiring longer anamnesis. Finally, the workflow-level impact metrics were collected in a specific clinical context; therefore, measured time savings and perceived burden reduction may differ across institutions, documentation practices, and patient populations.

Conclusion

The present work investigated the development and evaluation of an interactive platform for automated clinical triage, representing a workflow-oriented approach to supporting pre-ambulatory differential diagnosis. The primary objective was to determine whether high-capacity generic models used in a zero-shot setting or smaller, specialized models optimized through domain-specific fine-tuning are more suitable for this type of diagnostic decision support. In addition to accuracy, the study explicitly considered the quality of follow-up questioning as a core mechanism through which the platform reduces diagnostic uncertainty and prepares a clinician-ready record prior to examination.

Our findings support the hypothesis that for well-defined clinical tasks with constrained outputs, domain adaptation provides a measurable advantage over generalized model capabilities. While the generic model (GPT-4) demonstrated high linguistic competence and produced fluent responses, it was less consistent in generating discriminatory follow-up questions that effectively narrow the differential diagnosis under the fixed two-question interaction constraint. In contrast, the developed fine-tuned model – trained on curated medical data – improved diagnostic performance (macro F1-score) and more reliably generated targeted inquiries aligned with clinical reasoning patterns, thereby producing higher-yield information for subsequent synthesis into a structured clinical summary.

From a practical perspective, the platform demonstrates direct relevance to healthcare efficiency and usability. The observed 28% reduction in documentation time suggests that constrained, workflow-integrated automation can reduce clerical workload and support clinicians in high-throughput settings. Importantly, efficiency improvements were accompanied by positive subjective feedback, indicating that the system can reduce perceived administrative burden rather than merely shifting effort to a different step of the workflow. Together, these findings highlight that the clinical value of LLM-based systems is not solely determined by model scale, but by the interaction design, output constraints, and task-specific optimization strategy.

The proposed system architecture also illustrates a feasible path for governed deployment of locally hosted models in clinical environments. Serving approved model versions through a controlled API layer supports auditability, version tracking, and rollback, while reducing reliance on external cloud services for sensitive data processing. Although security and privacy depend on institutional implementation details, local deployment can provide a stronger foundation for data minimization and governance by keeping inference within the healthcare organization’s controlled infrastructure.

Future research will focus on extending the dataset and evaluation beyond the current diagnostic scope to cover a broader spectrum of conditions, including pediatric presentations and high-acuity emergency complaints. Further work will also include prospective validation in real clinical operation to assess long-term impact on diagnostic error rates, clinician workload, and patient experience, as well as monitoring robustness across diverse patient language, comorbidity profiles, and varying documentation practices. Overall, the platform represents a step toward hybrid medicine, where artificial intelligence does not replace the physician but becomes a bounded, reliable partner at the frontline of patient contact, supporting structured data collection and

accelerating early clinical decision-making. This is fully in line with the modern trend of precision or smart medicine [35].

Conflict of interest

The authors declare that they have no conflict of interest with respect to this research, including financial, personal, authorship, or other conflicts that could influence the research and its results presented in this article.

Funding

The study was conducted without financial support.

Data availability

The manuscript has no associated data.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technology in the creation of this work.

References

1. World Health Organization (2016), "Global strategy on human resources for health: Workforce 2030", *Geneva: WHO*, 64 p., available at: https://iris.who.int/handle/10665/250368
2. Zaitseva, E., Levashenko, V., Rabcan, J., Krsak, E. (2020), "Application of the structure function in the evaluation of the human factor in healthcare", *Symmetry*, Vol.12(1), 93 p. DOI: https://doi.org/10.3390/SYM12010093
3. Shanafelt, T. D. et al. (2010), "Burnout and medical errors among American surgeons", *Annals of Surgery*, Vol. 251(6), pp. 995-1000. DOI: https://doi.org/10.1097/SLA.0b013e3181bfdab3
4. Zaitseva, E., Levashenko, V. (2026), "Reliability engineering in healthcare: Opportunities and challenges", *Reliability Engineering and System Safety*, Vol. 267, 111933 p. DOI: https://doi.org/10.1016/j.ress.2025.111933
5. Zaitseva, E., Levashenko, V. (2025), "Reliability Analysis Based on Aleatory and Epistemic Uncertainty Using Binary Decision Diagrams", *International Journal of Intelligent Systems*, Vol. 1, 6471577 p. DOI: 10.1155/int/6471577
6. Sutton, R. T. et al. (2020), "An overview of clinical decision support systems: benefits, risks, and strategies for success", *NPJ Digital Medicine*, Vol. 3(17), pp. 1-10. DOI: https://doi.org/10.1038/s41746-020-0221-y
7. Chambers, D., Cantrell, A. J., Johnson, M. et al. (2019), "Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review", *BMJ Open*, Vol. 9(8): e027743. DOI: https://doi.org/10.1136/bmjopen-2018-027743
8. Wallace, W., Chan, C., Chidambaram, S. (2022), "The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review", *NPJ Digital Medicine*, Vol. 5(1):18. DOI: https://doi.org/10.1038/s41746-022-00667-w
9. Hajijama, S., Juneja, D., Nasa, P. (2024), "Large Language Model in Critical Care Medicine: Opportunities and Challenges", *Indian Journal of Critical Care Medicine*, Vol. 28(6), pp. 523–525. DOI: https://doi.org/10.5005/jp-journals-10071-24743
10. Asgari, E., Montaña-Brown, N., Dubois, M., et al. (2025), "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation", *NPJ Digital Medicine*, Vol. 8(274). DOI: https://doi.org/10.1038/s41746-025-01670-7
11. Chen, X., Xiang, J., Lu, S., Liu, Y., He, M., Shi, D. (2025), "Evaluating large language models and agents in healthcare: key challenges in clinical applications", *Intelligent Medicine*, Vol. 5(2), pp. 151–163. DOI: https://doi.org/10.1016/j.imed.2025.03.002
12. Singhal, K., et al. (2023), "Large language models encode clinical knowledge", *Nature*, Vol. 620, pp. 172-180. DOI: https://doi.org/10.1038/s41586-023-06291-2
13. Wang, G. et al. (2023), "ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation", *CC BY 4.0*. DOI: https://doi.org/10.48550/arXiv.2306.09968
14. Tran, H. et al. (2024), "BioInstruct: instruction tuning of large language models for biomedical natural language processing", *Journal of the American Medical Informatics Association (JAMIA)*, Vol. 31(9), pp. 1821–1832. DOI: https://doi.org/10.1093/jamia/ocae122
15. Hu, E. J. et al. (2021), "LoRA: Low-Rank Adaptation of Large Language Models", *arXiv preprint*, arXiv:2106.09685. DOI: https://doi.org/10.48550/arXiv.2106.09685
16. Bender, D., Sartipi, K. (2013), "HL7 FHIR: An Agile and RESTful approach to healthcare information exchange", *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 326–331. DOI: https://doi.org/10.1109/CBMS.2013.6627810
17. NIST (2001), "Advanced Encryption Standard (AES), "*Federal Information Processing Standards Publication (FIPS PUB)*" 197, 38 p. DOI: https://doi.org/10.6028/NIST.FIPS.197-upd1
18. Dwork, C. (2008), "Differential Privacy: A Survey of Results", *Theory and Applications of Models of Computation (TAMC)*, pp. 1–19. DOI: https://doi.org/10.1007/978-3-540-79228-4_1

19. Mu, O., Wang, W., Liu, W., et al. (2025), "Multimodal Large Language Model with LoRA Fine-Tuning for Multimodal Sentiment Analysis", *ACM Transactions on Intelligent Systems and Technology*, Vol. 16, No. 6, 139 p. DOI: https://doi.org/10.1145/3709147
20. Wei, J., Bosma, M., Zhao, V.Y. et al. (2021), "Finetuned Language Models Are Zero-Shot Learners", *ICLR*, DOI: https://doi.org/10.48550/arXiv.2109.01652
21. "GPT-4 Technical Report", *Computation and Language, arXiv preprint*, 2023. 100 p. DOI: https://doi.org/10.48550/arXiv.2303.08774
22. "Introducing Llama 3: The most capable openly available LLM to date", *Meta AI Blog*, 2024, available at: https://ai.meta.com/blog/meta-llama-3
23. Loshchilov, I., Hutter, F. (2019), "Decoupled Weight Decay Regularization", *International Conference on Learning Representations (ICLR)*, available at: https://openreview.net/forum?id=Bkg6RiCqY7
24. Semenov, S., Mozhaiev, O., Kuchuk, N., Mozhaiev, M., Tiulieniev, S., Gnusov, Yu., Yevstrat, D.,Chyrva, Y., Kuchuk, H. (2022), "Devising a procedure for defining the general criteria of abnormal behavior of a computer system based on the improved criterion of uniformity of input data samples", *Eastern-European Journal of Enterprise Technologies*, Vol. 6(4-120), pp. 40–49, DOI: https://doi.org/10.15587/1729-4061.2022.269128
25. Zaitseva, E., Rabcan, J., Levashenko, V., Kvassay, M. (2023), "Importance analysis of decision-making factors based on fuzzy decision trees", *Applied Soft Computing*, Vol. 134, 109988 p. DOI: https://doi.org/10.1016/j.asoc.2023.109988
26. Engeler, J., Stricker, D., Birrenbach, T. *et al.* (2025), "Designing and validating entrustable professional activities for emergency medicine: a stringent, construct-validity enhancing approach", *BMC Medical Education*, Vol. 25, 866 p. DOI: https://doi.org/10.1186/s12909-025-07449-4
27. Raschka, S. (2018), "Model evaluation, model selection, and algorithm selection in machine learning", *arXiv preprint*. DOI: https://doi.org/10.48550/arXiv.1811.12808
28. GOLD (2024), "Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease", *Global Initiative for Chronic Obstructive Lung Disease*, available at: https://goldcopd.org/2024-gold-report/
29. Dogru-Huzmeli, E., Moore-Vasram, S., Phadke, C. et al. (2025), "Evaluating ChatGPT's ability to simplify scientific abstracts for clinicians and the public", *Scientific Reports*, Vol. 15, 33466 p. DOI: https://doi.org/10.1038/s41598-025-11086-8
30. Gaber, F., Shaik, M., Allega, F. et al. (2025), "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis", *NPJ Digital Medicine*, Vol. 8, 263 p. DOI: https://doi.org/10.1038/s41746-025-01684-1
31. Sox, H. C. et al. (2013), *Medical Decision Making*, John Wiley & Sons, 328 p. DOI: https://doi.org/10.1002/9781118341544
32. Gatto, J., Seegmiller, P., Burdick, T. E. et al. (2025), "Follow-up Question Generation For Enhanced Patient-Provider Conversations", *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25222–25240. DOI: https://doi.org/10.18653/v1/2025.acl-long.1226
33. Sullivan, G. M., Artino Jr., A. R. (2013), "Analyzing and interpreting data from Likert-type scales", *Journal of Graduate Medical Education*, Vol. 5(4), pp. 541–542. DOI: https://doi.org/10.4300/JGME-5-4-18
34. Olson, K. D., Meeker, D., Troup, M. et al. (2025), "Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout", *JAMA Network Open*, Vol. 8(10):e2534976. DOI: https://doi.org/10.1001/jamanetworkopen.2025.34976
35. Zaitseva, E., Levashenko, V., Rabcan, J., Kvassay M. (2023), "A New Fuzzy-Based Classification Method for Use in Smart/Precision Medicine", *Bioengineering*, Vol. 10(7), 838 p. DOI: https://doi.org/10.3390/bioengineering10070838

Received (Надійшла) 06.02.2026
Accepted for publication (Прийнята до друку) 21.05.2026
Publication date (Дата публікації) 27.06.2026

Відомості про авторів / About the Authors

Фандакова Міріам – Жилінський університет, аспірантка кафедри інформатики, факультет управлінських наук та інформатики, Жиліна, Словаччина;

Miriam Fandakova – University of Zilina, Postgraduate Student of the Informatics Department, Faculty of Management Science and Informatics, Zilina, Slovakia;

e-mail: miriam.fandakova@st.fri.uniza.sk

ORCID ID: http://orcid.org/0009-0006-0763-3929

ISSN 2522-9818 (print) *Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 2 (36)*
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2026. No. 2 (36)*

Охотницкий Шимон – Жилінський університет, аспірант кафедри інформатики, факультет управлінських наук та інформатики, Жиліна, Словаччина;
Simon Ochotnicky – University of Zilina, Postgraduate Student of the Informatics Department, Faculty of Management Science and Informatics, Zilina, Slovakia;
e-mail: simon.ochotnicky@st.fri.uniza.sk
ORCID ID: http://orcid.org/0009-0002-2907-3687

Коваленко Андрій Анатолійович – доктор технічних наук, професор, Харківський національний університет радіоелектроніки, завідувач кафедри електронних обчислювальних машин, Харків, Україна;
Andriy Kovalenko – Doctor of Technical Sciences, Professor, Kharkiv National University of Radio Electronics, Head of the Electronic Computers Department, Kharkiv, Ukraine;
e-mail: andriy.kovalenko@nure.ua
ORCID ID: https://orcid.org/0000-0002-2817-9036
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=56423229200

ІНТЕРАКТИВНА ПЛАТФОРМА ПІДТРИМКИ КЛІНІЧНИХ РІШЕНЬ: АРХІТЕКТУРА ТА ПОРІВНЯЛЬНА ОЦІНКА ЗАГАЛЬНОЇ ТА ТОЧНО НАЛАШТОВЛЕНОЇ ДІАГНОСТИКИ ЗА СИМПТОМАМИ ЗА МЕТОДОЛОГІЄЮ (LLMS)

Сучасні системи охорони здоров'я стикаються зі зростаючим навантаженням, фрагментованою комунікацією та обтяжливістю документації, що призводить до затримок і діагностичних помилок на етапі первинної сортування пацієнтів. Запропоноване рішення покликане загалом підвищити узгодженість, швидкість та стандартизацію процесу первинної оцінки пацієнтів. У цьому дослідженні представлено інтерактивну платформу підтримки клінічних рішень, яка реалізує двоступеневий процес збору анамнезу з обмеженим робочим процесом для забезпечення диференціальної діагностики на основі симптомів на доамбулаторному етапі надання медичної допомоги. Система вирішує три завдання: (T1) автоматизований збір анамнезу пацієнта за допомогою обмеженого діалогу (точно два додаткових запитання з варіантами відповідей), (T2) формалізація описів симптомів у структуровану медичну (латинську) термінологію для документації, призначеної для клініцистів, та (T3) генерація рекомендацій щодо диференціальної діагностики за фіксованою та клінічно інтерпретованою схемою виводу. Ми порівнюємо базову загальну велику мовну модель (GPT-4, zero-shot) з моделлю, адаптованою до домену (Llama-3, налаштованою за допомогою LoRA), за однакових обмежень щодо взаємодії, підказок та форматування. Експерименти на 903 записах симптомів та діагнозів і виокремленому наборі з 200 контрольованих віньєток показують, що адаптація до домену забезпечує покращення макро-F1 на 10–12% та приблизно на 15% вищий показник Recall у складних випадках, водночас генеруючи клінічно більш дискримінативні та діагностично релевантні додаткові запитання, за оцінкою двох медичних експертів (κ = 0,88). У оцінці на рівні робочого процесу платформа скоротила час на документацію на 28% порівняно зі стандартним прийомом та зменшила адміністративні зусилля, необхідні для підготовки початкового резюме для лікаря. Ці результати свідчать про те, що спеціалізовані, налагоджені моделі, які можна розгортати локально, у поєднанні з обмеженим дизайном взаємодії, стандартизованими обмеженнями на вихідні дані та інтеграцією системи, орієнтованою на робочий процес, забезпечують безпечний, практичний та ефективний шлях для включення підтримки прийняття рішень на основі LLM у рутинні клінічні робочі процеси до амбулаторного лікування, зберігаючи при цьому безпеку, зручність використання та можливість аудиту на практиці.

Ключові слова: підтримка клінічних рішень; моделі з великою мовою програмування; точне налаштування; LoRA; діагностика на основі симптомів; медичний ШІ.

Бібліографічні описи / Bibliographic descriptions

Фандакова М., Охотницкий Ш., Коваленко А. А. Інтерактивна платформа підтримки клінічних рішень: архітектура та порівняльна оцінка загальної та точно налаштовленої діагностики за симптомами за методологією (LLMS). *Сучасний стан наукових досліджень та технологій в промисловості*. 2026. № 2 (36). С. 173–184. DOI: https://doi.org/10.30837/2522-9818.2026.2.173

Fandakova, M., Ochotnicky, S., Kovalenko, A. (2026), "Interactive platform for supporting clinical decision-making using large language models (LLMs)", *Innovative Technologies and Scientific Solutions for Industries*, No. 2 (36), P. 173–184. DOI: https://doi.org/10.30837/2522-9818.2026.2.173