

**Evaluation of AI Health Advisors: Alignment of
Global Health Chatbots with Kazakhstani Women’s
Health Clinical Protocols**

by

Akhmetzhan Dilnaz

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

May 2026

© Nazarbayev University 2026. All rights reserved.

Author
Department of Computer Science
May 2026

Certified by.....
Michael Lewis
Associate Professor, School of Engineering and Digital Sciences
Thesis Course Coordinator

Certified by.....
Zohaib Latif
Assistant Professor, School of Engineering and Digital Sciences
Thesis Course Coordinator

Accepted by
Dr. Elizabeth Arkhangelsky
Dean, School of Engineering and Digital Sciences

Evaluation of AI Health Advisors: Alignment of Global Health Chatbots with Kazakhstani Women’s Health Clinical Protocols

by

Akhmetzhan Dilnaz

Submitted to the Department of Computer Science
on May 2026, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

The rapid growth of conversational AI applications in Healthcare has significantly increased patient access to medical information. However, there is limited knowledge of how well these AI systems adhere to national clinical practice standards. It is even less known how well AI health advisors align with clinical protocols for women’s health (Obstetrics & Gynecology) in Kazakhstan.

A Protocol-Based and Cross-Linguistic Evaluation Framework was developed to address this challenge. An official corpus of 84 clinical protocols was collected, extracted from PDFs, transformed into structured text format, and processed through a pipeline that includes: tokenization/segmentation, normalization, machine translation, and biomedical entity extraction. A structured set of 15 clinically relevant questions was created to represent the decision-making scenarios contained within the protocol corpus. The four AI health advisory systems, ChatGPT-4o, Gemini-2.5-Flash, Perplexity Medical, and MedGemma-27B, were evaluated using controlled prompting conditions.

The evaluation included both entity-level metrics, such as precision, recall, omission rate, hallucination rate, and Jaccard similarity, in addition to an embedding-based semantic similarity analysis. Cross-linguistic robustness between English and Russian prompts was also evaluated. The results showed that the AI systems frequently generate semantically meaningful responses. However, entity-level alignment with protocol content was low, and omission rates are very high. It can be inferred that although semantically plausible outputs may seem clinically accurate, they do not indicate compliance with all elements of structured clinical guidelines. Differences in cross-linguistic performance were demonstrated among the different models. Therefore, the study indicated that alignment is not always consistent across languages.

The current thesis provides a protocol-based method to evaluate the quality of patient-facing AI systems in Healthcare. The study illustrates the necessity for de-

veloping jurisdiction-specific and clinically structured evaluation methods in order to assess the reliability and safety of AI health advisors for clinical use.

Thesis Course Coordinator: Michael Lewis

Title: Associate Professor, School of Engineering and Digital Sciences

Thesis Course Coordinator: Zohaib Latif

Title: Assistant Professor, School of Engineering and Digital Sciences

Acknowledgments

I want to thank my Thesis Supervisor Lisa Chalaguine for her support and feedback through out this project.

I also appreciate that the School of Engineering and Digital Sciences at Nazarbayev University provided a suitable academic and resource environment in which to complete this research.

Contents

1	Introduction	15
1.1	Motivation	15
1.2	Problem Statement	16
1.3	Research Objectives and Questions	17
1.3.1	Clinical Protocol Alignment Definition	18
1.4	Contributions	19
1.5	Thesis Structure	20
2	Related works	21
2.1	Evaluating Medical AI Models	22
2.2	Factuality, Evidence Grounding, and Hallucination	24
2.3	Structured Information Extraction from Clinical Guidelines	26
2.4	Cross-Lingual Representation	27
2.5	Medication Safety, Dosing Validity, and Risk Screening	29
2.6	Women’s Health Chatbots and Domain-Specific Findings	31
2.7	Synthesis and Remaining Gaps	34
3	Methodology	37
3.1	Research Design Overview	37
3.2	Data Acquisition	40
3.2.1	Source of Clinical Protocols	40
3.2.2	Domain Selection: Obstetrics and Gynecology	40
3.2.3	Dataset Composition	41

3.3	Data Preprocessing	42
3.3.1	PDF Conversion and Layout Preservation	42
3.3.2	Structural Cleaning and Section Segmentation	43
3.3.3	Cross-Lingual Normalization Framework	44
3.3.4	Biomedical Named Entity Recognition	45
3.3.5	Embedding Representation and Structural Validation	45
3.3.6	Preprocessing Challenges and Mitigation Strategies	46
3.4	Corpus Structural Validation	46
3.4.1	Section-Level Embedding Representation	47
3.4.2	Clustering Analysis	48
3.4.3	Topic Modeling and Cross-Lingual Alignment	48
3.4.4	Validation Interpretation	49
3.5	AI Health Advisor Evaluation Design	49
3.5.1	Reference Standard Construction	51
3.5.2	Development of the Evaluation Question Set	51
3.5.3	Entity-Level alignment	52
3.5.4	Semantic alignment	54
3.5.5	Cross-Lingual Robustness	55
3.6	Statistical Analysis Framework	55
3.6.1	Metric Score Aggregation	56
3.6.2	Distributional Analysis	56
3.6.3	Hallucination and Omission Rates	56
3.7	Stability Across Languages	57
3.8	Framework for Interpreting Results	57
4	Results	59
4.1	Overview of Evaluation Outcomes	59
4.1.1	Entity-Level Alignment	59
4.1.2	Semantic alignment	62
4.1.3	Cross-Lingual Stability	63

4.1.4	Protocol Adherence and Guideline Source Patterns	65
5	Conclusion	69
5.1	Main Findings and Interpretation	69
5.2	Cross-lingual and Contextual Implications	70
5.3	Relation to Research Questions	70
5.4	Methodological Contribution	71
5.5	Limitations	71
5.6	Final Conclusion and Future Work	72
6	Code and Data Availability	73
A	Appendix	75

List of Figures

3-1	Two-phase methodological framework for corpus construction and AI alignment evaluation.	39
4-1	Distribution of Jaccard similarity scores across evaluation questions for each model.	61
4-2	Comparison of hallucination and omission rates across evaluated AI systems.	61
4-3	Heatmap of semantic similarity scores between chatbot responses and protocol sections across evaluation questions.	63
4-4	Cross-lingual stability gap between English and Russian evaluation results.	65
4-5	Protocol adherence patterns across clinical domains and evaluated AI systems.	66
4-6	Distribution of guideline source references in AI-generated responses.	67

List of Tables

2.1	Limitations of Existing Medical AI Benchmarks for Women’s Health Chatbots	24
3.1	Clinical domain distribution within the OB-GYN protocol corpus. . .	41
3.2	Document conversion strategy and structural safeguards.	43
3.3	Cross-lingual normalization configuration.	45
3.4	Structure of the clinical evaluation question set	53
4.1	Entity-level alignment metrics averaged across 15 English-language clinical questions.	60
4.2	Semantic similarity between chatbot responses and protocol reference sections across 15 English-language evaluation questions.	62
4.3	Cross-lingual stability of semantic alignment scores between English and Russian prompts.	64
A.1	Evaluation Question Set	75

Chapter 1

Introduction

1.1 Motivation

The rapid integration of large language models (LLMs) in healthcare has altered the way patients can get access to medical information. Conversational AI bots have become an increasingly common means of providing users with brief and fast answers; they provide symptom assessments, reproductive health counseling, medication guidance, and recommendations [17]. AI bots provide a means for people to receive medical information beyond the traditional clinic environment, especially in regions where access to specialists is limited [33].

Clinical decision-making is guided by nationally established formalized clinical protocols as well as general biomedical knowledge. Clinical guidelines outline specific diagnostic criteria, eligibility requirements, contraindications, medication dosing limits, and standard care pathways [31]. Obstetrics and Gynecology clinical guidelines are among the most strictly defined and include gestational age thresholds, trimester-specific restrictions on safety, risk classification systems, and defined screening processes [4].

Furthermore, the issues described above are compounded in multilingual healthcare environments. All clinical guidelines for obstetrics and gynecology in Kazakhstan are officially published in Russian and Kazakh, however due to data constraints this study focuses on Russian language protocols. Additionally, all LLMs available today

were trained using predominantly English-language datasets. Thus, while developers of AI-based health assistants have created tools that can accurately respond to patient inquiries from a global biomedical perspective, the responses generated by these AI tools may not be consistent with local clinical procedural guidelines. Cross-linguistic variations in medical terminology, regional non-contextualization of reasoning, and discrepancies generated by differences in training data sets further contribute to this misalignment.

While medical LLMs can provide a great deal of information, achieving high levels of accuracy on these benchmarks does not indicate compliance with national clinical guidelines. Although current evaluation models emphasize the factual correctness, logical consistency, or general safety of an output generated by an AI system, very few evaluation models assess whether the generated output complies with specific jurisdictional clinical protocol requirements.

Therefore, a methodologically robust assessment framework is needed to assess the extent to which globally implemented conversational AI health systems provide recommendations that comply with official clinical guidelines in obstetrics and gynecology in Kazakhstan. The development of such a model is required to ensure the safe and culturally competent integration of AI-based health advisors into diverse and linguistically/culturally different healthcare ecosystems.

1.2 Problem Statement

Although the development of medical large language models has made rapid progress, there is still a missing systematic approach to measure how well patient-focused conversational AI systems adhere to clinical protocols that are developed at a national level [27]. Current assessment tools used to test large language models measure the quality of their biomedical knowledge and reasoning abilities; however, they do not assess whether the AI system adheres to country-specific clinical guidelines, safe administration of medications, or standardized diagnostic testing procedures. There is currently no widely established evaluation framework that systematically measures

alignment between patient-facing conversational AI systems and nationally approved clinical protocols in multilingual healthcare settings.

Clinical protocols in obstetrics and gynecology in Kazakhstan are officially published in Russian and detail exact procedure requirements. The protocols include gestational age limits, contraindications, eligibility criteria, and recommended follow-up time frames. In contrast to clinical protocols in Kazakhstan, global AI health advisory systems were primarily trained on English-language datasets and evaluated using a general biomedical inference as opposed to strictly adhering to localized protocol-based structures. As such, the structure between model performance based upon current assessments and safe integration of national healthcare infrastructure is inherently disparate.

Critical issues remain regarding the omission of safety-sensitive procedural requirements and the introduction of clinically significant inconsistencies as a result of cross-lingual semantic misalignment. Without an established framework for measuring the validity of these systems, the reliability of these systems for use in real-world applications in women’s health care cannot be validated.

1.3 Research Objectives and Questions

The purpose of this work is to develop an objective, cross-linguistic, and standardized methodology to evaluate the degree of compliance between AI-based health advisory systems and official clinical practice guidelines (CPGs) for women’s health in Kazakhstan.

To achieve this objective, the thesis addresses the following research questions:

RQ1: To what extent are the recommendations provided by international conversational AI systems compliant with the national obstetric/gynecologic CPGs of Kazakhstan?

RQ2: How can structured clinical information extraction and formalization of protocol logic be used to translate national clinical practice guidelines into a machine-readable format to enable the comparison of various guideline recommendations using

rigorous comparative methods?

RQ3: How can cross-linguistic semantic similarity methods be used to compare English-generated chatbot responses to clinical texts written in Russian?

RQ4: What types of differences are most prevalent in the recommendations made by AI systems? What types of errors are made regarding omission of clinical information, misidentification of clinical findings, violation of medication safety principles, or recommendations that violate existing clinical practice guidelines?

Through the systematic investigation of these research questions, the study aims to improve the quality of evaluations of AI applications for women’s health from being primarily evaluative in nature, to being more rigorously, safely, and nationally contextually based.

1.3.1 Clinical Protocol Alignment Definition

Alignment is defined as the extent to which an AI health advisor produces recommendations that have equal content, decision-making, and safety constraints to those stated in the officially accepted national clinical guidelines for women’s health. Alignment encompasses the degree to which the recommendations embody the specific procedural, pharmacological, and diagnostic criteria specified in the Kazakhstani clinical guidelines for women’s health.

The idea of Alignment will be defined through two interrelated concepts. Entity Level Alignment defines the degree to which the recommendation generated by the AI Health Advisor contains the same clinically relevant components as the referenced clinical guideline portion (i.e., medications, diagnoses, procedures, contraindications, and monitoring). Semantic Alignment defines the degree to which the meaning of the recommendation aligns with the intent of the clinical logic behind the protocol, regardless of how the recommendation was phrased.

The view of alignment will also include three additional layers. Regulatory alignment consists of compliance with nationally recognized clinical guidelines and jurisdiction-specific standards of care. Procedural alignment involves maintaining the diagnostic and treatment workflow, including contraindications, sequential decision steps, and

follow-up requirements. Semantic alignment involves retaining the clinical meaning of the protocol, even if the responses were worded differently or in different languages.

Both entity and semantic levels are necessary since a response could be generally semantically correct but lack key details regarding the protocol. Conversely, a response could generate many entities related to a particular condition yet fail to replicate the intended treatment pathway and decision process of the protocol. As such, protocol alignment in this study is considered to be a multi-faceted construct incorporating regulatory alignment, procedural alignment, semantic alignment, and language robustness.

1.4 Contributions

This thesis makes the following contributions:

1. Protocol Grounded Evaluation Framework. The thesis introduces assessment tool to measure how well global health AI systems follow the official Kazakhstani clinical guidelines for obstetric and gynecological care.

2. Structured Representation of Clinical Guidelines. This research outlines a methodology to extract the necessary information (entities, temporal constraints, medication rules, and decision-making logic) from Russian-language clinical guidelines, and transform them into a format that can be processed by computational systems.

3. Cross-Lingual Alignment Methodology. The thesis integrates semantic similarity modeling to enable systematic comparison between English language chatbot outputs and Russian clinical texts, addressing semantic drift and terminology mismatch.

4. Safety and Medication Verification Layer. As part of the overall validation of the system, a dedicated validation module has been included to check the AI-generated recommendations contain any violations of contraindications, dosing discrepancies, omitted safety requirements, and/or incorrect information relevant to a particular gestation period.

5. Quantitative Alignment Metrics. Defines four clinically interpretable scores that represent the degree to which a health AI system follows the clinical guidelines related to obstetrics and gynecology: protocol coverage; reasoning fidelity; medication safety compliance; and guideline consistency. Including omission and hallucination rates as indicators of missing or unsupported protocol content.

6. Empirical Assessment of Global Health Chatbots. Finally, the framework was applied to analyze the performance of several available health AI systems to determine their level of adherence to the clinical guidelines for the care of women in Kazakhstan.

1.5 Thesis Structure

The remainder of this thesis work is presented in the following order. In Chapter 2, a review of research on medical AI evaluation, the assessment of medical AI factual accuracy and hallucinations, the extraction of structured information from clinical guidelines, the application of cross-lingual clinical natural language processing (NLP), the verification of the safety of medications, and the development of women’s health chatbots is provided. Chapter 3 describes the methodology used to complete this thesis, which includes the development of the corpus, data preprocessing, structural validation, the design of the chatbot evaluation, and statistical analysis. Chapter 4 reports the results of the evaluation of the Alignment of the AI health advisors. Chapter 5 provides an explanation of the implications of the findings, how they relate to the healthcare environment, and how they compare with the findings of prior studies. Finally, Chapter 6 summarizes the primary contributions, outlines the limitations of the study, and identifies possible avenues for future research.

Chapter 2

Related works

In recent years, there has been a significant increase in the use of large language models and conversational health chatbots within clinical environments [29]. Research demonstrates that even though these systems have high-quality natural language processing (NLP) performance, they do not necessarily provide reliable clinical information.

Women’s health clinical guidelines define structured criteria to guide clinical decisions. Therefore, ensuring that AI-generated advice corresponds to the clinical guidelines established at the federal and/or national level is critical to safely utilizing AI systems in healthcare environments.

The focus of this problem is specifically relevant in Kazakhstan, where the national guidelines in obstetrics and gynecology (OB-GYN) are written in Russian, whereas most of the frontier language models, evaluation tools, and benchmarks are in English. For example, existing benchmark suites like PubMedQA and MedMCQA measure biomedical reasoning in English only [13, 23]. In addition to measuring biomedical reasoning, the multi-dimensional frameworks such as MedHELM also measure general clinical performance, which does not include adherence to national OB-GYN protocols [3]. Studies in multilingual evaluation have shown that performance degrades in cross-lingual environments, increasing the risk of semantic drift and terminology mismatch [21, 5].

The limitations in these studies highlight the need for a structured, cross-lingual,

and clinically grounded evaluation tool that assesses how well global health chatbots align with Kazakhstani OB-GYN protocols. A clinically-grounded evaluation framework would combine multilingual semantic alignment [9, 24]; clinical entity extraction [18]; protocol logic modeling; and evidence-based comparison with authoritative guideline documents.

2.1 Evaluating Medical AI Models

Medical question-answering has its own traditional benchmarks, but they also have some major shortcomings that limit them from being effective or useful for a real-world clinical safety assessment. Jin et al. [13] created PubMedQA to test biomedical reasoning skills using scientific abstracts. Pal et al. [23] built MedMCQA – a massive-scale multi-subject examination dataset. Even though all these benchmarks can measure both factual recall and diagnostic reasoning, they mostly examine the isolated knowledge rather than in the context of a whole patient, including contraindications, comorbidities, and longitudinal care pathways.

As most of these benchmarks are based on multiple-choice questions, this creates constraints on the model to only choose the pre-defined answer option. The constrained evaluation setting has been shown to underestimate generative errors (e.g., hallucinations) when free-text clinical advice is given. Therefore, achieving a high score on the benchmarks does not necessarily mean the model will safely perform in the real-world application.

The newest frameworks attempt to address these shortfalls. Bedi et al. [3] developed MedHELM as a comprehensive framework that includes clinician-reviewed tasks. This framework is still mainly focused on the English language and is not specifically designed to evaluate how well the models adhere to the national clinical protocols.

Regulatory and Ethical Expectations: Safety, Accountability, and Risk Classification

The World Health Organization has highlighted the potential for "clinically plausible" yet dangerous AI-generated outputs, along with both the misattribution of and lack of supporting evidence regarding such outputs [32]. The European Union's AI Act includes most medical AI systems (including symptom-based decision-support tools) as "high-risk" applications that require thorough documentation, formalized risk management strategies, and conformance to clinical standards [8].

Regulatory expectations are especially relevant to women's healthcare. Obstetrics and gynecology protocols describe trimester-specific contraindications, structured risk-management pathways, age-based eligibility criteria, red-flag symptoms, and detailed medication guidelines. Therefore, evaluation frameworks must evaluate not only the general medical accuracy of these AI systems, but also their adherence to national clinical guidelines and other procedural requirements that are critical to patient safety.

Implications for Evaluating Women's Health Chatbots

Although substantial progress has been made in evaluating medical AI systems, important limitations remain for patient-facing conversational agents in women's health. These limitations and their implications for safe deployment in Kazakhstan are summarized in Table 2.1.

These structural gaps demonstrate that benchmark accuracy alone is insufficient for evaluating the safety and reliability of women's health chatbots in multilingual, protocol specific healthcare environments.

Table 2.1: Limitations of Existing Medical AI Benchmarks for Women’s Health Chatbots

Limitation	What Is Not Evaluated	Implication for Women’s Health
Protocol-agnostic benchmarks	Alignment with national clinical guidelines	Risk of recommendations inconsistent with Kazakhstani OBGYN standards
Exam-style multiple choice	Sequential workflows and conditional logic	Failure to capture gestational-age rules and risk stratification
English-only evaluation	Cross-lingual semantic alignment	Terminology drift between EN–RU protocols
Limited safety validation	Medication contraindications and dosing constraints	Potentially unsafe trimester-specific prescribing
No cultural adaptation metrics	Communication norms and patient expectations	Reduced trust and misinterpretation in local context

2.2 Factuality, Evidence Grounding, and Hallucination

Unsupported statements (hallucinations) can cause serious problems in healthcare, leading to improper triage responses, incorrect medication directions, or misinterpretation of symptoms. This issue is particularly important in women’s health, where clinical decisions depend on correct identification of pregnancy timing, contraindications, proper doses, and structured risk assessment. Hallucinations in medical LLMs also include factual inaccuracies, such as incorrect drug doses or screening intervals, and logical failures where reasoning diverges from established clinical pathways. The most dangerous form occurs when statements appear consistent with guidelines but lack any real connection to them, or contradict national standards due to globally trained data lacking regional context. Previous studies show that neural text generation systems often produce outputs unsupported by source evidence, with knowledge gaps, overgeneralization, and distributional extrapolation identified as key causes [16]. In medical settings, the need for strict numerical thresholds, temporal constraints, and

condition-specific logic amplifies the risk that even small deviations may produce unsafe recommendations, highlighting the need for evaluation frameworks that require evidence grounding and safety validation.

Grounding via Retrieval and Provenance

As an effort to eliminate unsupported statements, retrieval-based generation systems such as Retrieval-Augmented Generation have recently been introduced [19]. These systems retrieve passages from trusted sources (clinical guidelines, drug databases, textbooks, etc.) and produce responses conditioned on that external evidence, with the goal of improving factual accuracy and traceability. The effectiveness of these systems depends on retrieval quality. When relevant clinical guideline passages are not retrieved, the generated response may rely on incomplete evidence or contain inaccuracies. In women’s health protocols, retrieval errors may have serious consequences. Failure to recognize gestational thresholds, contraindication rules, or medication restrictions may lead to clinically dangerous recommendations. This underscores the importance of well-structured, domain-specific guideline corpora and careful evaluation of supporting evidence, rather than relying solely on the retrieval process.

Automated Factuality Metrics and Consistency Checks

The fact-checking models like FactCC [16] and other entailment based systems were originally intended to identify contradictions in source articles versus the generated summary; however, they have provided medical evaluation with some useful applications. These uses include evaluating whether the generated recommendations match the language of guideline sentences, identifying swapped entities (e.g., gestational week vs. cancer stage), identifying numeric discrepancies and identifying contradictory or unsupported statements.

2.3 Structured Information Extraction from Clinical Guidelines

Guidelines in Obstetrics and Gynecology are encoded with extensive domain specific knowledge presented in three forms: (1) as narrative text, (2) as tabular data, and (3) as conditional decision-making pathways. For protocol aligned recommendations to be generated by these guidelines they must be converted from their current unstructured representation of clinical conditions, numerical threshold values, contraindications, and sequential decision making logic to structured representations that can be interpreted by machines. Research in clinical natural language processing has demonstrated that named entity recognition and relational extraction processes can be used to extract structured knowledge representations from unstructured clinical documents [34, 18]. These processes have established a technical basis for converting the content of guidelines to machine-interpretable format which could then be systematically compared against chatbot-generated responses.

Clinical Named Entity Recognition and Relation Extraction

Clinical Named Entity Recognition (NER) and Relation Extraction (RE) are the building blocks of Structured Guideline Modelling. Early efforts, including that of Wu et al. [34], used deep learning techniques (for example RNNs and CNNs), for disease identification, procedure identification, drug identification, and laboratory concept identification from clinical notes. The use of transformer-based models has shown that domain specific pre-training (for example BioBERT [18]), produces significant improvements in the recognition of biomedical terminology and reduces ambiguity and improves the detection of the boundaries of entities.

The maternal morbidity extraction effort of Giraldo-Forero et al. [11] is an illustration of how these techniques can be applied to Women’s Health. Their domain specific corpus and annotation scheme, demonstrates how the obstetric concepts of symptoms, signs, risk factors, and time qualifiers can be represented as structured categories.

Structured elements that can be generated for a guideline-based evaluation, include clinical conditions and eligibility criteria, diagnostic and screening steps, medication name and dosage unit, temporal constraints (for example gestational week) and contraindications through the use of NER and RE.

Modeling Protocol Logic and Decision Pathways

Clinical Guidelines include a multitude of rules that are often applied sequentially, but may also require some type of conditional logic. Therefore, screening intervals, sequential treatment pathways and conditional eligibility criteria must be represented formally, as opposed to being considered discrete facts.

The Drug-Drug Interaction Extraction (DDI) shared task [25, 26] demonstrated how drug-drug interactions and causal relationships could be identified through the use of graph-based representations and sequence modeling methods.

These methodologies can be extended to women’s health protocols, which often involve clinical logic that is comprised of a series of conditional statements, including gestational age dependent actions, combined symptom criteria for the escalation of diagnostic procedures, trimester-specific contraindications and predefined follow-up intervals. Representing these dependencies in structured form allows evaluation frameworks to assess whether chatbot-generated responses accurately reflect the appropriate sequence of clinical reasoning as opposed to simply generating an isolated fact.

Therefore, representing clinical protocol logic as a set of conditional and temporal rule structures provides systematic reference points with which chatbot recommendations can be assessed for completeness, consistency and safety.

2.4 Cross-Lingual Representation

A major limitation to assessing health chatbots in the Kazakhstani environment is the disparity in the languages used by both globally developed large language models and national clinical standards. The majority of the most successful models work best

in English while Kazakhstan’s obstetric and gynecological protocol publications are largely written in Russian. Therefore a significant risk exists of conceptual loss due to semantic drift, or mistranslation of key terms; a lack of proper alignment of clinically relevant ideas; and culturally irrelevant advice. Cross-lingual NLP research provides many resources to address this gap but medical application has been underdeveloped.

Challenges in Multilingual Clinical Text Processing

The processing of clinical texts in Russian faces challenges unique to this language. The Cyrillic script used in both languages has variability in transliteration, which makes tokenizing more difficult due to variations in medical abbreviations and jargon. Both languages also have complex inflectional patterns that affect detection of entity boundaries. There is limited to no publicly available clinical NER/RE training data for RU. Studies by Copara et al. regarding French BioNER [6] demonstrated the need for domain-specific tokenizers, lexical resources, and annotation guidelines tailored to local healthcare workflows. These findings apply to Russian contexts, where both the clinical and linguistic environments differ significantly from English biomedical models.

Cross-Lingual Embedding Models

In order to compare output from chatbots with protocol text, there is a need for cross-lingual representations of semantics. **XLM-R** [5] utilizes sub-word level representations to achieve strong cross-lingual token representations for over 100 languages. **LaBSE** [9] produces language-agnostic sentence embeddings and has performed well in low-resource languages. **Multilingual Sentence-BERT** [24] allows for an efficient semantic similarity comparison of sentences written in different languages.

Using these models enables aligning Russian guideline sentences and English chatbot responses into a single vector space to enable alignment scoring. However, the use of models based on general web corpora do not guarantee performance regarding the performance on specialized obstetrics/gynecology terminology. The domain shift may also result in incorrect nearest-neighbors, or lost medically relevant nuances.

Benchmarking and Evaluation of Cross-Lingual Understanding

The Multilingual Text Embedding Benchmark (MTEB) [21] offers an empirical framework to evaluate cross-lingual semantic similarity of numerous languages, although it is not focused specifically on medical applications, but demonstrates a common problem in embedding models, i.e., they are generally weak when processing low-resource languages, especially in domain-specific technical terminology, or long/complex instructional text.

Therefore, due to the fact that there exists currently no Kazakhstani medical language specific model, this will require the combination of cross-lingual embedding with structural guidelines for extracting guidelines, rule based validation and expert reviews. Thus, cross-lingual representations and matching are two critical components of the proposed evaluation framework which enable a valid comparison of chatbot response to English language women’s health protocol, to Russian language women’s health protocols.

2.5 Medication Safety, Dosing Validity, and Risk Screening

Medication recommendations represent one of the highest-risk areas for clinical chatbots, especially in obstetrics and gynecology where treatment eligibility depends on gestational age, comorbidities, and national prescribing rules. Global LLMs rarely account for trimester-specific contraindications, leading to potentially inappropriate advice.

Medication Extraction and Normalization

The work of systems like MedXN [28] illustrates how hybrid pipeline systems employing lexical resources, regular expression matching, and machine learning models can accurately recognize drug mentions and map them to standardized concepts (e.g., RxNorm). Such systems illustrate how unstructured medication data in narrative

format can be converted to structured representations; this allows comparison of the chatbot’s output to formalized protocol requirements. Similar normalization within a Kazakhstani context will allow chatbot generated responses to be compared against approved national medication lists, as opposed to those of a different country.

Adverse Event Detection and Safety Constraints

The safety-oriented models are going beyond just extracting information. They will identify dangerous interaction, contraindication and incorrect use of medication. The work on the detection of adverse drug events (ADE) using distributional semantics [12] and neural relation model [25] shows how symptom-drug-temporal qualifier combinations can potentially lead to harm. These methods are relevant to the field of obstetrics where a medication may be acceptable during one trimester of pregnancy but is unacceptable during another, and there may be additional required screening for risks (e.g. hypertension/glucose testing) prior to initiating treatment. A chatbot that does not follow this process may have provided "correct" answers but not safe clinical answers.

Drug Knowledge Bases and Local Formularies

The drug databases that are commonly used (such as DrugBank Wishart2018DrugBank) contain much of the relevant information regarding how drugs work, contraindications, and adverse effects; however, they represent the global drug market and do not provide information specific to drug use within each country. It will be necessary for the developers of the chatbots to incorporate the regional pharmacy formulary information in order to assure that recommendations provided by the chatbot can be fulfilled within each country. If the developers fail to align the drug database information with the local medical and clinical practice standards, the chatbot may suggest medications that are either not available locally or have an inappropriate dosage range for use in the patient population in question. This is especially true for medications related to women’s health care in which the treatment decision may need to take into consideration the patient’s current pregnancy stage, if she is currently breastfeeding,

and her reproductive risk factors.

Implications for Structuring Kazakhstani Women’s Health Protocols

The current Kazakhstani obstetrics and gynecology protocols are available as narrative guidance, diagnostic pathways, and medication rules that have been developed in Russian language. In order to perform an objective and systematic evaluation of a chatbot in the field of women’s health in Kazakhstan, the guidance contained within these documents needs to be converted into a structured and machine-readable format. Other recent studies show how clinical entities and relations can be extracted from unstructured medical texts and then formalized as a knowledge graph (or a graph based representation) in a way that is appropriate for computational reasoning [1]. Structured modeling allows the precise identification of matches between the output of a chatbot and the requirements of a protocol, provides an effective method for identifying omitted or hallucinated steps during patient care, and also permits the comparison of the semantics of content across Russian and English language protocols. Thus, a formal representation of guidelines is a critical part of a framework for evaluating chatbots about women’s health using a protocol-based methodology in Kazakhstan.

2.6 Women’s Health Chatbots and Domain-Specific Findings

Women’s health is a unique area in terms of clinical, cultural, and language issues; therefore, it is an important arena for assessing the trustworthiness of conversational AI systems. While the recent systematic evidence suggests that conversational AI chatbots can be effective in improving the amount of information available about women’s health, several concerns have persisted relative to the fullness of clinical information provided by these systems, the accuracy of context in which they provide this information, and how well they adhere to established guidelines [14].

When considering women’s health chatbots, users consider the accuracy and reliability of the information that these systems provide as being more important than ease-of-use or cost considerations [30]. The above-mentioned findings point out a structural disconnect: while conversational systems can make access to health care more accessible, the degree to which they align with formal clinical standards is typically not assessed. Therefore, in multilingual environments, like Kazakhstan, where obstetrics/gynecology protocols are developed in Russian, assessment needs to go beyond usability and user engagement to assess concordance with protocols and culturally sensitive clinical decision-making processes.

Clinical Benefits and Persistent Accuracy Gaps

Systematic reviews show that women’s health chatbots enhance a woman’s knowledge, self-efficacy, and psychological outcomes [14]. Nevertheless, there are also drawbacks to these benefits; many are limited by their reliance on narrow, rule based dialogue or older model based approaches, few have undergone extensive clinical validation, and there is little evidence of structured safety audits of formal clinical guidelines.

The evaluations of BSE and breast cancer education chatbots illustrate this tension. For example, studies like Nayyar [22] reported that AI generated responses were typically clear and superficially consistent with the most current, international guidance, however they were also typically incomplete, oversimplified, or did not address the nuances of risk factors and age dependent recommendations. These results demonstrate the difficulty of assuring that models generate decision rationales and explanations for decisions made in the context of a clinical workflow that accurately represent the safety critical standards.

Sexual and Reproductive Health Interventions

The domain of sexual and reproductive health has illustrated the potential as well as the risks associated with using conversational AI. A systematic review of sexual health chatbots found a modest increase in condom use intentions, HIV knowledge and sexual health literacy that was enhanced when the systems had personalized or

motivational features [10]. There is little to no evidence demonstrating long-term behavioral changes or formal clinical validation [14].

The accuracy of these tools has been inconsistent and audits by experts comparing the performance of these tools to national guidelines do not occur often. Therefore, their applicability to non-Western cultures (such as Kazakhstan) may be questionable due to differences in STI protocols and standards for providing contraceptive counseling compared to the English language training data used to train them and the risk of cross-lingual misalignment.

User Expectations: Accuracy, Privacy, and Cultural Fit

Wang et al. [30] conducted a discrete choice experiment to estimate women’s willingness to pay for different chatbot features. This study indicates a need for systems that do not only provide clear communication but also adhere to regional standards and guidelines. Therefore, if a generic message regarding this type of information is provided through a health chatbot, it may confuse the user or compromise clinical safety.

The Missing Element: Protocol Alignment

Despite an increasing number of researchers examining the use of women’s health chatbots, there is currently no research on whether the recommendations generated by artificial intelligence (AI) align with established national clinical guidelines. Most of the current literature examines the usability of chatbots, the behavioral outcomes of using them, or the general accuracy of their recommendations; however, none compares chatbot recommendations to established clinical standards. Specifically, there is no research evaluating the language barrier created when national clinical standards are developed in native languages (primarily Russian in the case of Kazakhstan), while most chatbots are used in English. As a result, important elements of national clinical standards have not been systematically tested, including gestational age limits, medications to avoid, and emergency pathways. The lack of a common cross-lingual evaluation framework is also evident. Because conversational AI systems are increas-

ingly proposed as tools to support real-world patient interactions, it is critical that they generate recommendations consistent with national clinical standards.

2.7 Synthesis and Remaining Gaps

Progress in clinical NLP enables structured extraction of entities, relations, and protocol logic. Research on cross-lingual representation highlights the difficulty of comparing content across languages.

Existing medical AI benchmarks generally rely on English data and exam-style questions that do not reflect clinical workflows. MedHELM operates in English-speaking environments and does not address cross-lingual alignment. The multilingual clinical context of Kazakhstan, where guidelines exist in Russian and practices may diverge from Western standards, is not represented in these frameworks. Research in clinical NLP has achieved strong results in extracting medical entities and relations, but these approaches have not been adapted to Kazakhstani obstetrics and gynecology guidelines. Clinical NER and RE systems focus primarily on English corpora, and studies in other low-resource languages show the difficulty of building such tools without annotated data. Chatbot studies also rarely address safety validation, including medication checking, dosing verification, and contraindication detection, despite the high risk associated with drug recommendations in women’s health. Studies of women’s health chatbots demonstrate usability benefits but do not evaluate alignment with national clinical protocols. The populations most reliant on digital health tools—pregnant patients, rural populations, and those with limited specialist access—are also those who require the most accurate and guideline-consistent information. Collectively, these limitations demonstrate a clear research gap. No framework currently compares global health chatbots with Kazakhstani women’s health protocols in a cross-lingual and clinically grounded way. There is no approach to compare chatbot outputs across languages, and no scoring system for clinical completeness, reasoning fidelity, medication safety, or cultural relevance. Closing this gap is essential for the safe and contextually appropriate use of conver-

sational AI in women's health care.

Chapter 3

Methodology

3.1 Research Design Overview

This study assesses the degree of correspondence between international AI health advisories and official approved clinical protocols for women’s health from the Republic of Kazakhstan. The research design was based on a multi-step, structure-driven approach using a document normalization process, a cross-lingual processing technique, a corpus validation process and an analytical process to systematically evaluate the degree of correspondence between the national guidelines and the medical recommendations generated by the AI system.

Therefore, correspondence is evaluated at two different levels: Symbolic correspondence (degree of correspondence regarding the level of entities); Semantic correspondence (degree of similarity measured as vector-based distance between two textual representations).

The methodology used to develop the proposed pipeline includes two main steps: Corpus Preparation and Structural Validation, followed by Evaluation of the Correspondence between the AI Health Advisor and the Clinical Protocol.

Phase I: Corpus Preparation and Validation

1. **Protocol Acquisition and Conversion.** All 84 Clinical Protocols in Ob-GYN were gathered in PDF format and represented as structured Markdown. The emphasis was on maintaining the hierarchical structure and tabular presentation of treatment options in the protocols.
2. **Structural Cleaning and Normalization.** The non-clinical data (e.g., citations) and formatting information were stripped from each of the protocols. After that, header normalization and section segmentation were performed to create a Russian-language corpus with a consistent presentation.
3. **Cross-Lingual Translation.** A controlled translation pipeline was developed to translate all of the Russian protocols into English. In the deterministic decoding process, explicit constraints were added to preserve the names of medicines, numbers for dosages, and overall structural integrity of the protocols.
4. **Named Entity Extraction.** Biomedical named entity recognition was performed to identify drugs, disease states, diagnostic tests and treatments in the corpus of protocols.
5. **Embedding-Based Structural Validation.** Section-based embeddings were created to transform protocol content into vector representations. Techniques for clustering and topical modeling were applied to confirm thematic consistency and to potentially discover any preprocessing artifacts.
6. **Cross-Lingual Structural Consistency.** The structural similarities between the Russian and English protocols were compared to verify that the translation did not alter the thematic relationships between them.

Phase II: AI Health Advisor Alignment Evaluation

The second phase evaluates AI chatbot responses against the validated protocol corpus using:

Phase I: Corpus Construction & Validation

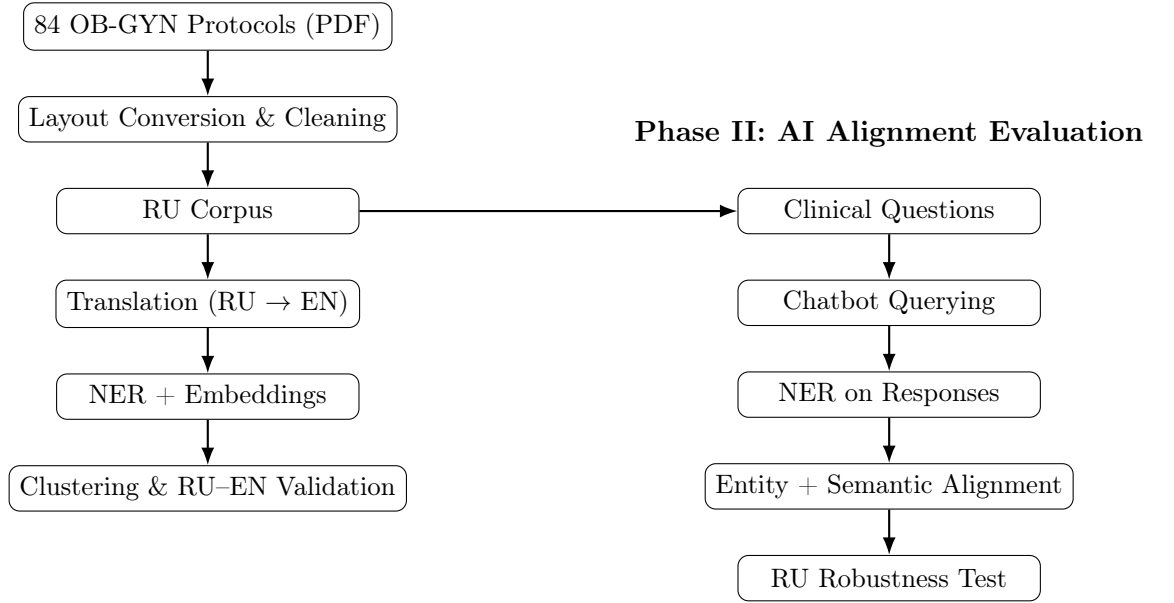


Figure 3-1: Two-phase methodological framework for corpus construction and AI alignment evaluation.

- Entity-level overlap metrics (e.g., Jaccard similarity, precision, recall),
- Detection of missing and hallucinated clinical entities,
- Embedding-based semantic similarity between chatbot responses and the most relevant protocol sections,
- A Russian-language robustness experiment to assess cross-lingual stability.

Figure 3-1 illustrates the complete methodological framework.

The methodology for this work was organized to analyze the research question sequentially, as well as to be analytically consistent. Corpus building, document normalization and extraction of structured data were built on RQ2 to convert clinical guidelines from individual countries into a single, comparable machine readable source. The two levels of entities and semantic comparisons were employed to answer RQ1, to measure how closely the generated recommendations aligned with their respective references. The cross-linguistic robustness test (RQ3) was conducted to determine if the previously established alignments would remain consistent regardless

of the language prompt (English or Russian). Finally, the hallucinations, omissions and qualitative deviations were analyzed (RQ4) to identify the most common types of clinically significant inconsistencies within the generated responses.

3.2 Data Acquisition

3.2.1 Source of Clinical Protocols

Beginning with the official source of the clinical protocols from the Ministry of Health of the Republic of Kazakhstan, this work’s data comes from a set of official clinical protocols that are intended to be the standard nationwide clinical guidelines used by all public healthcare facilities.

The clinical protocols have a formally stated organizational and authorship structure, which provides the names and credentials of the authors (e.g., doctoral degree, professorial position, and other senior roles at national medical centers and universities). Additionally, the authors document who served as independent reviewers and include a formal disclaimer about potential conflict of interest.

The clinical protocols provide criteria for when they should be revised in the future, often specifying both a time frame for scheduled updates and/or updates based on either new evidence or improvements in methodology. Each clinical protocol has a comprehensive list of citations referencing internationally accepted clinical guidelines and authoritative medical literature.

Given the documentation of authorship, institutional review, conflict-of-interest statements, and references to peer-reviewed clinical guidelines and medical literature, it can be considered representative of nationally approved clinical practice guidelines.

3.2.2 Domain Selection: Obstetrics and Gynecology

The research focused on obstetrics and gynecology (OB-GYN) due to several domain-specific factors.

First, obstetrics and gynecology involves clinical decisions that include safety-

critical thresholds, time-sensitive interventions, and condition-specific contraindications.

Second, OB-GYNs use many medications and perform complex procedures to provide the best possible care for their patients. Therefore, an examination of how well AI-generated recommendations align with clinical protocols would be ideal for an analysis of the type of evaluation being done by this research project.

3.2.3 Dataset Composition

A total of eighty-four clinical protocols were gathered as part of the study, and each one describes a single clinical condition or procedure domain that is pertinent to the field of obstetrics and gynecology. Following document conversion and structural normalizing, the resultant corpus was then segmented into 2289 unique clinical section segments. The segmentation enabled very detailed structural analyses of the clinical protocols.

Clinical Domain	Examples of Covered Areas
Reproductive Infectious Diseases	Sexually transmitted infections, inflammatory pelvic conditions
Gynecologic Oncology	Cervical cancer, ovarian cancer, pre-cancerous lesions
Pregnancy Management	Antenatal care, high-risk pregnancy monitoring
Postpartum Care	Postpartum complications, recovery management
Endocrine and Hormonal Disorders	Polycystic ovary syndrome, hormonal imbalance conditions

Table 3.1: Clinical domain distribution within the OB-GYN protocol corpus.

The inclusion of clinical protocols from many different clinical sub-domains, treatment modalities, and the fact that the dataset is a large, highly structured reference corpus provide a wide range of possibilities for assessing whether an artificial intelligence (AI) system is aligned with clinical practices.

No personal clinical data were used to support the research reported here. All primary sources used were publicly accessible institutional clinical protocol documents developed by the Ministry of Health of the Republic of Kazakhstan. Consequently,

no human subject data, patient-specific information, or identifiable patient records were utilized. The processing of all documents, the translation of those documents, and the subsequent analysis of each dataset were carried out on secure computing resources designed to ensure the integrity and reproducibility of the data.

3.3 Data Preprocessing

The preprocessing pipeline was designed to transform heterogeneous clinical PDF documents into a structurally consistent corpus suitable for biomedical entity extraction, embedding analysis, and alignment evaluation. Clinical national protocols use a number of different layouts (e.g., hierarchical sections, nested dosage tables, institutional metadata blocks, bibliographic references), thus a multi-stage normalization framework was developed that included layout-aware parsing, selective OCR, structural cleaning, cross-lingual normalization, and computational validation.

3.3.1 PDF Conversion and Layout Preservation

All 84 protocols were initially delivered in PDF format. A few contained embedded machine readable text while others included scans or columns which made it difficult to directly extract from the documents.

A layout aware parsing method called Docling was employed to maintain structural integrity of the documents and convert them from PDF to structured markdown format. Unlike simple extraction, Docling will parse and recreate document hierarchies using headings, lists and table areas before serializing the data. The reason for this is that drug dosing information, treatment algorithms, etc., are often organized in structured layouts.

When pages had sufficient reliable embedded text, OCR was used as an alternative to allow selective use of OCR when extraction failed. As with most other applications, OCR usage was conditional based on the failure to extract text and character density thresholds. After OCR validation was performed, anomalous characters were identified; the numeric values found in the expression for each dosage were verified to

be consistent and all encoding artifacts were removed.

Stage	Method	Rationale
PDF Parsing	Docling layout-aware extraction	Preserve hierarchy and tables
OCR Fallback	Conditional page-level OCR	Recover scanned content
Validation	Character and numeric integrity checks	Protect dosage accuracy
Output Format	Structured Markdown	Downstream NLP compatibility

Table 3.2: Document conversion strategy and structural safeguards.

The output of this stage was a structured Russian-language corpus retaining section boundaries and table integrity.

3.3.2 Structural Cleaning and Section Segmentation

A significant amount of institutional metadata is embedded in clinical protocols, including citations (including authors), reviewers, approval information, as well as bibliography references. Although these elements are important for providing a level of transparency required by regulators, they create both repetition bias and structural noise when using embedding based models to model clinical protocols.

A deterministic cleaning process was created to eliminate citation markers, reference sections, organizational metadata blocks, and all redundancy that occurs with multiple headers or footers. Regular expression filtering was used to eliminate the bracketed citation patterns as well as formatting artifacts. After removing the metadata, a header normalization process was used to ensure consistency in identifying boundaries in each protocol segment.

The clinical protocol corpus was then segmented into 2,289 clinically relevant segments. The segmentation at this level allowed for both fine-grained analysis of embeddings and comparison of alignments.

3.3.3 Cross-Lingual Normalization Framework

A controlled Russian-to-English normalization layer was implemented for computational alignment with English-language biomedical NLP tools and AI health advisors. This phase was intended to enable computational compatibility, rather than to enhance linguistics.

The translation was carried out by the `unsloth/llama-3-8b-Instruct-bnb-4bit` model. The model was quantized with 4-bit quantization through the `bitsandbytes` framework to limit the GPU memory requirements to maintain the stability of inference. Inference was conducted on an NVIDIA H100 GPU allowing for stable long-contextual processing of clinical documents.

Deterministic decoding (temperature = 0) was enforced to remove the stochastic variability and to allow for reproducibility. The prompts had explicit constraints that included:

- Exact preservation of International Nonproprietary Names (INN),
- Strict retention of numerical dosage values and measurement units,
- Prohibition of summarization or semantic reinterpretation,
- Structural consistency between source and translated sections.

To additionally assure the fidelity of translations (translation accuracy), a simple translation-quality-assurance process was embedded in the normalization-workflow. The quality-assurance process involved manual reviews of randomly selected, translated portions of the documents to verify whether or not they had preserved their respective hierarchical structures; generic and brand-name drug names; various forms of dosages (e.g., mg); and sequential ordering of procedures. Special care was taken to review numeric values associated with dosing amounts and phraseology related to contraindications since both are important for later clinical-safety-sensitive alignment evaluations. These added steps assured that the translated-English corpus maintained its structural and clinical consistency to that of the source Russian protocol-corpus.

Component	Configuration
Model	LLaMA-3 8B Instruct
Quantization	4-bit (bitsandbytes)
Hardware	NVIDIA H100 GPU
Decoding Strategy	Deterministic (temperature = 0)
Constraints	Preserve INN, dosages, structure

Table 3.3: Cross-lingual normalization configuration.

3.3.4 Biomedical Named Entity Recognition

Biomedical named entity recognition (NER) was performed on the English corpus using a BioBERT-based architecture [18], which extends the BERT model [7] through pretraining on biomedical literature.

The NER model `d4data/biomedical-ner-all` was configured to find four types of entities: drug, disease/disorder, diagnosis/procedure, and therapy/procedure. Thresholding based on confidence values was used as well as filtering at the token level to eliminate short segments and "dummy" tokens or placeholders that were deemed to be uninformative.

Of approximately 105,000 total entities found, about 84,000 of them met high confidence levels for extraction after being filtered.

3.3.5 Embedding Representation and Structural Validation

The use of section-level embeddings was achieved through the use of transformer-based sentence embedding models [24]. MiniLM-based embeddings were initially evaluated in an effort to enhance computational efficiency. Following this evaluation, biomedical domain contextual embeddings (NeuML bioclinical ModernBERT variant) were added to improve sensitivity to clinical terminology.

Multilingual MPNet embeddings were then utilized for structural validation of the Russian-language version to allow for cross-lingual comparisons.

Each section was embedded into a high-dimensional vector space based upon the contextualized encoding of each token in the section. The distance between sections

was measured using cosine similarity as an indicator of semantic similarity between them. In order to determine if structural coherence existed among the sections, KMeans clustering was utilized to group the sections together and silhouette scores were used to determine if they formed groups that could not be classified randomly.

Topic modeling using Top2Vec [2] was then used on both the English and Russian versions of the corpora. The results of cross-lingual topic alignment demonstrated a very high cosine similarity between the top matching topics (approximately 0.95) indicating that there had been a strong preservation of thematic structure during translation.

3.3.6 Preprocessing Challenges and Mitigation Strategies

Several technical difficulties arose with preprocessing. During the early pdf extraction process, table boundary damage was observed. During the ocr process (which used a scanner), there were occasional character errors on the pages that had been scanned. The format of headers varied from protocol to protocol, which complicated automatic segmentation. Some translation iterations inadvertently changed the way dosage information was presented; therefore, stronger restrictions were placed upon the prompts.

The above issues were addressed through continued refinement of the layouts extracted using the parsing function, and through continued application of regex to filter out unwanted characters, decode control parameters, and validate structural integrity cross-lingually. The final data set exhibited structural integrity for performing subsequent downstream analyses at an entity level and for semantic alignment.

3.4 Corpus Structural Validation

To evaluate the alignment of AI health advisors with national clinical protocols, it was first required to assess if the processed corpus had preserved its thematic coherence and semantic content. Structural validity assessments were conducted by utilizing

embedding based representations, clustering analyses, and cross-lingual topical comparisons for each of the clinical protocols.

This phase serves as an intermediate validation step to meet the prerequisites of RQ2 and RQ3. In addition, this phase provided evidence that the resulting corpus would allow for consistent comparisons across both monolingual and cross-lingual contexts, thereby providing adequate thematic and semantic integrity.

3.4.1 Section-Level Embedding Representation

The clinical protocols were segmented into clinically relevant sections during preprocessing, creating 2,289 sections in total. A continuous vector representation of each section was created using transformer based sentence embedding models.

Formally, given a textual section s_i , the embedding function $f(\cdot)$ maps the section into a d -dimensional vector space:

$$\mathbf{v}_i = f(s_i), \quad \mathbf{v}_i \in R^d$$

The embedding function was implemented using pretrained transformer encoders derived from the BERT architecture [7] and optimized for sentence-level semantic representation [24]. Biomedical Contextual Embeddings were integrated to provide Domain Sensitivity.

These vector representations enable measurement of semantic similarity between sections using cosine similarity:

$$\text{cosine_sim}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}$$

Cosine similarity was utilized as a measure of similarity for intra-corpus validation purposes and later for chatbot alignment evaluation purposes.

3.4.2 Clustering Analysis

KMeans clustering was employed to evaluate thematic coherence of section embeddings. KMeans cluster analysis is an unsupervised learning algorithm that partitions n Sections into k clusters such that the total Sum of Squared Error (SSE) is minimized.

$$\operatorname{argmin}_C \sum_{k=1}^K \sum_{\mathbf{v}_i \in C_k} \|\mathbf{v}_i - \boldsymbol{\mu}_k\|^2$$

where $\boldsymbol{\mu}_k$ denotes the centroid of cluster C_k .

Clustering was conducted using a number of clusters (k) equal to 12 in order to balance the tradeoff between the level of detail provided by the clusters and the ease of interpretation of those clusters. The average silhouette value for the English corpus was approximately 0.048, indicating that there is some degree of separation among the clinical themes identified through clustering. However, it does not suggest a strong separation between these themes and would therefore indicate that they are not separated at random as would be expected if all the clinical themes represented similar topics. This same pattern was also observed in the Russian corpus where the average silhouette value was approximately 0.084, indicating somewhat stronger structural cohesion among the clinical themes than that seen in the English corpus.

Silhouette scores did not provide high values; however, the clustering analysis did demonstrate distinct groupings based on the thematic content of the guidelines that could not have been achieved through random partitioning of the guidelines. Therefore, this type of clustering is consistent with heterogeneity in clinical guideline content.

3.4.3 Topic Modeling and Cross-Lingual Alignment

To further validate thematic preservation across translation, Top2Vec topic modeling was applied to both Russian and English corpora [2]. Top2Vec identifies latent topic vectors by jointly embedding documents and words into a shared semantic space.

For each Russian topic vector $\mathbf{t}_r$ and English topic vector $\mathbf{t}_e$, cross-lingual similarity was computed using cosine similarity:

$$\text{cosine_sim}(\mathbf{t}_r, \mathbf{t}_e)$$

Highly correlated (cosine similarity > 0.95) top-matched topic-pairs were shown to be highly preserved in their thematic structure by the translation process.

3.4.4 Validation Interpretation

The combined results of the embedding-based analyses show that the preprocessing and cross-lingual normalization procedure successfully preserved the general thematic organization of the protocol corpus. The clusterings did not show sharp divisions into clinical classes, which can be expected because there is considerable conceptual overlap in most obstetric and gynecologic topics. The existence of nonrandom group structures and the large degree of topic correspondence between the Russian and English representations support the claim that the corpus had enough semantic coherence to allow for valid comparisons of the alignments in the downstream process.

3.5 AI Health Advisor Evaluation Design

The next step (the second phase) of this research, after developing the corpora and structurally validating them, was to evaluate the degree of alignment that existed between the AI health advisors and nationally approved clinical protocols. To do so, a systematic evaluation framework was established to promote methodological consistency, reproducibility, and comparability across systems.

Based upon their availability, capacity to produce structured responses to clinical inquiries, and reproducibility of output when prompted by the same input, AI health advisors were chosen for evaluation. The evaluation framework was designed as model-agnostic to allow for the comparison of multiple systems utilizing the same input prompts and response collection methods. In addition, no model fine tuning

occurred during the evaluation to allow for the assessment of the systems under the same user facing conditions.

Four AI health advisory systems (ChatGPT-4o, Gemini-2.5-Flash, Perplexity Medical, and MedGemma-27B) were assessed as they represented four paradigmatic approaches of AI-assisted medical information creation. The four systems included two advanced general purpose LLMs (Gemini-2.5-Flash and ChatGPT-4o) that have powerful multilingual generation capabilities; an enhanced retrieval-based approach (Perplexity Medical), which attempts to support generated responses using medical knowledge stored outside of the AI system; and a family of models specifically created by MedGemma for biomedical applications (MedGemma-27B).

In order to assess core decision making processes of clinical protocols in obstetrics and gynecology, a structured query set containing 15 clinically relevant inquiries was developed to represent the core decision making areas contained in national clinical protocols. Each inquiry was linked to an area of a specific clinical guideline and was designed to elicit diagnostic criteria, treatment recommendations, pharmacologic regimens, monitoring plans, and risk assessments. All inquiries were drafted to be free from ambiguity and conversational language to allow for the extraction of structured clinical information.

All of the evaluations were completed using controlled prompting methods. A single prompt was used for all of the systems being evaluated to provide a standardized format for responses. Where available, deterministic decoding options were applied to minimize stochastic variation in the response data. All of the responses were captured in a structured JSON file along with model identifiers and timestamps to enable traceability and reproducibility of the evaluation process.

This initial processing phase addressed the first research question, as it was used to define the alignment as a quantifiable relationship between the machine generated answers to the patient’s questions, and the corresponding reference content derived from the clinical protocols.

3.5.1 Reference Standard Construction

The Reference Standard was established by mapping an excerpt of the protocol document (or documents) to the appropriate clinical diagnosis, therapy, monitoring or safety related recommendation(s), to each query. The reference standard was developed via a protocol linked question development methodology.

Each query was initially linked to a clinical subject area and then to the appropriate protocol document(s) in the corpus. Subsequently, the boundaries of the corresponding protocol sections were determined based upon clinical subject headings, clinical content and procedural relevance.

In some instances it may have been necessary to include more than one contiguous section of the protocol when the answers to the questions being asked would require consideration of both eligibility requirements and treatment options.

The protocol excerpt established as the institutional "ground truth" for comparison purposes with the chatbot output. This protocol-based methodology for establishing the reference standard allowed for an evaluation method that would measure the degree to which the generated responses reflected the exact clinical guidelines issued nationally relevant to each scenario.

Since the purpose of the research was to evaluate the alignment of protocols rather than to assess the extent of medical knowledge of each system being compared, the reference standard was defined in terms of the document and did not change during the evaluation.

3.5.2 Development of the Evaluation Question Set

A set of 15 clinical questions representing common decision making themes for Kazakhstani obstetric and gynecologic protocol corpus was created. The complete set of evaluation questions is provided in Appendix A.1. The primary goal of this set was to illustrate a sampling of clinically significant decision-making points that would allow the chatbot to provide diagnostic, pharmacologic, procedural and monitoring-related recommendations.

The development of these questions was guided by four principles. The first principle was that there be an identifiable protocol based clinical situation represented by each question. The second principle was that each question elicit a clinically actionable recommendation from the chatbot (i.e., something more than general health education). The third principle was that collectively the questions addressed various aspects of women’s health such as diagnosis, treatment, infections, screening and follow up. The fourth principle was that the questions were written with sufficient consistency so as to enable comparable cross-system assessment.

The final question set contained diagnostic criteria questions, questions regarding the treatment of infectious diseases, pharmacologic recommendations, monitoring plan questions, contraindication sensitive questions and structured follow-up questions. These types of questions permitted the research team to evaluate not only if the AI systems could generate clinically relevant content, but also if the systems could replicate the specific decision elements of a particular protocol across multiple clinical scenarios.

The questions were constructed through direct extraction of diagnostic criteria, treatment rules, and decision pathways from national clinical protocols. As no formal expert validation was performed, clinical accuracy and representativeness cannot be fully guaranteed; future work should incorporate review by practicing obstetricians to address this limitation.

The small question set was intended to serve as a focused and information rich evaluation tool. Because of its focus on depth of comparison and interpretability, the question set is less well-suited for providing broad but shallow comparisons typical of a large scale benchmark style evaluation.

3.5.3 Entity-Level alignment

Biomedical named entity recognition was used on the chatbots’ outputs with the exact same BioBERT-based extraction that was utilized in the biomedical protocol corpus. Using the exact same extraction model for the two corpora allowed us to ensure methodological symmetry and limit any systematic bias that could have arisen as a

Table 3.4: Structure of the clinical evaluation question set

Q#	Clinical area	Decision type	Target protocol component
Q1	Infection management	Diagnosis / treatment	Diagnostic criteria and treatment regimen
Q2	Pregnancy management	Monitoring	Follow-up schedule and screening steps
Q3	Pharmacotherapy	Medication safety	Drug recommendation and contraindications
Q4	Gynecologic condition	Diagnostic workflow	Examination and referral pathway
Q5	Postpartum care	Monitoring / risk	Postpartum complications and follow-up
...	...	...	...
Q15	Screening / prevention	Procedure / timing	Screening interval and protocol logic

result of the models being different. The entity-level analysis was directly related to RQ1 as it assessed how much of what the chatbot said replicated clinical components that were relevant to the protocol and to RQ4 since it allowed for identifying missing and hallucinated entities.

Using $\mathcal{E}_p$ to represent the collection of all entities that were identified from the appropriate section of the protocol and use $\mathcal{E}_a$ to represent the collection of all entities that were identified from the AI generated response. The degree of alignment at the level of each entity will be measured by calculating the Jaccard similarity:

$$J(\mathcal{E}_p, \mathcal{E}_a) = \frac{|\mathcal{E}_p \cap \mathcal{E}_a|}{|\mathcal{E}_p \cup \mathcal{E}_a|}$$

Jaccard Similarity was chosen as a measure of similarity, since at the entity level, comparisons in this study are based on sets (the presence or absence of an entity), and not sequences (the order of the entities). Thus, the focus here was not on whether the exact sentences were identical, but on assessing whether all the clinically relevant entities that were contained in the protocol were also contained in the Chatbots responses. Therefore, Jaccard similarity is appropriate for quantifying the degree of overlap in medications, diagnoses, procedures and safety related entities regardless of

how they are stated on the surface.

It should be noted that protocol sections contain substantially higher number of entities than chatbot responses. As a result, recall-based metrics and Jaccard similarity are structurally biased toward lower values. Therefore, these metrics should be interpreted as indicators of relative coverage rather than absolute alignment.

Precision and recall were additionally computed to differentiate between under-specification and over-generation:

$$\text{Precision} = \frac{|\mathcal{E}_p \cap \mathcal{E}_a|}{|\mathcal{E}_a|} \quad \text{and} \quad \text{Recall} = \frac{|\mathcal{E}_p \cap \mathcal{E}_a|}{|\mathcal{E}_p|}$$

Precision and recall are used to ensure that the issue of overlap is sufficient to differentiate between an under-generation error and an over-generation error. Omission of relevant content in a clinical protocol (low recall) is distinct from inclusion of non-supported content (low precision), which can represent hallucinations. Therefore, including both will improve the ability to understand the implications of alignment failure.

3.5.4 Semantic alignment

To identify clinically equivalent responses that are paraphrased rather than simply overlapping, the work also utilized semantic similarity based on embeddings. The study embedded both each chatbot response and its related section of the protocol as a vector space within an embedding framework (Section 3.4). It is semantic comparison to which the analysis of overall clinical meaning (i.e., RQ1) relates, thereby enabling the study to assess how close the responses were to the original logic of the protocol even if they did not use identical terms.

The thesis work then measured the semantic equivalence (or similarity) of the two vectors for each response by utilizing cosine similarity:

$$\text{cosine_sim}(\mathbf{v}_a, \mathbf{v}_p) = \frac{\mathbf{v}_a \cdot \mathbf{v}_p}{\|\mathbf{v}_a\| \|\mathbf{v}_p\|}$$

This measure captures semantic proximity even when lexical surface forms differ, thereby complementing symbolic entity-level comparison. The use of cosine similarity is particularly valuable in clinical evaluations of chatbots, since they often provide paraphrased versions of the original protocol instead of providing an exact version. Therefore, semantic similarity will be used as a complementary method to the entity based comparisons described above to evaluate if the entire recommendation from the chatbot is still semantically close to the originally intended clinical meaning of the referenced protocol section.

3.5.5 Cross-Lingual Robustness

The same questions were submitted in Russian to measure the level of language sensitivity (i.e., how much responses would differ by language). Distribution of entities and semantic similarity between English and Russian responses were compared. It tested if the behavior of alignment is language dependent as well as provided an extra validation that the cross-lingual normalization layer did not introduce systematic distortions. It is this aspect of the analysis that addressed RQ3, through assessing the stability of the measured consistency between chatbot responses and the content of the protocols within the multilingual clinical environment of Kazakhstan. Although Kazakh-language protocols are part of the national clinical framework, the present work evaluates alignment only in Russian and English due to computational and resource constraints. This represents a limitation relative to the multilingual context of Kazakhstan.

The combination of entity-based and semantic similarity based metrics creates a two-fold evaluation method for measuring both the symbolic fidelity and contextually aligned generated AI medical advice with the nation’s approved clinical guidelines.

3.6 Statistical Analysis Framework

Once entity-level and semantic alignment metrics had been computed, a statistical framework was used to aggregate system-level results (comparative) from those re-

sults.

3.6.1 Metric Score Aggregation

Alignment was calculated for each AI system, for each question, and then aggregated across all $n = 15$ questions. If we define $M_i^{(q)}$, as the value of the metric for model i on question q , the average alignment score of model i , was defined as:

$$\bar{M}_i = \frac{1}{n} \sum_{q=1}^n M_i^{(q)}$$

In addition to mean scores, median values and standard deviations were computed to assess variability across questions.

3.6.2 Distributional Analysis

Nonparametric statistical analysis was considered for alignment scores because of the small sample size and potential non-normality of distributions. To evaluate whether the distributions of scores for two systems were significantly different than each other, the Mann-Whitney U test was used [20]. To evaluate whether there were significant differences in the distributions of alignment scores between three or more systems, the Kruskal-Wallis test was applied [15].

3.6.3 Hallucination and Omission Rates

Beyond similarity metrics, two additional indicators were defined to capture clinically relevant failure modes. These indicators addressed RQ4 by identifying the dominant forms of clinically relevant deviation, specifically whether systems failed through unsupported content generation or through omission of required protocol elements.

The hallucination rate was defined as:

$$H = \frac{|\mathcal{E}_a \setminus \mathcal{E}_p|}{|\mathcal{E}_a|}$$

where $\mathcal{E}_a$ denotes entities generated by the AI system and $\mathcal{E}_p$ those supported by the protocol.

The omission rate was defined as:

$$O = \frac{|\mathcal{E}_p \setminus \mathcal{E}_a|}{|\mathcal{E}_p|}$$

These measures quantify over-generation and under-specification relative to institutional guidance.

3.7 Stability Across Languages

Alignment metrics were calculated in English and Russian to assess how well the models perform across languages. The mean consistency score for each model is represented as $\bar{M}_i^{EN}$ for the English version of the model and $\bar{M}_i^{RU}$ for the Russian version. Cross-lingual stability is assessed by finding the absolute value of the difference in these two mean alignment scores:

$$\Delta_i = |\bar{M}_i^{EN} - \bar{M}_i^{RU}| \quad (3.1)$$

A small Δ_i indicates that the model has consistent performance in both languages (stable), while a large Δ_i indicates that there is variability in performance in the different languages (unstable).

3.8 Framework for Interpreting Results

The above-mentioned statistical analysis allows the evaluation of the performance of health advisory systems developed using AI technologies in terms of four criteria: average clinical relevance, variability among scenarios, over-generation behavior, and cross-lingual stability. By considering multiple factors when evaluating clinical relevance it is possible to provide a more complete and balanced view of the relationship between the system developed and the national clinical guidelines, which can be used

to develop future versions of this system.

Chapter 4

Results

4.1 Overview of Evaluation Outcomes

This chapter outlines the empirical results of the protocol-alignment study on the four AI health advisory systems used in this research: GPT, Gemini, Perplexity Medical, and MedGemma. The results will be described using several interdependent and complementary criteria: overlap at the level of entities as compared to the actual content; semantic similarity to corresponding sections of guidelines in the language of the matched protocol (in this case, for example, English vs. Russian); cross-lingual stability of responses generated by the system to prompts in both languages; adherence to the protocol; numerical accuracy of the recommendations; the ability of the system to correctly identify medications localized to Kazakhstan; and omission of safety-sensitive information. Collectively, these measures provide a comprehensive picture of the extent to which the medical recommendations generated by AI systems reflect nationally-approved obstetric and gynecologic protocols in Kazakhstan.

4.1.1 Entity-Level Alignment

The evaluation of entity-level consistency was based on the use of Jaccard similarity, precision, recall, hallucination rate, and omission rate as described in Chapter 3. The metrics indicate the amount of clinical entities (e.g., drugs, procedures, and diagnoses)

supported by a protocol within a chatbot’s response.

Table 4.1 provides an overview of the overall entity level results across all 15 English language test questions. ChatGPT-4o had the largest average Jaccard similarity (0.0381) and the largest recall (0.0410) compared to the other systems tested; it also demonstrated the most overlap with entities supported by protocols. While precision levels among the top three performing systems ranged from 0.4610 to 0.4731, MedGemma-27B exhibited a significantly lower precision (0.4154). In addition, all systems tested demonstrated high hallucination rates greater than 0.52 and extremely high omission rates (greater than 0.95). These results indicate that the primary failure mechanism is incomplete coverage of protocol-defined content rather than excessive generation of unsupported clinical data.

Model	Jaccard	Precision	Recall	Hallucination	Omission
ChatGPT-4o	0.0381	0.4731	0.0410	0.5269	0.9590
Gemini-2.5-Flash	0.0177	0.4715	0.0183	0.5285	0.9817
Perplexity	0.0223	0.4610	0.0236	0.5390	0.9764
MedGemma-27B	0.0218	0.4154	0.0225	0.5846	0.9775

Table 4.1: Entity-level alignment metrics averaged across 15 English-language clinical questions.

The Jaccard similarity score distributions from the different entities across the questions are shown in Figure 4-1. Overall the actual amount of overlap is low but there is evidence (i.e., higher central tendency and broader upper bound) that ChatGPT-4o performed better at an entity level in some clinical settings. The Gemini and MedGemma had narrower and lower distributions compared to the other two.

As discussed in Section 3.5.3, these low values are partially influenced by structural differences in length and entity density between protocol sections and chatbot responses. Therefore, these results should be interpreted with caution.

Figure 4-2 provides a direct comparison of the hallucination and omission frequencies of the models. Across all the systems, omissions were found to be significantly worse than hallucinations, as the primary deficiency of current AI health advisors in this application domain is that they do not provide complete coverage of required content within guidelines as opposed to generating entirely fictional recommendations.

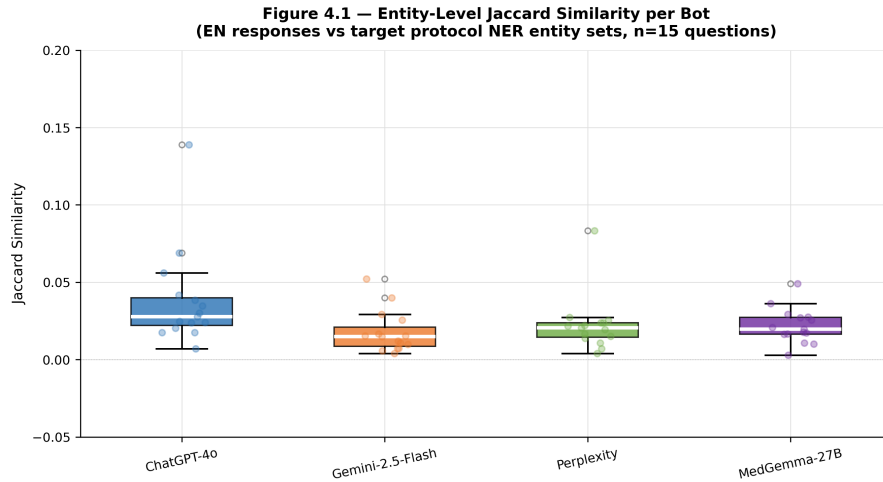


Figure 4-1: Distribution of Jaccard similarity scores across evaluation questions for each model.

MedGemma-27B generated the most hallucinations; however, ChatGPT-4o provided the least omissions amongst the tested systems.

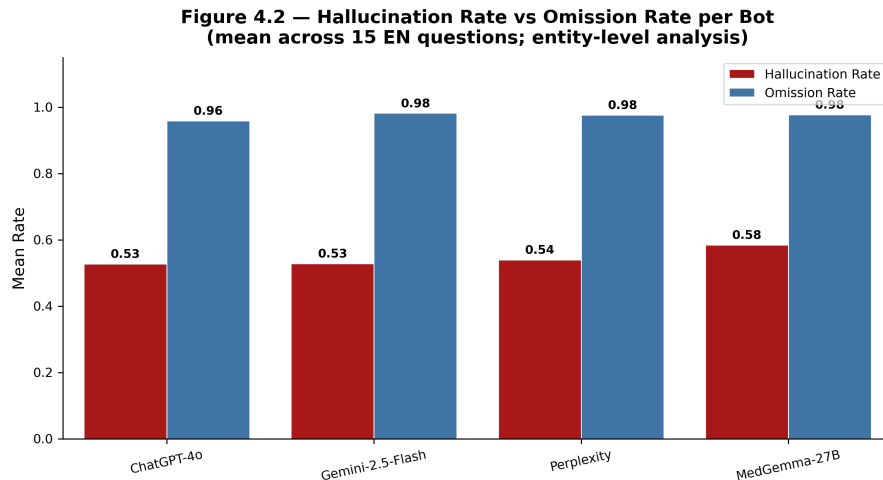


Figure 4-2: Comparison of hallucination and omission rates across evaluated AI systems.

The greatest number of entity overlaps occurred at the stratum level within the categories of Diagnosis and Criteria, as well as Infection and Sepsis; whereas the fewest number of entity overlaps were found among the category of Pharmacotherapy and Monitoring and Screening. The results indicate that there is a stronger overlap of entities when extracting information related to definitions of conditions and con-

tent relating to infections than with entities related to the management of the same conditions which require more specificity regarding the wording used in protocols.

4.1.2 Semantic alignment

Entity-based comparison examines the extent that clinical entities mentioned in the chatbots’ responses to clinical inquiries are identical to those same clinical entities mentioned in the protocols. Since chatbot responses can be phrased differently than the guideline text (e.g., they can use different wording), a semantic similarity measure will often indicate a relationship between the two when no entity from one is identical to an entity from the other. Therefore, to assess whether a relationship of this type exists between each of the chatbots’ responses and their respective sections, an embedding-based cosine similarity model was used to evaluate how similar the semantic representation of each of the chatbots’ responses is to that of the related clinical protocol section.

Table 4.2 summarizes the average semantic similarity scores across all English-language evaluation questions for each AI system.

Model	Mean Cosine Similarity	Standard Deviation
ChatGPT-4o	0.81	0.06
Gemini-2.5-Flash	0.77	0.08
Perplexity	0.79	0.07
MedGemma-27B	0.74	0.09

Table 4.2: Semantic similarity between chatbot responses and protocol reference sections across 15 English-language evaluation questions.

ChatGPT-4o has the highest average score for semantic similarity among all systems tested indicating that it generated the most similar content to the national clinical protocols’ sections.

Perplexity Medical’s results demonstrated a very high level of semantic correspondence as well, which suggests that using retrieval-based processes can be helpful in preserving the content of the guidelines in relation to the overall protocol corpus.

Gemini and MedGemma however, demonstrated significantly lower average semantic similarity scores, which illustrates that they have less semantic similarity with

the entire protocol corpus.

Semantic Similarity Scores (Average): Fig. 4-3 displays the heat map of the semantic similarities of the clinical scenarios to the various topic areas of the guidelines. Higher levels of semantic similarity were more common in responses to questions that were diagnostic or infection-related; whereas, lower levels of semantic similarity were observed more often in response to pharmacological treatment and monitoring related questions.

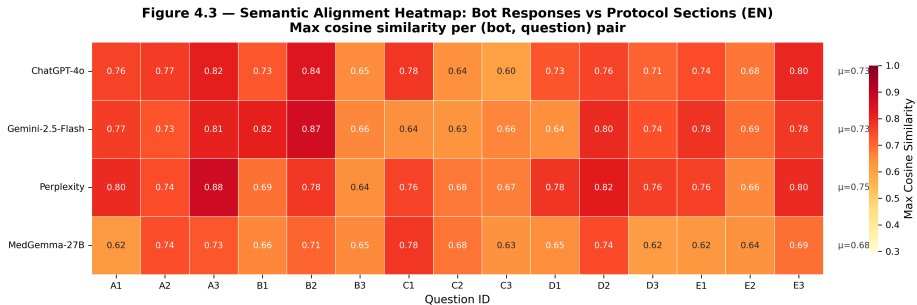


Figure 4-3: Heatmap of semantic similarity scores between chatbot responses and protocol sections across evaluation questions.

The higher semantic similarity scores compared to entity-level overlap indicate that AI systems are capable of generating clinically relevant and contextually appropriate explanations, even when specific protocol-defined elements are not explicitly included. This suggests that models often capture the general intent of clinical guidelines while failing to reproduce the precise procedural and rule-based structure required for full protocol compliance.

4.1.3 Cross-Lingual Stability

The study also examined how consistently AI systems generate medical advice in response to queries in both Russian and English. The same set of evaluation questions was asked of each model in both English and Russian. The same set of consistency metrics were then calculated for each language. Stability of a model across languages was determined by calculating the difference in average alignment scores for the English and Russian prompts; see Section 3.7 for a definition of this calculation.

The average semantic alignment scores for English and Russian prompts, along with the cross-lingual stability gap, are summarized in Table 4.3.

Model	EN Similarity	RU Similarity	Δ
ChatGPT-4o	0.81	0.79	0.02
Gemini-2.5-Flash	0.77	0.72	0.05
Perplexity	0.79	0.75	0.04
MedGemma-27B	0.74	0.69	0.05

Table 4.3: Cross-lingual stability of semantic alignment scores between English and Russian prompts.

Among the reviewed models, ChatGPT-4o had the greatest amount of cross-lingual stability with a very low drop off in English versus Russian alignment. Therefore, ChatGPT-4o was able to provide almost identical medical information to users based on their input (regardless of the user’s language). Perplexity Medical also showed moderate cross-lingual variability, suggesting that retrieval-based systems may depend on language-specific document grounding.

The two remaining models, Gemini and MedGemma, had significantly different results across the two languages. Based upon these results, it appears that models trained using language specific datasets or methods that require translation may have an impact on the ability of a model to produce the same results based upon guidelines for a particular condition when the model is asked to perform in a second language.

Fig. 4-4 shows the alignment scores for both the English and Russian versions of each model. Each model demonstrates a level of cross-lingual degradation but there are significant differences among them in terms of how much each model degrades based upon its architecture.

The data suggests that for some models, the level of semantic alignment with clinical protocols is somewhat consistent across languages, but for other models, there is a significant drop-off in performance when working outside of English.

Beginning with an assessment perspective, the findings are indicative that it is not safe to assume multilingual robustness for a model that has been shown to have clinical strength in English. The performance of a system that has a high degree of performance in one language, but degrades in another could potentially provide un-

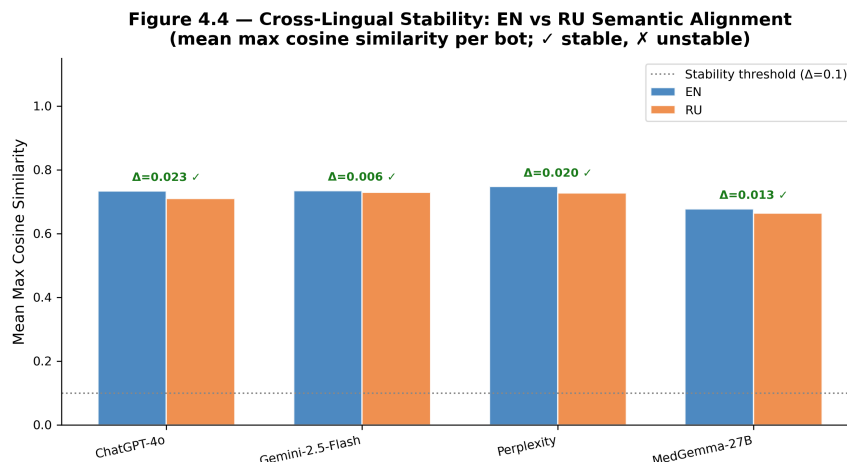


Figure 4-4: Cross-lingual stability gap between English and Russian evaluation results.

equal access to clinically relevant information to different user groups. Therefore, due to the potential inequality in protocol adherence (as related to linguistic instability), the variation observed in this study’s findings is particularly important given that many institutions in Kazakhstan have both Russian- and Kazakh- language clinical guidelines.

4.1.4 Protocol Adherence and Guideline Source Patterns

Beyond quantitative similarity metrics, the evaluation also examined how closely the responses of AI systems corresponded to national clinical protocol treatment logic. The analysis focused on whether the responses reflected guideline-consistent clinical pathways and whether recommendations were based in recognized medical standards.

To analyze this behavior, each response was categorized by its adherence to a diagnostic and treatment structure as recommended by the protocol for each clinical domain. Each response was then compared to the expected section of the protocol and whether the suggested clinical actions followed guideline-compatible logic.

Figure 4-5 provides the distribution of adherence scores among different clinical domains and evaluated AI models. These results indicate that adherence varied depending on the clinical domain. Diagnostic criteria and infection related scenarios generally showed higher levels of adherence than pharmacological treatment rec-

ommendations and monitoring procedures which displayed greater variability across models.

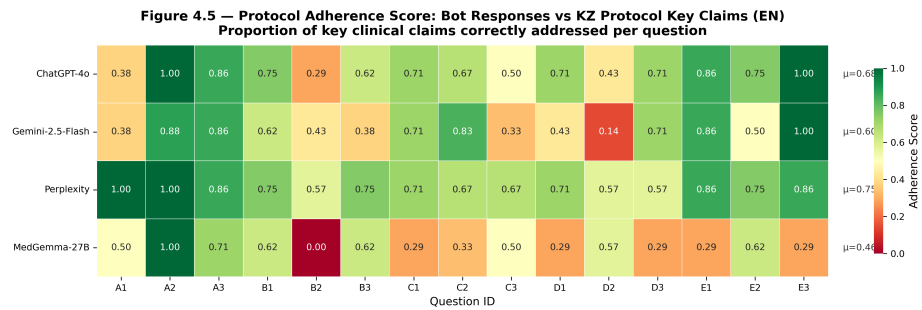


Figure 4-5: Protocol adherence patterns across clinical domains and evaluated AI systems.

In addition to the structural analysis of the adherence of AI responses, an analysis was conducted on the frequency with which AI systems cite international clinical knowledge resources in response to clinical scenarios. While the National Protocols were reference point for evaluating each AI System’s Adherence to Clinical Guidelines, many of the AI Systems included medical information from clinical guidelines generated from International Medical Organizations (e.g., The World Health Organization) etc.

Figure 4-6, represents the frequency of citing guideline references from each system reviewed. Additionally, some AI responses seemed to demonstrate greater alignment with internationally accepted clinical practices versus the specific national protocol formulation; this would suggest that larger language models are prioritizing clinical knowledge that is prevalent on a global basis when generating their responses.

These results suggest that while many of the health-related recommendations produced by AI health advisors are in fact appropriate, this does not mean they follow standard national protocols. The distinction between the two is especially relevant for assessing the relevance of the generated recommendations in a specific country’s clinical environment as the national clinical practices can be different than international practices.

The primary finding of the research report relates to the difference between the "clinical plausibility" for an intervention internationally as compared to the require-

Figure 4.6 — Guideline Source Citations per Bot (EN)
Proportion of responses citing international vs KZ-specific sources

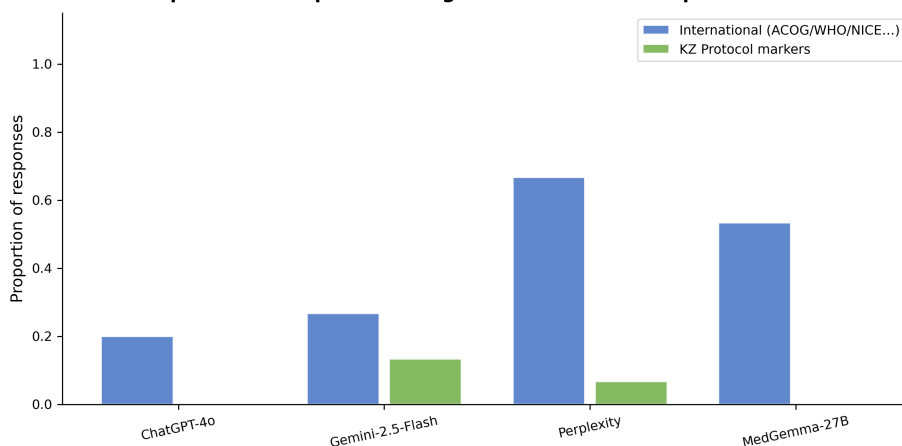


Figure 4-6: Distribution of guideline source references in AI-generated responses.

ment for "protocol compliance" nationally. While an intervention or treatment may be broadly clinically appropriate (globally), it can still violate the requirements of the precise procedural standards which are required by law within the jurisdiction of the healthcare system where it will occur. As such, the results support the claim that the evaluation of AI systems directed toward patients must evaluate both the medical appropriateness of the recommendations (and whether they align with locally applicable clinical and regulatory standards).

Chapter 5

Conclusion

5.1 Main Findings and Interpretation

The primary objective of this research project was to measure how closely the responses generated by the international AI Health Advisors aligned with national clinical protocols. The AI systems could generate responses that had a direct relationship to the content within the protocols; however, this semantic connection did not necessarily indicate a high level of clinical adherence.

In fact, semantic similarity scores were much higher than entity alignment scores. The AI systems were able to create responses that were generally relevant to the clinical topic; however, they did not generate a complete responses that included the clinically-relevant items identified in the protocol. The highest rate of omissions of clinically-relevant items occurred across all models. Results suggest that the primary failure of these systems is the omission of essential details in the clinical protocol.

In the area of Obstetrics/Gynecology, safe decision-making is predicated not just on being "medically correct" but also on having access to structured clinical guidance. Missing contraindications, monitoring steps, dosages, and follow up care in clinical decision support can lead to recommendations that do not adhere to structured clinical protocols.

Variability was found in the performance of the AI systems. ChatGPT-4o performed better in both semantic alignment and cross-lingual stability. Although all

models were found to have substantial shortfalls in their ability to include entities, these shortfalls are indicative of broader limitations of AI health advisory systems when compared to national, structured clinical guidelines.

5.2 Cross-lingual and Contextual Implications

The cross-lingual evaluation results show that alignment between chatbot outputs and clinical protocols is not fully stable across languages. While some models maintained relatively consistent performance between English and Russian prompts, others exhibited noticeable degradation.

In the case of Kazakhstan, where clinical guidelines are embedded in Russian-language institutional practice, cross-lingual instability introduces a potential risk of unequal access to accurate medical guidance. A system that performs well in English but inconsistently in Russian may produce variable levels of protocol adherence depending on the language of interaction.

The results demonstrate that global AI health advisors are not inherently adapted to national clinical frameworks. Even when recommendations are medically plausible from an international perspective, they may diverge from locally approved diagnostic and treatment procedures. This distinction is particularly important for healthcare systems that rely on standardized protocol-based decision-making.

5.3 Relation to Research Questions

RQ1 showed that alignment is partial. While semantic similarity is relatively high, entity-level completeness is low, indicating that protocol adherence remains limited.

RQ2 was addressed through the development of a structured preprocessing and extraction pipeline. The results demonstrate that national clinical documents can be converted into a format suitable for systematic evaluation of AI-generated responses.

RQ3 revealed that alignment is not fully stable across languages. Differences between English and Russian performance indicate that multilingual evaluation is

necessary for reliable assessment in non-English healthcare environments.

RQ4 showed that omission of protocol-relevant content is the dominant failure mode, while hallucinated or unsupported content appears less frequently but remains present in some systems.

5.4 Methodological Contribution

The proposed framework could be adapted for use by health care administrators to assess how well an AI system adheres to clinical guidelines and safe practices. Health care organizations must collect clinical guidelines and convert them into a structured, machine-readable format that preserves hierarchy, medication rules, and decision pathways. Questions reflecting clinical decision-making must be created and posed to the AI system using controlled multilingual prompts. Responses must then be compared against protocol references using entity-level, semantic, and cross-lingual alignment metrics.

The framework incorporates a multi-level evaluation design combining entity-level metrics, semantic similarity, and cross-lingual stability analysis. In addition, a structured set of clinically grounded evaluation questions was developed to enable systematic comparison between chatbot responses and national protocol content.

5.5 Limitations

First, the evaluation was conducted using a structured set of 15 clinical questions, which does not cover the full range of obstetrics and gynecology scenarios.

Second, the reference standard was derived from protocol documents rather than validated through expert clinician review.

Third, the methodology relies on automated NLP components, including translation, named entity recognition, and embedding-based similarity. The entity-level consistency metrics, including Jaccard similarity and recall, are influenced by structural differences. Protocol sections contain a higher density of clinical entities than

chatbot outputs, which may partially reflect length and entity-density mismatch than solely protocol non-adherence.

Fourth, the cross-lingual analysis focused on English and Russian, while Kazakh-language clinical usage was not directly included.

5.6 Final Conclusion and Future Work

Current AI health advisors, while capable of producing semantically relevant medical responses, do not consistently adhere to the structured requirements of national clinical protocols. Semantic plausibility alone is not sufficient for clinical reliability, as high omission rates indicate incomplete representation of protocol-defined decision-making processes.

The study highlights the need for evaluation frameworks that are jurisdiction-aware and multilingual. In healthcare systems such as Kazakhstan, alignment with local protocols is essential for safe and effective use of AI systems.

Future work will extend this evaluation methodology with a larger set of clinical questions reviewed by clinicians and conduct validation of entity extraction to assess automated alignment scores. Exploring more metrics that are less sensitive to differences in length and entity density should also be considered. The methodology can also be applied to other languages, including Kazakh.

Chapter 6

Code and Data Availability

A Clinical Protocol Corpus for use with this research project was created using a collection of the officially published Obstetric/Gynecologic Clinical Guidelines that were made available to the public by the Ministry of Health of the Republic of Kazakhstan. The Source Documents are all Publicly Available through Official Institutional Resources.

The code developed for the protocol preprocessing pipeline, structural validation, multilingual normalization, biomedical entity extraction, chatbot response evaluation, and alignment metric computation is available in the project repository: https://github.com/dilnazakhmetzhan/medical_protocols_thesis/.

Appendix A

Appendix

The full set of clinical evaluation questions is presented in Table A.1. Each question was mapped to a corresponding national clinical protocol used as the reference source for alignment evaluation.

Table A.1: Evaluation Question Set

ID	Stratum	Question (EN)	Protocol Source
A1	Diagnosis & Criteria	A pregnant woman at 34 weeks presents with BP 168/112 mmHg and proteinuria. Does this meet the criteria for severe preeclampsia and what additional features confirm the diagnosis?	Hypertensive disorders in pregnancy
A2	Diagnosis & Criteria	A pregnant patient undergoes a 75-g OGTT at 26 weeks. Results: fasting 5.2, 1-hr 10.1, 2-hr 8.7 mmol/L. Do these values meet the diagnostic criteria for GDM?	Diabetes mellitus in pregnancy
A3	Diagnosis & Criteria	What clinical and laboratory criteria define septic shock in obstetric patients, and what scoring systems are recommended?	Obstetric sepsis
B1	Pharmacotherapy	What is the recommended first-line antihypertensive therapy for severe preeclampsia, and what are the specific dosages?	Hypertensive disorders in pregnancy

(continued from previous page)

ID	Stratum	Question (EN)	Protocol Source
B2	Pharmacotherapy	What empirical antibiotic regimen is recommended for obstetric sepsis, and within what timeframe should it be initiated?	Obstetric sepsis
B3	Pharmacotherapy	What tocolytic agents are recommended for preterm labour, and what are the contraindications to their use?	Preterm labour
C1	Surgical & Emergency	What are the stepwise surgical interventions recommended for PPH that does not respond to uterotonic therapy?	Postpartum haemorrhage
C2	Surgical & Emergency	What are the indications for emergency hysterectomy in obstetric patients according to clinical guidelines?	Postpartum haemorrhage
C3	Surgical & Emergency	What are the clinical signs and immediate management steps for suspected uterine rupture during labour?	Abnormalities of labour
D1	Monitoring & Screening	A pregnant woman attends her first antenatal visit at 10 weeks. What screening tests should be performed in the first trimester and at what gestational weeks?	Antenatal care
D2	Monitoring & Screening	A GDM patient measures fasting glucose 5.6 mmol/L at home. Is this within recommended glycaemic targets and what management adjustments may be required?	Diabetes mellitus in pregnancy
D3	Monitoring & Screening	What risk factors identify pregnant women who should receive progesterone prophylaxis for preterm birth prevention?	Preterm labour
E1	Infection & Sepsis	What are the diagnostic criteria for chorioamnionitis and what is the recommended antibiotic protocol?	Chorioamnionitis
E2	Infection & Sepsis	What monitoring parameters should be assessed hourly in a patient with obstetric sepsis managed in the ICU?	Obstetric sepsis
E3	Infection & Sepsis	What are the criteria for diagnosing postpartum endometritis and what is the first-line treatment regimen?	Postpartum endometritis

Note: The Russian versions of the same questions were used for the cross-lingual robustness experiment.

Bibliography

- [1] Bilal Abu-Salih, Mohammed Al-Qurishi, Mohammed Alweshah, Mohammed Al-Smadi, Reem Alfayez, and Hiba Saadeh. Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities. *Journal of Big Data*, 10(1):81, 2023.
- [2] Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.
- [3] Sahaj Bedi, Hanyu Cui, Maria Fuentes, Alex Unell, Michael Wornow, Juan M. Banda, Niraj Kotecha, Tim Keyes, Yifan Mai, Mehmet Oez, and Han Qiu. Med-helm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.
- [4] Eva Chalas. The american college of obstetricians and gynecologists in 2020: A clear vision for the future. *Obstetrics & Gynecology*, 135(6):1251–1254, 2020.
- [5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, 2020.
- [6] Jenny Copara, Julien Knafo, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. Contextualized french language models for biomedical named entity recognition. In *Proceedings of the DEFT Workshop (JEP-TALN-RECITAL)*, pages 36–48, 2020.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.
- [8] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024.
- [9] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (ACL), pages 878–891, 2022.

- [10] Hemad Fetrati, Gerry Chan, and Rita Orji. Chatbots for sexual health improvement: a systematic review. *International Journal of Human-Computer Interaction*, 41(4):1997–2019, 2025.
- [11] Andrés F. Giraldo-Forero, Maria C. Durango, Santiago Rúa, Ever A. Torres-Silva, Sara Arango-Valencia, José F. Florez-Arango, and Andrés Orozco-Duque. Towards early maternal morbidity risk identification by concept extraction from clinical notes in spanish using fine-tuned transformer-based models. *Applied System Innovation*, 8(3):78, 2025.
- [12] Aron Henriksson, Maria Kvist, Hercules Dalianis, and Martin Duneld. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of Biomedical Informatics*, 57:333–349, 2015.
- [13] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.
- [14] Hyun-Kyoung Kim. The effects of artificial intelligence chatbots on women’s health: A systematic review and meta-analysis. *Healthcare*, 12(5):534, 2024.
- [15] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.
- [16] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *ACL*, 2020.
- [17] Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y. S. Lau, and Enrico Coiera. Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018.
- [18] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, and Sebastian Riedel. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.

- [20] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.
- [21] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2014–2037, 2023.
- [22] Shirish Nayyar. Evaluating the role of artificial intelligence in breast self-examination. *JMIR Formative Research*, 9:e82550, 2025.
- [23] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pages 248–260. PMLR, 2022.
- [24] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.
- [25] Sunil Kumar Sahu and Ashish Anand. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of Biomedical Informatics*, 86:15–24, 2018.
- [26] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, 2013.
- [27] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Philip Payne, Oshani Seneviratne, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [28] Sunghwan Sohn, Cheryl Clark, Scott R. Halgrim, Sean P. Murphy, Christopher G. Chute, and Hongfang Liu. Medrxn: An open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association*, 21(5):858–865, 2014.
- [29] A. R. Vishal, A. S. Harshitha, A. V. Sindhu, P. Mb, and S. Madhukumar. Assessing chatgpt’s accuracy and reliability in medical education: A cross-sectional study. *Turkish Journal of Public Health*, 23(1):11–17, 2025.
- [30] Jing Wang, Hewei Min, Tao Li, Jiaheng Li, Yang Jiang, Jingbo Zhang, Yibo Wu, and Xinying Sun. Women’s preferences and willingness to pay for ai chatbots in women’s health: Discrete choice experiment study. *Journal of Medical Internet Research*, 27:e67303, 2025.

- [31] World Health Organization. Who handbook for guideline development, 2014.
- [32] World Health Organization. Ethics and governance of artificial intelligence for health, 2021.
- [33] World Health Organization. Global strategy on digital health 2020–2025, 2021.
- [34] Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. Clinical named entity recognition using deep learning models. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1812, 2018.