STEAD-PE: A Structured Static-Dynamic PE Framework for Arabic Medical Response Generation

AHMED ALANSARY, Computer Science, October University for Modern Sciences and Arts, 6th of October City, Egypt
MOLHAM MAMDOUH, Computer Science, October University for Modern Sciences and Arts, 6th of October City, Egypt
MOHAMMAD MOFTAH, Computer Science, October University for Modern Sciences and Arts, 6th of October City, Egypt
ALI HAMDI, Computer Science, October University for Modern Sciences and Arts, 6th of October City, Egypt
KHALED SHABAN, Qatar University, Doha, Qatar

Medical response generation in low-resource and linguistically complex languages such as Arabic requires precise, context-aware advice, making Prompt Engineering a critical mechanism for controlling large language model behavior and improving output quality. The demand for scalable and reliable Arabic medical support systems has increased interest in prompt-based controllability, particularly where model fine-tuning and large annotated datasets are impractical. However, existing LLMs often show limited clinical relevance and semantic alignment when prompted with unstructured, patient-authored Arabic medical queries. In this work, we propose **ST**ructur**E**d st**A**tic-**D**ynamic PE (STEAD-PE), a prompt-only framework for Arabic medical response generation. STEAD-PE operates without model fine-tuning or parameter updates and leverages expert-curated Arabic medical posts paired with clinically validated treatment plans. The framework introduces case-specific structured metadata Domain, Severity, and Diagnosis—dynamically injected at inference time to enable clinically grounded and context-aware responses. We evaluate STEAD-PE across thirteen state-of-the-art LLMs, including Arabic-specialized, multilingual, and general-purpose models, using BERTScore for semantic similarity. Results show consistent and monotonic improvements in semantic and clinical relevance across all models, with STEAD-PE (Dynamic) achieving the highest BERTScore-measured semantic alignment in every case. Among evaluated models, LFM2-350M attains the best overall BERTScore (86.50%), while Qwen-32B improves from 80.89% (Zero-Shot) to 84.90% (Static) and 85.76% (Dynamic). These findings establish STEAD-PE as a robust, model-agnostic PE framework and highlight the importance of dynamically injected clinical context in reducing linguistic ambiguity and improving semantic fidelity in low-resource medical NLP settings.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; *Natural language processing*; Information extraction; • **Applied computing** → Health care information systems.

Additional Key Words and Phrases: Arabic Medical Text Generation, Prompt Engineering, Structured Dynamic Prompting, Structured Static Prompting, Natural Language Processing, Large Language Models

Authors' Contact Information: Ahmed Alansary, Computer Science, October University for Modern Sciences and Arts, 6th of October City, Egypt; e-mail: ahmed.mohamed406@msa.edu.eg; Molham Mamdouh, Computer Science, October University for Modern Sciences and Arts, 6th of October City, Giza Governorate, Egypt; e-mail: molham.mohamed@msa.edu.eg; Mohammad Moftah, Computer Science, October University for Modern Sciences and Arts, 6th of October City, Giza Governorate, Egypt; e-mail: mohammad.emad.cs@gmail.com; Ali Hamdi, Computer Science, October University for Modern Sciences and Arts, 6th of October City, Giza Governorate, Egypt; e-mail: ahamdi@msa.edu.eg; Khaled Shaban, Qatar University, Doha, Ad Dawhah, Qatar; e-mail: khaled.shaban@qu.edu.qa.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2375-4702/2026/10-ART
https://doi.org/10.1145/3859279

Just Accepted

1 Introduction

The rapid adoption of large language models (LLMs) across healthcare and clinical decision-support applications has elevated PE (PE) from a heuristic practice to a primary and systematic mechanism for controlling model behavior, reliability, and output quality. However, PE strategies are not universally transferable across languages or domains, creating a critical need for language- and domain-specific investigation, particularly in Arabic medical response generation. Arabic presents unique linguistic challenges, including rich morphology, complex word inflections, extensive dialectal variation, and informal patient-authored symptom narratives, all of which complicate semantic interpretation and effective prompt formulation. These challenges, coupled with the high-stakes nature of medical advice, underscore the importance of robust, context-aware PE frameworks tailored to Arabic medical settings.

Despite growing interest in Arabic clinical natural language processing (NLP), PE for Arabic medical applications remains underdeveloped and insufficiently standardized. The internal morphological complexity of Arabic, together with dialect-rich and colloquial symptom descriptions, often leads to inconsistent prompt performance and reduced clinical reliability [18]. Existing prompting methods frequently fail to capture latent medical semantics embedded in noisy or ambiguous patient narratives, resulting in factually incorrect, incomplete, or clinically vague responses [20]. Furthermore, the scarcity of high-quality, annotated Arabic medical datasets limits the systematic development, benchmarking, and evaluation of PE techniques [19]. Even when structured prompts are applied, LLMs remain vulnerable to hallucinations, unstable outputs, and sensitivity to minor lexical or structural prompt variations.[7, 20]. Recent studies further demonstrate that PE approaches primarily developed for English medical corpora do not generalize well to low-resource and linguistically complex settings such as Arabic, emphasizing the need for adaptive and context-aware prompting strategies [15]. Compounding these challenges is the lack of standardized evaluation frameworks for assessing Arabic medical response generation [22].

In this work, we investigate how structured prompt design directly influences semantic fidelity and clinical relevance in Arabic medical response generation, with a particular focus on low-resource scenarios where model fine-tuning is impractical or infeasible. Current LLMs often struggle to generate coherent, contextually appropriate, and medically grounded treatment plans when presented with informal patient-authored Arabic queries. Moreover, systematic comparisons between static and dynamic prompting strategies in Arabic medical contexts remain largely absent from the literature. To address this gap, we propose **ST**ructur**E**d st**A**tic-**D**ynamic PE (STEAD-PE), a prompt-only framework that examines how progressively enriching prompt structure and contextual specificity affects response quality.

The STEAD-PE framework evaluates three prompting configurations: STEAD-PE (Zero-Shot), which relies solely on raw patient symptom text; STEAD-PE (Static), which introduces a fixed instructional template with role specification and few-shot exemplars; and STEAD-PE (Dynamic), which further injects case-specific structured clinical metadata, namely Domain, Severity, and Diagnosis, directly into the prompt. Unlike prior work, STEAD-PE does not modify model parameters or training data, enabling fair and controlled comparisons across diverse LLM architectures. Figure 1 presents a high-level overview of the STEAD-PE framework, illustrating the systematic progression from Zero-Shot prompting to Static and Dynamic prompting, as well as the integration of structured clinical metadata during prompt construction process to enable adaptive, case-aware medical response generation.

We conduct a comprehensive evaluation of STEAD-PE across thirteen state-of-the-art LLMs, including Arabic-specialized, multilingual, and general-purpose models, using a Post-Plan Arabic medical dataset with expert-authored treatment recommendations. Performance is measured using BERTScore to quantify semantic alignment between generated responses and reference plans. Experimental results demonstrate that STEAD-PE yields consistent and monotonic improvements as prompt structure becomes more informative, with STEAD-PE (Dynamic) achieving the highest semantic similarity across all evaluated models. These findings validate the

Just Accepted

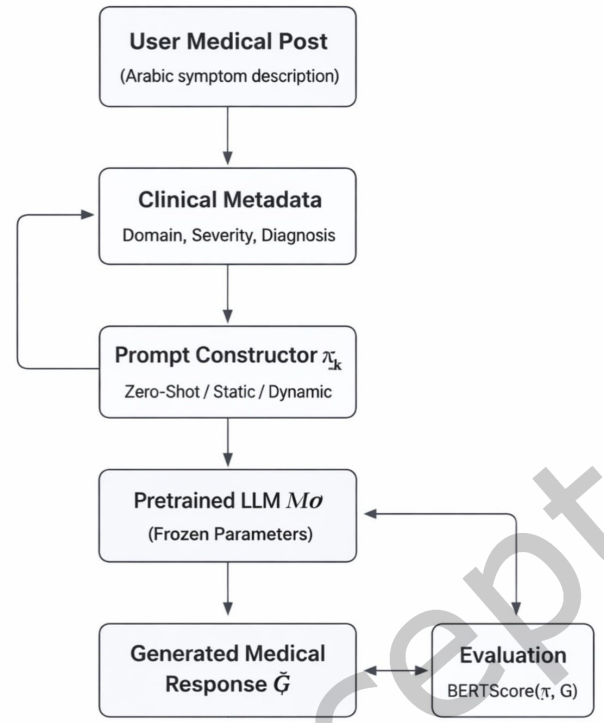


Fig. 1. High-level overview of the STEAD-PE framework for Arabic medical response generation

Just Accepted

effectiveness of structured and dynamically enriched PE as a robust, model-agnostic solution for Arabic medical response generation.

The main contributions of this work can be summarized as follows:

(1) We introduce STEAD-PE, a structured static-dynamic PE framework designed specifically for Arabic medical response generation, operating entirely in an inference-only setting without model fine-tuning.
(2) We propose the integration of clinically meaningful structured metadata—Domain, Severity, and Diagnosis—as dynamic prompt components, enabling adaptive, case-aware response generation.
(3) We conduct a large-scale, controlled evaluation across thirteen diverse LLMs, demonstrating consistent and monotonic performance gains from Zero-Shot to Static to Dynamic prompting.
(4) We provide a formal mathematical and information-theoretic formulation of static and dynamic prompting, offering theoretical insight into how structured clinical context reduces uncertainty and improves semantic alignment.
(5) We construct and utilize a context-enriched Arabic medical Post-Plan dataset that supports systematic benchmarking of PE strategies in low-resource medical NLP settings.

The remainder of this article is organized as follows. Section II reviews related work in Arabic medical NLP, medical text generation, and PE for LLMs. Section III describes the Arabic medical dataset used in this study, including its structure, annotations, and clinical attributes. Section IV presents the proposed STEAD-PE framework, detailing the prompting strategies, formal mathematical formulation, and the LLMs under evaluation. Section V reports and analyzes the experimental results across all models and prompting configurations. Finally,

Section VI concludes the paper by summarizing the key findings, discussing limitations, and outlining directions for future research.

2 Related Work

Research in medical language technologies has progressed across several interrelated directions, including Arabic clinical NLP, medical text generation, multilingual and domain-adapted LLMs, and PE for clinical reasoning. Early studies consistently emphasized the linguistic complexity of Arabic medical narratives and the limited availability of high-quality annotated resources, which remain major barriers to progress in Arabic medical NLP [2, 18]. Subsequent work explored transformer-based architectures, transfer learning, and data augmentation techniques to improve Arabic medical text classification and generation [3, 19]. More recent efforts introduced multilingual and domain-adapted medical LLMs, such as mT5-based models and MedSyn, aiming to enhance cross-lingual transferability and domain robustness [6, 13].

PE has emerged as a central paradigm for improving controllability, reasoning quality, and contextual alignment in LLM-based medical applications. Prior studies demonstrate that zero-shot, few-shot, and structured prompts can significantly influence diagnostic reasoning, summarization accuracy, and treatment recommendation quality [7, 21]. Complementary research has focused on evaluation frameworks for medical LLMs, particularly in sensitive domains such as mental healthcare, regulatory medical writing, and clinical summarization, where output reliability and factual correctness are critical [5, 16, 17]. While these works provide important methodological insights, they largely focus on English-language datasets and high-resource settings, limiting their applicability to Arabic medical contexts.

Despite these advances, Arabic medical text generation continues to face persistent challenges. Patient-authored medical posts are often informal, dialect-rich, and syntactically diverse, complicating both modeling and prompt design [18]. Even when leveraging web-scraped corpora and synthetic data augmentation, existing Arabic medical datasets remain imbalanced, inconsistently labeled, and sparsely annotated with clinically meaningful attributes [2]. In particular, the absence of explicit severity indicators, diagnostic summaries, and domain labels limits the effectiveness of context-aware and adaptive prompting strategies.

Medical text generation research has increasingly adopted synthetic data generation and multitask prompt-response corpora to address data scarcity [6, 27]. However, synthetic medical texts often suffer from semantic drift, hallucinated content, and reduced clinical validity, especially for rare conditions or region-specific treatment protocols [6]. Furthermore, models trained or evaluated primarily on synthetic data may inherit biases and factual inaccuracies, undermining their reliability in real-world clinical settings [27].

Multilingual and domain-adapted LLMs, including mT5 and large cross-lingual medical transformers, demonstrate strong transfer learning capabilities [13]. Nevertheless, their performance in Arabic medical contexts is constrained by cross-lingual interference, limited Arabic medical data during pretraining, and insufficient alignment with regional dialects and local clinical practices. Even specialized systems, such as MedSyn [6] and regulatory-text-focused LLMs [17], exhibit difficulties handling informal patient narratives and region-specific terminology.

While the effectiveness of PE for improving LLM controllability and inference quality is well established [14, 22], its application to Arabic medical response generation remains underexplored. Structured prompts can reduce ambiguity, yet they remain highly sensitive to template design and minor lexical variations [7]. Moreover, challenges such as hallucinations, illogical reasoning, and failure to infer implicit clinical meaning from noisy patient-authored text persist even under structured prompting regimes [20]. Recent work further highlights that PE strategies developed for English medical datasets do not generalize well to low-resource and linguistically complex languages, reinforcing the need for adaptive and context-aware prompting frameworks tailored to Arabic medical discourse [15, 21].

Just Accepted

A small number of concurrent efforts have begun to address prompt-engineering-based approaches to Arabic medical NLP. The MedArabiQ benchmark [1] evaluates LLMs on Arabic medical question-answering tasks spanning multiple clinical domains, while a related AraHealthQA shared-task submission [26] demonstrates that prompt engineering and ensemble-based prompting, without fine-tuning, can achieve competitive performance on Arabic clinical multiple-choice and open-ended question answering. These efforts corroborate the broader premise of the present work that structured prompting is an effective, fine-tuning-free strategy for Arabic clinical NLP but target discrete question-answering and classification formats rather than open-ended, case-specific treatment-recommendation generation, and do not incorporate severity- or diagnosis-aware structured metadata injection. STEAD-PE is, to the best of our knowledge, the first framework to formalize and empirically evaluate a rogressive static-dynamic prompting structure specifically for free-text Arabic medical response generation conditioned on case-specific clinical attributes.

Finally, evaluation methodologies for medical LLMs have evolved to include human-in-the-loop assessments, rubric-based scoring, and large-scale conversational benchmarks [5, 17]. However, such frameworks are rarely applied to Arabic medical text generation, often lack standardization, and fail to directly assess treatment appropriateness and clinical alignment. Moreover, many evaluations rely on exam-style or synthetic questions rather than authentic patient narratives, limiting their ecological validity [16]. These limitations motivate the need for a structured, prompt-centric evaluation framework, such as STEAD-PE, that enables controlled comparison of static and dynamic prompting strategies in Arabic medical contexts.

3 Dataset

The dataset used in this study consists of paired Arabic medical symptom descriptions and corresponding expert-authored treatment recommendations, structured into two primary fields: Post and Plan. Each Post represents a patient-authored medical query written in Arabic, while each Plan provides a clinically validated response authored by certified medical advisors. This pairing enables direct semantic comparison between generated and reference medical responses, which is essential for evaluating the effectiveness of the proposed STEAD-PE framework.

The dataset originates from a previously curated Arabic medical corpus collected from user-generated posts on an online social health platform, where patients shared medical complaints across a wide range of conditions [19]. The corpus spans multiple condition categories, including chronic diseases, dermatological conditions, and neurological symptoms, with the Severity attribute distributed across mild and severe cases in a broadly balanced proportion [19]. For the purposes of this study, the relevant subset of Post–Plan pairs annotated with Domain, Severity, and Diagnosis attributes was used; full size and per-class distribution statistics are reported in the original dataset description [19].

The Post entries are collected from Arabic-language medical discussion platforms and community-based question-answer forums, where users seek medical guidance across a diverse range of health concerns. These posts exhibit substantial linguistic variability, including dialectal expressions, informal phrasing, incomplete symptom descriptions, and non-standard medical terminology. In addition to symptom narratives, many posts include contextual details such as symptom duration, patient age, gender, and references to chronic conditions, reflecting the noisy and heterogeneous nature of real-world Arabic patient discourse.

The corresponding Plan entries consist of expert-authored treatment recommendations that are clinically grounded, medically coherent, and aligned with standard care practices. Each Plan is written to directly address the symptoms and contextual information described in the associated Post, serving as a reliable reference for semantic and clinical evaluation. Importantly, the Plan responses are intended for informational support rather than definitive diagnosis, ensuring ethical and medico-legal appropriateness.

Just Accepted

To support structured and dynamic prompt construction within STEAD-PE, each dataset instance is further annotated with three clinically meaningful attributes:

- **Domain:** the medical specialty most relevant to the case (e.g., dermatology, cardiology, otolaryngology);
- **Severity:** an ordinal assessment of symptom urgency, categorized as mild, moderate, or severe;
- **Diagnosis:** a high-level clinical interpretation summarizing the most likely medical condition inferred from the reported symptoms.

These annotations are assigned during the Plan creation process under the supervision of a qualified medical advisor, who oversaw the structuring and labelling of Domain, Severity, and Diagnosis to ensure clinical relevance and accuracy [19]. Severity is assigned according to an ordinal urgency scale (mild, moderate, severe) based on the symptom description and contextual details (e.g., duration, chronicity) reported in the Post, while Diagnosis is derived from the medical advisor's clinical interpretation of the reported symptom profile. As annotation was conducted under single-advisor medical supervision rather than by multiple independent annotators, no formal inter-annotator agreement score was computed; we acknowledge this as a limitation of the underlying dataset and identify the quantification of label consistency across independent annotators as a direction for future work. Unlike free-text symptom descriptions, Domain, Severity, and Diagnosis provide explicit and structured clinical signals that are not always reliably inferable from patient-authored text only. Within STEAD-PE (Dynamic), these attributes are programmatically injected into predefined prompt placeholders, enabling case-specific contextual adaptation at inference time.

Table 1 presents representative examples of Post-Plan pairs, along with their associated structured annotations. Collectively, these components form a context-enriched Arabic medical dataset designed to support controlled experimentation with static and dynamic PE strategies. The dataset's structure enables systematic analysis of how incremental prompt enrichment influences semantic alignment and clinical relevance, independent of model fine-tuning or parameter updates.

4 Methodology

In this study, we systematically examine how prompt structure and contextual enrichment influence Arabic medical response generation using the proposed STEAD-PE framework. All experiments are conducted in a strictly inference-only setting, where no model fine-tuning, parameter updates, or additional training data are used. This experimental design ensures that all observed performance differences are attributed solely to PE strategies, rather than to model adaptation effects. Multiple LLMs are evaluated under identical experimental conditions, enabling fair and controlled comparisons across architectures and parameter scales.

The evaluation is performed on a dataset of paired Arabic Post symptom descriptions and expert-authored Plan treatment recommendations, as described in Section 3. To quantify semantic alignment between generated and reference responses, we employ BERTScore, which is well suited for contextual, semantically rich, and lexically diverse medical text comparison, particularly in Arabic.

4.1 Overview of the STEAD-PE Framework

The STEAD-PE framework is designed to evaluate how progressively enriching prompt structure affects the quality, relevance, and stability of generated Arabic medical responses. As illustrated in Figure 1, STEAD-PE defines a three-stage prompting progression, moving from unstructured input to increasingly constrained and context-aware prompts:

- **STEAD-PE (Zero-Shot):** raw patient symptom text only;
- **STEAD-PE (Static):** a fixed instructional template with explicit role specification and few-shot exemplars;
- **STEAD-PE (Dynamic):** the static template augmented with case-specific structured clinical metadata.

Just Accepted

Table 1. Examples of patient post entries and their corresponding expert plan (Arabic, English translation, and Roman transliteration for each cell).

Post	Plan	Domain	Severity	Diagnosis
لو سمحتم زوجي الكوليسترول عالي و عمل رسم قلب بالمجهود وطلع كويس هل فيه اهميه للقسطرة الاستكشافيه	لو مازال موجودا احتمالية قصور بالشريان التاجي، فلا مانع من عمل القسطرة	قلب و صدر	متوسط	كوليسترول عالي
My husband has high cholesterol and underwent a cardiac stress test that was normal. Is diagnostic cardiac catheterization necessary?	If symptoms persist and coronary artery insufficiency is possible, diagnostic catheterization may be considered.	Cardiology / Chest	Moderate	High cholesterol
Law samahtum, zawgi ᶜinduhu kolesterol ᶜāli, wa ᶜamal rasm qalb bil-maghūd wa ṭalaᶜ kwayyis. Hal fī ahammiyya lil-qaṭṭara al-istikshāfiyya?	Law mā zāl mawjudan iḥtimāliyya quṣūr bil-shiryān al-tājī, fa-lā māniᶜ min ᶜamal al-qaṭṭara.	Qalb wa ṣadr	Mutawassiṭ	Kolesterol ᶜāli
بنت عندها ٩ سنوات بقالها فترة بتشرب مياة كتير وتتدخل الحمام كتير عاوزة اعرف المفروض اعمل تحاليل ايه واكشف تخصص ايه	د. باطنة أطفال في الأول لاختبار السكر. إن لم يكن فدكتور مسالك بولية.	باطنة و أطفال	متوسط	باطنة
A 9-year-old girl has been drinking a lot of water and urinating frequently. What tests should be done and which specialty should she visit?	Start with a pediatric internal medicine doctor to test blood sugar. If not related, see a urologist.	Internal Medicine / Pediatrics	Moderate	Internal medicine
Bint ᶜindahā tisᶜ sinīn, baʾālaha fatra bitishrab mayya kitīr wa bitidkhol il-ḥammām kitīr. ᶜāwza aᶜraf aᶜmil taḥālīl ēh wa akshif takhaṣṣuṣ ēh.	Duktūr bāṭina aṭfāl fil-awwal li-ikhtibār il-sukkar. In lam yakun, fa-duktūr masālik bawliyya.	Bāṭina wa aṭfāl	Mutawassiṭ	Bāṭina
السلام عليكم لو عضني فأر وخرج دم ليس بالكثير ماذا عليّ أن أفعل	روح على مستشفى وخذ إبرة تيتانوس	طوارئ	حرج	حمّى عضة فأر
Peace be upon you. If a rat bit me and there was slight bleeding, what should I do?	Go to a hospital and get a tetanus shot.	Emergency	Critical	Rat-bite fever
As-salāmu ᶜalaykum. Law ᶜaḍḍanī faʾr wa kharag dam mish kitīr, mādhā ᶜalayya an afᶜal?	Rūḥ ᶜalā mustashfā wa khudh ibra tītānūs.	Ṭawāriʾ	Ḥarij	Ḥummā ᶜaḍḍit faʾr
في كلكيعة صغيرة تحت الجلد ورا الأذن الشمال على العظمة كده ظهرت فجأة ولما اضغط عليها أحس بألم، ايه السبب؟	في التهاب في الأذن أو الحلق. نروح نكشف، ممكن تكون الغدة الليمفاوية ملتهبة.	غدد صماء	حرج	غدة ليمفاوية
There is a small lump under the skin behind the left ear on the bone. It appeared suddenly and hurts when pressed. What is the cause?	There may be inflammation in the ear or throat. Get examined; it could be an inflamed lymph node.	Endocrinology	Critical	Lymph node
Fī kalkīᶜa ṣughayyara taḥt il-gild wara il-udn il-shimāl ᶜala il-ᶜaẓma. Ẓahrit fagaʾa wa lammā aḍghaṭ ᶜalayhā aḥiss bi-alam. Ēh is-sabab?	Fī iltihāb fil-udn aw il-ḥalq. Nroh nikshif, mumkin tkon il-ghudda il-līmfāwiyya multahiba.	Ghudad ṣammāʾ	Ḥarij	Ghadda līmfāwiyya

Just Accepted

This progression enables systematic isolation of the contribution of structural guidance and dynamic clinical context, while keeping the underlying model parameters fixed.

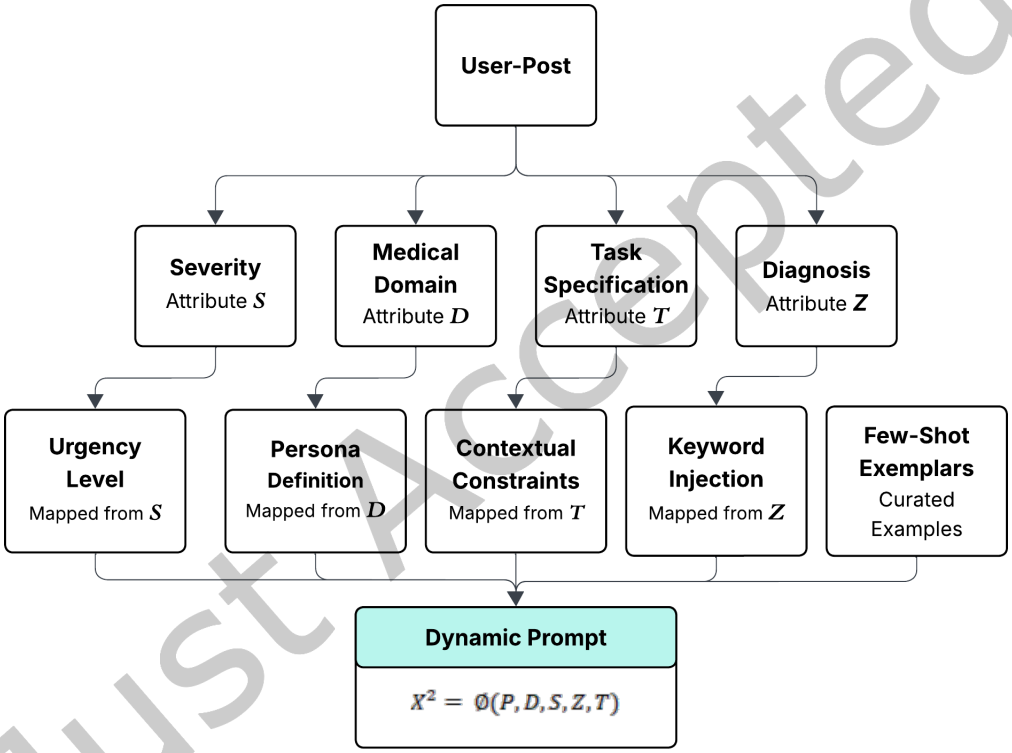
4.2 Prompting Strategies

4.2.1 ***STEAD-PE (Zero-Shot)***. In the STEAD-PE (Zero-Shot) configuration, the input to the LLM consists exclusively of the raw Post text, with no role definition, formatting constraints, or clinical guidance. This configuration serves as a baseline, reflecting each model's intrinsic ability to interpret informal, dialect-rich Arabic medical queries.

4.2.2 ***STEAD-PE (Static).*** The STEAD-PE (Static) configuration introduces a fixed instructional prompt template that remains constant across all dataset instances. This template:

- Assigns the model a clear medical advisory role;
- Specifies response style, tone, and structure;
- Includes few-shot examplars demonstrating the transformation from informal patient narratives to consise, clinically grounded recommendations.

By enforcing consistent role definition, reasoning constraints, and output formatting, the Static prompt reduces ambiguity and stabilizes model behavior, leading to more interpretable and semantically aligned responses compared to Zero-Shot prompting.



Just Accepted

Fig. 2. The design of the dynamic prompt, showing how the fields *Domain*, *Severity*, and *Diagnosis* are extracted from the dataset and inserted into the prompt template.

4.2.3 ***STEAD-PE (Dynamic).*** The STEAD-PE (Dynamic) configuration extends the Static template by injecting structured, case-specific clinical metadata directly into predefined prompt placeholders. Specifically, the following attributes are incorporated:

- `{Domain}`: relevant medical specialty;
- `{Severity}`: urgency level (mild, moderate, severe);

- `{Task}`: a contextual constraint specifying the clinical task type required of the response (e.g., diagnostic reasoning, treatment planning, or triage advice).
- `{Diagnosis}`: high-level clinical interpretation of the case.

These attributes are automatically extracted from the dataset annotations described in Section 3 and integrated into the prompt at inference time. The Dynamic prompt retains the same role definition and few-shot exemplars as the Static configuration, ensuring that any observed performance gains arise exclusively from the added structured context.

Figure 2 illustrates the design of the Dynamic prompt and the data-driven injection process.

4.3 Formal Modeling of STEAD-PE Dynamic Prompting

We formalize Arabic medical response generation under STEAD-PE as a conditional text generation problem with structured contextual constraints.

Let $\mathcal{V}$ denote the Arabic vocabulary and $\mathcal{Y} = \mathcal{V}^*$ the space of possible Arabic medical texts. Each dataset instance is modeled as a random tuple:

$$(P, G, D, S, Z, T) \sim \mathcal{D} \tag{1}$$

where

- $P \in \mathcal{V}^*$ is a patient-authored Post,
- $G \in \mathcal{Y}$ is the expert-authored Plan,
- $D \in \mathcal{D}_{dom}$ denotes the Domain,
- $S \in \mathcal{D}_{sev}$ represents the Severity,
- $Z \in \mathcal{D}_{diag}$ corresponds to Diagnosis.
- $T \in \mathcal{D}_{task}$ denotes the Task Specification, the contextual constraint identifying the type of clinical response required.

The objective is to generate a response $\hat{G}$ that maximizes semantic alignment with the reference plan G.

4.3.1 Prompt as a Conditioning Operator. Let $\mathcal{M}_\theta$ denote a pretrained LLM parameterized by θ, defining the conditional distribution:

$$p_\theta(G \mid X) \tag{2}$$

where X represents the constructed prompt.

PE is modeled as a deterministic conditioning operator:

$$\pi_k : (P, D, S, Z, T) \rightarrow X^{(k)} \tag{3}$$

with $k \in \{0, 1, 2\}$ corresponding to Zero-Shot, Static, and Dynamic prompting, respectively.

4.3.2 Prompting Strategies (Formalization).

Zero-Shot Prompting.

$$X^{(0)} = P \tag{4}$$

yielding:

$$p_\theta(G \mid P) \tag{5}$$

Just Accepted

Static Prompting.

$$X^{(1)} = \phi(P) \tag{6}$$

where $\phi(\cdot)$ is a fixed instructional template incorporating role definition and few-shot demonstrations.

Dynamic Prompting.

$$X^{(2)} = \phi(P, D, S, Z, T) \tag{7}$$

where structured clinical metadata are injected into predefined placeholders.

4.3.3 Information-Theoretic Interpretation. Dynamic prompting reduces uncertainty in the generated response by conditioning on additional medically relevant variables. This effect can be expressed as a reduction in conditional entropy:

$$H(G \mid P) \geq H(G \mid P, D, S, Z, T) \tag{8}$$

The information gain introduced by dynamic prompting is given by:

$$\mathcal{I}(G; D, S, Z, T \mid P) = H(G \mid P) - H(G \mid P, D, S, Z, T) \tag{9}$$

A positive mutual information value indicates that structured clinical metadata contribute explanatory power beyond raw symptom text.

4.3.4 Generation Process. The generated medical response under prompting strategy k is given by:

$$\hat{G}^{(k)} = \arg\max_{G} p_\theta(G \mid X^{(k)}) \tag{10}$$

Since model parameters θ remain fixed, performance differences arise exclusively from variations in prompt construction.

4.3.5 Evaluation as Expected Semantic Similarity. Let $\mathcal{S}(\cdot, \cdot)$ denote the BERTScore similarity function. The expected performance under strategy k is defined as:

$$\mathbb{E}_{(P,G)\sim\mathcal{D}}\left[\mathcal{S}(\hat{G}^{(k)}, G)\right] \tag{11}$$

The empirical estimate over N samples is:

$$\overline{\mathcal{S}}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{S}(\hat{G}_i^{(k)}, G_i) \tag{12}$$

Just Accepted

4.3.6 Marginal Gain Analysis. The contribution of structured prompting is quantified as:

$$\Delta_{\text{static}} = \overline{\mathcal{S}}^{(1)} - \overline{\mathcal{S}}^{(0)} \tag{13}$$

$$\Delta_{\text{dynamic}} = \overline{\mathcal{S}}^{(2)} - \overline{\mathcal{S}}^{(1)} \tag{14}$$

A strictly positive Δ_{dynamic} supports the hypothesis that structured clinical metadata improves semantic accuracy in Arabic medical response generation.

4.4 Experiment Setup

For each Post instance, all three prompting strategies are applied independently across all evaluated LLMs. Each run produces a generated Plan, which is then compared to the expert-authored reference using BERTScore. This design ensures a fair, controlled comparison of prompting strategies, while isolating the effect of prompt structure and contextual enrichment.

To ensure experimental consistency across all evaluated LLMs, the same inference hyperparameters were used throughout all experiments. Specifically, the temperature was set to 0.6, top-p was configured to 0.95, and the maximum number of generated tokens (`max_new_tokens`) was fixed at 100.

4.5 LLMs Under Evaluation

To assess the robustness and generalizability of STEAD-PE, experiments are conducted across thirteen state-of-the-art LLMs that vary in architecture, parameter scale, training objectives, and language coverage. The selected models include general-purpose, multilingual, and Arabic-specialized systems, enabling a comprehensive assessment of prompt responsiveness across diverse model families. Furthermore, the evaluated models span a wide range of sizes, from lightweight instruction-tuned systems to large-scale foundation models, enabling analysis of whether structured and dynamic PE benefits persist across different model capacities. Specifically, the evaluation includes Gemini-2.5-Flash [10], DeepSeek-V3 [11], Llama3.1-8B and Llama3.3-70B [25], Qwen-32B and Qwen2.5-7B [8], GPT-OSS-20B and GPT-OSS-120B [9], Kimi-K2 [4], Silma-9B-Instruct [24], c4ai-command-r7b-arabic [12], and LFM2 [23].

All models are accessed in inference-only mode, with no fine-tuning, parameter updates, or additional training performed. Each model was queried using identical prompt templates and input data under all STEAD-PE configurations. This ensures that observed performance differences are attributable solely to prompt design rather than model adaptation or training effects.

The inclusion of both Arabic-specialized and multilingual models allows for a systematic comparison of how language coverage and pretraining data influence responsiveness to structured and metadata-enriched prompts. This setup further enables an examination of whether STEAD-PE can compensate for limitations in Arabic medical knowledge within general-purpose LLMs.

Just Accepted

Table 2. Comparison of Zero-shot, Static, and Dynamic prompting strategies using a representative Arabic medical post. For each strategy, the table presents the prompt formulation and generated response in Arabic, along with the corresponding English translation and Roman transliteration. This comparison illustrates how increasing prompt structure and the injection of structured clinical metadata influence response clarity, clinical grounding, and semantic alignment.

Zero-shot	Static	Dynamic
Arabic: بعد ما قام من النوم مكانش قادر يمشي على رجله وعلى آخر اليوم ورمت شويه	**Arabic:** انت طبيب متخصص ودقيق في التشخيص. مهمتك هي قراءة المنشور بعناية وتحديد المشكلة الصحية، كن واضح ومختصر. أمثلة: ج: روح على س: لو عضني فأر وخرج دم ليس بالكثير ماذا علي أن أفعل؟ مستشفى خذ إبرة تيتانوس س: والدي ظهر عنده قرح فراش ويبرش في جسمه ويعوره، ازاي نعالجه؟ ج: شوفوا تحليل وظائف الكلى وظبطوا السكر أهم حاجة، ويدهن مكانها Fucidin، ومهم يتقلب كل شوية على السرير. س: في كلكيعة صغيرة تحت الجلد ورا الأذن الشمال ظهرت فجأة ولما ج: في التهاب في الأذن، نروح نكشف أضغط عليها أحس بألم ايه السبب؟ ممكن تكون الغدة الليمفاوية ملتهبة. بعد ما قام من النوم مكانش قادر يمشي على رجله وعلى آخر اليوم المنشور: ورمت شويه	**Arabic:** انت طبيب متخصص في مجال **عظام** ودقيق في التشخيص. مهمتك هي قراءة المنشور بعناية وتحديد المشكلة الصحية، كن واضح ومختصر. مع الأخذ في الاعتبار أن درجة خطورة المرض **غير حرجة**. مصطلحات هامة: **عظام**. أمثلة: ج: روح على س: لو عضني فأر وخرج دم ليس بالكثير ماذا علي أن أفعل؟ مستشفى خذ إبرة تيتانوس س: والدي ظهر عنده قرح فراش ويبرش في جسمه ويعوره، ازاي نعالجه؟ ج: شوفوا تحليل وظائف الكلى وظبطوا السكر أهم حاجة. س: في كلكيعة صغيرة تحت الجلد ورا الأذن الشمال ظهرت فجأة ولما ج: في التهاب في الأذن، نروح نكشف أضغط عليها أحس بألم ايه السبب؟ ممكن تكون الغدة الليمفاوية ملتهبة. بعد ما قام من النوم مكانش قادر يمشي على رجله وعلى آخر اليوم المنشور: ورمت شويه
English: After he woke up, he could not walk on his leg, and by the end of the day it became slightly swollen.	**English:** You are a specialist physician and accurate in diagnosis. Your task is to read the post carefully and identify the health problem; be clear and concise. Examples: Q: If a rat bit me and caused slight bleeding, what should I do? A: Go to a hospital and get a tetanus shot. Q: My father developed bedsores and has itching and pain—how can we treat it? A: Check kidney function tests and control blood sugar (most important), apply Fucidin, and change his position frequently. Q: A small lump under the skin behind the left ear appeared suddenly and hurts when pressed—what is the cause? A: There may be ear inflammation; get examined—could be an inflamed lymph node. Post: After he woke up, he could not walk on his leg, and by the end of the day it became slightly swollen.	**English:** You are a specialist physician in **Orthopedics** and accurate in diagnosis. Your task is to read the post carefully and identify the health problem; be clear and concise. Considering that the severity level is **Non-critical**. Key terms: **Orthopedics**. Examples: Q: If a rat bit me and caused slight bleeding, what should I do? A: Go to a hospital and get a tetanus shot. Q: My father developed pressure sores and has itching and pain—how can we treat it? A: Check kidney function tests and control blood sugar (most important). Q: A small lump under the skin behind the left ear appeared suddenly and hurts when pressed—what is the cause? A: There may be ear inflammation; get examined—could be an inflamed lymph node. Post: After he woke up, he could not walk on his leg, and by the end of the day it became slightly swollen.
Transliteration: Baᶜd mā qām min en-nōm, makānsh qādir yemshī ᶜalā riglu, wa ᶜalā ākhir en-nahār werrimit shwayya.	**Transliteration:** Enta ṭabīb mutakhaṣṣiṣ w daqīq fīt-tashkhīṣ. Muhimmitak hiya qirāᵓit el-manshūr bi-ᶜināya w taḥdīd el-mushkila eṣ-ṣiḥḥiyya; kun wāḍiḥ w mokhtaṣar. Amthila: S: Law ᶜaḍḍanī faᵓr w kharag dam lēs bil-katīr, mādhā ᶜalayya an afᶜal? G: Rūḥ ᶜalā mustashfā khud ibra tītānūs. S: Wālidī ẓahar ᶜinduh quraḥ firāsh w biyihirrish fī gismuh w biyyaᵓwarruh—izzāy niᶜāliguh? G: Shūfū taḥlīl waẓāyif el-kilya w ẓabṭū es-sukkar ahamm ḥāga, w yidhan makānhā Fucidin, w muhimm yitᵓallab kol shwayya ᶜalā es-sarīr. S: Fī kalkīᶜa ṣughayyara taḥt el-gild wara el-udn esh-shimāl ẓahrit fagaᵓa w lammā aḍghaṭ ᶜalayhā aḥiss bi-alam—ēh es-sabab? G: Fī iltihāb fī el-udn; nrūḥ n-kashif, mumkin tkon el-ghudda el-līm āwiyya multahiba. El-manshūr: Baᶜd mā qām min en-nōm, makānsh qādir yemshī ᶜalā riglu, wa ᶜalā ākhir en-nahār werrimit shwayya.	**Transliteration:** Enta ṭabīb mutakhaṣṣiṣ fī magal **ᶜiẓām** w daqīq fīt-tashkhīṣ. Muhimmitak hiya qirāᵓit el-manshūr bi-ᶜināya w taḥdīd el-mushkila eṣ-ṣiḥḥiyya; kun wāḍiḥ w mokhtaṣar. Maᶜa el-akhḍ fī el-iᶜtibār inna daragat khuṭūrat el-maraḍ **ghēr ḥarija**. Muṣṭalaḥāt hāmma: **ᶜiẓām**. Amthila: S: Law ᶜaḍḍanī faᵓr w kharag dam lēs bil-katīr, mādhā ᶜalayya an afᶜal? G: Rūḥ ᶜalā mustashfā khud ibra tītānūs. S: Wālidī ẓahar ᶜinduh quraḥ firāsh w biyihirrish fī gismuh w biyyaᵓwarruh—izzāy niᶜāliguh? G: Shūfū taḥlīl waẓāyif el-kilya w ẓabṭū es-sukkar ahamm ḥāga, w yidhan makānhā Fucidin, w muhimm yitᵓallab kol shwayya ᶜalā es-sarīr. S: Fī kalkīᶜa ṣughayyara taḥt el-gild wara el-udn esh-shimāl ẓahrit fagaᵓa w lammā aḍghaṭ ᶜalayhā aḥiss bi-alam—ēh es-sabab? G: Fī iltihāb fī el-udn; nrūḥ n-kashif, mumkin tkon el-ghudda el-līm āwiyya multahiba. El-manshūr: Baᶜd mā qām min en-nōm, makānsh qādir yemshī ᶜalā riglu, wa ᶜalā ākhir en-nahār werrimit shwayya.

Just Accepted

5 Results

This section reports a comprehensive evaluation of the STEAD-PE framework, analyzing how prompt structure and contextual enrichment affect Arabic medical response generation across multiple LLMs. Results are reported using BERTScore, which measures semantic similarity between generated responses and expert-authored reference plans, making it well suited for lexically diverse and context-rich Arabic medical text.

We first provide a qualitative comparison of responses generated under different prompting strategies, followed by a quantitative analysis across all evaluated models. We note that BERTScore improvements reported throughout this section indicate gains in semantic alignment with expert-authored reference Plans and should not, on their own, be interpreted as direct evidence of clinical correctness or safety; this distinction is discussed further in Section 7. Throughout this section, all models are evaluated in an inference-only setting, ensuring that observed performance differences arise exclusively from prompt design rather than model adaptation.

Table 3. Example instances of generated responses under the Zero-Shot, Dynamic prompting configurations and the ground truth with English translations, and Roman transliteration.

Question	Ground Truth	Zero-Shot Response	Dynamic Response
بعد ما قام من النوم مكانش قادر يمشي على رجله وعلى آخر اليوم ورمت شويه.	ممكن جراحة عظام أو جراحة عامة لتقييم الحالة وطلب التحاليل اللازمة.	قد يكون هناك خلل في القناة نصف الحلزونية أو الجهاز السمعي الداخلي، مما يؤدي إلي خلل في التوازن وصعوبة في المشي بعد النوم.	لازم يتم الكشف عند دكتور عظام عشان يتأكد من التشخيص من خلال الفحص السريري ويعض التحاليل المعملية زي حمض اليوريک ويبدأ العلاج المناسب.
ba'd ma 'am men cl-nom makansh ader yemshi 'ala regluh w 'ala akher cl-yom weremct shwayya.	momken gerahat 'ezam aw geraha 'amma lc-taqyim cl-hala w talab cl-tahalil cl-lazema.	qad yakun hunak khalal fi cl-qana ncsf cl-halazoniya aw cl-gchaz cl-sam'i cl-dakhili, mimma yu'addi ila khalal fi cl-tawazon wa su'uba fi cl-mashy ba'd cl-nom.	lazem yctem cl-kashf 'and doktor 'ezam 'ashan ycta'akkad men cl-tashkhis men khilal cl-fahs cl-scriri w ba'd cl-tahalil cl-ma'maliya zay hcmd cl-yorik we yabda' cl-'clag cl-monasib.
After waking up, he was unable to walk on his leg, and by the end of the day it became slightly swollen.	An orthopedic surgeon or a general surgeon may be needed to evaluate the condition and request the necessary tests.	There may be a problem in the semicircular canals or the inner ear system, which could lead to balance issues and difficulty walking after waking up.	The patient should be examined by an orthopedic doctor to confirm the diagnosis through a clinical examination and laboratory tests such as a uric acid test, then start the appropriate treatment.

5.1 Qualitative Analysis

Table 2 presents a representative qualitative comparison of medical responses generated using Zero-Shot, Static, and Dynamic prompting strategies. When provided only with raw patient-authored symptom text (Zero-Shot), models tend to produce generic, incomplete, or weakly grounded responses, often failing to infer symptom

Just Accepted

severity or provide structured guidance. Responses in this configuration frequently exhibit hedging language, missing precautionary advice, and limited clinical specificity.

Introducing a Static prompt leads to noticeable improvements in response organization, tone, and clinical coherence. The explicit role specification and few-shot exemplars encourage more structured recommendations, including basic precautionary measures and clearer follow-up guidance. However, Static prompting still relies solely on information inferable from the patient's informal narrative, which can result in ambiguous urgency assessment or overly conservative advice.

The STEAD-PE (Dynamic) configuration produces the most clinically grounded and context-aware responses. By injecting Domain, Severity, and Diagnosis metadata into the prompt, models consistently generate more targeted recommendations, including appropriate urgency cues, symptom prioritization, and clearer guidance on when to seek professional care. These qualitative observations suggest that structured clinical metadata provides critical contextual signals that are not reliably recoverable from noisy patient-authored text alone.

Table 3 presents representative examples comparing the ground truth with responses generated under the Zero-Shot and Dynamic prompting configurations using **LFM2-350M**. The Zero-Shot response demonstrates semantic drift, incorrectly associating the symptoms with an inner-ear balance disorder despite the presence of leg pain and swelling. In contrast, the Dynamic response aligns more closely with the expert-authored reference by correctly identifying the need for orthopedic evaluation and recommending clinically relevant follow-up procedures. These examples further illustrate how contextual enrichment improves the clinical relevance and semantic fidelity of generated Arabic medical responses.

Table 4. BERTScore of LLMs under the Zero-Shot setting.

Model	Zero-Shot BERTScore (%)
Gemini-2.5-Flash	81.56
DeepSeek-V3	80.80
Llama3.1-8B	82.41
Llama3.3-70B	81.90
Qwen2.5-7B	82.72
Qwen-32B	80.89
Gpt-oss-20B	79.98
Gpt-oss-120B	79.66
Kimi-K2	81.08
Silma-9B-Instruct	83.67
c4ai-command-r7b-arabic	83.65
LFM2-700M	85.46
LFM2-350M	**85.56**

Just Accepted

5.2 Zero-Shot Prompting Performance

Table 4 reports the baseline BERTScore results for all evaluated LLMs under the STEAD-PE (Zero-Shot) configuration. Overall performance in this setting is highly variable across models, reflecting differences in pretraining data, multilingual coverage, and intrinsic medical reasoning capability, with most BERTScore values ranging between 79% and 84%, indicating limited semantic alignment with expert medical recommendations. LFM2-350M achieves the highest Zero-Shot score (85.56%), suggesting stronger inherent instruction-following capabilities.

Despite reasonable fluency, Zero-Shot responses often suffer from semantic drift, where generated recommendations only partially align with the expert-authored plans. Models frequently omit important safety advice or misinterpret symptom severity, leading to lower semantic similarity scores. These results confirm that raw patient text alone provides insufficient structure for reliable Arabic medical response generation, particularly in informal and dialect-rich settings.

Table 5. BERTScore of LLMs under the Static prompting setting.

Model	Static BERTScore (%)
Gemini-2.5-Flash	84.90
DeepSeek-V3	84.61
Llama3.1-8B	82.73
Llama3.3-70B	84.46
Qwen2.5-7B	83.26
Qwen-32B	84.90
Gpt-oss-20B	80.83
Gpt-oss-120B	80.60
Kimi-K2	84.72
Silma-9B-Instruct	84.16
c4ai-command-r7b-arabic	84.15
LFM2-700M	85.82
LFM2-350M	**85.85**

Just Accepted

5.3 Impact of Static Prompting

As shown in Table 5, introducing a Static instructional prompt yields consistent and measurable improvements across all evaluated models. The average BERTScore increases indicate that explicit role definition, response constraints, and few-shot exemplars substantially enhance semantic alignment with reference plans.

The magnitude of improvement varies across models. For example, Qwen-32B improves from 80.89% to 84.90%, while DeepSeek-V3 increases from 80.80% to 84.61%. These gains demonstrate that prompt structure significantly enhances semantic alignment by limiting the generation space to clinically relevant responses. Importantly, no model experiences performance degradation under Static prompting, highlighting its reliability across architectures and model sizes.

Static prompting reduces response variability and ambiguity, leading to more stable and interpretable outputs. Notably, performance gains are observed even for models with weaker Arabic specialization, suggesting that prompt structure can partially compensate for limited domain exposure during pretraining. However, while Static prompting improves general response quality, it remains limited in its ability to adapt recommendations to case-specific clinical nuances.

5.4 Effectiveness of Dynamic Prompting

Table 6 demonstrates that STEAD-PE (Dynamic) achieves the highest BERTScore values across all models, outperforming both Zero-Shot and Static configurations. The observed gains indicate that injecting structured clinical metadata provides additional explanatory power beyond raw symptom text and fixed instructions.

Although the incremental gains over Static prompting are smaller than those observed when transitioning from Zero-Shot to Static, they are consistent and strictly positive. For instance, LFM2-350M improves from 85.85% to

Table 6. BERTScore of LLMs under the Dynamic prompt setting.

Model	Dynamic BERTScore (%)
Gemini-2.5-Flash	85.13
DeepSeek-V3	85.54
Llama3.1-8B	83.30
Llama3.3-70B	85.19
Qwen2.5-7B	83.67
Qwen-32B	85.76
Gpt-oss-20B	81.76
Gpt-oss-120B	81.04
Kimi-K2	85.15
Silma-9B-Instruct	85.38
c4ai-command-r7b-arabic	85.00
LFM2-700M	86.40
LFM2-350M	**86.50**

86.50%, and LFM2-700M increases from 85.82% to 86.40%. These results indicate that clearly encoding structured clinical context reduces ambiguity in patient posts and enables models to generate responses that more closely match expert medical reasoning.

Dynamic prompting enables models to explicitly reason about symptom urgency, medical domain relevance, and likely diagnoses, resulting in more precise and contextually appropriate recommendations. Importantly, these improvements are consistent across diverse model families, highlighting the model-agnostic nature of the STEAD-PE framework.

5.5 Dialect Robustness: Algerian and Moroccan Arabic

To assess the generalisability of STEAD-PE beyond the Egyptian dialect present in the primary dataset, we conducted a supplementary evaluation using patient posts written in Algerian and Moroccan Arabic. Five models were evaluated under all three prompting configurations: Zero-Shot, Static, and Dynamic. Table 7 reports the corresponding BERTScore values for each model–dialect–configuration combination.

The results indicate that the monotonic improvement pattern observed across prompting strategies in the primary Egyptian-dialect evaluation is largely preserved for both Algerian and Moroccan Arabic. Dynamic prompting consistently achieves the highest BERTScore values, followed by Static, and then Zero-Shot, across the majority of model–dialect pairs. These findings suggest that the structural and contextual enrichment provided by STEAD-PE generalizes across dialectal boundaries, even for dialects that are linguistically more distant from Modern Standard Arabic. Absolute BERTScore values vary modestly across dialects, which is consistent with the known variation in dialectal lexical coverage across pretrained multilingual models. Nevertheless, the relative ordering of prompting strategies remains stable, reinforcing the robustness of the STEAD-PE framework as a dialect-agnostic prompting methodology for Arabic medical response generation.

5.6 Cross-Model Observations

Table 8 summarizes the average BERTScore improvements across all evaluated models for each prompting strategy. Consistent with the per-model results in Tables 5- 6, both Static and Dynamic prompting yield positive performance gains, with STEAD-PE (Dynamic) achieving the largest average improvement. This aggregated

Just Accepted

Table 7. BERTScore of selected LLMs under Zero-Shot, Static, and Dynamic prompting for Algerian and Moroccan Arabic dialects.

Dialect	Model	Zero-Shot	Static	Dynamic
Algerian	LFM2-350M	82.30%	82.63%	83.42%
	LFM2-700M	82.45%	82.53%	83.57%
	Qwen2.5-7B	81.86%	81.97%	82.40%
	Llama3.1-8B	82.01%	82.80%	83.48%
	Silma-9B-Instruct	**83.76%**	**84.02%**	**84.24%**
Moroccan	**LFM2-350M**	**84.44%**	**84.73%**	**85.58%**
	LFM2-700M	83.50%	83.84%	84.60%
	Qwen2.5-7B	83.17%	83.56%	83.65%
	Llama3.1-8B	84.20%	84.63%	85.50%
	Silma-9B-Instruct	83.86%	84.60%	85.43%

Just Accepted

analysis confirms that the observed benefits are not driven by individual models but reflect a consistent trend across architectures.

Across all prompting configurations, larger and more multilingual models generally achieve higher absolute BERTScore values. However, the relative improvements introduced by Static and Dynamic prompting are consistent regardless of model scale, suggesting that prompt engineering benefits are largely orthogonal to parameter count.

Arabic-specialized and multilingual models exhibit greater absolute gains under Dynamic prompting, likely due to their enhanced capacity to integrate structured contextual cues with linguistic nuance. Nevertheless, even general-purpose models show substantial improvements, reinforcing the practical applicability of STEAD-PE across heterogeneous deployment scenarios.

A closer inspection of Table 8 suggests that the magnitude of improvement is related to a model's prior exposure to Arabic and to medical-domain text. Large general-purpose, multilingual models with comparatively limited Arabic medical pretraining such as Qwen-32B (+4.01), DeepSeek-V3 (+3.81), and Kimi-K2 (+3.64) exhibit the largest gains from Zero-Shot to Static prompting (Δ_{static}), indicating that explicit role definition and few-shot structuring substantially compensates for the absence of strong intrinsic Arabic medical grounding. In contrast, Arabic-specialized or lightweight instruction-tuned models, such as Silma-9B-Instruct and c4ai-command-r7b-arabic, show smaller Δ_{static} values but comparatively larger Δ_{dynamic} gains (+1.22 and +0.85, respectively), suggesting that these models are already reasonably well calibrated to Arabic phrasing and benefit proportionally more from the case-specific clinical signals—Domain, Severity, and Diagnosis—that raw text and generic structuring alone cannot supply. The GPT-OSS variants, which post the lowest absolute BERTScores across all configurations, show a similar pattern: modest Δ_{static} but relatively higher Δ_{dynamic} (+0.93 and +0.44), consistent with these models having narrower Arabic-domain pretraining coverage and therefore relying more heavily on injected metadata to recover clinically relevant signal. Taken together, these patterns indicate that Static prompting primarily benefits models lacking task-specific structure, whereas Dynamic prompting yields proportionally larger gains for models that already process Arabic competently but still require explicit clinical grounding to resolve case-specific ambiguity.

Table 8. BERTScore gains (%) introduced by Static and Dynamic prompting strategies.

Model	$\Delta_{\text{Static–Zero}}$	$\Delta_{\text{Dynamic–Static}}$
Gemini-2.5-Flash	+3.34	+0.23
DeepSeek-V3	+3.81	+0.93
Llama3.1-8B	+0.32	+0.57
Llama3.3-70B	+2.56	+0.73
Qwen2.5-7B	+0.54	+0.41
Qwen-32B	+4.01	+0.86
GPT-OSS-20B	+0.85	+0.93
GPT-OSS-120B	+0.94	+0.44
Kimi-K2	+3.64	+0.43
Silma-9B-Instruct	+0.49	+1.22
c4ai-command-r7b-arabic	+0.50	+0.85
LFM2-700M	+0.36	+0.58
LFM2-350M	+0.29	+0.65

Just Accepted

6 Discussion

The results presented in Section 5 demonstrate that prompt structure and contextual enrichment play a decisive role in Arabic medical response generation, even when using powerful pretrained LLMs in an inference-only setting. STEAD-PE provides empirical evidence that carefully designed prompt engineering can substantially improve semantic alignment and clinical relevance without any model fine-tuning, which is particularly important for low-resource and privacy-sensitive medical domains. The aggregated improvements reported in Table 8 further reinforce the robustness of STEAD-PE, demonstrating that Dynamic prompting consistently outperforms alternative strategies across all evaluated models.

One of the central findings is that raw patient-authored Arabic text alone is insufficient for reliable medical response generation. Informal phrasing, dialectal variation, and implicit symptom descriptions introduce substantial ambiguity, which zero-shot prompting fails to resolve consistently. This observation aligns with prior work in Arabic medical NLP, which highlights the limitations of relying solely on unstructured patient narratives [2, 18]. The results further confirm that LLMs do not reliably infer critical clinical attributes, such as urgency or diagnostic intent, without explicit guidance, even when semantic fluency appears high.

The introduction of Static prompting demonstrates that instructional structure and role specification meaningfully stabilize model behavior. Few-shot exemplars and response constraints reduce variance and improve coherence across models, corroborating findings from English-language medical prompt engineering studies [7, 21]. However, the observed performance ceiling under Static prompting suggests that fixed templates alone cannot fully compensate for missing clinical context, particularly in cases where urgency and diagnostic interpretation are only implicitly expressed in the patient text.

The consistent gains achieved by STEAD-PE (Dynamic) highlight the value of injecting structured clinical metadata into the prompt at inference time. By explicitly conditioning on Domain, Severity, and Diagnosis, models are better equipped to prioritize symptoms, adjust urgency cues, and tailor recommendations. From an information-theoretic perspective, this supports the argument that structured clinical attributes reduce conditional

uncertainty in response generation, as formalized in Section 4.3.3. Importantly, these gains are observed across all evaluated model families, underscoring the model-agnostic nature of the framework.

Beyond performance improvements, STEAD-PE offers practical advantages for real-world deployment. Because the framework operates without fine-tuning, it avoids challenges related to data privacy, regulatory compliance, and continual model retraining. This makes STEAD-PE particularly suitable for clinical decision support systems, telemedicine platforms, and Arabic-language health advisory services, where rapid adaptation and interpretability are critical.

The supplementary dialect evaluation reported in Section 5.5 provides further evidence for the generalizability of STEAD-PE. Despite the primary dataset being drawn from Egyptian-dialect patient posts, the framework produces consistent prompting-strategy gains for Algerian and Moroccan Arabic, two dialects with distinct phonological and lexical characteristics. This suggests that the mechanism underlying STEAD-PE's effectiveness-namely, the explicit injection of structured clinical metadata-operates at a semantic level that is largely invariant to surface-level dialectal variation. Accordingly, dialect coverage should be considered when curating metadata annotations for deployment in specific regional health contexts, though the core prompting architecture requires no dialect-specific modification.

At the same time, several limitations warrant discussion. First, the Dynamic configuration relies on accurate clinical metadata annotations, which are provided by medical experts in this study. In real-world systems, such metadata may need to be estimated automatically or provided through structured user input, potentially introducing noise. Second, while BERTScore captures semantic similarity, it does not directly measure clinical correctness, safety, or treatment appropriateness. Although qualitative analysis suggests improved grounding, human-in-the-loop evaluation remains essential for high-stakes medical applications.

Finally, while STEAD-PE significantly improves semantic alignment, it does not eliminate the risk of hallucinations or inappropriate recommendations inherent to generative models. Therefore, the framework should be viewed as a supportive decision-assistance mechanism rather than a substitute for professional medical judgment. Future work should explore hybrid evaluation protocols, uncertainty-aware prompting, adaptive metadata inference, as well as longitudinal user studies to assess real-world impact, and the consistency and stability of generated responses through repeated experimental runs across different LLMs, even under deterministic inference settings, to further validate the robustness and reproducibility of the framework, to further validate the robustness and reproducibility of the framework. In addition, future work should incorporate variance and confidence-interval reporting alongside formal statistical significance testing across repeated runs, and conduct a systematic ablation study that selectively removes or substitutes individual metadata attributes (Domain, Severity, Diagnosis) to isolate their respective contributions to the observed Dynamic-prompting gains.

7 Limitations

Several limitations of the proposed framework warrant discussion. First, the Dynamic configuration relies on accurate clinical metadata annotations, which are provided by medical experts in this study. In real-world systems, such metadata may need to be estimated automatically or provided through structured user input, potentially introducing noise and reducing prompt reliability. Second, this study relies exclusively on BERTScore as its quantitative evaluation metric. It is worth noting that this metric is not being applied against an arbitrary or unvalidated reference: each Plan is a clinically validated response authored by a certified medical advisor 3, so a high BERTScore reflects semantic alignment with an expert-vetted clinical answer rather than with generic fluent text. This partially mitigates, though does not eliminate, the risk of rewarding superficially fluent but clinically vacuous responses. Two residual concerns remain. First, BERTScore measures similarity to a single reference Plan per case; where multiple clinically valid treatment approaches exist for the same presentation, a generated response that is medically sound but phrased or structured differently from the specific reference may

Just Accepted

be penalized despite being clinically appropriate. Second, embedding-based similarity is not a safety check: a response that remains close to the reference in overall semantic content could still contain a locally incorrect or unsafe clinical detail without this being reliably reflected in the aggregate score. Consequently, while the use of expert-authored references lends BERTScore-based evaluation more clinical grounding than is typical for this metric, it remains an indirect proxy for clinical correctness and safety. Reliable deployment in high-stakes medical settings still requires human-in-the-loop evaluation by qualified clinicians across the relevant specialties which is not included in the present evaluation. We identify the design and incorporation of such metrics, alongside a clinician-rated evaluation protocol, as the most important direction for future work on this framework. Third, the framework operates entirely in an inference-only setting and does not eliminate the inherent risk of hallucinations or medically inappropriate recommendations associated with generative LLMs. Consequently, STEAD-PE should be considered a supportive decision-assistance framework rather than a replacement for professional medical judgment. Fourth, although the experiments demonstrate consistent improvements across multiple LLMs, the current study does not include repeated experimental runs to assess response stability and reproducibility under deterministic inference configurations. Future work should therefore investigate repeated-run evaluation protocols and robustness analysis across different models and prompting conditions. Fifth, while Section 2 situates STEAD-PE relative to concurrent Arabic medical prompting studies such as MedArabiQ [1] and the AraHealthQA shared task [26], no direct empirical comparison against these systems was conducted in this study, as they target classification and question-answering formats rather than the free-text treatment-recommendation generation task addressed here. Future work should benchmark STEAD-PE directly against these and other emerging Arabic medical PE pipelines under a shared task formulation to enable head-to-head comparison. Finally, the reported BERTScore values reflect single-run evaluations under fixed inference hyperparameters 4.4; the study does not report variance, confidence intervals, or formal statistical significance testing across repeated generations for the same input. Relatedly, the framework does not include an ablation study isolating the individual contribution of each injected metadata attribute (Domain, Severity, Diagnosis) to the observed Dynamic-prompting gains. Future work should incorporate repeated-run evaluation with variance and confidence-interval reporting, statistical significance testing, and a systematic per-attribute ablation analysis to disentangle these contributions.

8 Conclusion

This paper introduced STEAD-PE, a structured and dynamic prompt engineering framework for Arabic medical response generation using LLMs in a strictly inference-only setting. By systematically progressing from Zero-Shot to Static and Dynamic prompting, the study demonstrated that prompt design alone, without any model fine-tuning, can substantially improve semantic alignment and clinical grounding of generated medical responses.

Through comprehensive qualitative and quantitative evaluations across multiple state-of-the-art LLMs, the results showed that Static prompting enhances response stability and coherence, while Dynamic prompting consistently achieves the highest performance by explicitly injecting structured clinical metadata, including Domain, Severity, and Diagnosis. These improvements were consistent across diverse model architectures and parameter scales, underscoring the model-agnostic effectiveness of the STEAD-PE framework.

Beyond performance gains, STEAD-PE offers practical and ethical advantages for medical NLP applications. Its inference-only design avoids the need for sensitive data retention, costly retraining, or continuous model updates, making it particularly suitable for low-resource languages, privacy-constrained environments, and rapidly evolving clinical domains. Moreover, by separating clinical context provisioning from model parameters, the framework supports greater transparency and controllability in medical response generation.

While STEAD-PE improves semantic and contextual alignment, it is not intended as a replacement for professional medical judgment. Instead, it should be viewed as a decision-support mechanism that can assist clinicians

Just Accepted

and patients by providing more structured, context-aware informational guidance. Future work will explore automated metadata inference, uncertainty-aware prompting, and human-in-the-loop evaluation frameworks, as well as broader validation across additional medical specialties and real-world deployment scenarios.

References

[1] Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E Shamout. 2025. MedArabiQ: Benchmarking Large Language Models on Arabic Medical Tasks. *arXiv e-prints* (2025), arXiv–2505.

[2] Seif Ahmed, Abdulrahman Allam, Ali Hamdi, and Ahmed Mohamed. 2025. Arabic Large Language Models for Medical Text Generation. *Preprint* (2025).

[3] Seif Ahmed, Abdulrahman Allam, Ali Hamdi, and Ahmed Mohamed. 2025. Arabic Symptom Classification and Diagnosis Using Transformer Models and LLM-Based Augmentation. *Preprint* (2025).

[4] Moonshot AI. 2024. Kimi: Long-Context Large Language Models. *Technical Report* (2024).

[5] Ahmed Badawi, Ehsan Rahimi, Md. Tauhidul Laskar, Shubham Grach, Louis Bertrand, Lucas Danok, Jing Huang, Frank Rudzicz, and Ehsan Dolatabadi. 2025. When Can We Trust LLMs in Mental Health? Large-Scale Benchmarks for Reliable LLM Evaluation. *arXiv preprint* (October 2025).

[6] Pavel Blinov, Yulia Kuzkina, Vladimir Goncharov, Galina Zubkova, Nikita Zenovkin, Andrey Goncharov, and Andrey Savchenko. 2024. MedSyn: LLM-Based Synthetic Medical Text Generation Framework. *Preprint* (2024).

[7] Luke Cheptora. 2025. Why Prompt Engineering Is the Core Skill for the Modern AI Era. *Article* (August 2025).

[8] Alibaba Cloud. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609* (2023).

[9] Open-Source Community. 2024. GPT-OSS: Open-Source Large-Scale Language Models. *arXiv preprint* (2024).

[10] Google DeepMind. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805* (2023).

[11] DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. *arXiv preprint* (2024).

[12] Cohere for AI. 2024. Command Models for Multilingual and Enterprise NLP. *Technical Report* (2024).

[13] Iker García-Ferrero, Rodrigo Agerri, Aitor Atutxa, Elena Cabrio, Inigo de la Iglesia, Alberto Lavelli, Bernardo Magnini, Baptiste Molinet, Jose Ramirez-Romero, German Rigau, Jose M. Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. Medical mT5: An Open-Source Multilingual Text-to-Text Large Language Model for the Medical Domain. *arXiv preprint arXiv:2404.07613* (2024).

[14] Karthik Gopalakrishnan. 2024. Prompt Engineering: Revolutionizing Natural Language Processing. *Zenodo* (2024).

[15] Daniel Grabb. 2023. The Impact of Prompt Engineering in Large Language Model Performance: A Psychiatric Example. *Journal of Medical Artificial Intelligence* (2023).

[16] Ming-Yu Hsieh, Tzu-Ling Wang, Po-Hao Su, and Ming-Chi Chou. 2025. Impact of Prompt Engineering on the Performance of ChatGPT Variants Across Different Question Types in Medical Student Exams. *JMIR Medical Education* (2025).

[17] Christian Meyer, Daniel Adkins, Kaushik Pal, Riccardo Galici, Antonio G. Agundez, and Carsten Eickhoff. 2023. Neural Text Generation in Regulatory Medical Writing. *Frontiers in Pharmacology* (February 2023).

[18] Mohamed Mohamed, Rokaia Emad, and Ali Hamdi. 2025. Ensemble Transformers for Arabic Medical Text Classification. *Preprint* (2025).

[19] Mohamed Mohamed, Rokaia Emad, and Ali Hamdi. 2025. A Multi-Layered Large Language Model Framework for Disease Prediction. *Preprint* (2025).

[20] Nikhil Mungoli. 2023. Enhancing Control and Responsiveness in ChatGPT: A Study on Prompt Engineering and Reinforcement Learning Techniques. *Preprint* (2023).

[21] B. Naga. 2025. The Dynamics of Prompt Engineering: A Comprehensive Review. *Preprint* (2025).

[22] Sagar Patil and Yogesh Puranik. 2024. Importance of Effective Prompt Engineering. *International Journal of Research* 6, 10 (2024).

[23] LFM Team. 2024. Lightweight Foundation Models for Efficient Inference. *arXiv preprint* (2024).

[24] Sihma Team. 2024. Sihma: Instruction-Tuned Language Models. *Technical Report* (2024).

[25] Hugo Touvron et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[26] Mohamed Younes, Seif Ahmed, and Mohamed Basem. 2025. ! MSA at AraHealthQA 2025 Shared Task: Enhancing LLM Performance for Arabic Clinical Question Answering through Prompt Engineering and Ensemble Learning. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*. 176–183.

[27] Wei Zhu, Xuan Wang, Ming Chen, and Buzhou Tang. 2023. Overview of the PromptCBLUE Shared Task in CHIP2023. In *CHIP 2023 Conference*.

Received 5 January 2026; revised 1 July 2026; accepted 11 October 2026

Just accepted