DeCode: Decoupling Content and Delivery for Medical QA

Anonymous ACL submission

Abstract

Large language models (LLMs) exhibit strong medical knowledge and can generate factually accurate responses. However, existing models often fail to account for individual patient contexts, producing answers that are clinically correct yet poorly aligned with patients' needs. In this work, we introduce **DeCode** (**De**coupling **Co**ntent and **De**livery), a training-free, model-agnostic framework that adapts existing LLMs to produce contextualized answers in clinical settings. We evaluate DeCode on OpenAI HealthBench, a comprehensive and challenging benchmark designed to assess clinical relevance and validity of LLM responses. DeCode improves the previous state-of-the-art from 28.4% to 49.8%, corresponding to a 75% relative improvement. Experimental results suggest the effectiveness of DeCode in improving clinical question answering of LLMs.

1 Introduction

Large language models (LLMs) have recently achieved strong performance on a variety of medical natural language processing tasks, most notably medical question answering (QA), where models are evaluated on their ability to generate correct responses to clinically relevant questions (Singhal et al., 2025; Nori et al., 2023). This progress has been demonstrated across a growing collection of medical QA benchmarks, spanning multiple-choice and generative settings, professional examination-style questions, and open-domain clinical knowledge assessments (Jin et al., 2021; Pal et al., 2022). Collectively, results on these evaluations suggest that contemporary LLMs exhibit substantial medical knowledge and reasoning capability under standardized testing conditions (Saab et al., 2024; OpenAI, 2024).

Existing medical QA benchmarks, however, are predominantly designed to measure answer correctness or reasoning accuracy, often via exact-match, multiple-choice selection, or expert-graded factual validity. While these metrics are well-suited for assessing knowledge recall and clinical reasoning, they provide only a partial characterization of model behavior in patient-facing or clinical communication settings (Gong et al., 2025; Tu et al., 2025). In particular, such evaluations do not capture whether model responses are understandable, appropriately calibrated to patient context, or aligned with norms of safe and empathetic medical communication.

This limitation motivates the need for evaluation frameworks that extend beyond accuracy-based metrics. OpenAI HealthBench was introduced to address this gap by evaluating medical LLM outputs along multiple qualitative dimensions, including context seeking, emergency referrals, and responding under uncertainty, in addition to factual correctness (Arora et al., 2025). Unlike prior medical QA datasets, which typically assume a single correct answer independent of delivery style or audience, HealthBench explicitly models the interactional aspects of medical responses, enabling a more fine-grained analysis of clinically relevant response quality.

Empirical results on HealthBench further show that models with comparable accuracy on traditional medical QA benchmarks can exhibit substantial variation across other non-accuracy dimensions, revealing a misalignment between standardized QA performance and patient-centered context awareness (Arora et al., 2025). Together, these findings suggest that accuracy alone is insufficient as a proxy for real-world clinical readiness and underscore the importance of multidimensional evaluation for medical LLMs.

In this work, we introduce the **Decoupling Content and Delivery (DeCode)** framework, a modular approach for generating patient-specific medical responses from clinical conversations. DeCode decomposes an existing clinical interaction into multiple complementary analytical perspectives, each

implemented via a specialized LLM module. The outputs of these modules are subsequently synthesized to produce a final response that accounts for both medical correctness and patient context.

Importantly, DeCode operates in a training-free paradigm, orchestrating the generation process through explicit clinical formulation and structured discourse constraints. Empirically, we demonstrate that DeCode substantially improves performance on HealthBench, increasing the prior state-of-the-art score from 28.4% to 49.8%. Furthermore, we show that DeCode generalizes consistently across multiple leading LLMs, suggesting that the framework captures model-agnostic principles for personalized medical response generation.

The remainder of this paper is organized as follows. Related works are introduced in Section 2. The proposed method is presented in Section 3. Experimental setup and results are provided in Section 4 and Section 5, respectively. Finally, Section 6 concludes the paper.

2 Related Work

Early evaluations of large language models (LLMs) in medical question answering have primarily focused on standardized multiple-choice benchmarks, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019). These benchmarks have catalyzed substantial research on assessing and improving medical knowledge in LLMs (Singhal et al., 2025; Nori et al., 2023; Saab et al., 2024; Jeong et al., 2024; Li et al., 2024; Wu et al., 2025a). However, such evaluations remain inherently static and accuracy-centric, limiting their ability to assess communicative competence, contextual sensitivity, and patient-centered delivery beyond factual correctness (Gong et al., 2025).

HealthBench (Arora et al., 2025) introduces a multidimensional evaluation framework for medical QA based on open-ended, multi-turn clinical conversations. Unlike traditional multiple-choice benchmarks, HealthBench employs physician-authored rubrics to assess behavioral dimensions such as clinical accuracy, communication quality, and contextual awareness, enabling a more comprehensive evaluation of medical QA systems beyond factual correctness.

MuSeR (Zhou et al., 2025) targets HealthBench by proposing a self-refinement training framework in which a student LLM is supervised using high-quality responses from a reference teacher model. The student generates an initial response, performs structured self-assessment across multiple dimensions, and produces a refined final answer. While effective, this approach relies on computationally intensive data synthesis and additional training, limiting its applicability to trainable, open-source LLMs.

In parallel, multi-agent frameworks have been proposed to address complex medical QA by decomposing reasoning across specialized roles. MedAgents (Tang et al., 2024) employs role-playing specialists for debate-based hypothesis refinement, while MDAgents (Kim et al., 2024) dynamically configures expert teams based on query complexity. More recent approaches further extend this paradigm: KAMAC (Wu et al., 2025b) introduces on-demand expert recruitment to address knowledge gaps during generation, and AI Hospital (Fan et al., 2025) evaluates agent-based systems in interactive patient simulation environments. However, these methods primarily emphasize diagnostic reasoning and accuracy on traditional benchmarks, often overlooking how complex reasoning outcomes are translated into clear, user-aligned responses.

Building on these observations, we introduce DeCode, a modular framework that explicitly decouples medical content reasoning from response delivery. Unlike training-based or agent-centric approaches, DeCode requires no additional training and is model-agnostic, while emphasizing structured generation that supports contextualized and user-aligned medical responses. We present our implementation in the following section.

3 Method

Medical question answering with LLMs can be modeled as a form of conditional text generation $P(\mathcal{R} \mid \mathcal{H})$, where $\mathcal{R}$ denotes the response and $\mathcal{H}$ the conversation history. In practice, $\mathcal{H}$ contains rich patient-specific information—such as symptoms, risk factors, and health indicators—distributed across multiple dialogue turns. However, LLMs are typically trained to model this distribution directly, without mechanisms to explicitly aggregate these dispersed signals. As a result, specific patient details are frequently overlooked during response generation.

To address this limitation, we introduce DeCode, a framework that structures the generation process

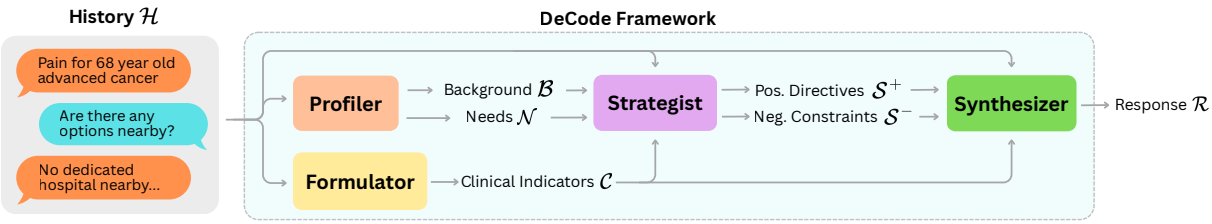

Figure 1: **The DeCode Framework Pipeline.** Given the conversation history $\mathcal{H}$, the system first employs the **Profiler** and **Formulator** to extract user context $(\mathcal{B}, \mathcal{N})$ and clinical indicators $\mathcal{C}$. These components are then synthesized by the **Strategist** to generate tailored directives $\mathcal{S}$ (consisting of positive strategies $\mathcal{S}^+$ and negative constraints $\mathcal{S}^-$). Finally, the **Synthesizer** constructs the response based on $\mathcal{C}$ and $\mathcal{S}$, ensuring both medical accuracy and user adaptability.

through four intermediate textual representations: user background $\mathcal{B}$, user needs $\mathcal{N}$, clinical indicators $\mathcal{C}$, and discourse strategy $\mathcal{S}$. As illustrated in Figure 1, these representations are orchestrated by four corresponding modules: Profiler $\mathcal{M}_{prof}$, Formulator $\mathcal{M}_{form}$, Strategist $\mathcal{M}_{strat}$, and Synthesizer $\mathcal{M}_{syn}$. By disentangling *content* from *delivery*, DeCode enables independent optimization of medical accuracy and communicative quality. The inference process is formalized as a sequential chain:

$$\mathcal{R} = \underbrace{\mathcal{M}_{syn}(\mathcal{S}, \mathcal{C}, \mathcal{H})}_{\text{Synthesis}} \circ \underbrace{\mathcal{M}_{strat}(\mathcal{B}, \mathcal{N}, \mathcal{C}, \mathcal{H})}_{\text{Strategy}} \circ \underbrace{\{\mathcal{M}_{prof}(\mathcal{H}), \mathcal{M}_{form}(\mathcal{H})\}}_{\text{Extraction}}$$

where $\mathcal{M}$ denotes the LLM modules tailored for specific sub-tasks. In the following sections, we detail the design of each module.

3.1 Profiler: User Context Disentanglement

Medical advice varies significantly across individuals. The same symptom may imply different risks depending on the user's background and lifestyle. To capture this nuance beyond surface-level queries, the Profiler $\mathcal{M}_{prof}$ extracts the user's specific context from the conversation history $\mathcal{H}$. We formalize this extraction as:

$$(\mathcal{B}, \mathcal{N}) = \mathcal{M}_{prof}(\mathcal{H}).$$

The user background $\mathcal{B}$ encapsulates critical attributes such as age, occupation, and living conditions that constrain actionable advice. Concurrently, the user needs $\mathcal{N}$ identifies the user's core intent by synthesizing the conversation history $\mathcal{H}$. By decoupling the user information $\mathcal{B}$ and $\mathcal{N}$ from the history $\mathcal{H}$, we allow improved clarity in identifying user-specific constraints during response formulation. The user background $\mathcal{B}$ and user needs $\mathcal{N}$ are then sent to the Strategist module.

3.2 Formulator: Clinical Distillation

A critical challenge in medical dialogue is that diagnostic cues are often dispersed throughout the conversation history $\mathcal{H}$, making it difficult to verify if the response covers all relevant medical aspects. To address this, the Formulator $\mathcal{M}_{form}$ functions as a clinical information distiller. It extracts and aggregates a structured set of clinical indicators $\mathcal{C}$ (e.g., symptoms, possible causes, and potential red flags) from the user statements in $\mathcal{H}$. We formalize this process as

$$\mathcal{C} = \mathcal{M}_{form}(\mathcal{H}).$$

Crucially, this module operates purely on a factual level, decoupling the medical substance from the delivery style. By explicitly manifesting $\mathcal{C}$ as an intermediate representation, the system provides a rigorous checklist for the downstream modules. This ensures that the final response is grounded in verified medical evidence and that high-stakes safety indicators are duly addressed, regardless of the chosen empathy level or conversation tone.

3.3 Strategist: Discourse Orchestration

Beyond factual accuracy, effective medical dialogue requires determining the optimal delivery strategy tailored to the user's cognitive and emotional context. The Strategist $\mathcal{M}_{strat}$ addresses this gap by synthesizing the conversation history $\mathcal{H}$, extracted user profile $(\mathcal{B}, \mathcal{N})$ and clinical indicators $(\mathcal{C})$ into a coherent strategy. We formalize this process as:

$$\mathcal{S} = \{\mathcal{S}^+, \mathcal{S}^-\} = \mathcal{M}_{strat}(\mathcal{B}, \mathcal{N}, \mathcal{C}, \mathcal{H}).$$

The resulting discourse strategy $\mathcal{S}$ comprises two complementary sets. Positive directives $\mathcal{S}^+$ prescribe the prioritization of clinical content and establish the appropriate level of technical detail, crucially instructing the model to actively seek clarification when information is insufficient. Conversely, negative constraints $\mathcal{S}^-$ serve as behavioral guardrails, preventing counterproductive styles (e.g., overly academic tones) and filtering out content that may be overwhelming or potentially misleading for the specific user. By enforcing these strategies, the module ensures that the final response is not only medically grounded but also empathetic and strictly aligned with the user's preferences.

3.4 Synthesizer: Controlled Generation

Finally, the Synthesizer $\mathcal{M}_{syn}$ generates the response R by integrating the clinical indicators $\mathcal{C}$ with the discourse strategy $\mathcal{S}$. We formalize this process as:

$$R = \mathcal{M}_{syn}(\mathcal{S}, \mathcal{C}, \mathcal{H}).$$

By separating content formulation from delivery planning, the Synthesizer operates as a constrained generator. It articulates the verified information in $\mathcal{C}$ while adhering to the directives defined in $\mathcal{S}$. This ensures that the generation process focuses on realization rather than reasoning, producing outputs that are clinically accurate and contextually appropriate.

Taken together, the Profiler, Formulator, Strategist, and Synthesizer form a coherent generation pipeline that transforms the conversation history $\mathcal{H}$ into a personalized and clinically grounded response R. Each module addresses a distinct stage of the reasoning–generation process, enabling explicit control over user understanding, clinical content, discourse planning, and surface realization. For reproducibility and clarity, the prompts corresponding to each module are provided in the appendix.

4 Experiments

4.1 Dataset and Evaluation

We evaluate on OpenAI HealthBench (Arora et al., 2025), which contains 5,000 simulated multi-turn patient–clinician conversations ending in a user query. Each conversation is annotated by medical professionals and assigned to one of seven themes: *emergency referrals*, *context seeking*, *global health*, *health data tasks*, *complex responses*, *hedging*, and *communication*. HealthBench additionally provides physician-authored rubrics per conversation, grouped into five evaluation axes: *accuracy*, *completeness*, *communication quality*, *context awareness*, and *instruction following*. For more details regarding the conversation themes and evaluation axes, please refer to the original paper (Arora et al., 2025).

Metric. We follow the official HealthBench protocol (Arora et al., 2025): each conversation is graded using its rubric and scored by GPT-4.1 (OpenAI, 2025a). Reported numbers are the mean normalized score over the evaluated set.

Splits. We report results on the full HealthBench dataset for the primary evaluation. Owing to the computational cost, all subsequent experiments are conducted on the Hard subset of 1,000 challenging conversations.

4.2 Implementation Details

Base LLMs. We use OpenAI o3 (OpenAI, 2025c) as the primary base model and additionally evaluate GPT-5 (OpenAI, 2025b), Claude-Sonnet 4.5 (Anthropic, 2025), and DeepSeek R1 (DeepSeek-AI et al., 2025) to assess generalization across model families.

Comparison Methods. We compare against: (i) **Zero-shot**, which directly prompts the base LLM to respond from the conversation history; (ii) **MDAgents** (Kim et al., 2024), which adapts team structure and recruits specialized experts based on query complexity, followed by consensus; and (iii) **KAMAC** (Wu et al., 2025b), which recruits experts on demand during generation. For KAMAC, we let the model choose the initial number of experts and fix the discussion to two rounds.

5 Results and Analysis

5.1 Main Results

In this experiment, we compare DeCode with a zero-shot baseline built on the same underlying LLM, OpenAI o3. Both methods are evaluated on the full HealthBench dataset as well as its hard subset, with results summarized in Table 1.

A clear performance gap emerges between the full dataset and the hard subset. On the full set, the zero-shot baseline performs weakest on *health data tasks*. Performance further degrades on the hard subset, where the baseline struggles across

Table 1: **Comparison between zero-shot prompting and DeCode.** Both methods use OpenAI o3 as the base LLM and are evaluated on the full HealthBench dataset and its hard subset. DeCode consistently outperforms the zero-shot baseline across most conversation themes and evaluation axes. Performance on the hard subset is substantially lower than on the full set, highlighting the increased difficulty of this evaluation setting. Deltas in DeCode columns are computed relative to the zero-shot baseline within the same split; deltas greater than 2 points are highlighted.

	Full Set		Hard Subset	
Metric (↑, %)	**Zero-shot**	**DeCode (Ours)**	**Zero-shot**	**DeCode (Ours)**
Overall Score	57.8	**67.8** (+10.0)	28.4	**49.8** (+21.4)
Themes				
Emergency Referrals	69.2	**80.3** (+11.1)	27.0	**59.1** (+32.1)
Context Seeking	51.2	**67.0** (+15.8)	30.0	**58.3** (+28.3)
Global Health	52.7	**65.2** (+12.5)	31.8	**49.8** (+18.0)
Health Data Tasks	44.3	**56.7** (+12.4)	17.0	**35.6** (+18.6)
Communications	67.9	**74.4** (+6.5)	29.2	**43.8** (+14.6)
Hedging	59.6	**69.5** (+9.9)	30.9	**54.6** (+23.7)
Complex Responses	**55.1**	52.6 (-2.5)	24.4	**41.3** (+16.9)
Axes				
Accuracy	66.4	**72.5** (+6.1)	45.6	**54.3** (+8.7)
Completeness	59.5	**74.0** (+14.5)	30.7	**58.8** (+28.1)
Communication Quality	**68.1**	61.9 (-6.2)	**55.5**	54.2 (-1.3)
Context Awareness	41.7	**53.4** (+11.7)	4.0	**40.5** (+36.5)
Instruction Following	**61.2**	59.4 (-1.8)	45.8	**46.5** (+0.7)

nearly all conversation themes; the highest score achieved is only 31.8% under the *global health* theme, highlighting the increased difficulty of this split.

In contrast, DeCode improves response quality across all conversation themes on the full set, with the exception of the *complex responses* category. Further analysis suggests that DeCode occasionally generates overly detailed responses for relatively simple or straightforward queries. While this behavior can enrich informational content, it may negatively affect perceived *communication quality*, contributing to the observed performance drop along this evaluation axis.

On the hard subset, DeCode yields substantial gains in overall performance. Notably, the lowest-scoring *health data* theme improves to 35.6%, while all remaining themes exceed 40%. These results underscore the effectiveness of DeCode in enhancing both the *content* and *delivery* of medical question answering under more challenging evaluation conditions.

5.2 Generalizability Across Backbone Models

In this experiment, we examine the generalizability of DeCode across different base LLMs. A key advantage of the proposed framework is its model-agnostic design, which allows it to be applied to a wide range of base LLMs while consistently improving medical question answering performance. We evaluate DeCode on the hard subset using several leading LLMs from different providers, with results reported in Table 2.

Based on the zero-shot performance of the base LLMs, *health data tasks* emerge as a particularly challenging category for GPT-5, OpenAI o3, and DeepSeek R1. In contrast, Claude-4.5 exhibits its weakest performance on *global health* tasks. Across evaluation axes, *context awareness* and *completeness* are especially challenging: OpenAI o3, Claude-4.5, and DeepSeek R1 all record single-digit scores on these dimensions in certain cases, indicating systematic deficiencies in handling complex contextual and informational requirements.

Consistent with the observations in Section 5.1, DeCode delivers substantial improvements over

Table 2: **DeCode performance across diverse base LLMs.** Comparison between zero-shot (ZS) and DeCode-enhanced performance on the HealthBench hard subset across multiple base LLMs. Inline deltas in the DeCode columns are computed relative to the ZS baseline for the same model; deltas larger than 2 points are highlighted.

	GPT-5		OpenAI o3		Claude 4.5		DeepSeek R1	
Metric (↑, %)	**ZS**	**DeCode**	**ZS**	**DeCode**	**ZS**	**DeCode**	**ZS**	**DeCode**
Overall Score	36.0	**50.7** (+14.7)	28.4	**49.8** (+21.4)	12.4	**40.0** (+27.6)	14.8	**25.7** (+10.9)
Themes								
Emergency Ref.	49.8	**60.8** (+11.0)	27.0	**59.1** (+32.1)	18.5	**50.2** (+31.7)	19.9	**33.0** (+13.1)
Context Seeking	41.1	**57.6** (+16.5)	30.0	**58.3** (+28.3)	12.1	**46.8** (+34.7)	15.0	**32.0** (+17.0)
Global Health	34.1	**48.4** (+14.3)	31.8	**49.8** (+18.0)	7.3	**33.9** (+26.6)	15.6	**22.3** (+6.7)
Health Data Tasks	25.8	**42.9** (+17.1)	17.0	**35.6** (+18.6)	15.0	**35.4** (+20.4)	3.1	**22.0** (+18.9)
Communications	36.2	**51.5** (+15.3)	29.2	**43.8** (+14.6)	16.6	**39.8** (+23.2)	16.1	**18.0** (+1.9)
Hedging	37.4	**53.1** (+15.7)	30.9	**54.6** (+23.7)	13.1	**48.7** (+35.6)	18.3	**34.1** (+15.8)
Complex Resp.	31.5	**40.3** (+8.8)	24.4	**41.3** (+16.9)	15.2	**27.5** (+12.3)	14.8	**15.9** (+1.1)
Axes								
Accuracy	46.1	**57.2** (+11.1)	45.6	**54.3** (+8.7)	28.6	**49.5** (+20.9)	30.5	**32.6** (+2.1)
Completeness	29.4	**56.6** (+27.2)	30.7	**58.8** (+28.1)	3.7	**44.1** (+40.4)	15.6	**27.8** (+12.2)
Comm. Quality	**61.0**	48.7 (-12.3)	**55.5**	54.2 (-1.3)	**67.3**	53.5 (-13.8)	**60.9**	58.4 (-2.5)
Cont. Awareness	31.0	**43.4** (+12.4)	4.0	**40.5** (+36.5)	1.5	**30.9** (+29.4)	0.0	**19.1** (+19.1)
Inst. Following	**51.8**	43.4 (-8.4)	45.8	**46.5** (+0.7)	**45.7**	43.6 (-2.1)	**44.8**	42.5 (-2.3)

the corresponding zero-shot baselines across all tested models. Notably, Claude-4.5, which attains an initial overall score of 12.4%, improves to 40.0% when integrated with DeCode. Similarly, the strongest baseline model, GPT-5, improves from 36.0% to 50.7%. Importantly, in all experiments the underlying base LLM remains unchanged. By explicitly decoupling *content* from *delivery*, DeCode systematically enhances the performance of diverse base LLMs across nearly all medical QA scenarios. These results demonstrate that the benefits of DeCode are robust and largely LLM-agnostic, extending across a wide range of model architectures and providers.

5.3 Comparison with Multi-Agent Frameworks

In this experiment, we evaluate representative multi-agent medical QA frameworks on the HealthBench hard subset using OpenAI o3 as the base LLM. Specifically, we consider MDAgents (Kim et al., 2024) and KAMAC (Wu et al., 2025b), with results summarized in Table 3.

Relative to the zero-shot baseline, MDAgents demonstrates consistent improvements across all conversation themes and four of the five evaluation axes. Notably, it achieves the strongest performance on *instruction following* among all compared methods. This behavior can be attributed to its *complexity-driven* orchestration strategy: MDAgents first estimates the difficulty of a given medical query and determines whether it can be handled by a single agent or requires a coordinated team of specialized *experts*. Once the team composition is selected, it remains fixed throughout the generation process. This upfront complexity assessment enables stable role assignment and coherent multi-agent collaboration, which appears particularly effective for instruction-heavy medical QA scenarios.

In contrast, KAMAC does not yield consistent gains over the zero-shot baseline and, in several cases, exhibits notable degradations in *context awareness*, *communication quality*, and *instruction following*. Unlike MDAgents, KAMAC follows a *knowledge-driven* strategy that dynamically recruits new specialists during the generation process when existing agents identify missing domain knowledge. While this adaptive recruitment mechanism is intended to enhance coverage, our analysis suggests that introducing new experts mid-stream can disrupt conversational coherence. Specifically, the newly added agents often generate responses that overlap with existing contributions or shift

Table 3: **Comparison with leading multi-agent frameworks.** We evaluate representative multi-agent methods, KAMAC (Wu et al., 2025b) and MDAgents (Kim et al., 2024), alongside standard zero-shot prompting and DeCode, all using OpenAI o3 on the HealthBench hard subset. Inline deltas are computed relative to the zero-shot baseline; deltas greater than 2 points are highlighted.

Metric (↑, %)	**Zero-Shot**	**MDAgents** (Kim et al., 2024)	**KAMAC** (Wu et al., 2025b)	**DeCode** (Ours)
Overall Score	28.4	36.2 (+7.8)	27.4 (-1.0)	**49.8** (+21.4)
Themes				
Emergency Referrals	27.0	36.3 (+9.3)	33.1 (+6.1)	**59.1** (+32.1)
Context Seeking	30.0	34.4 (+4.4)	27.4 (-2.6)	**58.3** (+28.3)
Global Health	31.8	43.0 (+11.2)	30.2 (-1.6)	**49.8** (+18.0)
Health Data Tasks	17.0	19.1 (+2.1)	7.7 (-9.3)	**35.6** (+18.6)
Communications	29.2	38.0 (+8.8)	32.9 (+3.7)	**43.8** (+14.6)
Hedging	30.9	39.2 (+8.3)	29.9 (-1.0)	**54.6** (+23.7)
Complex Responses	24.4	31.6 (+7.2)	28.5 (+4.1)	**41.3** (+16.9)
Axes				
Accuracy	45.6	52.9 (+7.3)	43.8 (-1.8)	**54.3** (+8.7)
Completeness	30.7	45.8 (+15.1)	36.1 (+5.4)	**58.8** (+28.1)
Communication Quality	**55.5**	49.5 (-6.0)	46.7 (-8.8)	54.2 (-1.3)
Context Awareness	4.0	4.6 (+0.6)	0.0 (-4.0)	**40.5** (+36.5)
Instruction Following	45.8	**53.5** (+7.7)	36.2 (-9.6)	46.5 (+0.7)

the discussion focus, leading to redundancy and task-level confusion. These effects are amplified in longer, multi-round discussions, ultimately degrading response quality.

Taken together, these results suggest that *when* and *how* experts are introduced plays a critical role in multi-agent medical QA performance. While complexity-driven, fixed-team orchestration promotes stability and coherent reasoning, dynamically expanding agent sets during generation may introduce coordination overhead that outweighs its potential benefits—particularly when all experts are instantiated from the same underlying base LLM.

5.4 Ablation Study

We conduct an ablation study to assess the individual contribution of each component in the DeCode framework. Specifically, we independently remove the Profiler, Formulator, and Strategist modules and compare their performance against the full DeCode model on the HealthBench hard subset. The results are summarized in Table 4.

Impact of the Profiler The Profiler module is designed to extract the user's background and underlying needs, enabling personalized response generation. Removing this component is therefore expected to reduce personalization in scenarios that require a deeper understanding of the user. As shown in Table 4, the absence of the Profiler leads to notable performance drops in *communications* and *complex responses*. These degradations align with our expectations, as removing the Profiler limits the model's ability to infer the appropriate level of detail and tailor responses to user-specific contexts.

Impact of the Formulator The Formulator module is responsible for identifying and structuring salient clinical indicators from the conversation history. Without this module, the model must rely on unstructured context, which can hinder coherent reasoning over clinical details. Consistent with this intuition, removing the Formulator results in substantial declines in *completeness* and *context awareness*, along with a modest reduction in *accuracy*. These findings highlight the importance of the Formulator in organizing clinical information and ensuring that relevant conditions are explicitly addressed during medical question answering.

Impact of the Strategist The Strategist module governs response delivery by shaping tone, framing, and discourse strategy. Its removal primarily

Table 4: **Ablation study on the HealthBench hard subset.** DeCode denotes the complete framework. Each ablation independently removes a single component (Profiler, Formulator, or Strategist), while keeping the remaining modules unchanged. Values report absolute scores, with inline deltas indicating changes relative to DeCode. Deltas larger than 2 points are highlighted.

Metric (↑, %)	**DeCode**	**w/o Profiler**	**w/o Formulator**	**w/o Strategist**
Overall Score	49.8	49.3 (-0.5)	39.7 **(-10.1)**	49.4 (-0.4)
Themes				
Emergency Referrals	59.1	61.3 **(+2.2)**	52.9 **(-6.2)**	57.9 (-1.2)
Context Seeking	58.3	57.6 (-0.7)	44.2 **(-14.1)**	59.4 (+1.1)
Global Health	49.8	50.7 (+0.9)	39.6 **(-10.2)**	51.5 (+1.7)
Health Data Tasks	35.6	34.2 (-1.4)	32.1 **(-3.5)**	40.0 **(+4.4)**
Communications	43.8	41.4 **(-2.4)**	33.6 **(-10.2)**	43.0 (-0.8)
Hedging	54.6	55.5 (+0.9)	43.9 **(-10.7)**	54.1 (-0.5)
Complex Responses	41.3	35.4 **(-5.9)**	30.8 **(-10.5)**	30.3 **(-11.0)**
Axes				
Accuracy	54.3	54.1 (-0.2)	47.5 **(-6.8)**	52.6 (-1.7)
Completeness	58.8	58.0 (-0.8)	43.9 **(-14.9)**	59.7 (+0.9)
Communication Quality	54.2	54.4 (+0.2)	59.0 **(+4.8)**	47.7 **(-6.5)**
Context Awareness	40.5	39.6 (-0.9)	29.6 **(-10.9)**	44.2 **(+3.7)**
Instruction Following	46.5	47.9 (+1.4)	49.2 **(+2.7)**	44.2 **(-2.3)**

affects how information is communicated rather than what information is presented. As observed in our results, ablating the Strategist leads to a pronounced drop in *communication quality* and a smaller but consistent decline in *instruction following*. Both axes reflect how effectively responses engage with and adapt to user expectations. These results underscore the role of the Strategist in ensuring that medically relevant content is conveyed in an appropriate, user-receptive manner.

6 Conclusion

In this work, we introduce **Decoupling Content and Delivery (DeCode)**, a modular framework for contextualized medical question answering. DeCode adapts a base LLM into four specialized components—**Profiler, Formulator, Strategist, and Synthesizer**—that jointly structure response generation by explicitly separating medical content reasoning from discourse and delivery. This design enables the model to produce medically accurate responses while remaining sensitive to user context and communication needs.

Experiments on the OpenAI HealthBench benchmark demonstrate that DeCode consistently outperforms a zero-shot baseline and remains competitive with leading multi-agent frameworks across both full and hard evaluation settings. Moreover, evaluations across multiple base LLMs show that DeCode generalizes well across different model families and architectures, highlighting its model-agnostic nature.

Future work may explore mechanisms for caching and updating patient-specific information across multi-round interactions, enabling stronger long-term personalization. Additionally, extending the DeCode paradigm beyond medical QA to other high-stakes, user-centered domains represents a promising direction. Taken together, these results suggest that DeCode provides a principled and extensible foundation for advancing contextualized medical question answering.

7 Limitations

Our evaluation is conducted on simulated patient–clinician conversations, which may not fully reflect the complexity, uncertainty, and risk profiles of real-world clinical settings. Although DeCode improves response quality without additional training, outputs generated by large language models may still contain errors or omissions and should not be used as a substitute for professional medical judgment. Validation and safeguards are necessary before deploying systems in clinical practice.

References

Anthropic. 2025. Claude 4.5 sonnet. https://www.anthropic.com/news/claude-sonnet-4-5.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. HealthBench: Evaluating large language models towards improved human health. *Preprint*, arXiv:2505.08775.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.

Eun Jeong Gong, Chang Seok Bang, Jae Jun Lee, and Gwang Ho Baik. 2025. Knowledge-Practice Performance Gap in Clinical Large Language Models: Systematic review of 39 benchmarks. *Journal of Medical Internet Research*, 27:e84120.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented Large Language Models. *Bioinformatics*, 40(Supplement 1):i119–i129.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease does this Patient Have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: an adaptive collaboration of llms for medical decision-making. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2024. Agent Hospital: A simulacrum of hospital with evolvable medical agents. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 53–62. ACM.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Preprint*, arXiv:2311.16452.

OpenAI. 2024. Openai o1 system card. Technical report, OpenAI.

OpenAI. 2025a. Gpt 4.1 system card. https://openai.com/index/gpt-4-1/.

OpenAI. 2025b. Gpt-5 system card. https://openai.com/index/gpt-5-system-card.

OpenAI. 2025c. Openai o3 system card. https://openai.com/index/introducing-o3-and-o4-mini/.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, and 48 others. 2024. Capabilities of gemini models in medicine. *Preprint*, arXiv:2404.18416.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomašev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, and 7 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450.

Sean Wu, Michael Koo, Fabien Scalzo, and Ira Kurtz. 2025a. AutoMedPrompt: A new framework for optimizing LLM medical prompts using textual gradients. *Preprint*, arXiv:2502.15944.

Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, Yanyuan Qiao, Imran Razzak, and Yutong Xie. 2025b. A knowledge-driven adaptive collaboration of LLMs for enhancing medical decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33483–33500, Suzhou, China. Association for Computational Linguistics.

Yuxuan Zhou, Yubin Wang, Bin Wang, Chen Ning, Xien Liu, Ji Wu, and Jianye Hao. 2025. Enhancing the medical context-awareness ability of llms via multifaceted self-refinement learning. *Preprint*, arXiv:2511.10067.

A Prompt Templates

The prompts for each module of **DeCode** are provided below. The Profiler module uses two independent prompts to extract the user background $\mathcal{B}$ and user needs $\mathcal{N}$. The prompt to extract the user background $\mathcal{B}$ and user needs $\mathcal{N}$ is provided in Figure 2 and Figure 3. The Formulator prompt to extract the clinical indicators $\mathcal{C}$ is listed in Figure 4. The Strategist prompt is given in Figure 5. Finally, the Synthesizer prompt is presented in Figure 6.

User Background ($\mathcal{B}$)

```
You are a medical intake specialist. Analyze the following conversation and extract the user's
background information.

CONVERSATION:
{conversation_history}

Extract and infer the following information about the user (if available in the conversation):
- Age or age group
- Career/Occupation
- Economic condition (inferred from context)
- Living place/location
- Living situation (alone, with family, etc.)
- Any other relevant personal context

IMPORTANT: Only include information that can be reasonably inferred from the conversation. Do NOT
make up information.

Respond in this EXACT format:

AGE: [age or age group, or "Not specified"]
CAREER: [occupation, or "Not specified"]
ECONOMIC_CONDITION: [economic status inferred from context, or "Not specified"]
LIVING_PLACE: [location/region, or "Not specified"]
LIVING_SITUATION: [living arrangement, or "Not specified"]
OTHER_CONTEXT: [any other relevant information, or "None"]

Be concise and factual. If information is not available, write "Not specified" or "None".
```

Figure 2: Prompt template for the User Background extraction.

User Need ($\mathcal{N}$)

```
You are analyzing a medical conversation to understand what the user needs.

CONVERSATION:
{conversation_history}

Identify what the user explicitly asks for or clearly needs. Be conservative - only include needs
that are:
1. Explicitly stated by the user
2. Clearly implied by the user's questions or concerns

DO NOT include:
- Things the user might need but didn't mention
- General medical advice that wasn't requested
- Assumptions about what the user should want

Respond in this EXACT format:

NEEDS:
1. [First explicit need]
2. [Second explicit need]
3. [Third explicit need]
...

If the user doesn't clearly state what they want, respond with:
NEEDS:
None specified

Be strict and conservative.
```

Figure 3: Prompt template for the User Need identification.

Clinical Indicators ($\mathcal{C}$)

```
You are a clinical safety and completeness planner.

Your ONLY job is to identify the medically important content that MUST be covered
for this case to be safe, accurate, and reasonably complete. You are NOT deciding tone or style.
You are optimizing for clinical accuracy and completeness, not brevity.

CONVERSATION:
{conversation_history}

Create a numbered list of key clinical content items that the final answer should try to cover,
such as:
- Important symptom details or history that should be addressed or clarified
- Key possible causes or differentials (described in a cautious, non-diagnostic way)
- Red-flag or emergency warning signs that should be mentioned if relevant
- What the user can monitor or do at home (if appropriate)
- When and how urgently they should seek in-person care
- Any important limitations or uncertainties of online advice

Rules:
- Focus on clinical content ONLY (WHAT to cover), not HOW to phrase it.
- Err on the side of including any clinically important point that might affect safety.
- Each item should be 1-2 sentences max.
- Avoid repeating the same content in multiple items.
- Do not invent new symptoms; only build on what is in the conversation.
- It is acceptable to mention reasonable possible causes or scenarios even if the user did not use
those exact words, as long as they logically follow from the described symptoms.

Respond in this EXACT format:
1. [Clinical content item]
2. [Clinical content item]
3. [Clinical content item]
...
```

Figure 4: Prompt template for the Formulator module ($\mathcal{M}_{form}$).

Discourse Strategy ($\mathcal{S}$)

```
You are a response-strategy planner for a medical assistant.

You receive:
- The original conversation
- A brief user background profile
- A list of what the user clearly needs
- A clinical content checklist (what should be covered for safety/completeness)

Your job is to design HOW the assistant should answer for THIS user: what to prioritize, how deep
to go, what style and structure to use, and what to avoid.

CONVERSATION:
{conversation_history}

USER BACKGROUND PROFILE:
{user_profile}

USER NEEDS (what the user clearly wants):
{needs_formatted}

CLINICAL CONTENT CHECKLIST (what should be covered):
{content_formatted}

Pay particular attention to:
- Whether the user's needs are clearly stated or vague/unspecified.
- Whether there is sufficient information available for a safe medical assessment.
- When needs or information are unclear, the plan should usually include a brief strategy for
clarifying key gaps (e.g., 1-2 focused questions), while still guiding the assistant to give the
best possible provisional answer based only on what is already known.

IMPORTANT:
- The assistant MUST still give concrete, practical, medically useful information even when
information is incomplete. Use conditional language (e.g., "If X..., then Y...") rather than
refusing to say anything.
- Do NOT tell the assistant to avoid discussing possible causes or next steps entirely.
- Clarification questions should be few (0-2 of the most important ones) and should not dominate
the answer.

Design a plan with TWO sections:

1. WHAT TO DO/COVER (TO DO):
- How the assistant should prioritize and present the content for THIS user.
- What level of technical detail is appropriate for this user.
- Whether to keep the answer short vs. more detailed.
- Whether to explicitly ask clarification questions (0-2 key questions only), and if so, in what
style and at what point (usually after giving main guidance).
- Which content items from the checklist are highest priority to cover explicitly.
- How to adapt the response to the user's apparent role, location, and constraints.

2. WHAT NOT TO DO/COVER (NOT TO DO):
- Things that would likely confuse, overwhelm, or frustrate THIS user.
- Styles to avoid (e.g., too technical, too casual, too vague, overly long).
- Types of content to avoid (e.g., extremely long, low-yield lists of differential diagnoses;
strong reassurance when red flags are possible; rigid instructions when access is limited).
- Any ways of answering that would clearly conflict with the user's instructions.

You are NOT writing the final medical answer. You are only writing the plan.

Respond in this EXACT format:

TO DO:
1. [Response strategy / priority tailored to user]
2. [Another response strategy / priority]
3. [Continue as needed]

NOT TO DO:
1. [Specific thing to avoid for this user]
2. [Another thing to avoid]
3. [Continue as needed]
```

Figure 5: Prompt template for the Strategist module ($\mathcal{M}_{strat}$).

Controlled Generation (R)

```
You are an experienced medical professional providing personalized advice.

Your highest priorities are:
1) Clinical accuracy and completeness of the information you provide.
2) Clear, practical guidance for the user.
3) Safe and appropriate communication.

ORIGINAL CONVERSATION:
{conversation_history}

PRESENTATION GUIDELINES (HOW TO ANSWER):

TO DO:
{to_do_formatted}

NOT TO DO:
{not_to_do_formatted}

CONTENT CHECKLIST (WHAT YOU MUST COVER CLINICALLY):
{content_formatted}

Your task:
1. Cover ALL items in the CONTENT CHECKLIST as clearly and concretely as possible. Aim for at
least one explicit sentence or short paragraph addressing each item.
2. Follow the TO DO / NOT TO DO guidelines for how to present the information in a way that fits
THIS user's background and needs.
3. Be explicit about uncertainty and information gaps, but still give the BEST POSSIBLE DIRECT
ANSWER based only on the conversation.
- Use conditional language (e.g., "If X..., then Y...") rather than refusing to answer.
4. You may ask up to 1-2 of the most important clarification questions, but they should be placed
near the end and should NOT replace giving guidance.
5. Keep the response user-centered and practical, and explain what the user can do next (e.g.,
monitor, self-care, when/where to seek in-person care).
6. End with a brief reminder that this information does not replace an in-person medical
evaluation and that the user should seek care if they are worried or if concerning symptoms arise.

Provide your response:
```

Figure 6: Prompt template for the Synthesizer module ($\mathcal{M}_{syn}$).