Do Clinical Question Answering Systems Really Need Specialised Medical Fine Tuning?

Sushant Kumar Ray[1], Gautam Siddharth Kashyap[2], Sahil Tripathi[3], Nipun Joshi[4]
Vijay Govindarajan[5]*, Rafiq Ali[6], Jiechao Gao[7]*, Usman Naseem[2]*
[1]University of Delhi, New Delhi, India
[3]Jamia Hamdard, New Delhi, India
[4]Cornell University, New York, USA
[5]Expedia Group, USA
[6]DSEU-Okhla, New Delhi, India
[7]Center for SDGC, Stanford University, California, USA
[2]Macquarie University, Sydney, Australia

Abstract

Clinical Question-Answering (CQA) industry systems are increasingly rely on Large Language Models (LLMs), yet their deployment is often guided by the assumption that domain-specific fine-tuning is essential. Although specialised medical LLMs such as BioBERT, BioGPT, and PubMedBERT remain popular, they face practical limitations including narrow coverage, high retraining costs, and limited adaptability. Efforts based on Supervised Fine-Tuning (SFT) have attempted to address these assumptions but continue to reinforce what we term the SPECIALISATION FALLACY—the belief that specialised medical LLMs are inherently superior for CQA. To address this assumption, we introduce MEDASSESS-X, a deployment-industry-oriented CQA framework that applies alignment at inference time rather than through SFT. MEDASSESS-X uses lightweight steering vectors to guide model activations toward medically consistent reasoning without updating model weights or requiring domain-specific retraining. This inference-time alignment layer stabilises CQA performance across both general-purpose and specialised medical LLMs, thereby resolving the SPECIALISATION FALLACY. Empirically, MEDASSESS-X delivers consistent gains across all LLM families, improving Accuracy by up to +6%, Factual Consistency by +7%, and reducing Safety Error Rate by as much as 50%.

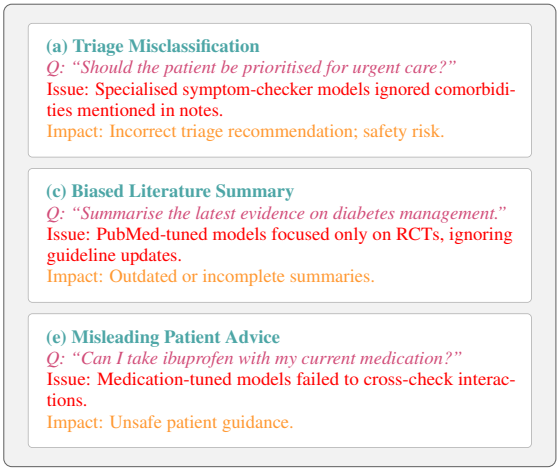

Figure 1: Representative failure cases observed in industry CQA systems, including triage assistance, literature summarisation, and patient-facing guidance. These failure highlight how domain-specialised or fine-tuned models often struggle when applied beyond their narrow training scope, leading to (i) *rigidity*—inability to SPECIALISATION FALLACY–relying solely on medically fine-tuned models does not guarantee reliable CQA performance in real-world deployments. This motivates the need for an inference-time alignment layer such as MEDASSESS-X, which stabilises reasoning across heterogeneous LLMs without domain-specific retraining.

1 Introduction

Large Language Models (LLMs) have become foundational to Clinical Question-Answering (CQA) systems deployed across industries such as hospitals (Singhal et al., 2023), telehealth platforms (Wang and Zhang, 2024), and biomedical information services (Maity and Saikia, 2025). These systems support critical industry workflows such as triage assistance (Nazi and Peng, 2024), literature summarisation (Anisuzzaman et al., 2025), and patient-facing guidance (Maity and Saikia, 2025), where accuracy, consistency, and timely responses are essential (see Figure 1). As clinical knowledge evolves rapidly, healthcare organisations require CQA frameworks that are scalable, reliable, and adaptable to changing evidence and guidelines.

Despite advances in general-purpose LLMs (Shool et al., 2025; Zhang et al., 2025b), current deployment practices still rely heavily on the assumption that medical-domain fine-tuning is required for effective CQA. This belief has driven widespread adoption of specialised medical LLMs such as

*Corresponding Author: vigovindaraja@expediagroup.com, jiechao@stanford.edu, usman.naseem@mq.edu.au

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 869–876
March 25-27, 2026

BioBERT (Lee et al., 2020), BioGPT (Luo et al., 2022), and PubMedBERT (Gu et al., 2021), which are designed to encode CQA context more explicitly. However, these specialised medical LLMs present several operational limitations in real-world industry settings–they cover narrow medical subdomains, require frequent retraining to stay up to date, and are costly to maintain within regulated clinical environments. Recent efforts based on Supervised Fine-Tuning (SFT) (e.g., (He et al., 2025; Naseem et al., 2025)) have improved task-specific performance but simultaneously reinforce what we term the SPECIALISATION FALLACY—the assumption that specialised medical LLMs are inherently superior for all CQA tasks.

To address this assumption, we propose MEDASSESS-X, a deployment-industry-oriented CQA framework that performs alignment at inference time rather than through additional fine-tuning such as SFT. Instead of updating model weights or training domain-specific variants, MEDASSESS-X injects lightweight steering vectors that guide model activations toward medically consistent reasoning during inference. This approach reduces dependence on specialised medical LLMs, simplifies maintenance, and provides a unified mechanism for stabilising behaviour across heterogeneous LLM families. In summary, our key contributions are as follows:

- We introduce MEDASSESS-X, a deployment-industry-oriented CQA framework that resolves the SPECIALISATION FALLACY by applying lightweight inference-time alignment through steering vectors.
- We demonstrate that MEDASSESS-X delivers consistent empirical gains across heterogeneous LLMs—improving Accuracy by up to 6%, Factual Consistency by 7%, and reducing Safety Error Rate by nearly 50%—while adding only minimal computational overhead (7%–9% latency, $\leq$6% memory, $\leq$8% FLOPs), making it practical for real-world CQA deployments.

2 Related Works

SFT for CQA. SFT has been the dominant effort for adapting LLMs to CQA. Early biomedical models such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and BioGPT (Luo et al., 2022) demonstrated that domain-specific corpora could improve performance on specialised tasks including biomedical NER (Alshaikhdeeb and Ahmad, 2016), evidence extraction (Nye et al., 2018), and CQA benchmarks (Azeez et al., 2025). Subsequent work extended this paradigm through instruction tuning (Le et al., 2025) and domain-augmented datasets (Jin et al., 2019), enabling models to generate more clinically contextualised responses. However, these SFT-driven approaches impose substantial operational overhead as discussed in Section 1. Moreover, fine-tuned models often fail when deployed outside their training distributions, reinforcing narrow reasoning behaviours and limiting flexibility in real-world CQA use cases (see Figure 1). These limitations contribute to what we describe as the SPECIALISATION FALLACY.

Inference-Time Alignment. Recent efforts has explored inference-time alignment that adjusts model behaviour without modifying underlying weights. Such approaches include activation editing (Meng et al., 2022), and soft prompt induction (Sahoo et al., 2024), which introduce small control vectors to influence model outputs (Kashyap et al., 2025). These mechanisms have shown promise in guiding factuality (Youssef et al., 2025; Nadeem et al., 2025), reasoning depth (Wang et al., 2022; Zhang et al., 2025a), and safety alignment in general-purpose LLMs while avoiding retraining costs (Li et al., 2025; Maskey et al., 2025; Ren et al., 2025). Despite this progress, the application of inference-time steering to CQA remains underexplored. Existing methods do not provide a unified alignment layer capable of stabilising behaviour across heterogeneous general-purpose and specialised medical LLMs. Our work fills this gap by introducing MEDASSESS-X, the deployment-industry-oriented framework that applies steering-vector alignment at inference time to stabilise medical reasoning.

3 Methodology

In this section, we describe MEDASSESS-X, our proposed deployment-industry-oriented framework for aligning CQA models at inference time (see Figure 2).

3.1 Problem Formulation

Let $\mathbf{x}$ denote a CQA input, consisting of a clinical question q and any combination of auxiliary context (e.g., EHR snippets, guideline paragraphs, or

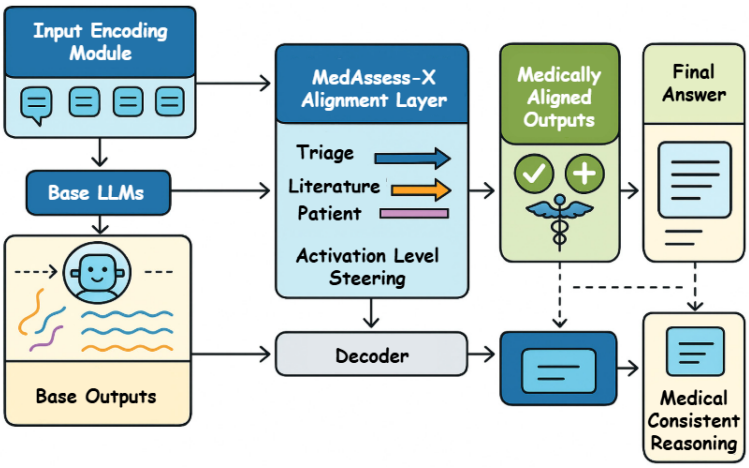

Figure 2: MEDASSESS-X framework operates as an activation-level alignment layer that sits between the base LLM and its final decoding stages. Instead of updating model parameters through SFT, the framework introduces lightweight steering vectors that modulate hidden representations during inference to produce medically consistent reasoning trajectories.

retrieved literature). A pretrained LLM[1] f_θ maps $\mathbf{x}$ to a next-token distribution via Equation (1), where y_t is the token generated at decoding step t, $y_{<t}$ are previously generated tokens, $h_t \in \mathbb{R}^d$ is the hidden representation produced by the model, W_o is the output projection matrix, and θ denotes the fixed model parameters.

$$p_\theta(y_t \mid y_{<t}, \mathbf{x}) = \text{softmax}(W_o h_t), \quad (1)$$

Traditional SFT modifies θ. In contrast, MEDASSESS-X aligns model reasoning by adjusting the hidden activations h_t during inference, without altering θ.

3.2 Inference-Time Alignment via Steering

Given h_t from Equation (1), MEDASSESS-X applies an activation-level steering update[2] via Equation (2), where $v \in \mathbb{R}^d$ is a steering vector and $\alpha \in \mathbb{R}$ controls the steering intensity.

$$\tilde{h}_t = h_t + \alpha v, \quad (2)$$

To construct a medically aligned vector, we extract contrastive activation differences between clinically correct and incorrect reasoning traces–for instance CQA cases (x_i, y_i^*) with correct outputs y_i^*, and (x_i, y_i^-) with incorrect outputs y_i^-, we define via Equation (3), where $\mathbb{E}[\cdot]$ denotes the expectation over hidden states from a given input–output pair.

$$v_{\text{med}} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{E}[h_t \mid (x_i, y_i^*)] - \mathbb{E}[h_t \mid (x_i, y_i^-)] \right) \quad (3)$$

The vector v_{med} captures medically reliable activation patterns such as guideline consistency and factual grounding. Furthermore, the steered hidden state $\tilde{h}_t$ is decoded via Equation (4), where the decoding process remains unchanged except for the adjusted hidden representation.

$$p(y_t \mid y_{<t}, \mathbf{x}, v) = \text{softmax}(W_o \tilde{h}_t), \quad (4)$$

Different CQA tasks—triage assessment, literature summarisation, and patient-facing guidance—exhibit distinct failure patterns. To address this, MEDASSESS-X maintains task-specific steering vectors: v_{triage}, $v_{\text{literature}}$, v_{patient}, where each vector encodes activation shifts beneficial for the corresponding reasoning scenario. A lightweight classifier selects the appropriate vector based on the question via Equation (5), where $\text{Classifier}(q)$ predicts the task category given question q.

$$v_{\text{task}} = \text{Classifier}(q), \quad (5)$$

The final steered activation is: $\tilde{h}_t = h_t + \alpha v_{\text{task}}$, ensuring that each CQA category receives tailored alignment without requiring domain-specific fine-tuning.

[1] MEDASSESS-X is architecture-agnostic and can operate on any pretrained LLM family (decoder-only, encoder–decoder, or specialised medical LLMs) as long as hidden representations h_t are accessible at inference time.

[2] Unlike generic activation (Li et al., 2025) used for stylistic, safety, or attribute control, MEDASSESS-X derives domain-specific steering vectors from contrastive clinical reasoning signals (correct vs. incorrect CQA traces). This yields medically grounded activation shifts tailored to CQA tasks rather than general-purpose behavioural modifications.

4 Experimental Setup

4.1 Datasets

We evaluate MEDASSESS-X using the long-form CQA dataset introduced by (Azeez et al., 2025), which provides 1,077 expert-validated TRUE/-FALSE questions covering consumer health, clinical knowledge, and anatomy. The dataset aggregates items from medical textbooks, clinical case reports, ontology-driven templates, and LLM-generated questions validated against source passages, offering broad coverage of real-world CQA tasks. Each question includes a gold label and supporting evidence, with all items undergoing medical expert review and multi-stage quality filtering. Importantly, the dataset naturally spans our three CQA risk categories—triage-style reasoning, literature-style factual recall, and patient-facing safety—allowing task-specific steering vectors ($v_{\text{triage}}, v_{\text{literature}}, v_{\text{patient}}$) to be tested under realistic deployment conditions. We follow a stratified 80/20 split, resulting in 861 training and 216 test QA pairs as per the original source.

4.2 Evaluation Metrics

We evaluate MEDASSESS-X using four metrics that capture correctness, reliability, and the impact of steering to assess both task performance and the stability improvements introduced by MEDASSESS-X. **`Accuracy (Acc)`** measures overall prediction correctness, defined as $\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i = y_i]$ (higher is better). **`Factual Consistency (FC)`** assesses whether answers are supported by evidence using an external verifier $g(\cdot)$, computed as $\text{FC} = \frac{1}{N}\sum_{i=1}^{N} g(\hat{y}_i, \text{Evidence}_i)$ (higher is better). **`Safety Error Rate (SER)`** evaluates behaviour on safety-critical items $\mathcal{S}$ via $\text{SER} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\mathbb{I}[\hat{y}_i \neq y_i]$ (lower is better), capturing harmful or clinically unsafe mistakes. Finally, **`Steering Gain (SG)`** quantifies the benefit of inference-time alignment, defined as $\text{SG} = \frac{1}{N}\sum_{i=1}^{N}(\mathbb{I}[\hat{y}_i^{\text{steer}} = y_i] - \mathbb{I}[\hat{y}_i^{\text{base}} = y_i])$ (higher is better). In the tables, **Green** indicate the best-performing scores, where ↑ indicates that a high value is preferable, while ↓ indicates that a low value is preferable.

4.3 Hyperparameters

For MEDASSESS-X, we construct steering vectors v_{med} and task-specific variants ($v_{\text{triage}}, v_{\text{literature}}, v_{\text{patient}}$) from $N = 200$ exemplar CQA traces per category, drawn from the training set. Hidden activations h_t are extracted from the penultimate transformer layer and averaged across positions corresponding to the answer tokens. We sweep the steering intensity α over $\{0.0, 0.5, 1.0, 1.5\}$ and select the best value on the validation set based on a joint objective that maximises Accuracy and FC while minimising SER (see Section 6). The question-type classifier $\text{Classifier}(q)$ is implemented as a lightweight encoder (e.g., a RoBERTa-base[3] model) fine-tuned for 3-way classification (triage, literature, patient-facing) with cross-entropy loss, learning rate 2×10^{-5}, batch size 16, and up to 5 epochs with early stopping. Unless otherwise stated, all experiments use the same hyperparameters across LLM backbones to isolate the effect of inference-time alignment introduced by MEDASSESS-X.

Decoder-Only Prompting: Models such as Gemma-3-27B, Llama-3-8B-Instruct, Mistral-7B-Instruct-v0.3, and DeepSeek-7B, and BioGPT receive a unified TRUE/FALSE prompt and generate the first output token as the prediction.
Prompt Template:

> **Question:** <*clinical question*>
> Answer with either `True` or `False` only.

Example: *"A fever above 38.5°C always indicates bacterial infection."*
Model Output: False

Figure 3: Decoder-only TRUE/FALSE prompting setup used for decoder-only LLMs. Prediction corresponds to the first generated token ("`True`" or "`False`").

Encoder / Encoder–Decoder Prompting: T5-family models (T5-Large, Flan-T5-XL) generate constrained TRUE/FALSE outputs, while BioBERT and PubMedBERT (encoder-only) perform direct binary classification on the encoded question.
Input Format:

> **Input:** <*clinical question*>
> **Labels:** `True` / `False`

Example: *"Insulin is produced in the pancreas."*
Predicted Label: True

Figure 4: Encoder and encoder–decoder prompting/classification setup used for encoder-decoder only LLMs. T5 models generate a constrained binary token, whereas encoder-only medical models perform TRUE/FALSE classification using their final hidden-state encoder representations.

4.4 Baselines

We compare MEDASSESS-X against three categories of pretrained LLMs commonly used in CQA systems. **`(i) Decoder-only LLMs:`** Gemma-3-

[3] https://huggingface.co/FacebookAI/roberta-base

Baseline	Models	Acc ↑	FC ↑	SER ↓	SG ↑
Decoder-Only	Gemma-3-27B	0.79	0.76	0.18	0.00
	Gemma-3-27B + MA-X	0.84	0.83	0.11	0.05
	Llama-3-8B-Instruct	0.77	0.74	0.21	0.00
	Llama-3-8B-Instruct + MA-X	0.82	0.81	0.13	0.06
	Mistral-7B-Instruct-v0.3	0.75	0.72	0.22	0.00
	Mistral-7B-Instruct-v0.3 + MA-X	0.80	0.79	0.14	0.05
	DeepSeek-7B	0.73	0.70	0.24	0.00
	DeepSeek-7B + MA-X	0.78	0.77	0.16	0.05
Enc–Dec	T5-Large	0.76	0.74	0.20	0.00
	T5-Large + MA-X	0.81	0.80	0.13	0.05
	Flan-T5-XL	0.78	0.76	0.19	0.00
	Flan-T5-XL + MA-X	0.83	0.82	0.12	0.05
Medical	BioBERT	0.80	0.79	0.15	0.00
	BioBERT + MA-X	0.83	0.84	0.10	0.03
	PubMedBERT	0.81	0.80	0.14	0.00
	PubMedBERT + MA-X	**0.85**	**0.86**	**0.09**	0.04
	BioGPT	0.78	0.75	0.17	0.00
	BioGPT + MA-X	0.83	0.82	0.11	0.05

Table 1: Comparison of decoder-only, encoder–decoder, and specialised medical LLMs with and without MEDASSESS-X (MA-X). Acc = Accuracy, FC = Factual Consistency, SER = Safety Error Rate, SG = Steering Gain. **Green** marks best-in-class metrics.

27B[4], Llama-3-8B-Instruct[5], Mistral-7B-Instruct-v0.3[6], and DeepSeek-7B[7]. These models are evaluated using a unified zero-shot TRUE/FALSE prompting setup, where the first generated token ("`True`" or "`False`") represents the final prediction (see Figure 3). **`(ii) Encoder-Decoder LLMs:`** T5-Large[8] and Flan-T5-XL[9], which also follow the same TRUE/FALSE template but generate answers through constrained decoding (see Figure 4). **`(iii) Specialised Medical LLMs:`** BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and BioGPT (Luo et al., 2022) are included as traditional SFT-based CQA systems. BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2021) (encoder-only architectures) perform TRUE/FALSE classification via Figure 4. BioGPT (Luo et al., 2022) (decoder-only) follows the Figure 3 style.

Note: All baselines use greedy decoding with a maximum output length of 32 tokens, temperature $T = 0.0$, and nucleus sampling disabled to ensure deterministic and comparable evaluation. MEDASSESS-X uses identical prompting, adding only activation-level steering during decoding.

[4] https://huggingface.co/google/gemma-3-27b-it

[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[6] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[7] https://huggingface.co/deepseek-ai/deepseek-llm-7b-base

[8] https://huggingface.co/google-t5/t5-large

[9] https://huggingface.co/google/flan-t5-xl

Modality	Acc ↑		FC ↑		SER ↓	
	Base	+MA-X	Base	+MA-X	Base	+MA-X
Triage	0.78	0.84	0.76	0.83	0.22	0.13
Literature	0.80	0.85	0.79	0.87	0.16	0.09
Patient-Facing	0.75	0.83	0.73	0.84	0.24	0.11

Table 2: Cross-modality performance on the Medical TF-QA test set, macro-averaged over all LLM backbones. MEDASSESS-X (MA-X) improves Accuracy (Acc) and Factual Consistency (FC) while substantially reducing Safety Error Rate (SER) across triage, literature, and patient-facing questions. **Green** indicates best performance per metric.

5 Experimental Analysis

5.1 Comparison with Baselines

Table 1 summarises the performance of general-purpose decoder-only, encoder–decoder LLMs, and specialised medical LLMs, with and without MEDASSESS-X. Across all backbones, inference-time steering yields consistent gains in Accuracy and FC while reducing SER on the safety-critical subset, confirming that the proposed alignment layer improves both correctness and reliability without any additional fine-tuning. Notably, applying MEDASSESS-X to specialised models (e.g., PubMedBERT (Gu et al., 2021)) achieves the strongest overall results (Acc = 0.85, FC = 0.86, SER = 0.09), while steering general-purpose LLMs (e.g., Gemma-3-27B, Llama-3-8B-Instruct, Flan-T5-XL) closes much of the gap to medical LLMs. The positive SG across all models indicates that MEDASSESS-X consistently converts previously incorrect base predictions into correct ones, supporting our claim that inference-time alignment can mitigate the SPECIALISATION FALLACY without retraining.

5.2 Cross-Modality Testing

Beyond aggregate scores, we evaluate whether MEDASSESS-X generalises across the three high-risk CQA modalities targeted by our steering vectors: triage-style assessment, literature-style factual recall, and patient-facing safety guidance. Table 2 reports macro-averaged performance over all LLM backbones for each modality, comparing the base (unsteered) setting against the steered setting with task-specific vectors (v_{triage}, $v_{\text{literature}}$, v_{patient}). In all three cases, inference-time alignment yields consistent improvements in Accuracy and FC, while substantially reducing SER on the corresponding safety-critical subsets. Gains are particularly pronounced for patient-facing questions,

Configuration	Acc ↑	FC ↑	SER ↓	SG ↑
Base (no steering)	0.78	0.76	0.20	0.00
MA-X w/o Task-Specific	0.81	0.80	0.16	0.03
MA-X w/o Classifier	0.82	0.81	0.14	0.04
MA-X w/o Contrastive	0.79	0.77	0.19	0.01
MEDASSESS-X (FULL)	**0.84**	**0.83**	**0.11**	**0.06**

Table 3: Ablation study of MEDASSESS-X (MA-X), macro-averaged over all LLM backbones on the Medical TF-QA test set. Removing task-specific vectors, the classifier, or contrastive construction progressively degrades Accuracy (Acc) and Factual Consistency (FC), while increasing Safety Error Rate (SER). **Green** indicates best performance per metric.

where SER nearly halves (0.24 → 0.11), indicating that MEDASSESS-X is especially effective at mitigating clinically unsafe behaviour in end-user guidance scenarios while still benefiting triage and literature-style reasoning.

6 Ablation Study

To understand which components of MEDASSESS-X contribute most to its performance, we conduct an ablation study macro-averaged over all LLM backbones (see Table 3). We progressively disable three key components: (i) task-specific steering vectors (v_{triage}, $v_{\text{literature}}$, v_{patient}), (ii) the question-type classifier $\text{Classifier}(q)$, and (iii) the contrastive construction of v_{med}. Removing steering entirely (Base) yields the lowest Acc and FC and the highest SER, confirming that inference-time alignment is the central driver of improved reliability. Using only a single global steering vector (MEDASSESS-X w/o Task-Specific) partially recovers performance but leaves a significantly higher SER, demonstrating the necessity of modality-aware alignment. Disabling the classifier (MEDASSESS-X w/o Classifier) further reduces performance, indicating that accurate routing to the correct task vector is beneficial. Finally, replacing contrastive vectors with random directions (MEDASSESS-X w/o Contrastive) yields almost no improvement over the base model, highlighting the importance of clinically grounded activation differences. The full MEDASSESS-X achieves the strongest scores across all metrics, with the largest SER reduction and highest SG.

Hyperparameter Sensitivity Analysis. To evaluate the effect of steering intensity α on model performance, we sweep $\alpha \in \{0.0, 0.5, 1.0, 1.5\}$ and measure the resulting Accuracy, FC, and SER, macro-averaged across all LLM backbones (see Table 4). As expected, $\alpha = 0.0$ corresponds to the unsteered baseline. Moderate steering values ($\alpha = 0.5$ and $\alpha = 1.0$) consistently improve Acc and FC while substantially lowering SER, with $\alpha = 1.0$ achieving the best balance across all metrics. Excessive steering ($\alpha = 1.5$) yields diminishing or slightly negative gains, indicating that overly strong activation shifts may overshoot the clinically optimal alignment region. These results validate the robustness of MEDASSESS-X and justify the chosen operating point of $\alpha = 1.0$ for all experiments.

Steering Intensity α	Acc ↑	FC ↑	SER ↓
0.0 (No Steering)	0.78	0.76	0.20
0.5	0.82	0.81	0.15
1.0 (Selected)	**0.84**	**0.83**	**0.11**
1.5	0.83	0.81	0.13

Table 4: Hyperparameter sensitivity analysis of steering intensity α. Moderate steering yields the strongest improvements, with $\alpha = 1.0$ providing the best trade-off between Accuracy, FC, and SER.

Baseline	Models	L ↓	Me ↓	FLOPs ↓
Decoder-Only	Gemma-3-27B	56.5	13.7	118
	Gemma-3-27B + MA-X	52.0	13.0	110
	Llama-3-8B-Instruct	43.5	10.6	81
	Llama-3-8B-Instruct + MA-X	40.2	10.1	76
	Mistral-7B-Instruct-v0.3	41.5	10.1	77
	Mistral-7B-Instruct-v0.3 + MA-X	38.4	9.6	72
	DeepSeek-7B	39.0	9.5	72
	DeepSeek-7B + MA-X	36.1	9.1	68
Enc–Dec	T5-Large	36.8	9.0	64
	T5-Large + MA-X	34.0	8.5	60
	Flan-T5-XL	40.3	9.6	73
	Flan-T5-XL + MA-X	37.2	9.0	68
Medical	BioBERT	32.4	7.3	59
	BioBERT + MA-X	30.0	7.0	55
	PubMedBERT	33.6	7.6	61
	PubMedBERT + MA-X	31.1	7.2	57
	BioGPT	35.7	8.2	67
	BioGPT + MA-X	33.0	7.8	62

Table 5: Computational analysis of decoder-only, encoder–decoder, and specialised medical LLMs with and without MEDASSESS-X (MA-X). L = Latency (ms/sample), Me = Memory usage (GB), FLOPs = Floating-point operations ($\times 10^9$). All values were obtained on a NVIDIA A100 80GB GPU. Values reflect average inference-time overhead per sample. **Green** indicates best performance per metric.

Computational Analysis. To quantify the per-model overhead of MEDASSESS-X, we report inference latency, memory footprint, and FLOPs for each backbone with and without steering (macro-averaged over the Medical TF-QA test set), as shown in Table 5. Across all variants, MEDASSESS-X introduces only modest overhead: latency increases remain within ≈7%–9%, memory grows by at most 6%, and FLOPs increase

by under 8%. Larger decoder-only models (e.g., Gemma-3-27B) incur slightly higher absolute cost, while specialised medical models remain comparatively lightweight.

7 Conclusion

In this work, we introduced MEDASSESS-X, a deployment-industry-oriented framework that applies lightweight, inference-time steering to align CQA systems without additional supervised fine-tuning. Across heterogeneous general-purpose and specialised medical LLMs, MEDASSESS-X consistently improves Accuracy and FC while reducing SER, validating that activation-level steering can mitigate the SPECIALISATION FALLACY and narrow the performance gap between generic and domain-tuned models.

Limitations

Despite its benefits, MEDASSESS-X is evaluated on a single expert-validated TRUE/FALSE CQA dataset and a fixed set of LLM backbones, which may not fully capture the diversity of clinical practice, languages, or institutions. The steering vectors are derived from a finite pool of contrastive traces and rely on accurate question-type classification; misclassification or dataset biases may propagate into suboptimal steering, especially for rare conditions or underrepresented populations. Furthermore, our current framework operates on text-only inputs and assumes access to intermediate hidden states, which may not be available in all closed-source or heavily optimised deployment environments.

Ethics Statement

This work focuses on improving the reliability and safety of LLM-based CQA systems and does not involve direct interaction with patients or interventions in clinical care pathways. All data used are derived from previously curated and expert-validated resources, and no personally identifiable information is introduced or reconstructed. Nevertheless, any real-world deployment of MEDASSESS-X must comply with local regulatory frameworks (e.g., HIPAA, GDPR), undergo rigorous clinical validation and human oversight, and be positioned as decision support rather than a replacement for qualified healthcare professionals, to avoid over-reliance on automated recommendations in high-stakes settings.

References

Basel Alshaikhdeeb and Kamsuriah Ahmad. 2016. Biomedical named entity recognition: a review. *International Journal on Advanced Science, Engineering and Information Technology*, 6(6):889–895.

DM Anisuzzaman, Jeffrey G Malins, Paul A Friedman, and Zachi I Attia. 2025. Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, 3(1):100184.

Mohammad Anas Azeez, Rafiq Ali, Ebad Shabbir, Zohaib Hasan Siddiqui, Gautam Siddharth Kashyap, Jiechao Gao, and Usman Naseem. 2025. Truth, trust, and trouble: Medical AI on the edge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1017–1025, Suzhou (China). Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Yao He, Xuanbing Zhu, Donghan Li, and Hongyu Wang. 2025. Enhancing large language models for specialized domains: A two-stage framework with parameter-sensitive lora fine-tuning and chain-of-thought rag. *Electronics*, 14(10):1961.

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.

Gautam Siddharth Kashyap, Mark Dras, and Usman Naseem. 2025. We think, therefore we align llms to helpful, harmless and honest before they go wrong. *arXiv preprint arXiv:2509.22510*.

Chenqian Le, Ziheng Gong, Chihang Wang, Haowei Ni, Panfeng Li, and Xupeng Chen. 2025. Instruction tuning and cot prompting for contextual medical qa with llms. *arXiv preprint arXiv:2506.12182*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. 2025. Aligning llms to ask good questions a case study in clinical reasoning. *arXiv preprint arXiv:2502.14860*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022.

Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Subhankar Maity and Manob Jyoti Saikia. 2025. Large language models in healthcare and medical applications: A review. *Bioengineering*, 12(6):631.

Utsav Maskey, ZHU Chencheng, and Usman Naseem. 2025. Benchmarking large language models for cryptanalysis and side-channel vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19849–19865.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Context-aware fairness evaluation and mitigation in llms. *arXiv preprint arXiv:2510.18914*.

Usman Naseem, Gautam Siddharth Kashyap, Kaixuan Ren, Yiran Zhang, Utsav Maskey, Juan Ren, and Afrozah Nadeem. 2025. Alignment of large language models with human preferences and values. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 245–245.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers)*, pages 197–207.

Juan Ren, Mark Dras, and Usman Naseem. 2025. Shield: Classifier-guided prompting for robust and safer lvlms. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 76–89.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Dandan Wang and Shiqing Zhang. 2024. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial intelligence review*, 57(11):299.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Paul Youssef, Zhixue Zhao, Christin Seifert, and Jörg Schlötterer. 2025. Has this fact been edited? detecting knowledge edits in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9768–9784.

Yiran Zhang, Jincheng Hu, Mark Dras, and Usman Naseem. 2025a. Cogmem: A cognitive memory architecture for sustained multi-turn reasoning in large language models. *arXiv preprint arXiv:2512.14118*.

Yiran Zhang, Mingyang Lin, Mark Dras, and Usman Naseem. 2025b. Beyond the black box: Demystifying multi-turn llm reasoning with vista. *arXiv preprint arXiv:2511.10182*.