The Multi-Persona Integration Pattern: Comparing Large Language Models Applied to Medical Advice

William Schreiber and Jules White
Department of Computer Science,
Vanderbilt University Nashville, TN, USA
{william.schreiber, jules.white}@vanderbilt.edu

Douglas C. Schmidt
Department of Computer Science
William & Mary, Williamsburg, VA, USA
dcschmidt@wm.edu

Abstract

This paper extends the **Multi-Persona Interaction** pattern, which is a structured method initially proposed in prior research [13] that assigns multiple expert roles to a single large language model (LLM) to generate integrated, multidisciplinary insights. Specifically, we employed a standardized medical scenario involving four distinct expert personas—a General Practitioner, Cardiologist, Endocrinologist, and Nutritionist— and presented this identical scenario to five state-of-the-art LLMs—*GPT-4, Claude, DeepSeek, Gemini*, and *Meta*. We then benchmarked their outputs against GPT-4 using the *LLM-as-a-Judge* [6] evaluation pattern.

Our assessment scores responses along the following eight criteria relevant to medical AI performance: *Medical Completeness, Role Fidelity, Structural Clarity, Patient Usability, Hallucination Risk, Prompt Compliance, Integration Quality*, and *Safety & Disclaimers*. Qualitative analysis reveals a common benefit: every LLM leveraged the *Multi-Persona Interaction* pattern to generate richer, cross-disciplinary recommendations than a monocular prompt would elicit. Yet the LLMs diverged in how well they balanced the pattern's forces: GPT-4 and Claude offered the most harmonious role coordination, Gemini maximized detail at the risk of verbosity, DeepSeek offered surprisingly consise, accurate assessments, while the Meta model blended personas so tightly that role boundaries blurred.

These observations refine our understanding of the *Multi-Persona Interaction* pattern, highlighting design consequences such as (1) the importance of explicit role reminders to curb boundary-bleed, and (2) the need for adaptive summary scaffolds to tame verbosity. We close with guidelines for selecting both prompts and LLMs when safety-critical, expert-level consultations are required. We also situate our findings within the broader study of patterns prompt engineering and LLM evaluation.

1 Introduction

Large language models (LLMs) can be guided to adopt specific personas or expert roles to generate more contextually relevant responses. This approach has been formalized in recent work as part of a persona-based prompt engineering framework [15]. [16] One pattern from this framework, the *Multi-Persona Interaction* pattern, enables an LLM to embody multiple expert personas simultaneously [13]. By integrating multiple perspectives in one interaction, the LLM can produce more comprehensive outputs than a single-perspective response. Our prior work posits that clearly defining multiple roles in the prompt and instructing the LLM to integrate their insights can yield well-rounded answers covering all relevant angles of a problem.

Despite these advances in multi-persona generation, systematically evaluating the quality of such multi-faceted responses remains challenging. Traditional evaluation methods—relying on human experts or simplistic automated metrics—struggle to capture the nuances of answers that span multiple specialties and perspectives. In this study, we address this challenge by introducing an *LLM-as-a-judge* [6] evaluation pattern as a novel evaluation technique. In essence, we leverage a strong LLM itself as an impartial evaluator to judge the responses of other LLMs. This approach allows us to assess rich, multi-dimensional outputs with greater consistency and depth, marking a significant innovation in how LLM performance can be compared in complex scenarios.

This paper applies the *Multi-Persona Interaction* pattern to a collaborative medical diagnosis scenario to examine how different LLMs perform under coordinated expert roles. As outlined in [13], a clinician might ask an LLM to act concurrently as a General Practitioner (GP), Cardiologist, Endocrinologist, and Nutritionist to obtain a holistic assessment and treatment plan for a patient. As shown in Table 1) **??**, this scenario represents a realistic use case where *Multi-Persona Interaction* can provide integrated, multidisciplinary medical advice. We use a prompt derived from this case study – instructing the LLM to adopt those four medical personas and collaborate on the patient's case – as the basis for evaluating multiple LLMs in practice.

We evaluate the performance of four contemporary LLMs against GPT-4 [11] (as a reference LLM) on the same multi-persona medical prompt. The LLMs include: Claude Sonnet 3.7 [1] (an AI assistant by Anthropic known for its lengthy,

detailed responses), DeepSeek V3 [3] [4] (a Chinese open-source LLM), Gemini 2.0 Flash [5] (Google's next-generation LLM), and a Meta AI LLM [8] (analogous to Llama-family LLMs). By analyzing each LLM's output, we identify how differences in LLM architecture and training manifest in a multi-role setting.

To ensure a rigorous comparison, we adopt GPT-4 not only as a participant in the exercise but also as an automated judge of quality. In our evaluation, GPT-4's own response is treated as an exemplary answer against which the others are compared, and we further harness GPT-4's advanced reasoning to provide consistent judgments on specific aspects of each LLM's output. In effect, we leverage one advanced LLM to critically evaluate and score the responses of other LLMs across a standardized set of criteria. This *LLM-as-a-judge* evaluation pattern enables a detailed, systematic comparison that would be difficult to achieve with manual review alone.

We evaluated the content of each LLM's final response along the following eight criteria relevant to medical AI performance [9], encompassing both the clinical content and the communication aspects of the response:

1. **Medical completeness** – The extent to which the response thoroughly covers relevant diagnostic considerations (possible conditions), explores symptom etiology, and outlines appropriate treatment strategies. A complete response would address cardiac, metabolic, and other potential causes of the patient's symptoms, list appropriate diagnostic evaluations, and suggest interventions from each specialty spanning lifestyle, medications, and further testing.
2. **Role fidelity** – How well the LLM adhered to the distinct roles assigned. We examined whether each section of the answer stayed "in character" (GP, cardiologist, etc.), providing information pertinent to that specialty without undue overlap or omission, and whether the LLM signaled role transitions clearly.
3. **Structural clarity** – The organization and coherence of the response. The clear segmentation by persona (*e.g.*, headings or labels for each specialist's input) and whether the LLM provided a synthesized final plan combining the insights, versus a disjointed or confusing structure.
4. **Patient usability** – The degree to which the advice is understandable and actionable for a layperson patient. This includes clarity of language (minimal unexplained jargon), concrete recommendations (*e.g.*, specific lifestyle changes or follow-up steps), and overall usefulness of the information provided to the patient.
5. **Hallucination risk** – Whether the LLM introduced any information that was clearly incorrect, unsupported, or medically unsound ("hallucinations") - the factual accuracy of medical statements and any speculative advice not justified by the scenario (which could pose risks if taken seriously). According to the pattern framework, a risk of multi-persona prompts is that the added complexity might increase the chance of the LLM making incorrect assumptions or fabricating details.
6. **Responsiveness to prompt** – How directly and completely the LLM followed the prompt's instructions. This involves whether the LLM indeed provided a comprehensive analysis and multidisciplinary plan as asked, and if it required additional prodding or user turns to do so.
7. **Integration quality** – How well the LLM merged the perspectives of different personas into a unified recommendation, whether the final output reconciled the advice from all four specialists, indicating collaboration, or if it left the advice fragmented.
8. **Safety and disclaimers** – Whether the LLM included appropriate cautions or disclaimers about the limitations of AI-provided medical guidance. This aspect is important for safe deployment, reminding the patient to consult professionals and not treat the responses as definitive medical advice.

Evaluating all five LLMs against these eight criteria provides a refined understanding of how well each LLM fulfills the requirements of a multi-persona consultation and where each might fall short.

This paper provides the following three contributions to research:

1. *LLM-as-a-Judge evaluation pattern* – We introduce an evaluation methodology that leverages an LLM as an automated "judge" of quality, using GPT-4 to systematically assess and compare LLM responses across the eight criteria defined above. This approach demonstrates an efficient way to evaluate complex LLM outputs and is a distinguishing feature of our study's methodology.
2. *Comparative analysis of multi-persona responses* – We present an in-depth, side-by-side comparison of five advanced LLMs (Claude, DeepSeek, Gemini, Meta, with GPT-4 as a reference) on the multi-persona medical consultation task, highlighting the unique strengths and limitations of each LLM's response. This analysis reveals how each LLM handles the demands of simultaneously balancing multiple expert roles.
3. *Theoretical framework application* – We apply a *Persona* pattern language [13] as a lens to interpret the results, examining how well the theoretical benefits of multi-persona prompting are realized in practice and where potential pitfalls emerge.

Anchoring our approach in a pattern language [2] provides a principled evaluation of persona-based interactions with LLMs. The findings inform both prompt designers (in understanding how various LLMs adhere to complex instructions) and LLM developers (in identifying areas where LLM behavior diverges in structured multi-role contexts).

The remainder of this paper is organized as follows: Section 2 explains our study methodology and describes the results of our comparative analysis of all five LLMs; Section 3 analyzes these results; Section 4 discusses related works; Section 5 presents concluding remarks and outlines future work; An **Appendix A** contains a copy of all the actual prompts and responses as recorded by the LLM interactions; It is located in a github repository:
https://github.com/ascwill/Original-Prompts-and-Responses.git

2 Experimentation Setup and Results

This section explains our study methodology and describes the results of our comparative analysis of all five LLMs.

2.1 Methodology

We employed a two-stage prompt based on the "Collaborative Medical Diagnosis and Treatment Planning" use case from the *Persona* pattern language [13]. We did not otherwise intervene or correct the LLMs. In the first stage, the system prompt instructed the LLM to adopt four specific medical professional personas and to outline a comprehensive diagnostic analysis and treatment plan from each perspective. In the second stage, the user (or scenario) provided the patient's symptoms and background ("I am experiencing symptoms such as fatigue, shortness of breath, and weight gain. . . elevated blood sugar levels. . . "), triggering the LLM to produce its full response. GPT-4's answer (treated as the baseline) and each of the four other LLMs' answers were analyzed.

Our analysis involved studying each LLM's response and benchmarking it against GPT-4. Direct quotes from the LLM transcripts show specific points of comparison [18]. GPT-4's response serves as a reference point: we note where other LLMs match, exceed, or fall short of the baseline. Below, we present the comparative results for each evaluation criterion, followed by a discussion interpreting these findings in a broader context [14].

2.2 Results from the Experiments

This section describes the results of our comparative analysis of all five LLMs in our study. Each LLM is scored on a 1–5 scale (5 = best) per criterion, based on the methodology described in Section **??**. Figure 1 shows the nuanced differences in the text. The results this figure show all LLMs are capable, but their responses have varying emphasis and

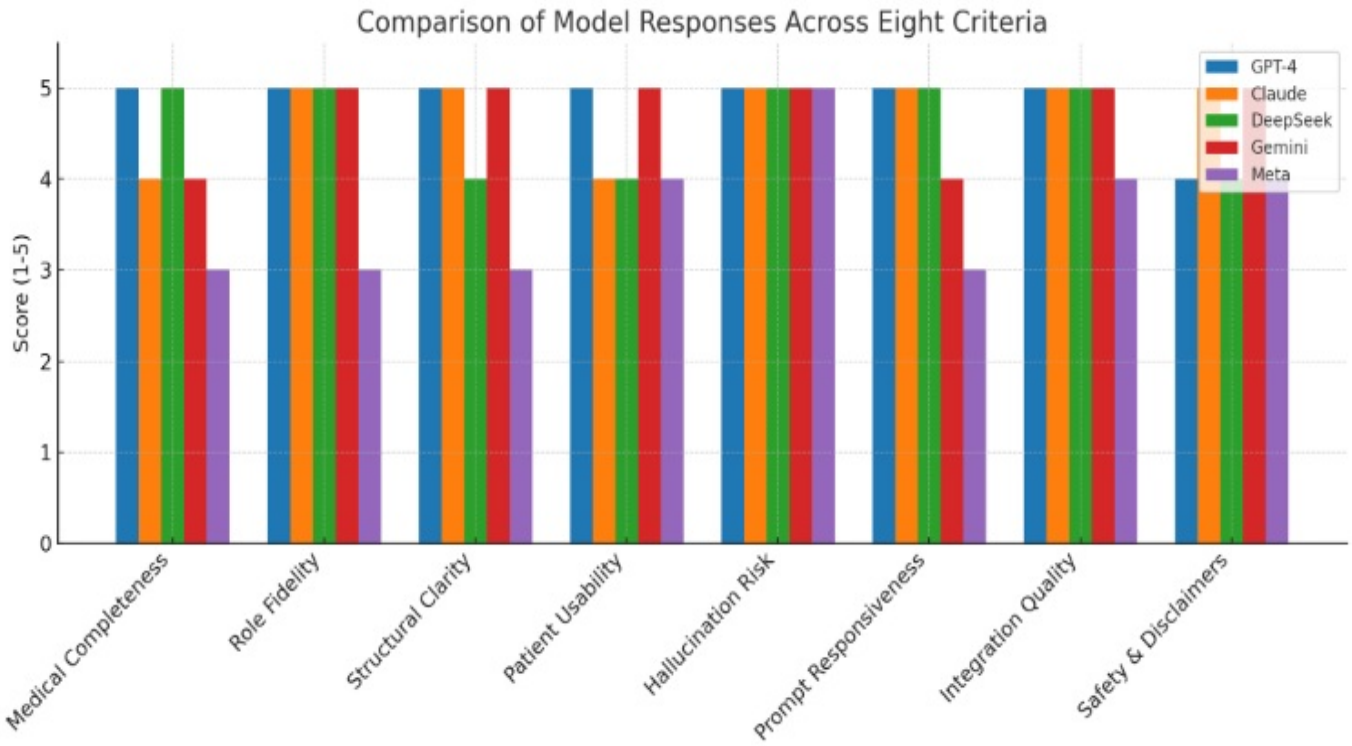

Figure 1: Comparative Summary of LLM Performance Across Evaluation Criteria.

adherence to the prompt's multifaceted demands.

In particular, Figure 1 shows that GPT-4 (blue) leads or ties for the top position in most criteria, reflecting its strong all-around performance. Claude (orange), DeepSeek (green), and Gemini (red) also perform well, with particular strengths in certain areas (Claude and Gemini excel in *Safety* due to disclaimers. DeepSeek and GPT-4 in Completeness).

Meta (purple) shows slightly lower scores in format-related criteria (*Completeness, Role Fidelity, Structural Clarity, Prompt Responsiveness*) but remains competitive in medical factuality (*Hallucination Risk*) and patient-centric measures (*Usability*).

2.2.1 Medical Completeness

We now look at how thoroughly the five LLMs we tested cover relevant diagnostic considerations, explore symptom etiology, and outline appropriate treatment strategies.

GPT-4 (Baseline) – GPT-4 delivered a thorough differential diagnosis and plan, covering a wide spectrum of potential issues and interventions. It identified multiple possible causes for the patient's symptoms, ranging from metabolic and endocrine disorders to cardiac and even respiratory factors. Notably, GPT-4's General Practitioner segment explicitly listed five key diagnostic considerations: metabolic syndrome, type 2 diabetes or pre-diabetes, elevated heart disease risk, anemia or thyroid dysfunction, and sleep apnea. This shows exceptional expanse, as it goes beyond the obvious (diabetes and heart disease) to include less apparent contributors like anemia and obstructive sleep apnea.

Correspondingly, GPT-4 recommended a comprehensive battery of tests and evaluations: *e.g.*, hemoglobin A1C for long-term glucose control, a full lipid panel, thyroid function tests, iron studies, an ECG, an echocardiogram, and even a coronary calcium scan for asymptomatic coronary artery disease screening. In terms of treatments, GPT-4 covered lifestyle modifications (diet, exercise, stress reduction) as well as possible medications (antihypertensives, statins, metformin) in its interdisciplinary plan. This depth in GPT-4's response indicates excellent medical completeness effectively acting like an entire medical board review of the case.

Claude – Claude's response was also comprehensive, though slightly more focused on the most likely systems (cardiovascular and metabolic) and somewhat less detailed on secondary considerations. Claude's General Practitioner section noted that the combination of symptoms *"suggest several possible underlying conditions"* including cardiovascular issues, metabolic disorders like pre-diabetes or diabetes, and thyroid dysfunction. This covers the core possibilities but omits an explicit mention of anemia or sleep apnea (conditions that GPT-4 did include). Claude's cardiologist perspective appropriately raised concerns about heart failure, coronary artery disease, or cardiomyopathy given the shortness of breath and family history.

Its endocrinologist perspective addressed insulin resistance/diabetes and hypothyroidism, aligning closely with GPT-4 on those points. For workups and treatments, Claude recommended many of the same investigations as GPT-4 (physical exam, HbA1c, thyroid tests, lipid profile, blood pressure checks, ECG, and even a sleep study if warranted.) The cardiology workup from Claude included an echocardiogram, stress test, and Holter monitoring, which is comparable to GPT-4's suggestions.

Overall, Claude's medical completeness was high – it covered all major organ systems pertinent to the case and suggested appropriate multidisciplinary interventions – but it was marginally less exhaustive than GPT-4. For instance, Claude did not explicitly mention less common etiologies (like Cushing's syndrome or anemia), focusing instead on the more probable diagnoses. This slightly narrower scope is evident in the integrated plan where Claude emphasizes urgent issues, lifestyle, medications, and follow-up, without listing specific differential diagnoses beyond the initial assessments.

DeepSeek – DeepSeek matched GPT-4 in many aspects of completeness and, in some respects, provided an even more itemized breakdown of the case. It systematically analyzed each symptom through the General Practitioner's lens, which led to a very granular consideration of causes: *e.g.*, DeepSeek noted fatigue could stem from hormonal imbalances (thyroid), cardiovascular issues, or metabolic disorders; shortness of breath could indicate cardiac or pulmonary issues or anemia; and weight gain might be due to hormonal factors, insulin resistance, or lifestyle causes. This symptom-by-symptom analysis demonstrates a comprehensive approach to differential diagnosis, ensuring that no major category of cause (endocrine, cardiac, pulmonary, hematologic, lifestyle) was overlooked.

DeepSeek's workup plan was extensive, mirroring GPT-4's breadth: it called for repeated blood glucose measurements, HbA1c, a full thyroid panel, lipid profile, iron studies, ECG, and even a chest X-ray to check for pulmonary contributors to shortness of breath. Like GPT-4, DeepSeek's cardiologist section recommended an echocardiogram, stress test, and Holter monitor, and its endocrinologist section focused on confirming diabetes and hypothyroidism with appropriate tests.

One area DeepSeek excelled was explicitly listing the most likely diagnoses after analyzing all perspectives. In a "Comprehensive Diagnosis" subsection, it summarized that pre-diabetes or type 2 diabetes, hypothyroidism, and cardiovascular risk (potential early heart disease) were the top conclusions given the data. This explicit diagnostic summary is very useful, as it directly answers the prompt's request for a diagnosis. In terms of treatment, DeepSeek outlined a multidisciplinary plan that included lifestyle changes (exercise, diet, weight loss, stress management), medications for diabetes (metformin), thyroid hormone if needed, and cardiovascular prevention (statins, antihypertensives). It even discussed patient education and regular follow-up.

In summary, DeepSeek's completeness was on par with GPT-4, covering all relevant ground. The only minor gap was that it did not mention sleep apnea in any section, an omission also seen in Claude and Gemini. However, DeepSeek uniquely brought up a chest X-ray for lung issues, showing initiative in exploring pulmonary causes that others (except

GPT-4's mention of sleep apnea) did not.

Gemini – Gemini's response was comprehensive in terms of standard-of-care considerations, though it stayed closer to the core cardio-metabolic issues and was less inclined to explore unlikely tangents. It immediately recognized the case as serious and requiring prompt evaluation, focusing on the interplay of heart disease risk and diabetes. Gemini's analysis from each specialist covered the essentials: the GP would do a thorough exam and baseline blood tests (CBC, metabolic panel, lipids, thyroid, HbA1c), the cardiologist would evaluate for heart failure or coronary disease with ECG, echocardiogram, stress testing, and even advanced imaging like a coronary CT angiogram or cardiac MRI if needed, and the endocrinologist would investigate diabetes and hormonal causes, including recommending an oral glucose tolerance test, insulin and C-peptide levels to gauge insulin.

These latter tests (OGTT, insulin, C-peptide) show that Gemini considered detailed endocrine evaluation, which not all other LLMs explicitly did. While Gemini did not list anemia or pulmonary issues explicitly, it did cover hypothyroidism and Cushing's syndrome as considerations in the endocrinologist's assessment for weight gain and fatigue. In treatment planning, Gemini gave a robust combined strategy: medications for blood sugar (metformin, sulfonylureas, and insulin), blood pressure (ACE inhibitors, beta-blockers, etc.), and cholesterol (statins), alongside lifestyle modifications (dietary changes, regular exercise, weight management, stress reduction, smoking cessation), and regular monitoring of blood sugar, blood pressure, and cholesterol. This plan is essentially as medically complete as GPT-4's, touching on every major intervention.

The slight shortfall in Gemini's completeness was the absence of an explicit list of differential diagnoses in its final answer. It did not directly state "the likely diagnoses are X, Y, Z," instead embedding its diagnostic reasoning within the specialist discussions (*e.g.*, noting the endocrinologist would check for diabetes or hypothyroidism, the cardiologist for heart failure, etc.). As a result, a lay user reading Gemini's response might not come away with a named diagnosis, although the information to conclude "pre-diabetes/diabetes and possible heart disease" is all there. Hence, Gemini thoroughly covered the needed evaluations and treatments, but its emphasis remained on the central problems (cardio-metabolic syndrome) without venturing into less common possibilities—hence it is comprehensive but a bit conservative in scope compared to GPT-4 or DeepSeek.

Meta LLM – Meta's answer was medically sound but comparatively narrower in scope and detail. It clearly identified the patient's issues as part of a metabolic and cardiovascular syndrome, outright diagnosing the patient with pre-diabetes, metabolic syndrome, and elevated cardiovascular risk. These diagnoses are very plausible given the scenario and demonstrate that Meta synthesized the case into the most likely big-picture answer. However, Meta's approach did not enumerate other possible diagnoses like hypothyroidism or sleep apnea – it focused only on the metabolic-cardiac axis.

For example, unlike others, the Meta response did not mention thyroid function at all in its final diagnosis or plan (even though its initial assumed profile noted mood swings and suggested checking cortisol and thyroid in a generic workup, those details did not carry through to the final answer after the specific prompt was given). In terms of recommended workup, Meta's final response actually skipped detailing many diagnostic tests, presumably because it considered the diagnoses already established by the given information.

It emphasized management: lifestyle changes and medications. Meta's multidisciplinary treatment plan stressed diet (low saturated fat, low sugar, low sodium, with whole foods) and exercise (150 minutes/week of moderate activity) as first-line measures. It also suggested starting metformin for blood sugar control and statins for cholesterol, as well as omega-3 supplements and a multivitamin. These recommendations are appropriate for metabolic syndrome and pre-diabetes, aligning with standard preventive care. Meta included a follow-up strategy (quarterly blood sugar checks, bi-annual lipid panels, annual cardiac risk assessments). What Meta did not do was itemize all the investigations a doctor might still perform to rule out other conditions—the answer assumed the problem was metabolic syndrome and treated it as such.

For example, no mention of an echocardiogram, sleep study, or expanded endocrine tests is found in Meta's final plan; it's implied that the primary concerns are already identified. So, while Meta's answer is medically correct and focused on the main issues, it is less exhaustive. It might miss some secondary diagnoses (*e.g.*, it did not discuss checking thyroid function or anemia, which several other LLMs did) and does not walk the reader through each specialist's thought process. In a sense, Meta gave the end result (the diagnosis and plan) without showing all the intermediate considerations. This makes it efficient but somewhat less comprehensive than the others.

2.2.2 Role Fidelity

We now examine whether each section of an LLM's answer stayed "in character," and whether the LLM signaled role transitions clearly.

GPT-4 – GPT-4 demonstrated strong role fidelity by clearly delineating the contributions of each persona. Its response was structured into labeled sections for each specialist: "1. General Practitioner Perspective," "2. Cardiologist's Perspective," "3. Endocrinologist's Perspective," and "4. Nutritionist's Perspective," each providing insights unique to that role. For example, the GP section dealt with general differential diagnosis and initial tests, the cardiologist focused strictly on heart-related issues (heart failure, CAD, hypertension effects) and cardiac tests, the endocrinologist addressed

blood sugar and hormonal imbalances (diabetes, thyroid, cortisol), and the nutritionist concentrated on diet plans and nutritional goals. There was minimal overlap between sections: each specialist spoke within their domain.

Where there was some cross-domain discussion, it was clearly framed by the specialist's perspective. For instance, the cardiologist mentioned lifestyle recommendations such as limiting sodium and doing exercise—advice that overlaps with nutrition and general health—but it was given in the context of improving cardiovascular health (thus still appropriate for the cardiologist to mention). Similarly, the endocrinologist suggested a low-glycemic diet and weight loss for blood sugar control, which overlaps with the nutritionist's domain, but again this was within the scope of endocrine management of diabetes. These slight overlaps are not "undue"; rather, they reflect realistic interdisciplinary thinking (specialists reinforcing each other's advice) and do not detract from role fidelity. GPT-4's clear sectioning and the hand-off to an integrated plan at the end indicate it fully honored the multi-persona format. Each role "stayed in its lane" for the analysis portion, then their recommendations were merged coherently.

Claude – Claude also adhered well to the assigned roles. Its answer was structured by specialist sections, labeled as General Practitioner Assessment, Cardiologist Perspective, Endocrinologist Assessment, and Nutritionist Recommendations, followed by an Integrated Treatment Approach. In each section, Claude maintained focus appropriate to that specialty.

For example, Claude's GP section provided an overview and general next steps (physical exam, broad initial tests, referrals) without encroaching on detailed cardiac or dietary advice. The Cardiologist section exclusively discussed cardiac concerns (heart failure, CAD risks) and recommended cardiac tests. The Endocrinologist section dealt with blood sugar and hormonal issues and suggested relevant tests like glucose tolerance and thyroid panels. The Nutritionist section dealt solely with diet: low-glycemic diet, heart-healthy foods, sodium moderation, hydration, etc., tailored to the patient's issues.

There was very little overlap: each role's advice was confined to its domain of expertise. Claude's answer reads almost as if four different specialists each wrote their portion and then a summary was provided, which is exactly the intended effect. This fidelity is further highlighted by Claude explicitly segregating recommendations (*e.g.*, it didn't have the GP giving dietary advice or the nutritionist diagnosing any conditions). The roles were well-respected, making the overall answer neatly compartmentalized by expertise.

DeepSeek – DeepSeek exhibited excellent role fidelity, with each persona's responsibilities clearly defined and followed. In its response, after a brief introduction, DeepSeek laid out separate sections for General Practitioner (GP) Overview, Cardiologist's Assessment, Endocrinologist's Evaluation, and Nutritionist's Recommendations, each numerically itemized internally.

The GP overview handled general patient evaluation steps (symptom review, medical history, physical exam, initial tests and explicitly stated the GP's role in coordinating care. The cardiologist's section was focused on heart-related symptoms and diagnostics, such as chest pain, shortness of breath, ECG, echocardiogram, stress test, and cardiac biomarkers like troponin and BNP for heart failure. The endocrinologist's section concentrated on metabolic issues, discussing blood sugar, insulin resistance, thyroid function tests, etc., and conditions like diabetes or hypothyroidism. Finally, the nutritionist's segment gave diet and lifestyle guidance (low-glycemic diet, portion control, omega-3 intake, weight management strategies). Each of these sections stayed tightly on topic.

DeepSeek's format was perhaps the most explicitly structured, reading almost like a checklist from each specialist, which prevented any blurring of roles. The GP did not suggest specific treatments better left to specialists, and the specialists largely didn't repeat each other's points. The integration of roles happened later in the "Multidisciplinary Treatment Plan" section, not within the role-specific sections, so fidelity was maintained up to the point of integration. DeepSeek's response behaved exactly as intended for the prompt: each persona contributed in its own domain, with the GP coordinating and no role overstepping its boundaries.

Gemini – Gemini's answer strongly followed the multi-persona format, with each role's perspective clearly demarcated and minimal overlap between them. In the initial part of its response, Gemini actually asked for patient details and outlined what each specialist would do, reinforcing the persona separation (this was before the final answer with the given patient info).

Upon providing the final answer, Gemini structured it by headings for each role: General Practitioner (GP) Overview, Cardiologist Assessment, Endocrinologist Assessment, Nutritionist Recommendations, followed by a Multidisciplinary Treatment Plan. Under each of these headings, the content was confined to that role's perspective. The GP section covered general history, exam, and initial tests (and then rightly referred the patient to cardiology/endocrinology). The Cardiologist's section listed specific cardiac tests and concerns (ECG, echocardiogram, stress test, coronary CT if needed), without straying into, say, dietary advic.

The Endocrinologist's section enumerated metabolic and hormonal evaluations (glucose tests, insulin and C-peptide, signs of diabetes or other endocrine disorders). The Nutritionist's section focused on diet and lifestyle (dietary assessment, personalized meal plan focusing on heart-healthy foods, portion control, cutting processed sugars, etc.). There was no

undue repetition: for example, while both the endocrinologist and nutritionist care about blood sugar, the endocrinologist's role was to diagnose and suggest medical management of diabetes, whereas the nutritionist provided the diet to manage it. Gemini kept those threads separate appropriately. The final integrated plan combined their advice, but prior to that, each persona's voice was distinct.

Gemini's fidelity to roles was essentially perfect – the text explicitly reads as if four specialists gave their independent recommendations which were then compiled. Gemini also took care not to let one persona pre-empt another: the GP in Gemini's answer explicitly says they would refer to the cardiologist and endocrinologist for the respective domains, underscoring that those matters will be handled by the other roles. This shows a clear understanding of keeping within the GP's generalist scope. Overall, Gemini adhered to the multi-persona structure rigorously.

Meta LLM – Meta's response showed the weakest role fidelity of the five, primarily because it did not explicitly maintain separate voices for each role in the final answer. In Meta's final output, the analysis was presented in an integrated narrative rather than divided by specialist. Meta's answer began with a "Patient Profile" and then directly provided a "Comprehensive Diagnosis" and "Multidisciplinary Treatment Plan". It did acknowledge multiple roles at the very end by listing a "Multidisciplinary Healthcare Team" (primary care physician, cardiologist, endocrinologist, dietitian, exercise physiologist) responsible for the ongoing care.

However, throughout the diagnostic reasoning and treatment plan, Meta did not use the distinct sections like "GP's perspective" or "Cardiologist says..." as the prompt had instructed. Essentially, Meta gave a unified response as a single voice, presumably of an authoritative medical source, rather than toggling between personas. This means that role separation was largely lost in its presentation.

There is overlap inherent in Meta's combined approach: *e.g.*, its treatment plan intermixes lifestyle (typically the nutritionist's domain), medical therapy (cardiologist/endocrinologist domain), and follow-up (GP domain) all in one list. Because it didn't label which specialist contributes what, the burden is on the reader to infer that "dietary changes" come from a nutrition standpoint, "metformin" from endocrinology, etc. While Meta did implicitly cover all these areas, the prompt's instruction to speak from four distinct perspectives was not explicitly followed in the structure of the answer.

Note that in Meta's initial response attempt, it did break out some content by role (GP, Cardiologist, etc.) in a generic way, suggesting it understood the idea. But when it came to the final answer with specifics, Meta chose a more consolidated format. In terms of role fidelity, Meta deviated from the expected pattern: it provided integrated expertise but without maintaining the persona boundaries. In a comparative sense, Meta's answer reads like a single doctor (or a fused "hive mind" of all specialists) giving recommendations, whereas all other LLMs clearly segmented the roles. This makes Meta an outlier on this criterion. It still delivered multidisciplinary content, but not in the explicitly role-partitioned manner requested.

2.2.3 Structural Clarity

We now analyze the organization and coherence of LLM responses, and whether the LLM provided a synthesized final plan combining the insights, versus a disjointed or confusing structure.

GPT-4 – GPT-4's response was highly organized, reflecting a logical flow that was easy to follow despite the depth of content. It used numbered sections and clear headings for each specialist role, which gave the answer a top-down structure. Each section had internally consistent formatting, often with bullet points and subheadings.

For example, within the GP section, GPT-4 used bullet points to enumerate possible conditions and "Next Steps (General Care Plan)" which listed recommended investigations and actions. Similarly, in the cardiologist section, after a brief narrative, GPT-4 listed "Potential Heart-Related Issues" as bullet points and then "Cardiologist's Recommended Tests & Interventions" as indented sub-points. This hierarchical structuring (major point followed by specific sub-points) made the information extremely digestible.

The endocrinologist and nutritionist sections followed suit with subheadings like "Potential Endocrine Issues" and "Endocrinologist's Recommended Tests", and for nutrition, "Nutritional Goals" and "Recommended Diet Plan". After covering all roles, GPT-4 explicitly provided a "Final Comprehensive Treatment Plan" which consolidated the advice into a single plan of action. This final section was also well-structured with sub-items: *e.g.*, lists of medical tests to do, lifestyle modifications, and follow-up steps.

Notably, GPT-4 ended with a numbered "Next Steps & Follow-Up" list for each specialist follow-up, which increases clarity by explicitly telling the user what to do first, second, and so on. In terms of coherence, the narrative flowed from explaining the problem and potential diagnoses, to investigations, to treatments, to follow-up, mirroring how a clinician would approach it. Each part referenced the previous appropriately (*e.g.*, the integrated plan clearly draws on the issues identified in each specialist section).

There were no contradictory or redundant parts in GPT-4, thanks to careful integration. GPT-4's structure can be described as textbook-like – it was segmented and labeled, which aids scanning. However, it maintained a single cohesive voice throughout the structured sections.

Claude – Claude's response was also well-structured, though slightly less granular in internal subheadings compared to

GPT-4. It separated the answer into specialist sections with clear headings (GP Assessment, Cardiologist, Endocrinologist, Nutritionist) and then provided an integrated summary. Each section in Claude's answer began with a brief explanatory paragraph followed by bullet-point recommendations, which gave it a clean layout.

For example, the GP section starts with a sentence or two about the combined significance of symptoms, then lists recommendations with bullet points (physical exam, blood tests, monitoring, etc.). The cardiologist section similarly had a short intro followed by bullet points of workup suggestions. Claude's structural strength lies in brevity and focus; it didn't enumerate as many sublevels as GPT-4, which made each role section concise.

The Integrated Treatment Approach was clearly marked and delivered as a numbered list of steps 1 through 5. These steps logically encapsulated the plan: (1) address urgent issues, (2) implement lifestyle changes, (3) consider medications, (4) schedule follow-ups, (5) track progress metrics. This final list provided an easy-to-follow roadmap and tied together the specialists' inputs cohesively.

Coherence in Claude's response was high: the flow from one specialist section to the next was smooth, and the integrated plan referred back to elements raised in the sections (*e.g.*, lifestyle modifications recommended by both GP and Nutritionist are consolidated in step 2 of the integrated approach). If there is a slight critique, it's that Claude's formatting was somewhat uniform – it didn't use nested bullet hierarchies as much, which occasionally made the text a block of bullet points after a paragraph. But the use of bold role labels and a final numbered plan mitigated any potential confusion. Claude's answer was structurally clear and methodical, presenting information in the expected order and format with obvious section breaks and a coherent conclusion.

DeepSeek – DeepSeek's response had a very explicit structure, albeit one that read a bit like an outline or checklist. It was organized by roles, each introduced by a heading line, and within each role section it frequently used numbered lists and sub-bullets to enumerate points. For instance, the GP Overview section listed items 1 through 4 (Symptom Analysis, Medical History, Physical Examination, Initial Tests) and within those used sub-points (*e.g.*, under Symptom Analysis, bulleting specific symptoms and considerations).

This pattern continued: the Cardiologist's Assessment had "1. Symptoms" and "2. Diagnostic Tests" and "3. Conditions to Consider" with sub-bullets under each, and similarly for Endocrinologist's Evaluation. The Nutritionist's section used a numbered list for different focuses (1. For Elevated Blood Sugar, 2. For Heart Health, 3. For Weight Management, 4. General Wellness) and bullet points under each with dietary recommendations. This highly segmented approach ensured no detail was lost, but it risked feeling a bit dense or formal. However, DeepSeek mitigated this with a final "Multidisciplinary Treatment Plan" that was more narrative, presented as another numbered list (1. Lifestyle Modifications, 2. Medications, 3. Monitoring, 4. Patient Education) each with sub-bullets explaining what to do.

After that, DeepSeek provided a brief "Comprehensive Diagnosis" section (numbering the likely diagnoses) and a concluding paragraph summarizing the approach. In terms of clarity, the outline form made it very easy to parse specific information (one could, for example, quickly find what the endocrinologist recommends by looking at that section and its numbered points). The coherence was also maintained: despite the segmented format, the pieces fit together logically.

One could argue that the sheer number of list levels made the answer look more like a report or protocol than a flowing explanation, which might be slightly harder for a layperson to read continuously (this touches on usability as well). But structurally, it was unambiguous and thorough. The transitions were essentially the headings – each persona heading signaled a shift in viewpoint, which worked given the prompt. DeepSeek's structure can be described as explicit and exhaustive, with a clear outline that covers everything methodically. The slight drawback is that it may read less fluidly than GPT-4 or Claude's more narrative style, but it compensates with absolute clarity on what each point is.

Gemini – Gemini's response was clearly structured and well-formatted for readability. It presented the answer in distinct sections for each role, each introduced by a bold heading, followed by an integrated plan and important notes. The layout included a mix of narrative sentences and bullet points that made the content approachable. For example, the GP Overview began with a bullet list of what information the GP would gather and do, and the Cardiologist Assessment similarly used bullets for each test or consideration.

Gemini frequently annotated the bullets with brief explanations in plain language (*e.g.*, describing what an ECG or echocardiogram is for), which improves clarity. After going through GP, Cardiology, Endocrine, and Nutrition sections in order, Gemini provided a "Multidisciplinary Treatment Plan" that aggregated recommendations in categories like Medications, Lifestyle Modifications, and Regular Monitoring. This section was organized with bullet points and sub-bullets, ensuring the different aspects of the plan (for diabetes, blood pressure, cholesterol control, etc.) were separated. Gemini concluded with an explicit disclaimer note.

The coherence of Gemini's response was excellent; it flowed logically from identifying the problem to analyzing it from each perspective to then solving it collectively. Structurally, one benefit of Gemini's output was the liberal use of new lines and whitespace around list items, making it visually easy to scan. Important points (like the initial disclaimer *"This is for informational purposes only..."* and the concluding *"Important Note"*) were separated out, so they stood out as separate blocks. This use of whitespace and bullet points prevented the answer from becoming a wall of text.

Gemini's strategy of prefacing the role analysis with a quick recap of the situation (*"This is a serious combination of symptoms... here's a breakdown of each specialist's approach"*) provided a nice structural intro to set the stage. Gemini's answer had a well-organized structure with clear sections and lists, maintaining coherence through consistent formatting and a final integration. It reads somewhat like an educational article, which is appropriate for the scenario.

Meta LLM – Meta's structural clarity was moderate; it was organized, but not in the multi-section way others were, and a bit less transparent in flow. The final response from Meta was divided into a few key parts: a brief patient profile assumption, a "Comprehensive Diagnosis" section, a "Multidisciplinary Treatment Plan," and a list of the multidisciplinary team roles. These were clearly labeled, which helps structure. Within the diagnosis section, Meta simply listed the diagnoses 1, 2, 3 (as bullet points), which is straightforward. The treatment plan was written in a narrative list form with numbered sub-sections (Lifestyle Modifications, Medications and Supplements, Regular Monitoring and Follow-up) and bullet points under each detailing recommendations.

The plan thus had internal organization similar to others, but with fewer categories. After the plan, Meta enumerated the healthcare team members 1–5 in a list. This is a logical way to conclude, as it reinforces the roles involved, but interestingly this list of team members appears somewhat abruptly, since Meta did not separate its initial analysis by these roles. In terms of coherence, Meta's answer flows as a single narrative: it identifies the diagnoses, then directly gives the plan for those diagnoses, and mentions the care team. This is coherent in a medical sense (problem followed by solution).

However, because Meta didn't step through each role's reasoning, the transition from "Comprehensive Diagnosis" to "Treatment Plan" might feel like a jump—essentially, Meta collapsed the analysis phase. Compared to others, the structure is less layered: others had an analysis per role and then a plan, whereas Meta had a short integrated analysis and an integrated plan. Despite that, the answer is not confusing; it's actually shorter and perhaps easier to follow for a reader who just wants the conclusion. The clarity might suffer only if the reader expected the role-by-role breakdown (which the prompt might have primed them to expect).

From a document design perspective, Meta's use of headings and lists was clear but minimalistic. It did not have as many subheadings or bullet groups as GPT-4 or DeepSeek. This simplicity can be double-edged: it's succinct, but some nuance is not immediately visible. For instance, a user scanning Meta's answer would see the diagnoses list and might assume those are the only considerations. Meta's structure was concise and direct, with clear labeling of diagnosis and plan, but it lacked the multi-perspective breakdown that was a structural hallmark of the prompt. It was coherent internally, but in a different format than the others, trading granular clarity for brevity.

2.2.4 Patient Usability

We now look at the degree to which the advice from LLMs is understandable and actionable for a layperson patient, as well as the overall usefulness of the information provided to the patient.

GPT-4 – GPT-4's answer, while comprehensive, was presented in a way that a diligent non-expert reader could follow and act upon. It provided a lot of information, which can be overwhelming, but it also gave concrete actionable steps and explanations that improve usability. For instance, GPT-4 concluded with a very clear "Next Steps & Follow-Up" section enumerating what the patient should do in sequence: "1 Schedule a GP appointment... 2 Consult a Cardiologist... 3 See an Endocrinologist... 4 Work with a Nutritionist...". This effectively translates the complex analysis into a simple checklist for the user, making the advice immediately actionable.

GPT-4 also interspersed its technical content with short explanations or parenthetical notes that aid understanding. It defined terms or reasons briefly, *e.g.*, noting that A1C is to check long-term blood sugar levels, or that fatigue could be linked to thyroid or anemia. The nutrition section gave specific diet examples (a sample daily meal plan with breakfast, lunch, snack, dinner), which is highly usable as it paints a practical picture for the patient. GPT-4's language was generally professional but accessible; it did use some medical terminology (like "metabolic syndrome" or "HOMA-IR score" for insulin resistance) that might be challenging for a layperson.

However, whenever a term was potentially unfamiliar, GPT-4 either explained it or the context made the meaning clear. For example, it introduced "metabolic syndrome" and immediately described it as "a combination of high blood sugar, excess weight, and cardiovascular risk factors", which demystifies the term. Similarly, when listing medications or interventions, GPT-4 did so with context ("medications such as ACE inhibitors, beta-blockers, or statins may be considered") without diving into details that a patient wouldn't need (it didn't, for instance, list dosages or mechanisms, just the category of drug and why it might be used).

Another aspect of GPT-4's usability is its reassuring and collaborative tone. It doesn't sound alarmist but instead systematically goes through issues and ends by offering further guidance. This invitational tone can make a patient feel supported rather than overwhelmed.

The main caution is that GPT-4's answer is quite lengthy. A non-expert might not digest it all in one reading. But thanks to its structured format, the user can easily break it down into sections. The explicit listing of what to do next and the inclusion of lifestyle suggestions (diet, exercise, stress management) give the patient plenty of tangible advice

to implement. GPT-4's response is highly usable for a patient: it's packed with information but clearly organized, with definitions, examples, and a step-by-step plan that a patient could follow up on with their doctors.

Claude – Claude's response was relatively concise and user-friendly. It provided needed information without extraneous detail, and it maintained a compassionate, cautionary tone that is appropriate for patient communication. Early in the interaction, Claude explicitly reminded the user that while it can provide information, it's not a substitute for professional medical advice, setting proper expectations. In the final answer, Claude's language was straightforward. It addressed the user in second person (*"Your symptoms... suggest..."*), which engages the reader directly without being overly technical.

Claude gave recommendations in plain terms: for example, *"I recommend: A complete physical examination; Comprehensive blood panel including HbA1c, thyroid function tests, and lipid profile; ... Possibly a sleep study..."*. Each bullet is a single, digestible action item or test, largely free of jargon (aside from necessary terms like HbA1c or ECG, which most patients with those concerns might recognize). If a term was technical, it was usually accompanied by a simple explanation or context (*e.g.*, noting the sleep study is to rule out sleep apnea, which is clearly linked to the symptoms).

The structural clarity we discussed also aids usability: Claude's response is segmented into moderate-length paragraphs and lists, which avoid overwhelming the reader. Importantly, Claude's integrated plan includes tracking metrics and follow-up scheduling, which is useful advice for a patient as it tells them how to monitor progress. Claude's tone is reassuring yet realistic. It neither downplays the issues nor overdramatizes them.

For instance, it notes shortness of breath and family history are *"concerning from a cardiac standpoint"*, which is honest but then it immediately talks about what can be done (tests and interventions) rather than leaving the user in fear. Claude included a polite disclaimer about not replacing consultation, which is actually helpful to the patient as it encourages them to seek real care. In terms of readability, Claude's sentences are relatively short and clear since doesn't dive into long-winded explanations. The use of bullet points makes it easier to skim and pick out key recommendations.

All these factors make Claude's response highly usable for a patient: the patient can easily extract a list of tests to ask their doctor about, lifestyle changes to start, and understand why those steps are suggested. The only small limitation is that Claude's answer is slightly less detailed than GPT-4's; on the flip side, that brevity might be an advantage for a patient who could be overwhelmed by too much information. Claude strikes a good balance between completeness and simplicity from a patient's perspective.

DeepSeek – DeepSeek's output was extremely detailed, which is a double-edged sword for patient usability. On one hand, it gave the patient a trove of information and a clear rationale for each recommendation, which can empower an engaged patient. On the other hand, the level of detail (like multiple nested lists and technical terms) might be challenging for some non-experts to digest fully.

DeepSeek did aim to make the content actionable: it provided lists of what to do (*e.g.*, a structured "Lifestyle Modifications" list including specific exercise targets and stress management techniques, a "Medications" list naming classes of meds for each condition). It also gave a concrete follow-up plan (regular follow-ups with GP and specialists, periodic blood tests, annual cardiac evaluations) which informs the patient how ongoing care should proceed. These are very useful because they essentially tell the patient what to expect or request in future medical visits.

DeepSeek's explanation style is thorough but sometimes reads like a medical document. For example, it mentioned *"assess for arrhythmias", "evaluate for exercise-induced cardiac issues", "HOMA-IR Score"* in the endocrinologist section, and listed specific tests like *"24-hour Holter monitoring", "OGTT", "antibodies"* for thyroid, etc. These terms might not be immediately clear to a layperson. Unlike Gemini, DeepSeek didn't always provide a lay explanation for each test (Gemini, for example, explained an OGTT as checking how the body processes sugar over time, which DeepSeek didn't explicitly do).

However, DeepSeek's concluding section helps usability: its Conclusion paragraph stepped back and summarized in plain language that the multidisciplinary approach addresses the symptoms and risk factors, and encourages working with the healthcare team. That statement reinforces the message in simpler terms and likely leaves the patient feeling that this plan is comprehensive and positive. The tone of DeepSeek is professional and encouraging: *"By working with your GP, Cardiologist, Endocrinologist, and Nutritionist, you can achieve better health outcomes..."*. DeepSeek's collaborative, optimistic tone is good for a patient reading it.

So, is DeepSeek's response usable for patients? Yes, particularly for patients who appreciate detail and rationale. It effectively arms them with knowledge (perhaps even to bring to their doctor). But it might be a bit overwhelming for those not inclined to parse a structured outline. The readability is a bit dense due to heavy use of bullet points and technical listing.

DeepSeek could be seen as providing a "manual" for the patient's condition. If the patient takes the time to go through it, they would find every recommendation clearly spelled out (*e.g.*, exactly what diet changes to make, how much exercise to aim for, which tests to get). In fact, some patients might prefer this level of detail. Others could be intimidated by it. Comparatively, GPT-4 and Claude packaged information in more narrative form at times, which can be easier to read.

DeepSeek is slightly more formal.

In general, DeepSeek is highly actionable but somewhat less immediately readable for the average patient due to its exhaustive, outline-style presentation. Patients willing to engage deeply with their care plan would find it very useful, while those looking for a quick summary might find it heavy.

Gemini – Gemini's response appears to have been crafted with patient readability and safety in mind, arguably more so than any other LLM. It begins with a very clear disclaimer that frames the information as for informational purposes and urges consulting a professional, which not only protects from misuse but also signals to the patient that they should treat the information as guidance to discuss with their doctor. The content itself is presented in friendly, non-alarming language.

For example, rather than simply listing tests, Gemini often adds why a test is done in everyday terms: *"Complete Blood Count (CBC): To assess for anemia or infection.", "Comprehensive Metabolic Panel: To evaluate kidney and liver function, electrolyte balance, and blood sugar levels."*. This explanatory style demystifies medical jargon on the fly. A patient reading Gemini's response would understand not just what the doctors will do, but why each step is important, which is empowering and educational.

Gemini also emphasizes important lifestyle advice in a motivating way: *e.g.*, *"Losing even a small amount of weight can significantly improve heart health and blood sugar control."* – a statement like this is encouraging and grounded in something a patient can do. The integrated treatment plan bullet points in Gemini's answer read almost like a checklist for the patient: take medications if prescribed (with examples given), exercise regularly (with a specific goal of 30 minutes most days), manage stress (with methods like yoga or meditation), stop smoking, etc. Each is an actionable item, and they're phrased as positive recommendations (*"Aim for at least 30 minutes of exercise..."*, *"Techniques such as yoga... can help reduce stress"*) rather than imperatives, which is a gentle approach likely to be well-received by patients.

Gemini's patient usability is also enhanced by it explicitly drawing boundaries: reminding the user this is general and that they must follow up with healthcare professionals for personalized care. This manages patient expectations and safety. The tone throughout is supportive and cautionary at the same time. It warns the patient appropriately (for example, *"this warrants prompt medical evaluation"*, which rightly instills a sense of urgency to see a doctor), yet also provides a full plan so the patient feels informed about what that evaluation and treatment might entail.

In terms of structure for usability, Gemini's use of bullet points with short explanatory phrases is ideal for someone who doesn't have a medical background. They break down complex procedures into understandable pieces. Gemini's response could almost serve as a patient education handout. Gemini scored excellently on patient usability. A patient would likely come away with both a better understanding of their health situation and a clear idea of next steps.

Meta LLM – Meta's response was relatively succinct and focused, which generally aids patient usability, but it omitted some explanatory detail that might leave less-informed patients with questions. On the positive side, Meta clearly identified the issues (pre-diabetes, metabolic syndrome, cardiovascular risk) and then gave a numbered plan of lifestyle changes and medications. The advice itself was practical and actionable: *e.g.*, the lifestyle modification section says *"Focus on a balanced diet that is low in saturated fats, added sugars, and sodium. Emphasize whole, unprocessed foods... Aim for at least 150 minutes of moderate-intensity aerobic exercise per week... Engage in stress-reducing activities such as meditation, yoga, or deep breathing."*.

Meta generally provided clear guidance a patient can attempt to follow immediately. It's specific enough (the exercise guideline is quantified, the diet advice gives categories of foods to eat or avoid, and even stress management techniques are suggested) without being overly prescriptive or technical. The medication recommendations are also straightforward: *"Metformin: to improve insulin sensitivity... Statins: to reduce cholesterol and cardiovascular risk... Omega-3 Fatty Acids: to reduce inflammation..."*.

Meta even included *"Multivitamin: to ensure adequate nutrient intake."*, which might be a bit generic but is something a layperson can easily do and might appreciate. This shows Meta was thinking in terms of tangible interventions. Furthermore, Meta laid out follow-up needs (quarterly blood sugar monitoring, bi-annual lipid checks, annual cardiac risk assessments), which informs the patient how their progress will be tracked. This responses are quite useful for planning and emphasizes that these conditions require regular check-ins.

The structure, with numbered lists, made it easy for a patient to see the main categories of action (Lifestyle, Medications, Monitoring, Team). However, compared to some others, Meta's explanation of terms was minimal. It used the term "metabolic syndrome" but did provide a brief definition (*"a cluster of conditions that increase the risk of heart disease, stroke, and type 2 diabetes"*). This is good. It assumed the user could understand "pre-diabetes" and did not mention hypothyroidism at all, so no need to explain it. One potential gap is that Meta didn't explicitly encourage the user to see a doctor in its text (no direct disclaimer or statement like "consult your physician").

It listed the multidisciplinary team members that should be involved in care, which implies that the patient should be seeing these professionals, but it didn't phrase it as advice to the patient. Had Meta said, "You should work with your primary care physician, cardiologist, etc." it would be clearer. Nevertheless, by including that team list, it gave the

patient a clear idea of who should be on their healthcare team, which is itself actionable information (the patient might realize they need to get referrals to those specialists).

The tone of Meta's answer is factual and neutral, perhaps less comforting or conversational than some others, but not harsh. It reads like a doctor's summary: clear and matter-of-fact. Many patients do appreciate that direct style, especially when it's not too long.

One additional safety element: at the end of Meta's output, there was a system-added line: *"Messages are generated by AI and may be inaccurate or inappropriate."*. If the patient sees that, it would hopefully prompt them to verify information with a professional. But this was not a deliberate part of the LLM's response; it's likely a platform warning.

Excluding that, Meta itself didn't provide a cautionary disclaimer in its text, which is a slight negative for patient safety (it treated its information as a definitive plan). Meta's answer is usable in that it is concise and full of clear recommendations, but it might assume a bit of knowledge (it doesn't explain each medical term or explicitly guide the patient to next steps beyond lifestyle and monitoring). It's a solid outline for care, and a proactive patient could follow it (diet, exercise, possibly ask their doctor about metformin and statins, etc.). It's just a bit less personalized or hand-holding than Gemini or GPT-4. So, in terms of patient-friendliness: Meta is straightforward and actionable, but slightly less explanatory and cautious compared to the best examples.

2.2.5 Hallucination Risk

We now see if the LLMs introduced any information that was clearly incorrect, unsupported, or medically unsound.

GPT-4 – GPT-4's response did not exhibit any hallucinations or factual inaccuracies; on the contrary, it was medically well-founded and all claims were substantiated by standard medical knowledge. Every condition and test GPT-4 mentioned is relevant to the patient's presenting symptoms.

For example, suggesting that fatigue and weight gain could implicate hypothyroidism or that shortness of breath and weight gain could be signs of congestive heart failure are textbook interpretations. The diagnostic tests recommended (A1C, lipid panel, TSH, ECG, etc.) are all standard and appropriate. GPT-4 did not introduce any irrelevant conditions or outlandish therapies. All recommended treatments (like potentially using metformin for diabetes or statins for cholesterol) align with established guidelines.

Even its mention of a coronary calcium scan, while somewhat advanced, is a real test that might be considered in someone with risk factors to evaluate silent coronary artery disease. There was no point in GPT-4's answer where it invented a fact or gave an incorrect explanation. Importantly, GPT-4 stuck to evidence-based reasoning: it identified metabolic syndrome as a unifying diagnosis, which is well-supported by the combination of obesity, high blood sugar, and hypertension risk. It listed plausible differentials like sleep apnea for fatigue and weight gain, which is reasonable in an overweight patient and commonly co-occurs with metabolic syndrome.

In contrast, GPT-4 did not mention anything that would be out of left field (*e.g.*, it did not say "maybe you have a rare disease" without cause). The level of detail (like mentioning HOMA-IR for insulin resistance or the dexamethasone suppression test for Cushing's syndrome) shows GPT-4 had extensive medical knowledge, but those are real concepts/tests, not hallucinated ones. They might be beyond what a typical GP would immediately consider, but they are not wrong or made-up; they indicate thoroughness. GPT-4's response had zero factual hallucinations. Everything can be cross-verified in medical literature. If anything, GPT-4's largest "risk" might be overloading the user with information, not giving misinformation.

Claude – Claude's answer was factually accurate and free of hallucinations as well. It provided a conservative, sensible analysis that aligns with standard medical practice. Claude did not assert anything beyond established medical knowledge. For instance, it correctly noted that the symptom combination could indicate early heart failure or coronary artery disease, which is true. It also pointed out the possibility of sleep apnea contributing to these symptoms, which is a valid consideration.

The tests and interventions Claude recommended (ECG, Holter monitor, echocardiogram, HbA1c, etc.) are all real and appropriate. Claude did not invent any non-existent tests or treatments. It also avoided giving any specific numerical results or false data. Essentially, Claude played it straight with the information given. Each role's perspective was grounded in how a real clinician from that specialty would think.

For example, Claude's endocrinologist section said the symptoms and lab results suggest possible insulin resistance/-diabetes and hypothyroidism – both are common in that scenario and in no way far-fetched or incorrect. One hallmark of hallucination is if an AI provides a very specific piece of information that wasn't given, like a lab value or a medication name that wasn't mentioned. Claude did not do this; it worked only with the provided symptom set and general medical knowledge. It also inserted a proper disclaimer, which shows it was being cautious.

There were no unsubstantiated claims: every claim (like *"shortness of breath...could indicate early heart failure"*) is substantiated by the clinical context. Claude did not, for example, diagnose the patient definitively with a condition without evidence; it merely raised possibilities to be investigated, which is exactly what a doctor would do. Therefore, we

can say that Claude's response has an extremely low hallucination risk. It was factual and careful, reflecting mainstream medical understanding.

DeepSeek – DeepSeek likewise maintained a very factual and evidence-based approach in its answer. It did not hallucinate any information or stray into unsupported territory. Everything it stated about symptoms, tests, and treatments is grounded in standard medical practice or plausible clinical reasoning. For example, DeepSeek mentions checking for cardiac biomarkers (troponin, BNP) if cardiac issues are suspected – these are indeed the markers used for heart attacks and heart failure, respectively, and are appropriate in context.

DeepSeek recommended an OGTT for further evaluation of elevated blood sugar, which is a real test to confirm diabetes. The breadth of DeepSeek's discussion remained within plausible limits; it didn't jump to any rare diagnoses or off-label treatments. One notable aspect: DeepSeek's "Comprehensive Diagnosis" listed Hypothyroidism as one likely diagnosis.

While the patient's data did not explicitly mention thyroid levels, hypothyroid is a legitimate possible cause of fatigue and weight gain, so including it is not a hallucination but part of differential diagnosis. It is consistent with others (GPT-4 and Claude also brought up thyroid dysfunction). DeepSeek's mention of "metabolic syndrome" and distinguishing pre-diabetes vs. type 2 diabetes, etc., are all factual interpretations of the scenario. No factual errors were observed.

DeepSeek effectively served as a compilation of correct medical advice. The only potential issue could be if the user misinterpreted something due to the heavy detail, but that's not a hallucination issue—that's a communication issue. DeepSeek did well not to overstep what was known: it used the family history of heart disease appropriately (to heighten concern for heart issues, but it did not, for example, claim the patient has a heart disease without further tests). It consistently phrased things as possibilities or needs for further evaluation, which is medically correct. DeepSeek's output had no hallucinated content; it was thorough and accurate.

Gemini – Gemini was very grounded in its response, with no inaccuracies or invented information. All medical information given by Gemini aligns with standard practice and the facts presented by the user. For instance, Gemini correctly associated the symptoms with the possibility of heart failure or coronary disease and did not introduce any irrelevant conditions.

Gemini also accurately described the purpose of various diagnostic tests (*e.g.*, explaining an OGTT or insulin test), and these explanations were accurate. The treatments it suggested (metformin, insulin for diabetes; ACE inhibitors, beta blockers for blood pressure; statins for cholesterol) are precisely the medications that would be used in such a scenario; no unusual or incorrect therapies were given.

Gemini even included a mention of *"in some cases, medications may be needed to improve heart function and reduce symptoms of heart failure"*, which is a cautious and correct note given the shortness of breath symptom – if the patient did have early heart failure, that would indeed be treated with specific medications. Gemini did not present any made-up statistics or unverifiable claims; rather, it tended to stick with general principles (which is appropriate given the context).

Gemini also provided realistic advice such as smoking cessation and stress management, which are generic but evidence-based recommendations for improving health outcomes. Because Gemini was somewhat verbose in giving justifications, it actually lowered any risk of misunderstanding or hallucinatory statements; it showed its reasoning which was logical. It's also worth noting that Gemini repeatedly clarified that its information is general and not personalized medical advice. This meta-commentary is truthful about the limitations of the information, further reducing the risk that any particular statement would mislead – the user is reminded to cross-check with a professional. Gemini had no hallucinatory content; its response was factual, standard, and prudent.

Meta LLM – Meta's response did not contain hallucinations; it was concise and focused on likely diagnoses and appropriate recommendations. The diagnoses Meta gave (pre-diabetes, metabolic syndrome, cardiovascular risk factors) are all directly supported by the user's input (elevated blood sugar leads to pre-diabetes, combination of issues leads to metabolic syndrome, family history and symptoms leads to cardiovascular risk). It did not assert something like "you have X disease" without basis; metabolic syndrome is a conceptual grouping of risk factors which the patient clearly has, so that is a fair "diagnosis" to state.

The treatment plan items Meta listed are all real and commonly advised: dietary changes (low salt, low sugar, etc.), regular exercise with specific guidelines (150 minutes/week, which is the standard recommendation by health organizations), medications like metformin and statins (first-line pharmacotherapy for pre-diabetes and hyperlipidemia), and follow-up intervals (quarterly A1c checks, etc. – perhaps more frequent than some might do, but not unrealistic). Meta did not fabricate any new tests or cures.

One thing to examine: Meta assumed the patient to be "45-year-old" in the Patient Profile. This detail was not provided by the user. While making up an age could be considered a form of guess, Meta explicitly said it will "assume" this to build the profile, which indicates it was creating a scenario in absence of provided detail.

This assumption is a minor point: it's not a hallucination in the sense of a false fact presented as truth, but rather a reasonable hypothetical fill-in. It might or might not match the real patient, but Meta needed an age to assess risk, and

choosing middle-age is logical given the scenario. This assumption does not lead to any incorrect advice – it likely guided Meta to emphasize metabolic syndrome (common in middle age).

If the patient were much younger, some recommendations might differ, but since age wasn't specified, this assumption is not egregious. Other than that, Meta's entire answer was consistent with medical standards. It didn't mention thyroid issues, but omission is not a hallucination; it simply focused on the key issues it chose.

Meta's advice about using omega-3 and a multivitamin is mild and common—omega-3 supplements do have evidence for cardiovascular benefits and a multivitamin is a generic precaution. Crucially, Meta did not claim these supplements would do something unsupported; it just listed them as recommendations, which many doctors do as well. There were no invented drug names or anything strange. Meta's response had negligible hallucination risk. It stuck to relevant facts and mainstream medical advice. The small liberty it took in assuming an age was transparently communicated and does not constitute a dangerous or misleading fabrication.

2.2.6 Responsiveness To Prompt

We now observe how directly and completely the LLMs followed the prompt's instructions, and if they require additional prodding or user turns to do so.

GPT-4 – GPT-4 was highly responsive to the prompt, arguably exceeding expectations in following both the letter and spirit of the instructions. The user's request had two main parts: (1) adopt four roles (GP, Cardiologist, Endocrinologist, Nutritionist) and (2) provide a comprehensive analysis (diagnosis and multidisciplinary treatment plan) for the symptoms provided. GPT-4 did exactly this. It clearly acted as each of the four specialists, dedicating a substantial section to each perspective and labeling them accordingly. GPT-4's output shows it followed the multi-persona format precisely. Each role's analysis was thorough, indicating GPT-4 embraced the instruction to provide a "comprehensive analysis" from each angle.

In addition, GPT-4 delivered a "Final Comprehensive Treatment Plan", which fulfills the instruction to suggest a multidisciplinary treatment plan. In doing so, it integrated the roles' insights rather than keeping them disjointed, which is exactly what "multidisciplinary" implies. Importantly, GPT-4's answer was complete: it addressed diagnosis (by discussing likely conditions and next diagnostic steps) and treatment (covering lifestyle, medications, follow-ups).

At no point did GPT-4 ignore or omit part of the user's query. The user specifically asked for both a diagnosis and a plan, and GPT-4's very first lines of the answer set the stage for delivering that (*"Below is a multidisciplinary assessment..."*). By the end of the answer, GPT-4 had indeed given a comprehensive diagnosis (in the form of identifying metabolic syndrome, possible diabetes, etc., in the GP section and summarizing all likely issues) and a comprehensive plan (the final section with tests, lifestyle changes, and follow-ups).

GPT-4 even offered to continue helping (asking if the user wants more guidance on tests or meal prep). While this help was not requested, it shows GPT-4's responsiveness to any implicit need for clarification or further assistance, which is a bonus in terms of completeness and user-friendliness. GPT-4 exactly followed the prompt's structured requirement and delivered a thorough solution. It used the multi-persona format effectively and made sure every element was covered. GPT-4's performance on prompt responsiveness is excellent – it can be considered a gold standard in this comparison.

Claude – Claude was quite faithful to the prompt after obtaining the needed patient information. In the initial exchange, Claude needed the user to provide specific symptoms before it could give a comprehensive answer, which is a fair request for clarification. Once the user did, Claude fully complied with the instructions. It assumed the four roles distinctly and provided insights accordingly, and it compiled those insights into a treatment plan.

Claude's final answer explicitly thanked the user for the information and said it will provide insights from multiple perspectives, directly echoing the multi-role instruction. It then gave each specialist's findings (GP, cardiology, endocrine, nutrition) and an integrated approach. The answer was complete: it addressed possible diagnoses and the multidisciplinary plan. One could argue Claude was slightly less explicit about stating "the comprehensive diagnosis is X" than, say, DeepSeek or Meta. It didn't itemize diagnoses in a list; instead, it wove them into the narrative (*e.g.*, mentioning possible conditions under each specialist and in the GP's overview).

However, the diagnoses were certainly covered (metabolic disorder, heart failure risk, etc.), just not in a single list. The prompt didn't demand a bullet list of diagnoses, only a "comprehensive diagnosis," which Claude provided in prose form through its analysis. Importantly, Claude's answer was indeed multidisciplinary: it never skipped one of the roles or merged them. Each role was given attention, and each did their part as instructed. The treatment plan was multidisciplinary by including lifestyle, specialist follow-ups, and medication considerations. There was also faithfulness to the instruction about roles not overlapping – although this is covered under Role Fidelity, it's relevant that Claude respected the format.

In terms of completeness relative to the user's needs, Claude delivered a fully fleshed-out answer that covered everything from tests to lifestyle to monitoring. There were no prompt instructions ignored. Claude added appropriate notes (like disclaimers) even though not explicitly asked; this doesn't detract from responsiveness, it actually enhances the quality of response to a sensitive prompt like medical advice.

Claude was therefore very responsive. If we are nitpicking, it might not have spelled out some less obvious diagnoses

(*e.g.*, anemia) explicitly by name as GPT-4 did, but it did hint at "other conditions" and did include needed tests (CBC, etc.). Overall, Claude gave the user exactly what they asked for: a thorough diagnosis consideration and plan from four specialists. Thus, it meets the prompt requirements solidly.

DeepSeek – DeepSeek was highly responsive to the prompt, fulfilling every aspect of the request. It clearly followed the multi-persona directive, thoroughly analyzing the case from each of the four specified roles. DeepSeek even took initiative to ask for patient details or context in its initial response (it provided a "general framework" pending specifics), showing it was keen to tailor the answer properly – a good practice in responsiveness.

Upon receiving the details, DeepSeek dived deeply into each role's perspective and did not stop at just listing recommendations; it also provided the "Comprehensive Diagnosis" section explicitly, labeling the diagnoses it concluded, and a "Conclusion" that sums up the approach. This explicit listing of diagnoses ("pre-diabetes or Type 2 Diabetes... Hypothyroidism... Cardiovascular Risk") directly addresses the "provide a comprehensive diagnosis" part of the user's query. Some other LLMs, like Claude and Gemini, did this implicitly; DeepSeek did it explicitly, which leaves no ambiguity about the diagnostic answer. As for the treatment plan, DeepSeek's "Multidisciplinary Treatment Plan" was one of the most detailed, covering lifestyle, medications, monitoring, education.

DeepSeek's behavior shows it thoroughly addressed the "multidisciplinary treatment plan" request, leaving virtually no stone unturned. If anything, DeepSeek might have given more than strictly asked by adding a distinct conclusion section, but that only enhances completeness. There's clear evidence that DeepSeek was mindful of every element of the prompt. It respected the structure exactly. The user's prompt asked for the GP, cardiologist, etc., to each do their part and indeed DeepSeek's answer reads like a collaboration among them culminating in a plan.

In general, there was no part of the user's instructions that DeepSeek missed or contradicted. In terms of completeness, it might have even gone slightly beyond by providing some educational content like why certain tests are needed, but that's within the scope of "comprehensive analysis". Overall, DeepSeek's response was fully aligned with the query – it provided a comprehensive diagnosis and a robust plan that is multidisciplinary, all structured through the lens of the four roles. Thus, it was completely responsive and comprehensive in its answer.

Gemini – Gemini followed the prompt closely, but with a slightly different approach in framing the final answer. It certainly embraced the multi-persona format, giving each specialist's viewpoint and then a combined plan, so it was responsive in structure and content. The answer is thorough about analysis and plan, but one could note a subtle point: Gemini's final response did not explicitly state the final diagnoses in a single summary statement. For example, it did not say "The comprehensive diagnosis is metabolic syndrome manifesting as pre-diabetes and risk of heart failure" or similar. Instead, it embedded the diagnostic reasoning within each section: the GP and endocrinologist sections discuss the possibility of diabetes and hypothyroidism, the cardiologist section discusses possible heart disease.

However, Gemini stopped short of synthesizing those into a named diagnosis section in its final answer. Thus, did Gemini provide a comprehensive diagnosis as requested? Indirectly yes, but explicitly maybe not. A patient reading it might understand the likely problems (the content strongly points to *"you likely have early diabetes and need to check your heart"*), but the answer itself, perhaps out of caution, doesn't label the patient with a specific diagnosis outright. It instead focuses on what each specialist would do to reach a diagnosis or to treat. This may be due to a design choice to avoid making definitive diagnostic statements as an AI. However, the prompt wording asks for a diagnosis.

In terms of fulfilling the user's request to the letter, Gemini could be seen as slightly hedging on the "diagnosis" part by not stating it in one place. That being said, Gemini certainly did not ignore any part of the question: it clearly acted as GP, cardiologist, etc., and it did give a multidisciplinary plan. The plan was very detailed and actionable, satisfying that half of the query completely. And the analysis by each specialist covers the diagnostic possibilities in detail, which arguably satisfies the diagnostic requirement in a distributed way.

Overall, Gemini is responsive to the prompt. It gave the user exactly what was asked – just in a cautious tone. It also adhered to additional prompt nuances like offering role-specific info and then an integrated approach, demonstrating a strong alignment with instructions. The presence of disclaimers and careful language might have tempered direct diagnostic statements, but quality-wise that's a prudent approach. From a prompt compliance standpoint, all four roles were addressed, and a combined plan was delivered, which means Gemini did as instructed. So, aside from the minor note on how it phrased the diagnosis, Gemini's answer was comprehensive and on-point with respect to the user's request.

Meta LLM – Meta's answer was somewhat mixed in prompt responsiveness. It definitely addressed the user's query by providing a diagnosis and a treatment plan, but it did not strictly follow the exact format requested (*i.e.*, the *Multi Persona Interaction* pattern) in its final output. The prompt specifically said to act as GP, Cardiologist, etc., implying that the response should be structured through those personas.

Meta's final answer did not present information separated by those roles; it instead provided an integrated diagnosis and plan from a single narrator perspective. In doing so, it still gave the content the user needed (the likely conditions identified and how to treat them multidisciplinary), but it didn't explicitly show the GP's overview, the cardiologist's findings, etc., as the user might have expected from the prompt wording. This is a deviation from the prompt format.

However, Meta did acknowledge multiple disciplines in the solution by listing the multidisciplinary team members who should be involved. So, it implicitly covered the "multidisciplinary" aspect, but not in the role-playing style. As far as content completeness, Meta's answer was quite focused. It identified three main diagnoses (which is the "comprehensive diagnosis" it chose to present) and provided a corresponding plan. One could argue it left out some possible diagnoses like hypothyroid, so in that sense the diagnostic exploration was not as exhaustive. But it did hit the main ones given the data (pre-diabetes/metabolic syndrome).

The user's prompt likely expected identification of any relevant conditions, and Meta did include the key relevant ones but possibly missed a minor one or two that the prompt didn't explicitly mention either. The treatment plan Meta gave covered lifestyle, medications, and monitoring – that's certainly multidisciplinary and comprehensive. Meta's conciseness means it didn't talk about, say, how each specialist would contribute in detail. If the user expected a speaking-from-each-role style answer, Meta's approach might not fully meet that expectation.

That said, from a pure requirement standpoint, the user wanted a diagnosis and plan from a multi-expert perspective; Meta delivered that, albeit in a synthesized format. It might have partially "ignored" the instruction to explicitly have the GP, cardiologist, etc., speak separately, but it did not ignore the need for multiple domains of expertise – those were reflected in the content (diet, exercise from nutritionist, metformin from endocrinologist, statin from cardiologist's concerns, etc., all present in the plan). So the information content matches what a multidisciplinary team would produce; it's just not formatted as a dialogue between them.

In terms of completeness, Meta's answer is a bit shorter than others, but it still covers the crucial points. Therefore, one might rate Meta as mostly responsive – it answered the question with a comprehensive diagnosis and plan (so the user's problem was addressed), but it did not strictly adhere to the requested interaction pattern format, which was a key aspect of the prompt. This makes its responsiveness slightly lower than the others in a comparative sense. It's a good answer, but not delivered in the exact style the user asked for.

2.2.7 Integration Quality

We now review how well the LLMs merged the perspectives of different personas into unified recommendations.

GPT-4 – GPT-4 showed excellent integration quality. After providing detailed analysis from each specialist, GPT-4 didn't leave the advice fragmented; it brought everything together in a way that made sense as a unified care approach. The "Final Comprehensive Treatment Plan" in GPT-4's answer synthesized recommendations across disciplines. For example, it combined the need for further tests in one list, and it combined lifestyle and dietary changes (GP and Nutritionist inputs) into a holistic set of modifications.

GPT-4 also aligned the specialists by ordering the follow-ups: first see GP to coordinate, then cardiologist, then endocrinologist, then nutritionist. This ordering reflects a thoughtful plan – typically one would start with the GP and then go to specialists, which GPT-4 explicitly laid out. Such sequencing indicates it integrated the roles practically. Moreover, GPT-4's integrated plan avoided contradictions and reinforced synergies.

For instance, the GP and Cardiologist both stress exercise; GPT-4's integration just lists exercise once as a key lifestyle intervention for both cardiovascular and metabolic health. The Nutritionist and Endocrinologist stress diet changes; GPT-4's plan merges them into a single dietary recommendation focusing on low-glycemic, heart-healthy foods (which covers both blood sugar and heart needs). This cohesive blending is precisely what one would hope a multidisciplinary team's final recommendations look like.

GPT-4 also addressed any overlapping issues by combining perspectives: *e.g.*, stress management was recommended by possibly GP/endocrinologist (for cortisol) and that made it into the integrated plan once. The continuity of the plan (from immediate tests to long-term lifestyle and follow-up) was smooth, showing each specialist's role at the right stage.

In essence, GPT-4's output reads as if a team of doctors actually convened and agreed on a single plan. The presence of a unified voice in the final plan, as opposed to four separate voices, indicates strong integration. GPT-4's integration quality is outstanding – it leveraged all the insights from separate domains and weaved them into a comprehensive, coordinated management strategy.

Claude – Claude effectively integrated the specialists' perspectives into a single, coherent plan. After outlining each discipline's input, it presented an "Integrated Treatment Approach" as a numbered list. The plan is concise yet comprehensive: it begins with urgent issues—any acute findings flagged by a specialist (*e.g.*, a critical cardiology concern) must be addressed first.

After acute findings come lifestyle modifications, incorporating dietary and exercise guidance from the nutritionist and GP. Medication adjustments follow, reflecting pharmacologic recommendations from cardiology and endocrinology. Finally, the plan schedules follow-up visits with the relevant specialists and specifies metrics—such as weight, blood pressure, and blood glucose—to monitor progress and keep both general and specialty care aligned.

This list shows that Claude combined the information effectively: each step is broad enough to encompass multiple specialists' contributions. For instance, "lifestyle modifications" would include the Nutritionist's diet plan and the GP's general lifestyle advice in one step.

"Consider medication if necessary for blood sugar, cardiovascular health, or thyroid function" explicitly references integrating the Endocrinologist's (blood sugar, thyroid) and Cardiologist's (cardiovascular) domains under one umbrella of pharmacotherapy. That single bullet implies that once test results are in, the relevant medications from different specialties should be started together in a cohesive manner. Another integration point is the follow-up schedule: Claude says *"regular follow-up with appropriate specialists"* rather than leaving it implicit. The GP should therefore ensure the patient sees cardiology, endo, etc., which is an integration of care continuity – the GP coordinates, as expected.

The mention of tracking metrics like weight, BP, blood sugar is also integrative because it addresses outcomes relevant to all specialties (nutrition/endo care about weight and blood sugar, cardio cares about BP and weight, etc.). Claude's integration was slightly less granular than GPT-4's in terms of combining specifics, but it definitely provided a unified roadmap. There were no contradictory pieces: nothing the cardiologist said was left hanging or conflicting with what the endocrinologist said – the plan harmonized them.

The brevity of Claude's integrated plan actually suggests it synthesized overlapping advice well, because it didn't need to be long to cover everything. Claude's ability to unify the insights was strong; the resulting plan is balanced and ensures that recommendations from each domain support each other in a timed sequence (urgent issues first, lifestyle universally, meds as needed, continuous monitoring). It truly sounds like one plan rather than four separate ones pasted together.

DeepSeek – DeepSeek's integration of the specialist insights was comprehensive and explicit. After detailing each specialist's recommendations thoroughly, it introduced a "Multidisciplinary Treatment Plan" section which systematically merged those recommendations. DeepSeek's integrated plan was structured by category (Lifestyle, Medications, Monitoring, Education), and each category clearly draws from multiple specialties:

- The *Lifestyle Modifications* section combined general advice and nutrition advice (exercise, weight loss, stress management) that involve GP, Cardiology (exercise benefits heart), Endocrinology (weight loss aids diabetes), and Nutrition (diet) all together as foundational changes. It even quantifies exercise goals and weight loss targets, reflecting an integrated priority for all conditions.
- The *Medications* section listed medications for Diabetes (metformin), Thyroid (levothyroxine if needed), Cardiovascular (statins, antihypertensives). This is a direct integration of the Endocrinologist's domain (diabetes, thyroid) and Cardiologist's domain (heart meds) into one treatment regimen, ensuring the patient's various issues are pharmacologically addressed in parallel.
- The *Monitoring* section proposed follow-ups that span all areas: regular follow-ups with GP and specialists, periodic blood tests (HbA1c for diabetes, lipid profile for heart, thyroid tests for endocrine), and even annual cardiac evaluations if risk is high. This shows integrated long-term care, scheduling all necessary check-ups for different conditions in a cohesive surveillance plan.
- The *Patient Education* section emphasizes the patient's understanding and adherence, which is something all specialists would agree on – integrating the human aspect of care, not just medical orders.

DeepSeek didn't have a separate bullet for dietary advice in the integrated plan, presumably because it considered diet under lifestyle modifications (the nutritionist's recommendations were already quite detailed above, and "Lifestyle Modifications" implicitly includes diet).

In the integrated plan's text, diet isn't re-listed, but weight loss and exercise are, which indirectly covers diet since weight loss typically involves dietary changes. Perhaps it could have mentioned diet explicitly again for clarity, but given the nutritionist section above was exhaustive, it might have felt redundant. The integration is still evident because the nutritionist's goal (weight management) is reflected. The integrated plan aligns with the diagnoses DeepSeek identified, as it addresses each one (diabetes/pre-diabetes, hypothyroid, heart risk) with appropriate measures.

Moreover, DeepSeek provided a short conclusion after the plan, reinforcing that all these measures together will help the patient. That conclusion itself is a sign of integration – it doesn't speak from any one specialist's view, but as a collective statement of improved outcomes by multidisciplinary approach. This final touch emphasizes that the plan was considered as a whole.

DeepSeek essentially produced a coordinated care plan that a primary doctor might give the patient after consulting with all specialists. It's integrated in content and in presentation. DeepSeek's integration quality is excellent; it successfully unified the multiple expert contributions into a single, organized plan covering immediate, intermediate, and long-term management across conditions.

Gemini – Gemini demonstrated high integration quality by taking the varied recommendations from each persona and merging them into a cohesive management strategy in the "Multidisciplinary Treatment Plan" section. In that section, Gemini organized the plan by types of intervention (Medications, Lifestyle Modifications, Regular Monitoring), which inherently mixes concerns of different specialists:

- Under Medications, Gemini didn't segregate by specialty, but rather by problem: it listed medications for blood sugar control (endocrinologist domain), blood pressure control (cardiology domain), cholesterol control (cardiology),

and even heart failure medications if needed (cardiology). By listing them together, it indicates these could all be part of the patient's regimen, addressing each major issue. This integrated list helps ensure that, for instance, while treating blood sugar, one doesn't forget to also treat blood pressure and cholesterol – a common multidisciplinary approach to metabolic syndrome.

- Under Lifestyle Modifications, Gemini combined general advice (exercise, weight management, stress management, smoking cessation) that apply across endocrine, cardiac, and general health. The recommendations here are not labeled by which specialist said them originally; they come as a unified encouragement for the patient to exercise, lose weight, reduce stress, and stop smoking – all actions that simultaneously benefit the heart, metabolism, and overall health. By listing them once, Gemini integrated overlapping advice (for instance, both the cardiologist and endocrinologist would want exercise; it's listed once).
- Under Regular Monitoring, it listed all parameters to keep an eye on: blood sugar, blood pressure, cholesterol, plus scheduling follow-up appointments with GP and each specialist. This demonstrates integration by creating a single follow-up plan covering different facets of the patient's condition. Instead of, say, telling the patient separately to monitor blood sugar and to monitor BP, it's all in one monitoring plan.

These integrated plan bullets strongly reflect that Gemini synthesized the team's input. Earlier in the answer, moreover, the roles themselves sometimes foreshadow integration, *e.g.*, the GP section ends by saying they would refer to cardiologist and endocrinologist, showing coordination. The final plan fulfills that by indeed involving those specialists in follow-ups.

Gemini provided no conflicting advice, *i.e.*, its integration resolved any differences by emphasizing consistent messages (like all specialists in effect want diet and exercise improvements, which the integrated plan pushes heavily). The presence of disclaimers doesn't interfere with integration; in fact, Gemini's warning that it's not a substitute for professional help implicitly encourages integrated real-life care (the user should engage their whole healthcare team).

Gemini's integrated outcome was very cohesive. If one reads just the "Multidisciplinary Treatment Plan" of Gemini's answer, it stands alone as a comprehensive plan touching on all needed areas (medical therapy, lifestyle, monitoring), which is a sign of effective integration. Gemini merged the specialized recommendations into a well-rounded, unified plan for the patient.

Meta LLM – Meta's entire answer was effectively presented as an integrated perspective from the start, which has pros and cons for integration quality. On the positive side, since Meta's final response wasn't split by roles, the plan it gave was inherently integrated – it came out as one holistic plan rather than needing an explicit merging step. The multidisciplinary treatment plan section lists interventions that cover the domains of different specialists in one sequence: lifestyle changes (diet/exercise – GP/Nutritionist), medications (Endocrinologist/Cardiologist), and then monitoring and follow-up with each relevant specialist.

For example, Meta's plan recommended diet and exercise to tackle metabolic syndrome (which aligns the advice of GP, Endo, and Nutritionist), then suggested metformin (Endo) and statins (Cardio) concurrently to address the identified issues, and set up a follow-up routine that involves checking cardiovascular risk and blood sugar regularly (Cardio and Endo). Finally, Meta explicitly listed the healthcare team members who will implement this plan: PCP, Cardiologist, Endocrinologist, Dietitian, Exercise Physiologist. Listing those roles at the end is a way of saying "all these experts will work together on this plan," which underscores integration.

The plan itself did not break out who does what step (unlike GPT-4 which said first GP then Cardio, etc.), but listing the team implies the tasks are divided among them. The cohesive nature of Meta's plan is evident because it clearly targets the combined diagnosis it made (metabolic syndrome and preventing heart disease). There were no contradictory elements: *e.g.*, it didn't advise anything that would negatively affect another condition. All pieces (diet, exercise, metformin, statin) complement each other in reducing cardiovascular and diabetes risk.

One minor integration issue in Meta's response might be that, because it assumed the patient is 45 and framed everything around metabolic syndrome, it might have under-integrated any consideration outside that frame (like it didn't integrate a thyroid check into the plan, because it chose to ignore that possibility entirely). Since Meta skipped an explicit separate analysis per role, we didn't see an intermediate step of integration (like an "integrated plan" heading separate from analysis).

Meta therefore essentially gave an answer that was already integrated. The risk here is if integration wasn't done well, it would show as something missing in the plan. Arguably, Meta's plan missed addressing thyroid or sleep apnea, which others might have integrated. So its integration was complete for the issues it acknowledged. It integrated cardiology and endocrinology nicely (diet, exercise, metformin, statin covers both), and also considered dietary support and exercise training (hence including dietitian and physiologist in team).

Including an exercise physiologist is a unique integrative suggestion – none others explicitly mentioned that, but it makes sense for executing the exercise plan. That indicates Meta thought of involving another professional role for better integration of lifestyle changes. Meta's integration quality is good: the plan is unified and covers multi-domain

interventions without conflict. It might be slightly less comprehensive in integration simply because it narrowed the scope to fewer issues, but within that scope, all relevant elements across disciplines were present.

2.2.8 Safety and Disclaimers

We now study whether the LLMs included appropriate cautions or disclaimers about the limitations of AI-provided medical guidance.

GPT-4 – GPT-4's final answer did not explicitly include a standard medical disclaimer (*e.g.*, "I am not a doctor" or "this does not replace seeing a physician"). It dived into the answer assuming the role of medical professionals and gave the advice as such. However, GPT-4 did incorporate numerous safety-conscious elements within the content.

For example, its entire approach was to steer the user towards proper medical evaluation: it recommended a battery of diagnostic tests and specialist consultations, effectively instructing the user to engage thoroughly with healthcare providers (GP, cardiologist, endocrinologist, etc.). By saying, for example, "Schedule a GP appointment... Consult a Cardiologist... See an Endocrinologist... Work with a Nutritionist," GPT-4 strongly encourages professional involvement.

GPT-4's answers serve a similar purpose to a disclaimer by ensuring the user will be under medical supervision rather than self-managing solely based on AI advice. Moreover, GPT-4's tone was careful not to give absolute guarantees or commands that could be dangerous. It used language like "suggest" and "consider," and provided ranges and options rather than one-size-fits-all answers. This approach reduces risk because it avoids definitively diagnosing or prescribing without context; instead it provides information to be used in a clinical setting.

GPT-4 asked for more information initially before giving the comprehensive answer, indicating a cautious approach (needing data to avoid giving misguided info). While this isn't a disclaimer, it shows a safety-oriented process. What GPT-4 did not do is explicitly remind the user to treat its advice as informational only. So if we strictly consider the presence of disclaimers, GPT-4's response lacks an overt statement of limitation. Nonetheless, GPT-4 did in practice direct the user to get care in-person at each step, which is arguably more effective than a generic disclaimer.

If a user followed GPT-4's plan, they would definitely be in contact with real doctors for tests and treatments, thus mitigating risks. So, in summary: GPT-4's answer was safe in content and guidance (steering to real medical follow-up), but it didn't provide a formal disclaimer or an explicit caution that "this is not a substitute for professional advice." In a legalistic sense, that's a minor omission. But given the context, GPT-4's entire answer essentially functions as extended advice to see professionals. Thus, safety-wise, GPT-4 was very responsible, even if implicit. One could suggest GPT-4 could have added a line about confirming everything with a doctor, but its heavy emphasis on involving doctors basically covers that.

Claude – Claude took clear steps to ensure user safety and include appropriate caveats. Right from the initial interaction, Claude inserted a statement: *"Please note that while I can offer general medical information, this shouldn't replace consultation with licensed healthcare professionals who can perform proper examinations and tests."* This is an explicit disclaimer delivered before even giving the final answer.

At the start of the answer after receiving the patient details, Claude again reminded: *"please remember this isn't a substitute for an in-person medical evaluation."* These disclaimers directly address the safety criterion – they set expectations that the advice is not definitive and that professional care is essential. This language is very clear and user-facing, likely reducing the chance someone would take the AI's answer as a final diagnosis without seeing a doctor.

Beyond disclaimers, Claude's answer encourages safe practices: it suggests lots of tests and specifically a follow-up with professionals. The integrated plan includes "regular follow-up schedule with appropriate specialists", which is guiding the user to continue under medical supervision.

Claude also avoided any risky or fringe advice: all recommendations were standard-of-care, which in itself is a safety feature (no encouragement of unproven supplements or delay of needed treatment). The tone was measured – *e.g.*, acknowledging concern but not causing panic, which helps the user respond appropriately. Moreover, in asking for more information at first, Claude showed caution against giving incomplete or possibly wrong advice without details. That demonstrates a safety mindset (though it was also just logically necessary to know symptoms).

Overall, Claude excelled at providing disclaimers and cautionary notes. It explicitly told the user not to treat the AI as a doctor replacement and to see professionals. It did this more than once. Therefore, any user reading Claude's response would be aware that they need to consult a real doctor and that this advice is general. This makes Claude's answer very safe from a usage standpoint, likely preventing misuse of the information or over-reliance on it.

DeepSeek – DeepSeek's final answer did not include a strong, explicit disclaimer of the sort Claude and Gemini provided. It didn't outright say "I am not a doctor" or "consult a professional for personalized advice." That said, DeepSeek did incorporate the concept of working with healthcare professionals throughout the answer, which indirectly addresses safety.

For example, DeepSeek's conclusion states: *"By working with your GP, Cardiologist, Endocrinologist, and Nutritionist, you can achieve better health outcomes and reduce the risk of complications."*. This is a clear prompt for the user to involve all those professionals in their care, implying that the user should indeed be engaging with actual doctors (GP, etc.) to

carry out the plan. It doesn't position the AI as the final authority; instead it pushes the user to collaborate with healthcare providers. Moreover, the integrated plan said *"Regular follow-ups with your GP and specialists"*, making it evident that the advice must be implemented in tandem with professional oversight.

These elements encourage safe use. However, DeepSeek did not explicitly warn the user that the information might be incomplete or that it's not a diagnosis. It spoke confidently as a care plan. While everything it said was correct, from a safety perspective, usually an AI should include a note about "please consult a healthcare provider" if giving medical advice.

DeepSeek's perspective was basically role-playing the providers rather than an assistant outside the scenario, so it might have foregone a direct disclaimer to maintain immersion. It's approach is a slight safety drawback because a naive user might not be reminded of the limitations of AI advice. That being said, the content of DeepSeek's advice itself is safe: it doesn't suggest anything harmful or that the user manage on their own. If anything, it's likely to result in the user seeking more medical care (given all the tests and referrals recommended). So the safety risk is low in terms of content.

One risk with DeepSeek's approach is if the user somehow thought this plan was certain and didn't need actual doctor's input – but since the plan requires doctors to do tests and prescribe meds, the user inevitably has to see them. DeepSeek could have added a line like "This plan is general and you should see your doctor to tailor it," but implicitly it designed the answer assuming that would happen. So, safety-wise: it's quite responsible in content, moderate in explicit disclaimer.

One area to note is that at the very start of DeepSeek's generic response, it could say it provides a general framework pending more info. This response would shows DeepSeek wouldn't give a final plan without specifics, which is a safety measure to avoid wrong advice. Once specifics were given, it did not add a disclaimer. So overall, DeepSeek's answer prioritizes actual medical follow-up, but lacks an overt disclaimer telling the user to treat the info with caution as AI output.

Gemini – Gemini put a very strong emphasis on safety and disclaimers. It began its final response with a bold disclaimer: *"This is for informational purposes only. For medical advice or diagnosis, consult a professional."*. This one sentence upfront addresses exactly the concern of misuse, making it clear that the user should not treat the answer as personal medical advice and needs to see a doctor.

After delivering the content, Gemini ended with another explicit caution: *"Important Note: This information is for general knowledge and educational purposes only, and does not constitute medical advice. It is essential to consult with a qualified healthcare professional for any health concerns or before making any decisions related to your health or treatment."*. This closing disclaimer is comprehensive and leaves no ambiguity. It uses emphatic wording to drive the point home. Between these two statements, the user is repeatedly reminded about the limitations and the need for real medical consultation.

Gemini's behavior is exemplary in terms of adhering to ethical AI practices for medical queries. In addition to formal disclaimers, Gemini's content itself was safety-oriented. In particular, it urged "prompt medical evaluation" given the serious combination of symptoms, which is a way of saying "don't just rely on this text, go get checked soon."

Gemini also gently warned what signs to be mindful of (like signs of diabetes it listed) which helps the user gauge urgency. And its integrated plan always included follow-up appointments with the user's healthcare team, ensuring that none of this plan is to be executed in isolation from doctors. Gemini essentially fenced the advice with disclaimers, from start to finish. Thus, Gemini clearly prioritized user safety by doing as much as possible to not let the user misinterpret the AI as a doctor.

Gemini is thus the best of the five LLMs in terms of explicit safety messaging. The presence of both an opening and closing disclaimer is almost redundant but in a good way – it errs on the side of caution. Gemini's answer is very safe for the user: it educates, but strongly encourages professional guidance and does not present itself as a final authority.

Meta LLM – Meta's answer did not explicitly include a self-referential disclaimer in the text of its advice. Unlike Claude or Gemini, Meta did not say "this is not medical advice" or "consult a doctor" in its narrative. The absence of any first-person or assistant perspective commentary (since it wrote as if it were the medical team) meant it didn't break character to warn the user. At the very end of the provided PDF content, however, Gemini generated the following text: *"Messages are generated by AI and may be inaccurate or inappropriate."*. This looks like a platform-generated disclaimer rather than something Meta wrote, but if the user saw it, they'd be alerted.

Putting that aside, within Meta's answer, the encouragement for professional consultation is implicit through its plan: it lists a multidisciplinary team including a primary care physician, cardiologist, etc., that should be involved, and it calls for "Regular Follow-up Appointments... with the healthcare team". So, like the others, it's pushing the user to be under the care of real healthcare providers. But Meta didn't explicitly address the user with a cautionary note about AI limitations or the need to confirm things with a doctor.

It assumed the role of the doctor giving a plan, which might make a lay user think "this is my treatment plan." That is somewhat mitigated by how the plan inherently requires doctors to prescribe metformin or statins and perform tests; the user cannot do those alone. So practically, the user will have to consult a doctor to carry out the plan. In that sense,

the plan enforces seeking medical assistance. Nonetheless, from a safety communication perspective, Meta's answer is less explicitly cautious than Claude or Gemini. It's more similar to GPT-4 and DeepSeek in that it doesn't say "I'm not a doctor" in so many words, but it heavily implies doctor involvement by listing the medical team and follow-ups.

Meta's straightforward authoritative tone might be slightly risky if a user misunderstood it as conclusive without needing verification. But the content is accurate and standard, so risk of direct harm is low. It also didn't provide any advice that would be dangerous if taken in isolation (like telling the user to take a medication without seeing a doctor – it just names medications that a doctor would likely prescribe). So safety is preserved through the nature of the advice.

Meta's answer ensures professional oversight through its recommendations but lacks an actual disclaimer or explicit safety note from the AI. It falls in the middle: not as safe-guarded in wording as Claude/Gemini, but its plan inherently requires safe behaviors (doctor visits, etc.). Given best practices, Meta could have been more direct in cautioning the user about the AI aspect, but at least it didn't encourage any unsafe self-management.

3 Analysis of the Results

Our experiments in Section 2.2 examined how five advanced LLMs tackled a complex medical query using a multi-persona approach. This section analyzes these results in terms of the eight criteria relevant for medical AI performance: *Medical Completeness, Role Fidelity, Structural Clarity, Patient Usability, Hallucination Risk, Prompt Compliance, Integration Quality*, and *Safety & Disclaimers*. GPT-4's response served as a baseline, exhibiting exemplary performance with extensive medical detail, clear role segmentation, and a well-synthesized care plan. The other LLMs—Claude, DeepSeek, Gemini, and Meta—each brought unique strengths and minor weaknesses to the task, as summarized in Figure 1.

Our comparative analysis showed that all LLMs were medically competent, correctly identifying the central health issues and proposing appropriate interventions. Notably, none hallucinated irrelevant or false medical information, which speaks to the maturity of these LLMs in the medical domain. However, differences emerged in how they structured and delivered the information, as well as the precautions taken to ensure user safety.

All five LLMs were able to interpret the multi-persona prompt and produce clinically relevant answers, but with differing emphases:

- GPT-4 delivered a gold-standard comprehensive consultation, excelling in detail and integration, with only minor critique in not explicitly self-disclaiming.
- Claude provided a well-balanced, safe, and organized response that a patient or doctor could readily use as a guideline, closely adhering to instructions.
- DeepSeek offered maximal detail and thoroughness, essentially a full blueprint of care, which is excellent for completeness though potentially requiring interpretation for lay use.
- Gemini prioritized clarity, usability, and safety, making it arguably the best choice for direct patient-facing communications among the LLMs, while still covering the medical bases.
- Meta produced a focused and correct plan with high integration, though it somewhat abstracted away the persona format and cautionary tone, which in certain use cases could be a limitation.

Each LLM acted as a General Practitioner (GP), Cardiologist, Endocrinologist, and Nutritionist to assess a patient's symptoms (fatigue, shortness of breath, weight gain, high blood sugar) and propose a multidisciplinary treatment plan. We used GPT-4's response as the baseline, assessing the LLMs across eight criteria. The results show that all LLMs benefited from the multi-persona approach, but they differed in the depth of their analysis, clarity, and adherence to assigned roles.

Medical Completeness – GPT-4 exhibited exceptional thoroughness, addressing a broad range of potential conditions and recommending a wide array of diagnostic tests. For example, GPT-4's GP section included five possible diagnostic considerations, from metabolic syndrome to thyroid dysfunction. This breadth was particularly valuable, covering both common and less apparent conditions, such as anemia and sleep apnea.

By comparison, Claude was slightly less exhaustive, focusing more on the most likely diagnoses but omitting less common possibilities. DeepSeek and Gemini provided similarly comprehensive analyses, but DeepSeek stood out for its detailed breakdown of symptoms and its explicit summary of likely diagnoses. Meta, while concise, provided fewer diagnostic suggestions, missing some secondary considerations like hypothyroidism.

Role Fidelity – GPT-4 demonstrated clear role fidelity, ensuring that each specialist's input stayed within their domain without unnecessary overlap. The response was structured into distinct sections—each reflecting the perspective of the GP, Cardiologist, Endocrinologist, and Nutritionist—before integrating the advice into a unified treatment plan. Claude also adhered well to the roles, presenting them in a structured format with clear labels and minimal role overlap.

DeepSeek provided a well-defined structure, though it used a more formal outline style, which could be less engaging for some users. Gemini followed the persona structure effectively, with minimal overlap between specialists. However,

Meta diverged from the format, providing an integrated, singular response rather than maintaining separate sections for each role.

Structural Clarity – GPT-4's response was the most organized, with each section clearly labeled and presented in a logical, easy-to-follow format. The use of bullet points, subheadings, and a final integrated treatment plan ensured that the information was digestible and accessible to both medical professionals and patients. Claude's structure was similarly effective but slightly less granular in terms of subheadings, making it less detailed than GPT-4.

DeepSeek's use of nested lists was comprehensive but could be seen as dense and formal. Gemini's approach was user-friendly, with a mix of narrative and bullet points, making it easy to scan while maintaining clarity. Meta's structural clarity was moderate, but its integrated approach made it less explicit in its adherence to the persona structure.

Patient Usability – GPT-4 excelled in patient usability, presenting its advice in a clear, actionable format. It offered specific lifestyle recommendations (such as sample meal plans) and clinical steps, ensuring that a layperson could easily follow the guidance. Claude also performed well, providing concise and understandable recommendations with clear instructions.

While DeepSeek was thorough, patients could be overwhelmed with its level of detail. Gemini offered a user-friendly approach, using plain language to explain medical terminology and giving actionable steps. Meta's response, though clear and actionable, was less detailed and lacked some explanatory context that could have improved patient understanding.

Hallucination Risk – All LLMs, including GPT-4, displayed a low risk of hallucinations. GPT-4's response was grounded in evidence-based medical knowledge, with no unsupported claims or fabricated details. Claude, DeepSeek, and Gemini also provided medically sound advice without introducing any incorrect or speculative information. Meta's response, while concise, occasionally assumed additional details (such as the patient's age) that were not provided in the prompt, but these assumptions did not lead to false medical advice.

Responsiveness to Prompt – GPT-4's treatment plan was the most realistic and comprehensive, incorporating diagnostic tests, medications, lifestyle changes, and follow-up steps in a well-rounded manner. It also included a sequence of actions, which mirrored how a healthcare team would typically approach the patient's care. Claude and DeepSeek both provided realistic and medically sound treatment plans,

However, DeepSeek's format was more formal, potentially making it harder for non-experts to digest. Gemini's treatment plan, while thorough, was more focused on core issues like diabetes and heart disease, missing some secondary considerations. Meta's plan, though efficient, lacked the depth and comprehensive nature of the other LLMs.

Integration Quality – GPT-4 demonstrated superior integration, blending the recommendations from each specialist into a coherent, multidisciplinary treatment plan. For example, lifestyle changes suggested by the GP and Nutritionist were unified into a single recommendation, and medications for diabetes and cardiovascular health were presented together, reflecting the combined expertise of all roles. Claude's integration was also strong but less detailed than GPT-4's, focusing more on general categories of recommendations.

DeepSeek excelled in its detailed integration but risked overwhelming the patient with its exhaustive approach. Gemini's integration was clear but conservative, focusing primarily on cardiovascular and metabolic concerns. Meta's integration was the weakest, lacking the explicit role boundaries required by the prompt.

Safety and Disclaimers – All LLMs, including GPT-4, displayed many safety references, GPT-4 provided a comprehensive, actionable treatment plan, but it did not explicitly include a standard medical disclaimer, while Geminiexcelled in providing safety-oriented messaging, starting with a strong disclaimer. Both Claude and DeepSeek took proactive steps to ensure user safety by including clear and explicit disclaimers both before and after the treatment plan. Meta provided sound medical advice but did not include explicit disclaimers within the body of the response. Instead, the disclaimer appeared in a platform-generated warning at the end.

Why ChatGPT (GPT-4) is the Gold Standard for Comparison – GPT-4 stands out as the gold standard in this comparison for several reasons. First, its performance on medical challenge problems is exceptional, as demonstrated in the paper "Capabilities of GPT-4 on Medical Challenge Problems" [9]. GPT-4 outperformed its predecessors, such as GPT-3.5 and specialized LLMs like Med-PaLM, in both zero-shot and five-shot evaluations on official exams like the USMLE. It achieved an average score of 86.65% on the USMLE Self-Assessment, surpassing the passing threshold by a significant margin.

Moreover, GPT-4's calibration and reasoning abilities—especially in complex, multi-modal tasks—make it a reliable and consistent choice for medical applications. In addition to its performance, GPT-4's natural language generation capabilities, such as its ability to explain medical reasoning, personalize content for educational purposes, and handle counterfactual scenarios, showcase its potential for clinical decision support, patient education, and medical training. GPT-4's careful balance between depth and clarity, coupled with its understanding of medical context, positions it as the ideal candidate for multi-persona medical consultations, where the need for both accuracy and accessibility is paramount.

4 Related Work

Recent work on LLMs has explored persona-based prompting, where models are instructed to adopt specific roles or perspectives to guide their responses. Beyond single-role prompts, multi-persona prompting techniques have been introduced to allow an LLM to embody multiple roles simultaneously, integrating expertise from various angles within one interaction, demonstrating better performance on high-openness tasks [10]. The *Multi-Persona Interaction* pattern formalizes this approach by defining distinct personas (*e.g.* different expert roles) in the prompt and requiring the model to produce a cohesive answer that combines these perspectives. Such multi-role prompting has been shown to enrich the depth and breadth of responses in complex tasks that demand interdisciplinary knowledge [13].

Several frameworks also leverage multiple LLM agents or personas cooperating or debating to improve reasoning. For example, multi-LLM debate and collaboration methods have been used to simulate artificial "societies" of agents and to boost logical reasoning performance. In particular, Sandwar et al. [12] propose a "Town Hall" debate prompt that splits one LLM into multiple personas who argue and vote on a solution, yielding significant accuracy gains on a reasoning benchmark compared to standard chain-of-thought baselines. These advances show how multi-persona and multi-agent interactions can elicit more robust problem-solving behavior from LLMs.

Similarly, multi-agent collaboration frameworks, such as those explored by Liang et al. [7], separate LLM instances via coordinated expert roles. Distinct LLM agents assume roles collectively analyzing cases. These multi-agent teams have shown marked improvements in diagnostic accuracy, with collaborative outputs surpassing single-agent benchmarks. Our study builds upon these findings by applying structured multi-persona prompting specifically to medical consultations, comparing multiple leading LLMs systematically.

Evaluating the outputs of LLMs — especially open-ended dialogues or multi-role conversations — remains hard. Traditional automated metrics (*e.g.*, string overlap or accuracy against a single reference answer) often fail to capture the quality of complex, free-form responses. To address this problem, the *LLM-as-a-Judge* paradigm has gained traction by using a powerful LLM as an evaluator to score the outputs of other models, providing a reference-free, scalable alternative to costly human evaluation.

Recent studies suggest that a well-designed LLM judge can approximate human evaluators closely. For instance, GPT-4-based judgments achieved over 80% agreement with human preference rankings, on par with inter-annotator agreement levels [17]. In a multi-turn medical dialogue benchmark, an automatic GPT-4 evaluation aligned with expert human ratings to within a few percentage points, further supporting the viability of LLM-based evaluation.

In light of the developments outlined above, our study extends prior work by uniting these threads. In particular, we apply a structured *Multi-Persona Interaction* pattern to a realistic multi-party medical consultation scenario and benchmark several state-of-the-art LLMs in this context. While others have explored multi-agent medical AI systems and general LLM benchmarks, to our knowledge no prior work has qualitatively compared multiple top-tier models by having them assume coordinated doctor personas in the same consultation. Moreover, we leverage GPT-4 as an impartial judge to evaluate each model's performance under consistent criteria, drawing on the *LLM-as-a-Judge* paradigm within a multi-role setting. This design allows us to systematically assess how different advanced LLMs handle a complex, role-structured medical dialogue.

5 Concluding Remarks

The *Multi-Persona Integration* pattern described in this paper harnesses LLMs as an AI "medical roundtable," enabling a single prompt to summon coordinated expertise from multiple specialist personas and deliver well-rounded advice. This study uses that pattern to measure how today's leading LLMs behave when the stakes include patient safety and clinical accuracy. We learned the following lessons from our work on this paper:

- **Prompt patterns are powerful, but handle with care** – Multi-persona prompting reliably expands diagnostic depth and treatment breadth, yet without explicit role reminders some LLMs merge voices and dilute specialist value.
- ***LLM-as-a-Judge* = scalable quality control** – This evaluation pattern gives researchers a ready-made framework to compare future LLMs or new medical scenarios fairly and at scale, *e.g.*, employing GPT-4 to grade peers across eight dimensions offers a reproducible, cost-effective benchmark that surfaces nuances human reviewers might miss.
- **Role fidelity is fragile** – Meta's occasional persona "collapse" shows that system prompts or stronger in-prompt scaffolds may be required to keep specialists from blending into a generic narrator.
- **Safety and user clarity are non-negotiable** – Completeness must be balanced with patient-friendly language and explicit disclaimers; LLMs that don't volunteer these cues need extra prompting or fine-tuning.
- **No single LLM is ideal—just different strengths** – GPT-4 and Claude coordinate personas most gracefully, Gemini trades brevity for exhaustive detail, DeepSeek presents its information in a clear organized way, and Meta offers fluent synthesis but sometimes blurs boundaries.

- **A blueprint for high-stakes evaluation** – By establishing this pattern compound (*Multi-Persona Integration*), our study contributes a blueprint for evaluating LLMs in high-stakes decision-making contexts like healthcare, where thoroughness, accuracy, and safety are paramount. The confidence and clarity of the results – obtained through an AI-assisted yet unbiased adjudication – reinforce the viability of LLM-as-a-judge as a general strategy for AI evaluation. We believe that this work lays a foundation for reproducible, extensible, and clinically relevant benchmarking of AI systems.

The following are research questions we plan to address in future work:

1. **Ethical and bias evaluation** – How do various LLMs differ in their ethical decision-making processes and biases when simulating multiple medical roles, and what methods can be employed to systematically measure and mitigate these biases?
2. **Explainability and transparency** – To what extent do explanations provided by LLMs in multi-persona medical consultations align with clinical reasoning expected from human medical professionals, and how can this alignment be quantitatively evaluated and improved?
3. **Large-scale benchmark testing for accuracy** – Test multi-persona LLM responses on large-scale medical question-answering benchmarks (*e.g.*, board exam question banks or biomedical QA challenges) to quantify factual accuracy and breadth of knowledge. Systematically evaluating the multi-persona approach on established datasets – from straightforward diagnostic queries to complex clinical vignettes – would provide statistically robust evidence of where the approach improves correctness and where errors persist. Such evaluation would also help identify any knowledge gaps in current LLMs when confronted with wide-ranging medical scenarios.
4. **Real-world clinical utility and acceptance** – What factors determine healthcare professionals' acceptance and integration of multi-persona LLM consultations into real-world medical practice, and how can the effectiveness of these LLMs be evaluated through clinical trials or observational studies in actual clinical settings?
5. **Diverse LLMs as evaluation judges** – Investigate using different strong LLMs as judges (or an ensemble of judges) to assess consistency and potential bias in automated evaluation. While our study relied on GPT-4 as the sole arbiter, future work could examine if swapping in another top-tier LLM (or a consensus of several LLMs) yields similar scoring outcomes. Analyzing agreement between AI and human judges, as well as between different AI judges, will be crucial to validate the reliability and fairness of the *LLM-as-a-judge* pattern. This direction will also shed light on any bias introduced by a single evaluator LLM and how it might be mitigated.

References

[1] Anthropic. Claude 3.7 sonnet, 2024.

[2] Frank Bushmann, Kevin Henney, and Douglas C Schmidt. *Pattern-Oriented Software Architecture Volume 5: On Patterns and Pattern Languages*. John Wiley & Sons, 2007.

[3] DeepSeek-AI. Deepseek-v3 github, 2024.

[4] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan

Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025.

[5] Google-AI. Gemini 2.0 flash, 2024.

[6] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.

[7] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, 2024.

[8] Meta-AI. Meta ai, 2024.

[9] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[10] Carlos Olea, Holly Tucker, Jessica Phelan, Cameron Pattison, Shen Zhang, Maxwell Lieb, Doug Schmidt, and Jules White. Evaluating persona prompting for question answering tasks. In *Proceedings of th e 10th international conference on artificial intelligence and soft computing, Sydney, Australia*, 2024.

[11] Open-AI. Chatgpt-4, 2024.

[12] Vivaan Sandwar, Bhav Jain, Rishan Thangaraj, Ishaan Garg, Michael Lam, and Kevin Zhu. Town hall debate prompting: Enhancing logical reasoning in llms through multi-persona interaction. *arXiv preprint arXiv:2502.15725*, 2025.

[13] William Schreiber, Jules White, and Douglas C Schmidt. Toward a pattern language for persona-based interactions with llms. October 2024.

[14] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

[15] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. In *Proceedings of the 30th Pattern Languages of Programming (PLoP) conference*, Allerton Park, IL, October 2023.

[16] Jules White, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C Schmidt. ChatGPT Prompt Patterns for Improving Code Quality, Refactoring, Requirements Elicitation, and Software Design. In *Generative AI for Effective Software Development*, pages 71–108, 2024.

[17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[18] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large Language Models Are Human-Level Prompt Engineers. https://arxiv.org/abs/2211.01910, 2023.