Diagnostic Performance and Cost-Efficiency of Large Language Models in Secondary Hypertension: A Blinded Comparative Study

Asena Gökçay Canpolat *, Özge Baş Aksu, Rıfat Emral, Uğur Canpolat

Posted Date: 18 March 2026

doi: 10.20944/preprints202603.1486.v1

Keywords: secondary hypertension; large language models; artificial intelligence; clinical decision support; diagnostic reasoning; chatbot comparison; hallucination; cost-efficiency

 doi:10.20944/preprints202603.1486.v1

Article

Diagnostic Performance and Cost-Efficiency of Large Language Models in Secondary Hypertension: A Blinded Comparative Study

Asena Gökçay Canpolat [1,*], Özge Baş Aksu [1], Rıfat Emral [1] and Uğur Canpolat [2]

[1] Department of Endocrinology and Metabolism, Ankara University School of Medicine, Ankara, Turkey

[2] Department of Cardiology, Hacettepe University School of Medicine, Ankara, Turkey

* Correspondence: asena-gokcay@hotmail.com; Tel: 90312 5082108; Fax: 90312 309 45 05

Abstract

Background/Objectives: Secondary hypertension requires complex diagnostic reasoning and guideline-based management, posing challenges for artificial intelligence–based clinical decision-support systems. This study aimed to comparatively evaluate the performance of three large language models (LLMs) in diagnostic reasoning, clinical management, follow-up planning, and patient-oriented communication related to secondary hypertension. **Methods**: In this cross-sectional blinded study, three LLMs (ChatGPT-5.2, Claude Sonnet 4.6, and Gemini 3.0 Pro) were evaluated using 10 expert-developed clinical case vignettes representing major etiologies of secondary hypertension. Model outputs were anonymized and independently assessed by three senior clinicians (two endocrinologists and one cardiologist) using a 7-point Likert scale across five domains: (1) accuracy and hallucination control, (2) quality and comprehensiveness, (3) reliability and clinical guidance, (4) cost-efficiency, and (5) clinical usability. Group differences were analyzed using Kruskal–Wallis tests with Bonferroni-corrected pairwise comparisons. Inter-rater agreement was evaluated using two-way mixed-effects intraclass correlation coefficients with absolute agreement. **Results**: A total of 90 blinded expert ratings were analyzed. Claude Sonnet 4.6 achieved the highest composite performance score (6.63 ± 0.45), followed by ChatGPT-5.2 (5.82 ± 0.55) and Gemini 3.0 Pro (5.27 ± 0.89) ($H = 40.055$, $p < 0.001$). Claude Sonnet 4.6 significantly outperformed both models across all evaluation domains. ChatGPT-5.2 demonstrated intermediate performance and significantly exceeded Gemini 3.0 Pro in reliability and clinical usability. Performance differences were most pronounced in domains requiring complex clinical reasoning, whereas cost-efficiency scores were relatively comparable among models. Claude Sonnet 4.6 ranked first in nine of ten clinical vignettes. Inter-rater agreement demonstrated consistent ranking patterns among evaluators. **Conclusions**: Large language models exhibit heterogeneous performance in secondary hypertension–related clinical tasks. Although advanced models show promising capabilities as clinical decision-support tools, performance remains model-dependent, particularly in complex endocrine–metabolic scenarios. Domain-specific validation and prospective clinical studies are required before routine clinical implementation.

Keywords: secondary hypertension; large language models; artificial intelligence; clinical decision support; diagnostic reasoning; chatbot comparison; hallucination; cost-efficiency

1. Introduction

Secondary hypertension (SH) represents a heterogeneous group of disorders in which elevated blood pressure arises from an identifiable and potentially reversible underlying cause. It accounts for approximately 5–10% of all hypertension cases in the general population and up to 20–30% among patients with resistant hypertension, early-onset hypertension, or abrupt deterioration of previously stable blood pressure control (1, 2). The most common etiologies include primary aldosteronism,

 doi:10.20944/preprints202603.1486.v1

renovascular disease, chronic kidney disease, obstructive sleep apnea, pheochromocytoma, Cushing's syndrome, thyroid disorders, and medication- or substance-induced hypertension. Early recognition of these conditions is of major clinical importance because targeted therapy may substantially improve blood pressure control, reduce long-term cardiovascular and renal morbidity, and in selected cases provide definitive cure (3, 4).

Despite its clinical relevance, the diagnostic workup of SH remains complex and frequently underutilized in routine practice. Clinical manifestations are often nonspecific, and distinguishing secondary forms from essential hypertension can be difficult, particularly in patients with mild or moderate blood pressure elevations. Furthermore, screening recommendations vary across international guidelines, creating uncertainty regarding patient selection and optimal testing strategies (2, 5). iochemical evaluation is complicated by pre-analytical variability, assay limitations, and the confounding effects of commonly prescribed antihypertensive agents on renin–angiotensin–aldosterone axis measurements. Imaging modalities, while informative, may reveal incidental findings that do not necessarily indicate causality. These diagnostic ambiguities increase the cognitive burden on clinicians and may lead to delayed diagnosis, inappropriate testing, or missed opportunities for curative interventions. Consequently, SH represents a clinical domain in which structured reasoning, guideline familiarity, and integrative decision-making are essential.

Recent advances in artificial intelligence have introduced large language models (LLMs) as emerging tools with potential applications in medical education, diagnostic reasoning, clinical decision support, and patient-oriented health communication. LLMs are trained on vast corpora of text data and are capable of generating context-aware responses that simulate human-like reasoning. In healthcare settings, these systems have demonstrated the ability to synthesize medical knowledge, produce differential diagnoses, interpret clinical scenarios, and assist in management planning. Systematic reviews and simulation-based assessments suggest that LLMs may achieve near-expert performance in selected medical tasks, particularly when structured prompts and well-defined clinical contexts are provided (6-10). Comparative evaluations of major LLM-based systems—including ChatGPT, Gemini, and Claude—have revealed substantial variability in diagnostic accuracy, reasoning transparency, consistency of responses, and safety profiles. Differences in training data composition, alignment strategies, reinforcement learning frameworks, and model architectures may contribute to heterogeneous clinical performance (11-14). While some studies report high levels of diagnostic concordance with expert clinicians, others highlight risks related to hallucinated information, incomplete guideline adherence, overconfident recommendations, and variability across repeated queries. These limitations underscore the importance of systematic benchmarking of LLMs in clinically realistic scenarios before their integration into decision-support workflows.

Notably, existing literature has primarily focused on general internal medicine cases, examination-style question banks, or single-step diagnostic tasks. Data remain limited regarding LLM performance in complex endocrine disorders that require multistep reasoning, biochemical interpretation, and adherence to evidence-based algorithms. Secondary hypertension constitutes an ideal test framework because its evaluation demands integration of clinical features, laboratory interpretation, medication effects, imaging strategies, and longitudinal management planning. Moreover, appropriate care requires careful distinction between screening, confirmatory testing, subtype classification, and therapeutic decision-making—processes that challenge both human clinicians and AI-based systems.

To our knowledge, no blinded comparative investigation has systematically evaluated LLM performance across multiple domains of clinical reasoning within the context of secondary hypertension. Furthermore, few studies have incorporated clinician-rated assessments, structured scoring systems, and cost-efficiency analyses to reflect real-world applicability.

Therefore, the present study aimed to perform a blinded head-to-head comparison of three widely used large language models across key domains of endocrine clinical reasoning related to secondary hypertension. Specifically, we evaluated their performance in diagnostic reasoning,

 doi:10.20944/preprints202603.1486.v1

follow-up planning, therapeutic decision-making, and patient-oriented clinical communication. By assessing clinical accuracy, inter-model consistency, and cost-efficiency, this study seeks to clarify the potential role and current limitations of LLMs as supportive tools in real-world endocrine and hypertension practice.

2. Materials and Methods

2.1. Study Design and Ethical Considerations

This cross-sectional, blinded, head-to-head comparative study was designed to evaluate the clinical performance of three state-of-the-art large language models (LLMs) in the diagnostic evaluation and management of secondary hypertension (SH). The study framework was structured to simulate real-world clinical reasoning tasks encountered in endocrine and hypertension practice. All assessments were performed using standardized hypothetical case vignettes rather than real patient data.

Because the study did not involve human participants, identifiable patient information, biological materials, or animal experimentation, formal ethical committee approval and informed consent were not required. This approach is consistent with previously published simulation-based evaluations of artificial intelligence systems in clinical medicine. The study adhered to principles of transparency, reproducibility, and methodological neutrality in AI benchmarking.

2.2. Model Selection and Anonymization

Three widely used and commercially available large language models representing different architectural frameworks and training methodologies were selected for evaluation:

- ChatGPT-5.2 (OpenAI)
- Claude Sonnet 4.6 (Anthropic)
- Gemini 3.0 Pro (Google DeepMind)

To reduce potential evaluator bias associated with brand recognition, model identities were anonymized prior to scoring. Each model was assigned a neutral elemental code name (Water, Earth, Air), and all outputs were reformatted to remove stylistic identifiers (Supplementary Table 1). Evaluators remained fully blinded to model identity throughout the assessment process.

Model selection was based on:

1. Widespread clinical and academic usage
2. Advanced multimodal reasoning capabilities
3. Public accessibility and reproducibility
4. Representation of distinct AI training paradigms.

2.3. Case Vignette Development

Ten high-fidelity clinical case vignettes were developed collaboratively by a multidisciplinary panel consisting of two endocrinologists and one cardiologist with extensive expertise in hypertension and endocrine disorders. Cases were designed to reflect realistic diagnostic scenarios encountered in tertiary endocrine referral centers.

Each vignette was constructed to include:

- Demographic characteristics
- Presenting symptoms and physical examination findings
- Relevant laboratory results
- Imaging findings where appropriate
- Medication history
- Key clinical decision points

Vignettes collectively represented major etiological categories of SH:

 doi:10.20944/preprints202603.1486.v1

1. Primary aldosteronism
2. Pheochromocytoma/paraganglioma
3. Atherosclerotic renal artery stenosis
4. Fibromuscular dysplasia
5. Primary hyperparathyroidism
6. Obstructive sleep apnea
7. Coarctation of the aorta
8. Cushing's syndrome
9. Renal parenchymal disease
10. Mixed or atypical presentations requiring complex differential diagnosis

Case complexity was intentionally varied to test model performance across straightforward, intermediate, and diagnostically challenging scenarios requiring multistep reasoning.

2.4. Prompt Standardization and Model Querying

To ensure methodological consistency, all LLMs were queried within the same time window using default model configurations without temperature adjustment or external tool augmentation. A standardized, high-fidelity English prompt was developed to minimize prompt-induced variability. The prompt instructed each model to assume the role of a "board-certified specialist in endocrinology and hypertension" and to provide structured clinical reasoning. No iterative prompting, clarification requests, or response regeneration were permitted. Each vignette was submitted once to each model in an independent session to prevent memory contamination across cases.

Model responses were required to follow a predefined structured format consisting of five clinical reasoning domains:

1. Diagnosis and Differential Diagnosis
2. Diagnostic Workup Strategy
3. Acute and Long-Term Management Plan
4. Follow-up and Monitoring Strategy
5. Patient-Oriented Education and Counseling

Outputs exceeding predefined length thresholds were truncated to ensure comparable evaluation conditions.

2.5. Evaluation Framework and Scoring System

2.5.1. Evaluator Panel

Blinded evaluations were performed independently by three senior clinicians:

- Two board-certified endocrinologists
- One board-certified cardiologist

Clinical experience ranged from 10 to 30 years in tertiary referral centers. All evaluators routinely manage patients with complex hypertension and endocrine disorders. Before scoring, evaluators underwent calibration using two pilot vignettes to harmonize scoring interpretation and reduce inter-observer variability.

2.5.2. Scoring Domains

Each model response was assessed using a 7-point Likert scale (1 = poorest performance; 7 = excellent performance) across five predefined domains:

1. Accuracy and Hallucination Control

Assessed factual correctness, internal consistency, and concordance with contemporary international guidelines (ESH/ESC hypertension guidelines and Endocrine Society recommendations).

 doi:10.20944/preprints202603.1486.v1

2. Quality and Comprehensiveness of Clinical Reasoning

Evaluated logical structure, pathophysiological reasoning, completeness of differential diagnosis, and integration of clinical data.

3. Reliability and Safety of Clinical Guidance

Assessed whether recommendations were safe, clinically appropriate, and free of potentially harmful or misleading suggestions.

4. Cost-Efficiency of Diagnostic Strategy

Evaluated prioritization of high-yield tests, avoidance of unnecessary investigations, and consideration of healthcare resource utilization.

5. Clinical Usability and Practical Applicability

Assessed clarity, organization, and suitability for real-world clinical implementation by practicing physicians.

2.5.3. Global Preference Assessment

In addition to domain scoring, evaluators selected a "Global Favorite" model for each vignette. This forced-choice metric identified the model providing the most trustworthy and clinically useful overall decision support.

2.6. Statistical Analysis

All statistical analyses were conducted using IBM SPSS Statistics version 25 (IBM Corp., Armonk, NY, USA). Because evaluation metrics consisted of ordinal Likert-scale data, distribution normality was assessed using the Kolmogorov–Smirnov test. Non-normal distributions were observed in at least one comparison group for each domain; therefore, non-parametric statistical methods were applied throughout. Differences in median domain scores among the three LLMs were evaluated using the Kruskal–Wallis H test. When significant omnibus differences were detected, pairwise post hoc comparisons were performed using the Mann–Whitney U test. To control for Type I error inflation due to multiple comparisons, Bonferroni correction was applied, establishing an adjusted significance threshold of: $\alpha = 0.05 / 3 = 0.017$. A composite performance index was calculated for each observation by averaging scores across the five evaluation domains. Inter-model differences in composite scores were analyzed using the same non-parametric procedures. Agreement among evaluators was assessed using a two-way mixed-effects intraclass correlation coefficient (ICC) model with absolute agreement definition. ICC values were interpreted as:

- <0.50→Poor reliability
- 0.50–0.75→Moderate reliability
- 0.75–0.90→Good reliability
- >0.90 → Excellent reliability

In addition to domain-level comparisons, case-level composite performance scores were analyzed to evaluate model consistency across distinct clinical scenarios. Mean composite scores for each vignette were summarized descriptively and visualized using a heatmap to facilitate pattern recognition across diagnostic categories. Domain-specific performance profiles were further illustrated using radar plots, while inter-model differences in mean domain scores were displayed using forest plots. These visualizations were generated to enhance interpretability of multidimensional performance patterns but were not used for inferential statistical testing.

Statistical significance thresholds were defined as: $p < 0.05$ for omnibus comparisons and $p < 0.017$ for Bonferroni-adjusted pairwise tests. All tests were two-tailed.

3. Results

A total of 90 blinded expert evaluations (three LLMs × 10 clinical vignettes × three independent evaluators) were analyzed across five predefined domains using a 7-point Likert scoring system. All responses were independently rated by three senior clinicians with extensive experience in tertiary

 doi:10.20944/preprints202603.1486.v1

endocrine and cardiovascular care, including two endocrinologists (A.G.C., R.E.) and one cardiologist (U.C.).

3.1. Overall Model Performance

Composite performance scores demonstrated statistically significant differences among models (Kruskal–Wallis H = 40.055, p < 0.001), indicating heterogeneous overall clinical performance.

Claude Sonnet 4.6 achieved the highest composite score (mean ± SD: **6.63 ± 0.45**), followed by **ChatGPT-5.2** (5.82 ± 0.55) and **Gemini 3.0 Pro** (5.27 ± 0.89).

Post hoc pairwise comparisons revealed:

- Claude Sonnet 4.6 significantly outperformed ChatGPT-5.2 (U = 785, p < 0.001)
- Claude Sonnet 4.6 significantly outperformed Gemini 3.0 Pro (U = 822, p < 0.001)
- ChatGPT-5.2 showed numerically higher scores than Gemini 3.0 Pro (U = 296, p = 0.023), though this difference did not remain significant after Bonferroni correction.

These findings establish a clear performance hierarchy:

Claude Sonnet 4.6 > ChatGPT-5.2 > Gemini 3.0 Pro

Domain-specific analyses demonstrated statistically significant differences across all five evaluation domains (Table 1).

Table 1. Criterion-Based Performance.

Criterion	Claude Sonnet 4.6	ChatGPT-5.2	Gemini 3.0 Pro	*p* (ANOVA)
Accuracy & Hallucination Control	**6.80 ± 0.41**	5.87 ± 0.57	5.40 ± 0.93	<0.001
Quality & Comprehensiveness	**6.63 ± 0.56**	5.50 ± 0.86	4.97 ± 1.00	<0.001
Reliability & Clinical Guidance	**6.67 ± 0.48**	5.80 ± 0.66	5.17 ± 0.91	<0.001
Cost-Efficiency	**6.43 ± 0.82**	6.10 ± 0.66	5.60 ± 1.22	0.003
Clinical Usability	**6.60 ± 0.50**	5.83 ± 0.59	5.20 ± 0.93	<0.001

Bold values indicate the highest-scoring model for each criterion.

Accuracy and Hallucination Control: Significant inter-model differences were observed (H = 42.443, p < 0.001). Claude Sonnet 4.6 demonstrated superior factual accuracy and guideline concordance, significantly outperforming both comparators after Bonferroni adjustment.

Although ChatGPT-5.2 scored higher than Gemini 3.0 Pro, this difference did not remain statistically significant after correction (p = 0.037).

Quality and Comprehensiveness of Clinical Reasoning: This domain demonstrated one of the largest effect sizes (H = 39.075, p < 0.001). Claude Sonnet 4.6 consistently provided more structured differential diagnoses, deeper pathophysiological integration, and more complete diagnostic strategies.

Pairwise differences between ChatGPT-5.2 and Gemini 3.0 Pro did not remain significant after correction (p = 0.022).

Reliability and Safety of Clinical Guidance: Claude Sonnet 4.6 again ranked highest (H = 40.314, p < 0.001), reflecting safer therapeutic recommendations and better adherence to guideline-based management pathways.

Notably, ChatGPT-5.2 significantly outperformed Gemini 3.0 Pro after Bonferroni correction (U = 284, p = 0.009), indicating comparatively greater reliability in treatment planning and risk avoidance.

Cost-Efficiency of Diagnostic Strategy: This domain showed the smallest between-group effect size (H = 9.148, p = 0.010), suggesting relatively comparable performance across models.

After Bonferroni adjustment:

- Claude Sonnet 4.6 significantly outperformed Gemini 3.0 Pro (U = 623, p = 0.007)
- Claude Sonnet 4.6 and ChatGPT-5.2 demonstrated statistically comparable performance (p = 0.032)
- ChatGPT-5.2 and Gemini 3.0 Pro showed no significant difference

 doi:10.20944/preprints202603.1486.v1

Overall, all models demonstrated relatively stronger performance in test prioritization compared with other reasoning domains.

Visual inspection of multidimensional performance profiles supported quantitative findings. Radar plot visualization demonstrated consistently superior domain-wide performance of Claude Sonnet 4.6, with more homogeneous scoring patterns compared with the greater variability observed for Gemini 3.0 Pro (Figure 1). Forest plot comparisons further illustrated the magnitude of inter-model differences across domains, highlighting the largest effect sizes in accuracy, comprehensiveness, reliability, and clinical usability, whereas cost-efficiency showed smaller between-model separation (Figure 2).

Clinical Usability and Practical Applicability: Clinical usability differed markedly among models ($H = 37.252$, $p < 0.001$). Claude Sonnet 4.6 produced clearer, better organized, and more clinically actionable outputs.

All pairwise comparisons remained significant after Bonferroni correction:

- Claude Sonnet 4.6 vs Gemini 3.0 Pro: $U = 798$, $p < 0.001$
- Claude Sonnet 4.6 vs ChatGPT-5.2: $U = 723$, $p < 0.001$
- ChatGPT-5.2 vs Gemini 3.0 Pro: $U = 272$, $p = 0.005$

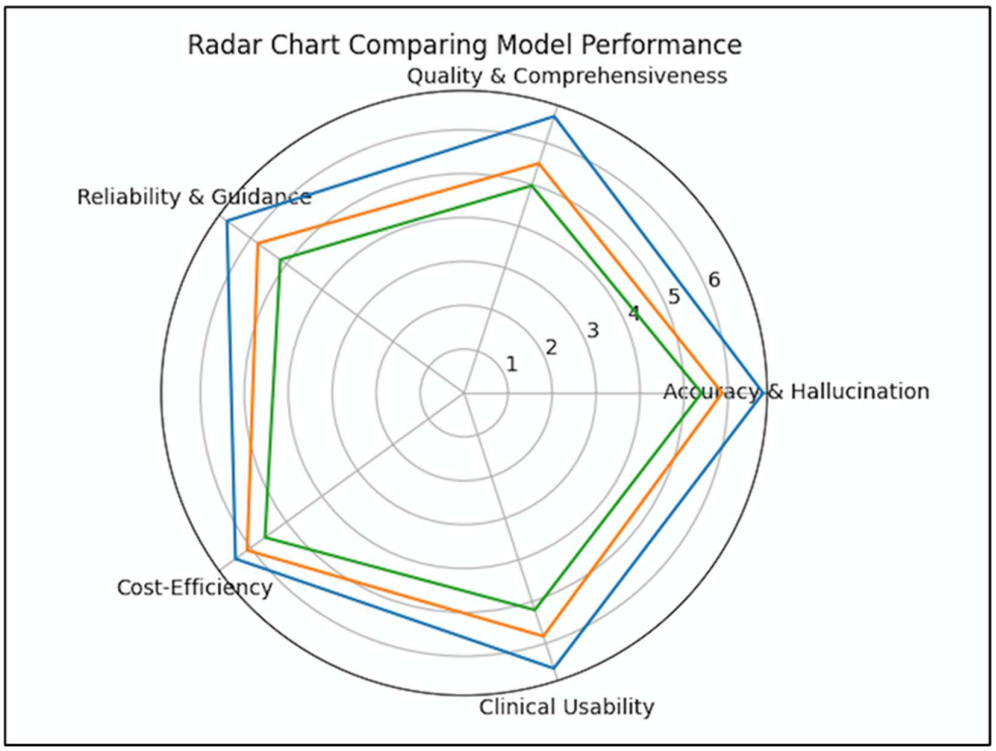

Figure 1. Forest plot showing mean domain scores across five clinical evaluation criteria. Claude Sonnet 4.6 consistently achieved the highest scores across all domains, followed by ChatGPT-5.2 and Gemini 3.0 Pro.

 doi:10.20944/preprints202603.1486.v1

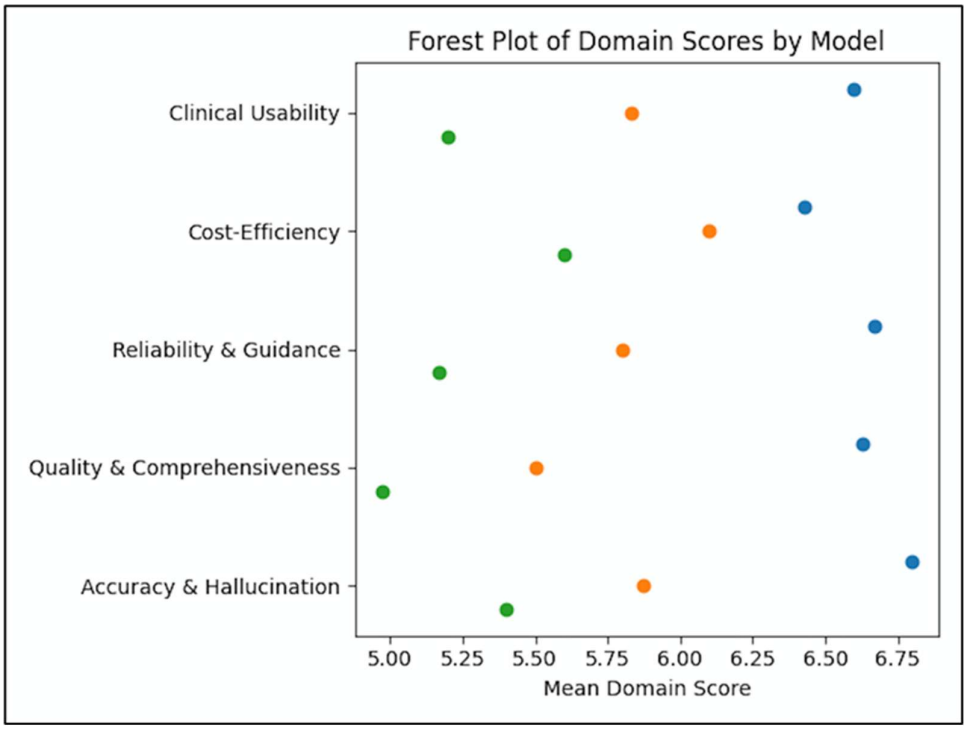

Figure 2. Radar chart illustrating multidimensional performance profiles. Claude Sonnet 4.6 demonstrates uniformly superior performance, while ChatGPT-5.2 shows intermediate performance and Gemini 3.0 Pro exhibits greater variability across domains.

3.2. Inter-Rater Agreement

Scoring distributions were consistent across evaluators, demonstrating strong inter-rater reliability.

Although evaluator R.E. assigned slightly higher absolute scores (overall mean ≈6.2) compared with U.C. and A.G.C. (≈5.9), all evaluators demonstrated identical model ranking patterns across all domains, confirming robust agreement in relative performance assessment.

This consistency supports the objectivity and reproducibility of the evaluation framework.

3.3. Case-Level Performance

Claude Sonnet 4.6 achieved the highest composite score in **9 of 10 clinical cases**, demonstrating consistent superiority across diverse etiologies of secondary hypertension.

Table 2. Composite mean scores per model across 10 clinical vignettes.

Case	Diagnosis	Claude Sonnet 4.6	ChatGPT-5.2	Gemini 3.0 Pro
1	Pheochromocytoma	**6.93**	5.40	6.00
2	Primary Hyperaldosteronism	6.07	5.87	**6.33**
3	Renal Artery Stenosis (Atherosclerotic)	**6.87**	5.80	5.93
4	Primary Hyperparathyroidism	**6.00**	5.33	5.80
5	White Coat / Early Essential Hypertension	**6.87**	5.20	5.87
6	Obstructive Sleep Apnea	**6.47**	5.73	5.93
7	Coarctation of the Aorta	**6.53**	5.20	6.00
8	Cushing Syndrome	**6.73**	4.27	5.20
9	Diabetic Nephropathy	**6.93**	4.73	5.53
10	Fibromuscular Dysplasia (Renal Artery)	**6.87**	5.13	5.60

Bold values indicate the highest-scoring model for each case.

 doi:10.20944/preprints202603.1486.v1

The only exception was **Case 2 (Primary Hyperaldosteronism)**, where Gemini 3.0 Pro achieved the highest score. However, score dispersion among models was minimal, likely reflecting the structured and algorithm-driven nature of this diagnosis.

Largest performance gaps were observed in complex endocrine–metabolic cases:

- **Cushing's syndrome (Case 8)**: $\Delta = 2.46$ (Claude vs ChatGPT)
- **Diabetic nephropathy (Case 9)**: $\Delta = 2.20$

These findings suggest increased vulnerability of some models when multistep hormonal interpretation and systemic metabolic reasoning are required.

In contrast, Claude Sonnet 4.6 demonstrated near-ceiling performance (≥6.87) in:

- Pheochromocytoma
- Renal artery stenosis
- White coat hypertension
- Fibromuscular dysplasia

This pattern indicates strong adaptability across vascular, endocrine, and mixed-etiology hypertension scenarios.

Descriptive analysis of vignette-specific composite scores demonstrated stable performance patterns across heterogeneous etiologies of secondary hypertension. Heatmap visualization revealed consistently high performance of Claude Sonnet 4.6 across nearly all clinical scenarios, whereas performance variability was more pronounced for ChatGPT-5.2 and Gemini 3.0 Pro, particularly in complex endocrine-metabolic conditions (Figure 3). These graphical representations facilitated identification of scenario-specific strengths and weaknesses that were not fully captured by aggregated statistical comparisons.

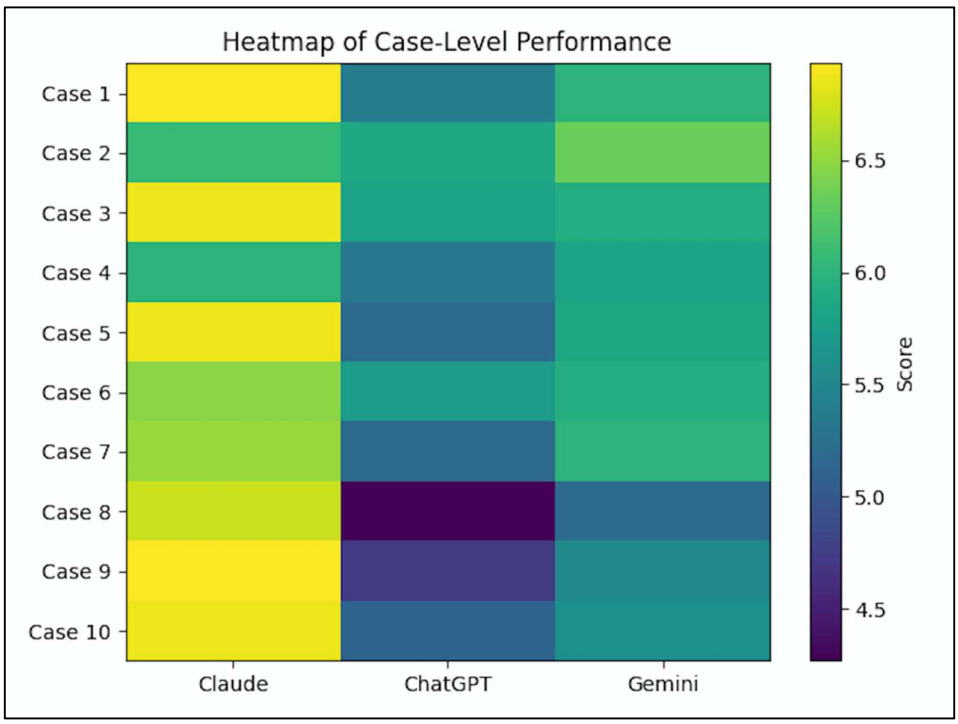

Figure 3. Heatmap depicting case-level composite scores across ten clinical scenarios. Warmer colors indicate higher performance. Claude Sonnet 4.6 ranked highest in nine of ten cases, with the largest inter-model performance gaps observed in complex endocrine disorders.

 doi:10.20944/preprints202603.1486.v1

4. Discussion

In this blinded head-to-head comparative study, we evaluated three state-of-the-art large language models (LLMs) across clinically realistic secondary hypertension (SH) scenarios encompassing diagnostic reasoning, diagnostic workup, management planning, longitudinal follow-up, and patient education. Using expert-developed case vignettes and a structured, guideline-aligned evaluation framework, a clear performance gradient emerged. Claude Sonnet 4.6 demonstrated consistently superior performance across all evaluation domains and ranked first in nine of ten clinical scenarios. ChatGPT-5.2 showed intermediate performance, while Gemini 3.0 Pro exhibited greater variability across domains and clinical contexts. Interestingly, cost-efficiency scores were comparatively similar among models, suggesting convergence in broadly guideline-consistent diagnostic prioritization despite differences in overall reasoning depth and reliability. Collectively, these findings underscore both the rapidly expanding clinical capabilities of LLMs and the substantial inter-model variability that currently limits uniform clinical deployment.

Secondary hypertension represents a particularly demanding test environment for AI-assisted reasoning because accurate evaluation requires integration of multidisciplinary knowledge, probabilistic thinking, and guideline-based decision-making rather than simple factual recall (15). Distinguishing primary from secondary etiologies, interpreting hormonal pathophysiology, accounting for medication interference, and sequencing confirmatory tests require nuanced multistep reasoning that challenges both trainees and experienced clinicians. Our findings suggest that advanced LLMs can approximate structured specialist reasoning when prompts explicitly define professional roles and output structure, consistent with prior studies demonstrating improved diagnostic performance using structured prompting frameworks (9, 16). The superior performance of the highest-ranked model—particularly in guideline concordance—supports emerging evidence that newer LLM architectures may better internalize evidence-based clinical reasoning patterns (12, 13, 17).

Hallucinations remain a principal barrier to safe implementation of generative artificial intelligence in clinical medicine (18). Performance differences in this study were most evident in hallucination control and reliability domains, where lower-performing models occasionally introduced unnecessary investigations, incomplete diagnostic hierarchies, or insufficient risk stratification strategies. This observation aligns with previous reports indicating that hallucinations are more likely to arise during complex, multistep clinical reasoning tasks rather than simple knowledge retrieval (19). Importantly, even the highest-performing model identified in this study did not achieve uniformly optimal outputs across all cases, reinforcing the prevailing consensus that LLMs should currently function as clinical decision-support adjuncts under physician supervision rather than autonomous decision-makers (20-22).

A distinctive contribution of this study is the inclusion of cost-efficiency as an evaluation domain. Over-testing is a recognized challenge in SH evaluation, where indiscriminate imaging and excessive biochemical investigations increase healthcare costs and patient burden. The relatively modest differences observed between models suggest that economic reasoning remains inconsistently represented within LLM outputs, emphasizing the importance of evaluating artificial intelligence systems not only for diagnostic accuracy but also for health-system applicability and value-based care alignment.

Across models, patient education responses were generally strong, consistent with previous findings that LLMs effectively translate complex medical concepts into accessible language (23). This capability may represent one of the earliest safe clinical applications of LLMs, particularly for chronic endocrine conditions requiring sustained patient engagement and shared decision-making. Nevertheless, variability in nuance, contextual framing, and risk communication underscores the continued necessity of clinician oversight (22).

The observed performance hierarchy indicates that LLM capability is rapidly evolving but remains model-dependent. Blinded anonymization minimized brand bias, demonstrating that perceived technological prominence does not necessarily predict clinical reasoning performance. To

 doi:10.20944/preprints202603.1486.v1

our knowledge, this is the first comparative chatbot evaluation specifically focused on secondary hypertension and assessed by experienced clinician-researchers using a structured, guideline-oriented framework. Study strengths include expert-generated vignettes reflecting real clinical complexity, blinded model evaluation, structured scoring aligned with evidence-based practice, strong inter-rater agreement, and incorporation of cost-efficiency as a novel clinical metric extending beyond accuracy-centered benchmarking.

Several limitations should be acknowledged. Case vignette designs cannot fully replicate real-world clinical uncertainty, comorbidity interactions, or longitudinal patient trajectories. Model outputs were assessed at a single time point despite rapid iterative updates in LLM architecture and training pipelines, and prompting strategies may influence performance outcomes (24). Additionally, evaluation was limited to English-language prompts, which may restrict generalizability across healthcare systems and linguistic settings.

Future research should evaluate LLM performance in prospective clinical workflows, incorporate real patient data streams, and examine longitudinal decision consistency. Domain-specific fine-tuning using endocrine-focused datasets and guideline-informed reinforcement learning strategies may further improve safety and reduce hallucination frequency (25-27). Development of benchmarking frameworks tailored to complex subspecialty conditions such as secondary hypertension will be essential as AI systems transition from experimental tools toward clinically integrated decision-support platforms.

Clinical Implications

The findings of this study suggest that contemporary large language models (LLMs) are approaching a level of performance that may support selected aspects of clinical decision-making in complex endocrine disorders such as secondary hypertension. Although none of the evaluated models demonstrated fully autonomous clinical reliability, higher-performing systems showed substantial capability in structured diagnostic reasoning, guideline-consistent investigation planning, and synthesis of management strategies. These characteristics indicate that LLMs may serve as supportive cognitive aids for clinicians, particularly in settings where access to subspecialty expertise is limited.

One potential near-term application lies in **decision-support augmentation**. In secondary hypertension, appropriate evaluation requires careful sequencing of biochemical tests, medication adjustments, and imaging modalities. LLMs capable of summarizing guideline-based pathways may assist clinicians in verifying diagnostic algorithms, reducing omissions, and improving adherence to evidence-based workflows. This support may be particularly valuable for general practitioners, trainees, and physicians practicing in resource-constrained environments.

A second practical implication involves **clinical documentation and information synthesis**. The ability of LLMs to integrate laboratory data, imaging findings, and clinical histories into structured summaries may reduce cognitive burden and improve efficiency in multidisciplinary care. In complex cases requiring coordination among endocrinologists, cardiologists, nephrologists, and radiologists, structured AI-assisted summaries could facilitate communication and reduce diagnostic delays.

Third, the consistently strong performance observed in **patient-oriented education** highlights an area of relatively low clinical risk and high potential utility. Secondary hypertension often involves chronic disease monitoring and complex hormonal evaluations that patients may find difficult to understand. LLMs can translate technical medical information into accessible language, potentially improving patient engagement, treatment adherence, and shared decision-making processes. However, clinician oversight remains essential to ensure contextual accuracy and appropriate risk communication.

The incorporation of **cost-efficiency considerations** in model evaluation further underscores the potential role of LLMs in promoting value-based care. By prioritizing high-yield diagnostic strategies and discouraging unnecessary testing, AI-assisted tools may contribute to more efficient resource

 doi:10.20944/preprints202603.1486.v1

utilization. Nevertheless, the modest differences observed among models suggest that economic reasoning remains insufficiently developed and should be strengthened through targeted training on health-system stewardship principles.

Importantly, these findings do not support replacement of clinician judgment. Instead, LLMs should currently be conceptualized as **adjunctive tools** that complement, rather than substitute, physician expertise. Human oversight remains indispensable for contextual interpretation, ethical decision-making, and individualized patient care.

As LLM architectures continue to evolve, integration into electronic health systems, guideline databases, and specialty-specific knowledge repositories may further enhance their reliability. Establishing regulatory standards, validation frameworks, and medico-legal guidance will be essential before widespread clinical implementation.

5. Conclusions

In this blinded comparative study, large language models demonstrated significantly different performance profiles in the evaluation and management of secondary hypertension. Claude Sonnet 4.6 showed the most consistent and highest overall performance, ChatGPT-5.2 demonstrated intermediate capability, and Gemini 3.0 Pro exhibited greater variability, particularly in complex endocrine–metabolic scenarios. Differences were most pronounced in domains requiring nuanced clinical reasoning, whereas cost-efficiency performance was relatively comparable. Although LLMs show promise as clinical decision-support tools, their reliability remains model-dependent, underscoring the need for domain-specific validation before routine clinical implementation.

Key Points

- This blinded comparative study evaluated three large language models (LLMs) using clinically realistic secondary hypertension scenarios and expert-guided scoring across multiple domains of clinical reasoning.
- Significant performance variability exists among LLMs, with Claude Sonnet 4.6 demonstrating superior diagnostic accuracy, reliability, clinical usability, and overall reasoning quality.
- ChatGPT-5.2 showed intermediate performance with relatively strong clinical guidance, while Gemini 3.0 Pro exhibited greater variability, particularly in complex endocrine–metabolic scenarios.
- Cost-efficiency scores were more comparable across models, suggesting that guideline-consistent test prioritization may be more uniformly learned than advanced diagnostic reasoning.
- LLMs demonstrated strong capability in patient-oriented communication, supporting their potential role in patient education and shared decision-making.
- Despite improving performance, hallucinations and variability in multistep reasoning remain key barriers to autonomous clinical use, reinforcing the need for clinician oversight.
- This study introduces a structured, guideline-aligned benchmarking framework for evaluating LLM performance in complex endocrine disorders.

Take-Home Messages

- Large language models are emerging as promising decision-support tools in complex hypertension management but currently demonstrate substantial variability in clinical reasoning performance.
- Advanced LLMs can approximate structured specialist thinking when guided by role-specific prompts and standardized output frameworks.
- Diagnostic safety remains a concern, particularly in multistep endocrine reasoning where hallucinations and incomplete risk assessment may occur.
- Patient education represents one of the most immediate and lower-risk applications of LLMs in endocrine care.

 doi:10.20944/preprints202603.1486.v1

- Artificial intelligence should currently augment—not replace—clinician expertise in secondary hypertension management.
- Future progress depends on domain-specific training, guideline-informed reinforcement learning, and prospective clinical validation.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Table S1. LLM code names and model correspondence used in the blinded evaluation.

Author Contributions: Conceptualization, AGC. and UC.; methodology, ÖBA.; software, ÖBA.; validation, AGC, UC .; formal analysis, ÖBA.; investigation, AGC.; resources, AGC.; data curation, AGC, ÖBA.; writing—original draft preparation, AGC, UC,RE .; writing—review and editing, UC, RE.; visualization, UC.; supervision, UC.; project administration, UC. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Acknowledgments: During the preparation of this manuscript/study, the authors used [NotebookLM] for the purposes of creating [Graphical abstract] and [ChatGpt 5.2] for the language editing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rimoldi SF, Scherrer U, Messerli FH. Secondary arterial hypertension: when, who, and how to screen? Eur Heart J. 2014;35(19):1245–54.
2. Carey RM, Calhoun DA, Bakris GL, Brook RD, Daugherty SL, Dennison-Himmelfarb CR, et al. Resistant Hypertension: Detection, Evaluation, and Management: A Scientific Statement From the American Heart Association. Hypertension. 2018;72(5):e53–e90.
3. Funder JW, Carey RM, Mantero F, Murad MH, Reincke M, Shibata H, et al. The Management of Primary Aldosteronism: Case Detection, Diagnosis, and Treatment: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab. 2016;101(5):1889–916.
4. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J. 2018;39(33):3021–104.
5. Whelton PK, Carey RM, Aronow WS, Casey DE, Jr., Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension. 2018;71(6):e13–e115.
6. Maity S, Saikia MJ. Large Language Models in Healthcare and Medical Applications: A Review. Bioengineering (Basel). 2025;12(6).
7. Su H, Sun Y, Li R, Zhang A, Yang Y, Xiao F, et al. Large Language Models in Medical Diagnostics: Scoping Review With Bibliometric Analysis. J Med Internet Res. 2025;27:e72062.
8. Chen P, Li Y, Zhang X, Feng X, Sun X. The acceptability and effectiveness of artificial intelligence-based chatbot for hypertensive patients in community: protocol for a mixed-methods study. BMC Public Health. 2024;24(1):2266.
9. Dinc MT, Bardak AE, Bahar F, Noronha C. Comparative analysis of large language models in clinical diagnosis: performance evaluation across common and complex medical cases. JAMIA Open. 2025;8(3):ooaf055.

 doi:10.20944/preprints202603.1486.v1

10. Pagano S, Strumolo L, Michalk K, Schiegl J, Pulido LC, Reinhard J, et al. Evaluating ChatGPT, Gemini and other Large Language Models (LLMs) in orthopaedic diagnostics: A prospective clinical study. Comput Struct Biotechnol J. 2025;28:9–15.
11. Madfa AA, Alshammari AF, Anazi BA, Alenezi YE, Alkurdi KA. Accuracy and reliability of Manus, ChatGPT, and Claude in case-based dental diagnosis. Front Oral Health. 2025;6:1686090.
12. Wojcik D, Adamiak O, Czerepak G, Tokarczuk O, Szalewski L. A bi-linguistic comparative analysis of ChatGPT-4, Gemini, and Claude performance on Polish medical-dental final examinations. Sci Rep. 2025;15(1):33083.
13. Idan D, Ben-Shitrit I, Volevich M, Binyamin Y, Nassar R, Nassar M, et al. Evaluating the performance of large language models versus human researchers on real world complex medical queries. Sci Rep. 2025;15(1):37824.
14. Tukur Jido J, Al-Wizni A, Aung SL. Readability of AI-Generated Patient Information Leaflets on Alzheimer's, Vascular Dementia, and Delirium. Cureus. 2025;17(6):e85463.
15. Wu L, Huang L, Li M, Xiong Z, Liu D, Liu Y, et al. Differential diagnosis of secondary hypertension based on deep learning. Artif Intell Med. 2023;141:102554.
16. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.
17. Guvel MC, Kiyak YS, Varan HD, Sezenoz B, Coskun O, Uluoglu C. Generative AI vs. human expertise: a comparative analysis of case-based rational pharmacotherapy question generation. Eur J Clin Pharmacol. 2025;81(6):875–83.
18. Shah SV. Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records. JAMA Netw Open. 2024;7(8):e2425953.
19. Wang D, Ye J, Li J, Liang J, Zhang Q, Hu Q, et al. Enhancing Large Language Models for Improved Accuracy and Safety in Medical Question Answering: Comparative Study. JMIR Med Educ. 2025;11:e70190.
20. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stuber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 2024;34(5):2817–25.
21. Bas Aksu O, Aydin RF, Gokcay Canpolat A, Demir O, Sahin M, Emral R, et al. Artificial intelligence in endocrine practice: comparing ChatGPT, Gemini, and Claude for adrenal incidentaloma care. J Endocrinol Invest. 2026;49(1):69–79.
22. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med. 2025;8(1):175.
23. Almagazzachi A, Mustafa A, Eighaei Sedeh A, Vazquez Gonzalez AE, Polianovskaia A, Abood M, et al. Generative Artificial Intelligence in Patient Education: ChatGPT Takes on Hypertension Questions. Cureus. 2024;16(2):e53441.
24. Liu D, Long Y, Zuoqiu S, Liu D, Li K, Lin Y, et al. Reliability of Large Language Model Generated Clinical Reasoning in Assisted Reproductive Technology: Blinded Comparative Evaluation Study. J Med Internet Res. 2026;28:e85206.
25. Asgari E, Montana-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025;8(1):274.
26. Yu E, Chu X, Zhang W, Meng X, Yang Y, Ji X, et al. Large Language Models in Medicine: Applications, Challenges, and Future Directions. Int J Med Sci. 2025;22(11):2792–801.
27. Qiu P, Wu C, Liu S, Fan Y, Zhao W, Chen Z, et al. Quantifying the reasoning abilities of LLMs on clinical cases. Nat Commun. 2025;16(1):9799.