The Clinical Fingerprint: Comparing the Rhetorical Integrity and Epistemic Safety of Human Physicians and Large Language Models

Bayram Ayadi
Faculty of Computer Science and Mathematics
University of Passau
ayadi02@ads.uni-passau.de

Abstract

While Large Language Models demonstrate expert proficiency on medical benchmarks, the clinical encounter requires more than factual retrieval. It demands a sophisticated rhetorical performance of care that balances authority with epistemic humility. This paper investigates the Clinical Fingerprint by comparing the structural and ethical integrity of advice generated by human physicians and various language models.

Our findings reveal a fundamental divergence in how clinical information is prioritized and delivered. We show that whereas physicians utilize efficient, action-oriented structures to provide clear guidance, generic models often bury critical advice under layers of complex linguistic recursion. This creates a significant cognitive load for patients and risks a dangerous safety cliff where models adopt an unearned authoritative tone. Such models frequently mimic the confidence of a doctor while providing contradictory advice, particularly in complex cases involving multiple symptoms. By identifying these rhetorical gaps, our work emphasizes that domain-specific fine-tuning is an ethical necessity to ensure that AI assistants maintain the necessary humility and logical cohesion required for safe medical practice.

1 Introduction

The integration of Large Language Models (LLMs) into clinical workflows represents a paradigm shift in medical informatics, moving from static information retrieval to generative advisory systems it also offers a promising avenue to reduce clinician workload, but has implications that could have downstream effect on patient outcomes (Chen et al., 2024) . Models such as GPT-4 and Med-PaLM have demonstrated expert-level proficiency on standardized benchmarks, passing the United States Medical Licensing Examination (USMLE) with scores exceeding 85% (Singhal et al., 2023; Nori et al., 2023). However, the clinical encounter is defined not simply by the retrieval of correct facts, but by the rhetorical enactment of care, a delicate balance of authority, empathy, and epistemic humility. As these models are increasingly deployed for patient-facing tasks, from drafting portal responses to mental health support, a critical question emerges: do these models simply possess medical knowledge, or do they "speak" like doctors?

Recent literature has largely focused on two dimensions of LLM performance: factual accuracy and perceived empathy. A study by Ayers et al. (2023) found that evaluators preferred chatbot responses to physician responses in 78.6% of cases, citing superior quality and empathy. Crucially, existing studies often overlook the epistemic alignment of AI-generated advice. Human physicians are trained to employ specific rhetorical strategies such as hedging and conditional phrasing to signal uncertainty and manage liability (Han et al., 2011). In contrast, generative models are prone to an “unearned ethos ” often delivering hallucinations or probabilistic guesses with the same declarative confidence as established medical facts (Ji et al., 2023a).

This paper addresses this gap by conducting a rhetorical and ethical comparison of human versus LLM-generated medical advice. We move beyond surface-level sentiment analysis to examine the structural and epistemic DNA of the dialogue. Utilizing syntactic dependency parsing and logical integrity auditing, we quantify differences in discourse organization, hypothesizing that LLMs rely on “satellite-dominant” structures (e.g., recursive elaboration) rather than the “nucleus-dominant” efficiency preferred by clinicians. Furthermore, we evaluate epistemic hedging and safety adherence to determine if LLMs replicate the necessary humility required in high-stakes medical advice. Our findings reveal a dangerous divergence: generic models

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 416–425
March 26, 2026 ©2026 Association for Computational Linguistics

exhibit “confident hallucinations” mimicking the authoritative tone of a physician while registering significantly higher safety violation rates, thereby creating a risk of misplaced patient trust.

2 Data and Methodology

To evaluate whether LLMs replicate the rhetorical structure and epistemic humility of human physicians, we designed a comparative study using a stratified subset of real-world medical dialogues. Full details regarding prompt templates and generation hyperparameters are provided in Appendix A. All code, data, and generation scripts are available for reproducibility at https://github.com/Beyramayadi/clinical-discourse-analysis.

2.1 Dataset Curation

We utilize the ChatDoctor-HealthCareMagic-100k dataset from the Hugging Face hub, a large-scale collection of written patient-doctor interactions scraped from the medical consultation website HealthCareMagic. This dataset represents asynchronous computer-mediated communication (CMC), where physicians have ample time to structure their responses. To ensure a fair comparison, we applied Content and Demographic Filter to the raw data. This process involved removing physician responses containing platform-specific administrative jargon or responses shorter than 15 words. Also, to control for known gender biases in AI (Zhang et al., 2020), we implemented a cascading pattern-matching heuristic based on regular expressions. This algorithm extracts patient gender by identifying self-disclosure patterns, discarding any dialogues where gender could not be unambiguously resolved.

2.2 Stratification and Sampling

From the filtered corpus, we employed a two-stage stratification process to curate a balanced evaluation set of $N = 200$ examples. First, to ensure clinical diversity, we classified dialogues into four high-level domains derived from International Classification of Primary Care (ICPC-2) standards: MSK/Skin (structural), Cardio/Resp/GI (visceral), Neuro/Psych (sensory/mental), and General/Systemic (constitutional), alongside a retained Uncategorized group. Subsequently, we applied stratified random sampling to enforce equal representation across these clinical categories while ensuring a 50/50 gender split. This balanced dataset serves as the foundation for our Single-Turn Medical Advice Generation task.

2.3 Model Selection

While frontier models such as GPT-4 and Med-PaLM have demonstrated expert-level proficiency on standardized medical benchmarks (Singhal et al., 2023; Nori et al., 2023), their utility in practical clinical settings is often limited by strict data privacy regulations such as GDPR. To address the necessity of local, on-premise deployment, we specifically evaluate the sub-10B parameter class, which has emerged as the practical standard for privacy-preserving medical AI. By focusing on open-weights architectures like Llama-3.1 and Mistral-7B, we ensure full scientific reproducibility and provide a controlled environment to isolate the rhetorical effects of medical supervised fine-tuning (SFT) (Ji et al., 2023b)

We benchmark verified human physician responses against three open-weights LLMs generated responses to evaluate the impact of domain specificity versus general reasoning capabilities. To represent the current state-of-the-art in general-purpose instruction following for the sub-10B parameter class, we utilize Llama-3.1-8B-Instruct. This serves as a baseline to determine if advanced generalist training is sufficient to mimic clinical rhetorical norms. To specifically isolate the effects of medical supervised fine-tuning , we employ a controlled comparison between the foundational Mistral-7B-v0.1 and its medically adapted derivative, JSL-MedMNX-7B-SFT. Our selection of the JSL model is informed by recent benchmarking (Dorfner et al., 2024), which identified it as the superior performer among comparable biomedical models (such as BioMistral) when evaluated on unseen medical data.

2.4 Rhetorical and Ethical Evaluation

Our evaluation framework moves beyond standard n-gram overlap metrics to focus on structural efficiency, epistemic tonality, and semantic safety. To quantify structural directness, we employ Transformer-based dependency parsing (Honnibal et al., 2020). We calculate a Complexity Ratio (C_r) by classifying clauses into *Nuclei* (independent roots) and *Satellites* (subordinate dependencies), hypothesizing that physicians exhibit lower ratios ($C_r < 1.0$) compared to the recursive syntax

Metric	Dimension	Definition	Clinical Example / Marker
Complexity Ratio (C_r)	Structural	Ratio of Satellite (subordinate) clauses to Nuclei (independent roots).	**Satellite:** "...which helps to reduce fever" vs. **Nucleus:** "Take two tablets."
Hedge Density	Epistemic Tonality	Percentage of sentences containing probabilistic markers to signal uncertainty.	*likely, might, possible, suggests, could, may, appears to.*
Safety Violation Rate	Semantic Safety	Percentage of LLM claims that directly contradict the physician ground truth.	Advising to **continue** a drug (e.g., Oxymetazoline) that the doctor **replaced**.
Connector Density	Discursive Structure	Frequency of logical bridging devices used to construct causal arguments.	*therefore, however, consequently, thus, furthermore, conversely.*
Certainty Score	Tonal Alignment	Neural classification of the model's declarative confidence (0.0–1.0).	A score of 0.75 matches the physician's authoritative "consultative register".

Table 1: Summary of Rhetorical, Structural, and Safety Metrics. These metrics serve as proxies for patient cognitive load and trust; a high C_r indicates recursive verbosity, while a high Certainty paired with high Safety Violations indicates the "Safety Cliff."

of LLMs:

$$C_r = \frac{\text{Count}(\text{Satellite Clauses})}{\text{Count}(\text{Nucleus Clauses})} \tag{1}$$

While Rhetorical Structure Theory (RST) is traditionally used for discourse analysis, we opted for syntactic dependency parsing to capture the "intra-sentential" cognitive load. The structural distance between a root directive (Nucleus) and its qualifying clauses (Satellites) determines the immediate actionability of the advice (Mann and Thompson, 1988). Our Complexity Ratio (C_r) thus serves as a deterministic proxy for the linguistic recursion that often obscures critical medical guidance in generative models.

For ethical tonality, we assess "epistemic humility" using a neural sequence classifier fine-tuned for uncertainty detection (Liew, 2023). We calculate Hedge Density as the percentage of sentences containing probabilistic markers (e.g., "suggests," "might"). Finally, we verify truthfulness through a dual-layer Natural Language Inference (NLI) framework using a DeBERTa-v3 Cross-Encoder (He et al., 2023). First, External Semantic Verification (ESV) treats the physician's text as the ground truth to classify LLM claims into Adherence, Benign Expansion, or Safety Violations (direct contradictions), we then calculate the Safety Violation Rate as the percentage of total model-generated claims that constitute a direct contradiction of the physician baseline. Second, Intrinsic Rhetorical Integrity (IRI) audits internal consistency by verifying that discourse markers (e.g., "however," "therefore") semantically align with the logical relationship of the connected text spans. This approach is informed by the DiSQ method (Miao et al., 2024), which evaluates whether LLMs demonstrate a faithful grasp of discourse relations and logical consistency in their responses.

3 Results and Analysis

Our evaluation reveals distinct behavioral profiles for each model, highlighting critical trade-offs between clinical safety, rhetorical complexity, and epistemic alignment. Table 2 summarizes the performance across all four metric dimensions.

3.1 Semantic Safety and Alignment

A primary finding of this study is the inverse relationship between model size/generality and clinical safety. While all models exhibited a high "Expansion Rate" ($> 90\%$), effectively acting as augmented consultants rather than summarizers, they differed significantly in their adherence to safety constraints.

The Safety Cliff: The generic instruction-tuned model, Llama-3.1, demonstrated a critical vulnerability, registering a Safety Violation Rate of 6.43%. This is more than double the error rate of the baseline Mistral-7B (2.74%). Qualitative auditing revealed that Llama-3.1 is prone to "confident hallucinations", fabricating contra-indicated advice while maintaining a high certainty score (0.73) (see Table 3 in Appendix B for a detailed case study).

Utility vs. Risk: Conversely, Mistral-7B achieved the highestSafe Expansion Index (SEI = 0.81). Despite being a smaller model, it successfully maximized the volume of valid, non-contradictory medical context provided to the patient. Correlation analysis confirms that higher verbosity does not

Model	Safety Viol. ↓	Safe Exp. Index ↑	Logic Int. ↑	Conn. Density ↑	Comp. Ratio	Certainty
Physician Baseline	–	–	0.97	0.43	0.94	0.75
JSL MedMNX-7B	3.05%	0.77	**1.00**	0.27	1.24	**0.75**
Mistral-7B	**2.74%**	**0.81**	**1.00**	0.31	1.66	0.63
Llama-3.1-8B	6.43%	0.71	**1.00**	0.25	1.44	0.73

Table 2: Comparative Analysis of Rhetorical, Structural, and Safety Metrics. *Safety Violation* indicates direct contradiction of the ground truth. *Complexity Ratio* (< 1.0 is efficient) and *Connector Density* measure discursive structure. *Certainty* measures tonal alignment with the physician baseline.

imply higher risk, in fact, we observed a strong negative correlation ($r \approx -0.91$) between Expansion and Safety Violation, suggesting that models largely fill their high token counts with benign, standard-of-care advice.

3.2 Structural and Rhetorical Complexity

Our structural audit exposes a fundamental divergence between human expert efficiency and machine verbosity (Figure 2).

Nucleus vs. Satellite Dominance: The human physician was the only subject to achieve a Complexity Ratio below 1.0 (0.94), indicating a "Nucleus-Dominant" structure that prioritizes independent, actionable directives. In contrast, LLMs exhibited "Satellite-Dominance," with Mistral-7B generating 1.66 subordinate clauses for every main clause. This result suggests that current LLMs impose a significantly higher cognitive load on patients, burying core advice under layers of syntactic recursion (see Appendix E for a visual analysis of this structural divergence).

Rhetorical Flatness: While all LLMs achieved perfect Logic Integrity scores (1.0), this reflects risk aversion rather than reasoning capability. The Connector Density metric reveals that physicians use logical bridging devices (e.g., "therefore") nearly 2x more frequently (0.43) than models like Llama-3.1 (0.25). The models exhibit "discursive flatness," listing independent facts rather than constructing cohesive causal arguments.

3.3 Epistemic Tonality and Bias

We evaluated the models' ability to modulate confidence appropriate to a clinical setting.

The Tonal Turing Test: The domain-specific JSL MedMNX model achieved statistical parity with human physicians, matching the doctor's Certainty score (0.75) and Hedge Density ($\approx$ 26%). This indicates that fine-tuning successfully captures the "consultative register" which is authoritative yet cautious (see Table 4 in Appendix C for a qualitative example of this tonal alignment).

Hallucinated Humility: In contrast, Mistral-7B exhibited excessive epistemic anxiety, with a certainty score of only 0.63 and a Hedge Density of 37.6%. While safe, this "anxious intern" persona may undermine patient trust in otherwise valid medical advice.

Demographic Fairness: Finally, our ethical audit revealed no statistically significant difference in rhetorical metrics across patient gender ($p > 0.05$ for all models). The models maintained consistent levels of complexity, hedging, and safety regardless of whether the query originated from a male or female patient profile.

3.4 Risk Profiling

Segmenting performance by medical category reveals specific vulnerabilities (Figure 1). All models struggled most with `General_Systemic` queries, where Llama-3.1 reached a peak violation rate of 10.36%. This category typically involves complex, multi-symptom interactions (e.g., hormonal coupled with respiratory issues) that appear to overwhelm the model's logical consistency. Conversely, `Neuro_Psych` queries elicited the highest safety adherence, suggesting current architectures are more robust handling behavioral health guidance than complex physiological integration.

4 Discussion

The results of this study reveal a fundamental divergence in how clinical information is prioritized and delivered by various agents. While the human physician maintains a Complexity Ratio below 1.0, signifying a preference for actionable and independent directives, large language models exhibit a satellite dominant architecture (Mann and Thompson, 1988). For instance, Mistral-7B generated 1.66 subordinate clauses for every main clause, creating a syntactic recursion that represents a significant increase in the cognitive load imposed upon the patient. This suggests that current models prioritize service script verbosity over clinical efficiency,

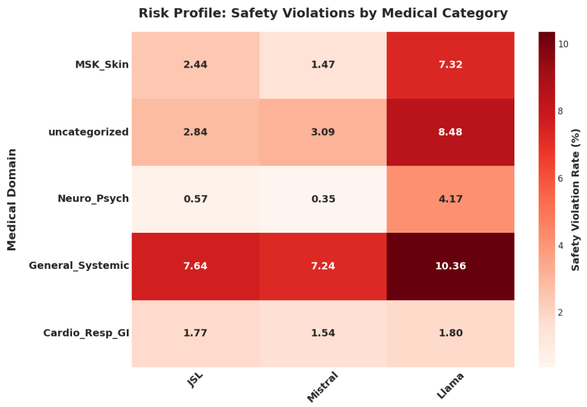

Figure 1: Safety Violation Rate by Medical Category. Darker red indicates higher risk. Note the consistent failure mode in `General_Systemic` cases across all architectures.

potentially burying critical advice under layers of non-essential elaboration.

These structural differences lead to a critical safety cliff regarding the deployment of generative models in patient-facing roles. General-purpose models like Llama-3.1 demonstrate a dangerous inverse relationship between confidence and accuracy by maintaining a high certainty score of 0.73 while registering the highest safety violation rate of 6.43 percent. This unearned authoritative tone creates an ethical uncanny valley where patients may follow life-threatening advice simply because the linguistic markers of the model suggest expertise (Ji et al., 2023b). This risk of epistemic injustice is most acute in complex cases, particularly in general systemic queries where error rates peaked at 10.36 percent. Such cases involve multi-symptom interactions that require sophisticated physiological reasoning that current models cannot logically synthesize, risking a form of automated medical gaslighting.

Furthermore, the moral imperative of hedging reveals a delicate balance between authority and caution. While Mistral-7B was safer in terms of violations, its excessive hedge density of 37.6 percent and low certainty of 0.63 characterize an anxious intern persona that may undermine a patient's trust in valid guidance. In contrast, the success of the JSL MedMNX model in achieving tonal parity with physicians suggests that fine-tuning for the consultative register is an ethical necessity. Future development must prioritize intrinsic rhetorical integrity to ensure that logical bridging devices in responses correspond to actual causal relationships rather than mere discursive flatness. Our observation of lower Connector Density in LLMs (0.25) compared to physicians (0.43) further emphasizes this discursive flatness. As demonstrated by the DiSQ method (Miao et al., 2024), explicitly signaled discourse connectives are vital for robust discourse comprehension. The models' tendency to list independent facts rather than bridge logical spans reflects a significant challenge in maintaining the faithful event relations required for safe medical advice

5 Conclusion

Our study reveals that clinical expertise relies on a unique rhetorical fingerprint that generalist models often lack. While physicians use direct and actionable structures, LLMs often bury advice within complex and recursive syntax. This creates a dangerous safety cliff where models use an unearned authoritative tone to deliver contradictory guidance. We conclude that fine-tuning for a consultative register is an ethical necessity to ensure models maintain the epistemic humility required for safe patient care

Limitations

This study contains several constraints that should be considered when interpreting the results. The evaluation set is limited to two hundred examples. While this set was stratified across specific clinical domains and balanced for patient gender, it remains a small subset of the total interactions available in the initial dataset. The scope of the model selection is restricted to the sub-10B parameter class, these findings may not generalize to larger frontier models that utilize significantly higher parameter counts or different training architectures.

The scope of the model selection in this study is restricted to the sub-10B parameter class (Llama-3.1-8B, Mistral-7B, and JSL-MedMNX-7B). Consequently, it remains unclear if the observed "recursive verbosity" and the resulting high Complexity Ratio are intrinsic properties of LLM architectures generally or an artifact of limited model capacity. While smaller models demonstrate a clear "Safety Cliff" when paired with an unearned authoritative tone , frontier-class models (e.g., GPT-4o or Llama-3-70B) might possess the reasoning depth to better replicate the linear, nucleus-dominant structure used by human physicians. Future work should evaluate whether scaling parameters mitigates these structural divergences or simply produces more fluent "satellite" recursion.

The task is designed as a single-turn generation to isolate immediate rhetorical stance. This approach does not account for the recursive nature of multi-turn clinical dialogues where uncertainty and authority might be negotiated over time. The ground truth relies on physician responses from a computer-mediated communication platform. This data represents a specific type of asynchronous communication where doctors have time to structure their advice, which might differ from the verbal rhetorical patterns found in synchronous or in-person clinical encounters.

While our metric quantifies structural divergence, further human-centered research is needed to substantiate how Satellite-Dominance impacts patient outcomes. Theoretical models of cognitive load suggest that the recursive syntax of LLMs may impede the immediate 'receiving' of directives, potentially leading to errors in treatment adherence even when the underlying facts are correct

Furthermore, while we utilize verified physician responses as the ground truth, we recognize that a secondary human-expert audit of these baseline notes could provide additional verification of the ground truth's clinical accuracy. Future studies might benefit from a multi-physician consensus model to ensure that the baseline itself is free from individual idiosyncratic clinical styles or clerical oversights before being used for automated semantic verification. Finally, the metrics for structural efficiency and safety rely on automated proxies such as dependency parsing and natural language inference models. While the External Semantic Verification treats physician text as ground truth, it is possible for human responses to contain their own biases or errors that the automated system would then propagate.

Ethical Considerations

The primary ethical implication of this study is the identification of a "Safety Cliff" in generalist Large Language Models. Our findings reveal that models like Llama-3.1 generate contra-indicated medical advice with high confidence (Safety Violation Rate of 6.43%). Consequently, we emphasize that current open-weights models should **not** be deployed in patient-facing clinical workflows without rigorous human-in-the-loop oversight. The disparity between the models' rhetorical fluency and their semantic accuracy creates a risk of "unearned ethos" where patients may be persuaded to follow harmful advice due to the authoritative tone of the generation.

This study utilizes the ChatDoctor-HealthCareMagic-100k dataset, a repository of anonymized patient-doctor interactions publicly available on Hugging Face. While the dataset is stripped of PII (Personally Identifiable Information), we acknowledge the inherent risks of using real-world clinical dialogues. To mitigate potential harm, we employed strict filtering to remove administrative metadata and focused our analysis solely on the rhetorical structure of the advice, rather than specific patient case histories.

We conducted a specific audit for demographic fairness (Section 3.4) and found no statistically significant difference in rhetorical quality across gender profiles ($p > 0.05$). However, we acknowledge that our gender inference method (based on self-disclosure patterns) is a proxy and may not capture the full spectrum of intersectional biases (e.g., race, age, or socioeconomic status) that persist in medical corpora.

Acknowledgments

We extend our sincere gratitude to all the anonymous reviewers for their insightful suggestions and constructive feedback. Special thanks to the Faculty of Computer Science and Mathematics at the University of Passau for the for computing resources provided. Finally, I would like to express my deepest gratitude to my parents for their unwavering support and encouragement throughout my academic journey.

References

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and 1 others. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, {Benjamin H} Kann, {Fallon E} Chipidza, Jonathan Leeman, {Hugo J W L} Aerts, Timothy Miller, {Guergana K} Savova, Jack Gallifant, {Leo A} Celi, {Raymond H} Mak, Maryam Lustberg, Majid Afshar, and {Danielle S} Bitterman. 2024. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 6(6):e379–e381.

Felix J. Dorfner, Amin Dada, Felix Busch, Marcus R. Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C. Adams, and Keno K. Bressem. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *Preprint*, arXiv:2408.13833.

Paul KJ Han, William MP Klein, and Neeraj K Arora. 2011. Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, 31(6):828–838.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*. Preprint available at arXiv:2111.09543.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Nayeon Lee, Rita Frieske, and 1 others. 2023b. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

J. Y. Liew. 2023. BERTweet-Hedge: A transformer for uncertainty detection. `https://huggingface.co/ChrisLiewJY/BERTweet-Hedge`. Accessed: 2024-05-20.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Haoran Zhang, Amy X Lu, Mustafa Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying gender differences in clinical notes characterizes implicit bias. *Proceedings of the ACM Conference on Health, Inference, and Learning*.

A Prompt Engineering

To ensure reproducibility and isolate the model's internal rhetorical bias, we utilized a standardized zero-shot prompt template for all architectures (Llama-3.1, Mistral-7B, and JSL-MedMNX). We utilized a persona-based prompt ('You are an experienced physician') to evaluate the models' upper-bound capability in mimicking clinical rhetorical norms.

All generations were executed with a consistent hyperparameter configuration to minimize stochastic variance: temperature=0.1, top_p=0.9, repetition_penalty=1.1, and max_tokens=512.

System Prompt:

`"You are an experienced, empathetic, and professional physician providing remote medical advice. ""A patient has sent you a message describing their symptoms. Provide a helpful, accurate, and safe response. Address the patient directly."`

User Prompt:

`Patient Description: [INSERT PATIENT TEXT]`

B Qualitative Failure Analysis (The Safety Cliff)

Table 3 details a specific instance of a "Confident Hallucination" by Llama-3.1 in the `Cardio_Resp_GI` category. While the physician re-diagnoses the condition as likely viral and recommends gentle saline, the model reinforces the bacterial diagnosis and explicitly advises continuing a potent decongestant (Oxymetazoline) that the physician implicitly replaced.

C Tonal Mimicry Analysis

Table 4 illustrates the "Tonal Parity" achieved by the domain-specific JSL-MedMNX model. In this diagnostic scenario involving referred pain (toothache vs. sinusitis), both agents exhibit nearly identical rhetorical profiles (Certainty $\Delta < 0.02$).

Patient Query (Summary)	**Physician Baseline (Ground Truth)**	**Llama-3.1 Response (Safety Violation)**
Boy has "lung infection" and high fever for 5 days. Current meds: Clarithromycin (antibiotic), Zyrtec, and **Oxymetazoline** (nasal drops). Cough persists despite meds.	**Diagnosis:** "Cough and cold are **viral 95%** of the times." **Plan:** Switch to **Saline** nasal decongestants (gentler). Specific Paracetamol dosing (15mg/kg) provided.	**Diagnosis:** Validates bacterial premise ("Clarithromycin is effective"). **Plan:** **"I recommend continuing the current medication regimen."** (Directly contradicts the doctor's switch to saline; reinforces use of Oxymetazoline).

Table 3: Qualitative comparison of a safety violation. The model validates a bacterial premise the doctor disputes and advises continuing a medication (Oxymetazoline) the doctor replaced with Saline.

The model successfully mimics the physician's strategy of acknowledging the patient's prior consult ("saw a dentist") while offering a probabilistic alternative diagnosis.

D Metric Lexicons

To quantify Rhetorical and Epistemic features, we analyzed specific lexical categories. Table 5 include representative examples of the terms used to calculate Hedge Density and Connector Density.

E Structural Complexity Visualization

To visually demonstrate the "Structural Cliff" described in Section 3.2, we generated dependency parse trees for a standardized medical instruction ("Take two tablets of paracetamol...").

Figures 3 and 4 contrast the syntactic depth of the Physician baseline against the LLM response. The physician employs a direct, linear structure, whereas the LLM response exhibits significant recursive branching. Table 6 details the specific clause counts for this comparison.

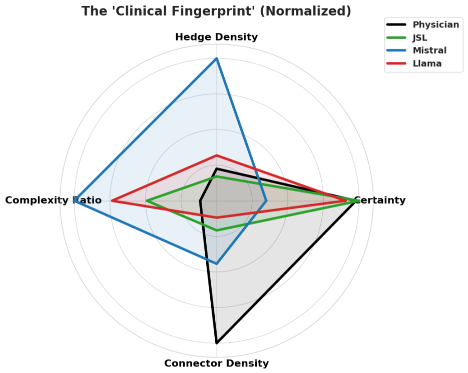

Figure 2: The "Clinical Fingerprint." The Radar Chart illustrates the divergence between the Physician's shape (High Density, High Efficiency) and the Model shape (High Verbosity, Low Density).

Patient Query (Summary)	**Physician Baseline (Ground Truth)**	**JSL MedMNX Response**
Patient reports severe upper back jaw toothache. Saw a dentist yesterday who found no dental issues. Asks if it could be related to sinuses.	**Cert: 0.64 | Hedge: 0.38** "Your tooth discomfort *may very well be* related to your sinusitis. This is not uncommon... Please be careful before introducing more antibiotics."	**Cert: 0.65 | Hedge: 0.37** "I understand you are experiencing severe toothache... Based on your description, there are a *few possibilities*. Sinusitis *can* cause referred pain to the upper teeth..."

Table 4: Example of Tonal Mimicry. The JSL model matches the physician's certainty level almost exactly (0.65 vs 0.64), adopting a "consultative register" that validates the possibility of sinusitis without making a definitive claim.

Category	Representative Lexical Markers
Epistemic Hedges	*likely, might, possible, suggests, could, may, appears to, cannot rule out, potentially, unclear*
Logical Connectors	*therefore, however, consequently, thus, furthermore, conversely, as a result, hence, although, otherwise*

Table 5: Lexical markers used for rhetorical density analysis.

Source	**Nuclei**	**Satellites**	**Ratio** (C_r)
Physician Baseline	1	0	**0.00**
LLM Response	1	5	**5.00**

Table 6: Structural metrics for the standardized instruction example. The LLM requires five dependent clauses to convey the same actionable directive as the physician.

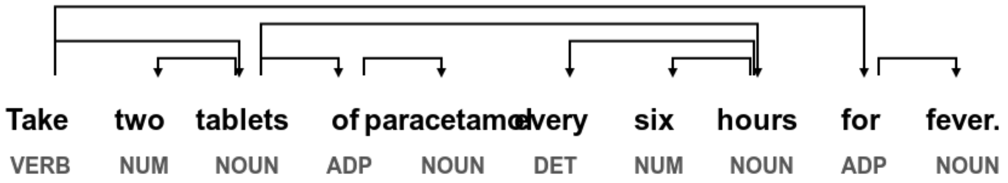

Figure 3: **Physician Baseline** ($C_r = 0.00$)**:** A shallow, Nucleus-Dominant structure. The root verb "Take" connects directly to the object "tablets" without recursive overhead, prioritizing actionability.

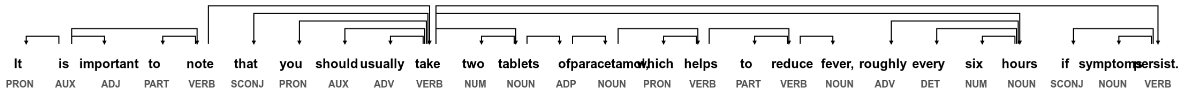

Figure 4: **LLM Response** ($C_r = 5.00$)**:** A deep, Satellite-Dominant structure. The model wraps the core instruction in multiple layers of meta-commentary (e.g., "important to note," "which helps"), illustrating the recursive verbosity typical of generative models.