Title: *Temperature-Driven Variability in Emergency Diagnostic Accuracy by a Leading Language Model*

Authors:

Philip C. Jarrett, MD[1] (philip.jarrett@utsouthwestern.edu), Jared Hill, BA[1] (jared.hill@utsouthwestern.edu), Marshall Howell, MD[1], Kristen Grabow Moore, MD[2], Joby J. Thoppil, MD PhD[1], Laura Vargas Ortiz, BS[1], Samuel T. Parnell, MD[1], D. Mark Courtney, MD[1], Samuel A. McDonald, MD[1], Deborah B. Diercks, MD[1], Andrew R. Jamieson, PhD[3], Dazhe Cao, MD[1]

Author Affiliations:

1. Department of Emergency Medicine, The University of Texas Southwestern Medical Center, Dallas, TX, USA
2. Department of Emergency Medicine, Emory University School of Medicine, Atlanta, GA, USA
3. Lyda Hill Department of Bioinformatics, The University of Texas Southwestern Medical Center, Dallas, TX, USA

Corresponding Author*:

Philip Jarrett, MD MBA
E-mail: philip.jarrett@utsouthwestern.edu
Phone: (817) 542-9646
UT Southwestern Medical Center, 5323 Harry Hines Blvd E4.300, Dallas, TX 75390-8579

Running Title: Temperature Effects on LLM Diagnosis

Keywords: Artificial Intelligence; Diagnosis, Computer-Assisted; Emergency Service, Hospital; Clinical Decision Support Systems

Financial Support: This work was unfunded.

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

Author Contributions: PJ, DC, and SP conceived of the research question and study design. PJ and DC contributed to the statistical methods and data analysis. MH and KM contributed to FoEM case selection and validation of a priori critical test selection. JH, LVO, and PJ contributed to blinded human review of critical test scoring. PJ, JH, MH, KM, JT, SM, SP, DD, MC, and DC contributed to writing the original draft, while MC, SM, LVO and AJ provided extensive editorial review. PJ, DC, SM, and AJ contributed to study supervision. PJ takes responsibility for the paper as a whole.

Acknowledgments: Many thanks are owed to Dr. Jeremy Berberian, Dr. Alison G. Marshall, and Dr. Rory Merritt for the original authorship of the FoEM cases used in this study, as well as to the numerous editors who peer-reviewed the cases for inclusion in the FoEM curriculum. We are also grateful to Katherine Riley Martin, M.S., for her expertise in the regulatory approval of this research. The Cursor Integrated Development Environment was used for code development and execution using Anthropic's Claude 3.5 Sonnet and Claude 3.7 Sonnet +/- Thinking. GPT-4o was used to support manuscript drafting after development and execution of all study methods. The authors thoroughly reviewed and revised all code and manuscript content. The lead author assumes all responsibility for the content of this manuscript.

Title: *Temperature-Driven Variability in Emergency Diagnostic Accuracy by a Leading Language Model*

ABSTRACT

Objectives

Large language models (LLMs) are being deployed in healthcare to improve clinical decision support. Yet, the "temperature" parameter, which controls LLM randomness, may significantly influence diagnostic performance and is often left at the default value. Understanding this core parameter's impact on automated diagnostic tasks will be crucial to the safe application of LLMs in high-risk settings. We evaluated the impact of temperature on diagnostic accuracy and breadth through large-scale iterative sampling when assessing high-risk conditions.

Methods

We conducted a simulation-based diagnostic accuracy experiment using four emergent medical cases from a widely adopted, login-protected emergency medicine (EM) curriculum. Each case was presented to GPT-4o under two conditions (with and without physical exam). The cases (myxedema coma, ascending cholangitis, carbon monoxide poisoning, cryptococcal meningitis) were selected for their discrete, unambiguous diagnoses. Each case was presented to GPT-4o 250 times at five distinct temperature settings (0.0 - 1.0) for each data state (with/without physical exam), resulting in 10,000 LLM diagnostic outputs. Diagnostic accuracy and breadth were reported across all cases and temperatures.

Results

With inclusion of physical exam, GPT-4o achieved perfect diagnostic accuracy at temperature 0.0 for all cases. As temperature increased, overall diagnostic accuracy decreased from 100% to 89.4% (95% Confidence Interval [CI]: 87.3-91.2%). For ascending cholangitis, diagnostic accuracy fell to 70.4% at max temperature (95% CI: 64.5-75.7%). Across all cases, unique diagnoses rose from 18 (temp 0.0) to 105 (temp 1.0), representing a 583% increase in breadth as temperature increased. Physical exam was critical to diagnostic performance in some cases while non-contributory to others.

Conclusions

Temperature impacts LLM diagnostic accuracy and breadth. Lower temperatures yielded accurate diagnoses, while higher temperatures increased diagnostic breadth at the cost of occasional misdiagnosis. Appropriate tuning of temperature is critical for reliable application of LLMs; these findings are relevant for LLM-based diagnostic decision support tools, emphasizing the need for transparency in temperature reporting.

Keywords: Artificial Intelligence; Diagnosis, Computer-Assisted; Emergency Service, Hospital; Clinical Decision Support Systems; Temperature Parameter; Model Randomness

INTRODUCTION

Large language models (LLMs) such as OpenAI's GPT-4o have demonstrated significant capabilities in medical reasoning, with some studies suggesting diagnostic accuracy comparable to, or even exceeding, that of physicians[1-6]. These models are being applied to clinical decision support, particularly in high-stakes environments like emergency medicine[7-12]. LLMs are also integral to the ambient dictation solutions in early deployment that formulate clinical notes on behalf of physicians[13]. However, LLM performance is non-deterministic; it can vary with repetitions of the same task and can be influenced by various parameters, one of the most critical yet often underdiscussed being "temperature"[14-18].

Temperature, in the context of LLMs, is a parameter that controls the randomness of a model's output[15-18]. When generating text, an LLM assigns probabilities to potential next words, i.e. tokens. A low temperature setting (e.g., 0.0 or 0.2) makes the model more deterministic: it will consistently choose the highest probability words, leading to more focused, predictable, and often repetitive outputs. Conversely, a high temperature (e.g., 0.8 or 1.0) increases randomness, allowing the model to select less likely words. This can result in more diverse, creative, or novel outputs, but also increases the risk of deviation from the most accurate response to clinical questions or tasks.

Temperature may directly impact the reliability and variability of LLM-generated medical advice[17,18]. If an LLM suggests differential diagnoses, such as in the case of clinical notes drafted by ambient dictation services, a low temperature might consistently produce the most common or textbook diagnoses, potentially missing rarer conditions. A high temperature might suggest a wider range of possibilities, including less common ones, but could also introduce "noise" or medicolegal risk in the form of less relevant, incorrect or unevaluated

diagnoses. The choice of temperature setting, therefore, represents a trade-off between precision and recall, or between consistency and exploration.

Despite its importance, temperature settings are underreported in LLM clinical research, limiting the reproducibility of findings[1-3,16-19]. Additionally, much of the literature on LLM diagnostic performance has focused on challenging models with rare and obscure conditions, which may misrepresent daily practice[1-3]. Furthermore, the default temperature settings in widely used consumer-facing chatbots can be variable or opaque to the user, eroding trust in outputs intended for clinical use[14]. There is a need to systematically evaluate how temperature affects LLM performance on medical tasks, particularly in scenarios that mimic real-world data limitations, such as an early patient encounter in the emergency department where only initial history of present illness, vitals, and perhaps a physical exam, are available.

The primary objective of this study was to determine the impact of the temperature parameter on GPT-4o's diagnostic accuracy when evaluating high-risk emergency medicine cases based on early-encounter clinical information. Secondary objectives included assessing the effect of temperature on the breadth of differential diagnoses generated and understanding the consistency of model outputs across a range of temperatures.

METHODS

Study Design and Case Selection

We conducted a simulation-based diagnostic accuracy experiment using GPT-4o. *Figure 1* provides a graphical representation of the study procedures. Four challenging, high-risk emergency medicine cases were adapted from the Foundations of Emergency Medicine (FoEM) curriculum, a widely adopted training resource protected by login, minimizing the likelihood of its inclusion in GPT-4o's training data[20-21]. The selection of these specific cases was guided by several criteria, as determined by editors of the FoEM curriculum:

1. Each case must lead to a single, discrete, and unambiguous final diagnosis.
2. The clinical information provided must represent a cross-sectional snapshot of an early patient encounter (i.e., chief complaint, vitals, brief history of present illness (HPI), past medical/surgical history (PMH/PSH), and physical examination).

3. The cases must reflect a set of diagnoses that span a range of organ systems.

The diagnoses associated with these cases, as determined by the FoEM authors during historical curriculum development, served as the gold standard for assessing diagnostic accuracy. The selected cases were:

- Myxedema Coma (Endocrinology)
- Ascending Cholangitis (Infectious Disease/Gastroenterology)
- Carbon Monoxide Poisoning (Toxicology)
- Cryptococcal Meningitis (Neurology/Infectious Disease)

The information curated from FoEM for each case is provided in *Table 1*. No data was available to estimate case difficulty, though the FoEM editors who chose the cases intentionally targeted vignettes they perceived as challenging or subtle.

Because LLMs cannot conduct physical exams and such findings are often unavailable, incomplete or undocumented during early clinical encounters, we conducted a sensitivity analysis excluding physical exam data to assess its impact on diagnostic performance.

Language Model Selection

GPT-4o (model version `gpt-4o-2024-05-13` via the OpenAI Application Programming Interface [API]) was chosen as the primary LLM for this study due to its widespread availability and strong performance in early medical benchmark studies[14]. Future work will evaluate the role of temperature in slower but more performant reasoning models.

Iterative Sampling and Prompting

The temperature parameter typically spans a range of 0.0 (low randomness) – 2.0 (high randomness). Each of the four cases was submitted to GPT-4o 250 times at five distinct temperature settings (0.0, 0.25, 0.50, 0.75, and 1.0), both with and without the inclusion of physical exam findings. Temperatures above 1.0 were not used in this study because preliminary testing indicated impaired adherence to prompt instructions at higher values. Iterative prompting across all case-temperature pairs resulted in 1,250 outputs per case for each condition (with exam, without exam), leading to 2,500 outputs per case overall, and a total of 10,000 outputs for the entire study.

All prompts followed a structured format designed to simulate consultation with an artificially intelligent (AI) diagnostic assistant. To standardize evaluation and simulate realistic clinical prioritization, the model was restricted to producing *exactly three* differential diagnoses per case, including one leading diagnosis. This constraint served several purposes. First, limiting output length reduced the risk of inflated diagnostic accuracy that can occur when evaluating correctness within long, unranked lists of conditions. This issue has been previously noted in LLM performance research related to patterns of overtesting, where sensitivity is maximized at the cost of specificity[22]. Second, it better reflects real-world diagnostic decision-making, where clinicians often apply heuristics to focus on a prioritized list of possibilities when determining workup or disposition. Third, a three-item limit helped isolate the model's prioritization behavior, forcing it to make judgment calls about what diagnoses warrant inclusion under constraint, an important consideration for clinical deployment where LLM outputs are often interpreted without repetition or extensive human curation. This structure also enabled more meaningful assessment of diagnostic breadth across iterations, by observing variability in top-ranked differentials across multiple runs, rather than simply tallying every possible condition the model might eventually list when unconstrained.

Each prompt began with a standardized header (see *Supplement 1* for full prompt structure), which included: (1) instructions about the diagnostic task, (2) expected output formatting, and (3) clinical information about a single case (chief complaint, vital signs, history of present illness, past medical history, and—if applicable—physical exam findings). In the sensitivity analysis excluding physical exam data, this section was replaced with the phrase "unavailable at this time." Prompts were designed to be stateless and were submitted in parallel, ensuring that the model had no memory of prior case iterations. Additionally, internet search and external reference capabilities were disabled during all interactions, requiring GPT-4o to rely exclusively on its embedded knowledge to generate diagnostic responses. To promote clarity, the model was also instructed to avoid hedging language such as "probable" or "possible."

Outcome Measures

The primary outcomes for this study were specified a priori:

1. Diagnostic Accuracy: This refers to the correctness of the diagnoses proposed by GPT-4o when compared against a pre-specified, definitive diagnosis for each case, known as the "gold standard."

The gold standard diagnosis for each of the four scenarios was established by the original authors of the FoEM curriculum cases, which were selected for this study by FoEM editors. Diagnostic accuracy was assessed in two ways:

a. Leading Diagnosis: Whether the single diagnosis listed by GPT-4o as its first and most likely diagnosis matched the gold standard diagnosis for that case.
b. Differential Diagnosis: Whether the gold standard diagnosis for that case appeared anywhere within the three-item list of differential diagnoses provided by GPT-4o.

2. Diagnostic Breadth: The total number of unique diagnoses proposed by the LLM for each case, tallied at each temperature setting.

Diagnostic Accuracy Scoring

A keyword search was used to score diagnostic accuracy for each of the 10,000 outputs. For each case, specific keyword phrases representative of the correct diagnosis were predefined:

- Myxedema Coma: "myxedema"
- Ascending Cholangitis: "cholangitis"
- Carbon Monoxide Poisoning: "carbon monoxide", "co"
- Cryptococcal Meningitis: "cryptococcal meningitis", "cryptococcal meningoencephalitis", "cryptococcal encephalitis"

The listed diagnoses (leading and the full three-item differential) provided by GPT-4o were searched for these keywords (case-insensitive). If a keyword was found in the leading diagnosis, it was scored as correct for "Leading Diagnosis". If found in any of the three differential diagnoses, it was scored as correct for "Differential Diagnosis." If the leading diagnosis was correct, the differential diagnosis was always correct by extension.

To validate this keyword scoring method, a subsequent human review process was performed. Two medical students (JH, LVO), blinded to the keyword scores and temperature settings, independently reviewed 1,000 LLM outputs (250 from each case, sampled evenly across all temperatures and physical exam inclusion states). The reviewers scored the leading and differential diagnoses as correct or incorrect based on the gold standard diagnosis for each case. Disagreements between the two student reviewers were adjudicated by a

board-certified emergency medicine physician (PJ). The results of this validation were used to confirm the reliability of the algorithmic approach.

Diagnostic Breadth Analysis

Each individual LLM output was constrained to a three-item differential diagnosis, representing the model's top-priority conditions based on the presented case. However, when the same case is presented repeatedly, the model may generate different sets of diagnoses each time due to the non-deterministic nature of LLMs. This *across-iteration variability*—despite identical input—may mimic diagnostic uncertainty, implying model indecision or shear randomness exaggerated by adjustment of the temperature parameter.

To quantify this variability, we measured diagnostic breadth as the total number of *unique* diagnoses generated across all iterations of a given case at each temperature setting. To isolate the role of temperature in this analysis, only cases that included physical exam data were included (n = 5,000). Because diagnoses were often expressed with subtle differences in terminology, all outputs were standardized to group synonymous or conceptually equivalent terms (e.g., "MI" and "Myocardial Infarction," or "Occult Sepsis" and "Sepsis of Unknown Source"). This allowed us to calculate the total number of distinct, standardized diagnostic categories per case-temperature pair. In doing so, we aimed to characterize how increasing temperature influences the diversity—and potential inconsistency—of diagnostic outputs over repeated interactions.

Statistical Analysis

Diagnostic accuracy (leading and differential) was calculated as a percentage with corresponding 95% confidence intervals (CIs) using the Wilson score interval method. These metrics were calculated for each case-temperature pair. Accuracy trends and CIs were plotted across cumulative iterations to visualize performance stability. The number of unique diagnoses (diagnostic breadth) was tabulated. All analyses were performed using Python (Version 3.9), with libraries including `pandas` for data manipulation, `statsmodels` for CIs, and `matplotlib`/`seaborn` for visualizations.

Ethics and Reporting

This study was approved by the Institutional Review Board of the author's academic institution during expedited review of a parent protocol. The present study does not involve human subjects. The authors

adhered to the TRIPOD-LLM Reporting Guidelines for studies using or evaluating LLMs. During the preparation of this work, the authors used GPT-4o (OpenAI) to assist with writing and revision. After using this tool, the authors extensively reviewed and edited the content as needed and take full responsibility for the final content.

RESULTS

Diagnostic Scoring Procedure

The keyword search was applied to all 10,000 LLM outputs to score leading and differential diagnoses as correct or incorrect. To validate this approach, a random 10% sample (n=1,000) was independently reviewed by two medical students who were successfully blinded to the scores of the keyword search function. Reviewer agreement was high, with only three discrepancies between humans across all 1,000 cases (Cohen's κ = 0.99). These three cases were escalated to a board-certified emergency physician, also blinded to the keyword-based scores and human reviewer scores. In all instances, the adjudicator's decision aligned with the keyword search. Following adjudication, there was perfect concordance between the keyword search and human review (Cohen's κ = 1.0), confirming the reliability of this systematic scoring method.

Diagnostic Accuracy

When provided physical exam data, GPT-4o demonstrated high diagnostic accuracy, particularly at lower temperature settings. *Figure 2* demonstrates the diagnostic accuracy trend of GPT-4o with iterative sampling for each case-temperature pair. Across all cases, the model achieved 100% accuracy for the leading diagnosis (temp = 0.0).

As temperature increased, a discernible trend of decreasing accuracy was observed for the leading diagnosis. Overall accuracy of the leading diagnosis was 100% at temperature 0.0, 99.7% at 0.25, 98.4% at 0.50, 93.6% at 0.75, and fell to 89.4% at temperature 1.0. Accuracy of the three-item differential diagnosis remained higher but showed a similar decline: 100% at temperatures 0.0, 0.25, and 0.50, 98.0% at 0.75, and 92.6% at 1.0.

The impact of temperature on diagnostic accuracy varied by case:

- Myxedema Coma: Leading diagnosis accuracy was 100% at temperatures 0.0, 0.25, and 0.50, 99.6% at 0.75, and 98.4% at 1.0. Differential accuracy was 100% at temperatures up to 0.75, and 99.2% at 1.0.
- Ascending Cholangitis: This case showed the most sensitivity to temperature. Leading diagnosis accuracy was 100% at 0.0, 98.8% at 0.25, 94.0% at 0.50, 81.6% at 0.75, and dropped to 70.4% at 1.0. Differential accuracy followed a similar pattern: 100% (0.0), 98.8% (0.25), 94.0% (0.50), 85.2% (0.75), and 78.8% (1.0).
- Carbon Monoxide Poisoning: GPT-4o maintained 100% leading and differential diagnosis accuracy across all temperature settings for this case.
- Cryptococcal Meningitis: Leading diagnosis accuracy was 100% at 0.0 and 0.25, 99.6% at 0.50, 93.2% at 0.75, and 88.8% at 1.0. Differential accuracy was 100% (0.0, 0.25), 99.6% (0.50), 97.2% (0.75), and 92.4% (1.0).

The diagnostic performance of GPT-4o at each case-temperature pair and overall, with and without physical exam data, is tabulated in *Table 2*.

Sensitivity Analysis of Physical Exam

Excluding physical exam data led to a substantial reduction in diagnostic performance overall, with accuracy for the leading diagnosis dropping to 49.6% (95% CI: 48.2–51.0) and the three-item differential to 72.0% (95% CI: 70.7–73.2). The effect varied by case:

- Myxedema Coma was particularly sensitive in terms of the leading diagnosis, which dropped to 3.3% (95% CI: 2.4–4.4), although the correct diagnosis still appeared within the differential list in 92.2% (95% CI: 90.6–93.6) of iterations.
- Ascending Cholangitis showed complete dependence on physical exam findings—diagnostic accuracy fell to 0% for both the leading and differential diagnoses in their absence.
- Carbon Monoxide Poisoning was unaffected by the removal of physical exam data, with diagnostic accuracy remaining at 100% across all temperature settings.

- Cryptococcal Meningitis showed only minimal sensitivity to the physical exam, with leading diagnosis accuracy at 95.0% (95% CI: 93.6–96.0) and differential diagnosis accuracy at 95.8% (95% CI: 94.5–96.7).

A full visualization of diagnostic accuracy trends without physical exam data, including 95% CIs, is available in *Figure 3*.

Diagnostic Breadth Analysis

To isolate the impact of temperature on diagnostic breadth, this analysis was limited to the 5,000 cases that included physical exam data. Across all cases and temperatures, GPT-4o generated 371 unique raw diagnoses (case-insensitive). These were subsequently consolidated into 91 standardized diagnostic categories through human review. Many of the raw outputs reflected minor variations in phrasing, abbreviation, or specificity (e.g., “Tuberculous Meningitis”, “CNS Tuberculosis”, “Tuberculosis with CNS Involvement”), which were grouped under a single representative diagnosis. The final total of 91 reflects the number of unique, clinically distinct diagnoses generated across the 5,000 iterations of model output. This consolidation allowed for more accurate quantification of diagnostic breadth while minimizing inflation from superficial lexical variation.

Temperature had a significant effect on the variety of diagnoses generated by GPT-4o across iterations. Increasing temperature consistently led to a greater number of proposed diagnoses for each case.

- Myxedema Coma: Unique diagnoses increased from 3 at temperature 0.0 to 27 at temperature 1.0.
- Ascending Cholangitis: Unique diagnoses rose from 6 at temperature 0.0 to 29 at temperature 1.0.
- Carbon Monoxide Poisoning: Unique diagnoses increased from 5 at temperature 0.0 to 25 at temperature 1.0.
- Cryptococcal Meningitis: Unique diagnoses grew from 4 at temperature 0.0 to 24 at temperature 1.0.

Overall, the average number of unique diagnoses was 4.5 at temperature 0.0, increasing to an average of 26.25 at temperature 1.0. This represents a 583% increase in diagnostic breadth across iterations of the same case. Higher temperatures led to inconsistent three-item differentials across case iterations. *Figure 4* depicts a heatmap of unique diagnoses at each case-temperature pair.

As temperature increased to 1.0, the correct diagnosis often remained one of the most frequent, but its overall frequency typically decreased, and a larger set of alternative diagnoses appeared more often. For instance, in the case of ascending cholangitis at temperature 1.0, the correct diagnosis was still a common response, but other conditions like "Acute Hepatitis," "Cholecystitis," and "Hypovolemia" appeared with increasing frequency, diluting the consensus around the underlying condition. A heatmap of the standardized diagnoses in the three-item differential organized by frequency of inclusion are provided by case in *Figure 5*.

DISCUSSION

This study systematically evaluated the influence of the temperature parameter on GPT-4o's diagnostic performance using challenging emergency medicine cases and data inputs typical of an initial patient encounter. Our findings demonstrate that temperature augments both the accuracy and the variability of diagnostic outputs generated by GPT-4o.

The primary observation is the trade-off mediated by temperature: lower temperatures (e.g., 0.0) yielded deterministic, consistent, and—within our study design—highly accurate diagnostic outputs. At these settings, GPT-4o reliably identified the correct diagnosis as its leading choice. This suggests that low-temperature settings constrain the model to prioritize high-probability conditions, which may be desirable in clinical workflows where precision and reliability are paramount. The study also demonstrates that the correct diagnosis for the selected cases was usually reflected by the token(s) of highest probability for GPT-4o when constrained to structured outputs.

Conversely, higher temperatures led to greater variability in diagnostic output—manifesting as increased “diagnostic breadth.” While this broader differential could reflect desirable diagnostic exploration, it also introduced inconsistency across iterations and degraded the accuracy of the leading diagnosis. For some cases, like ascending cholangitis, the decline in leading diagnosis accuracy at higher temperatures was substantial. These findings underscore a critical consideration for clinical use: in settings where LLMs are queried only once (e.g., through consumer-facing chatbots or synchronous CDS tools), high temperature settings may introduce an unacceptable level of diagnostic uncertainty. Importantly, this finding does not indicate that GPT-4o lacked the capacity to generate the correct diagnosis. Instead, it illustrates how higher

temperature settings increase stochasticity during token sampling, causing the model to occasionally select lower-probability (and potentially incorrect) diagnostic tokens—even when the correct answer remains well-represented within the underlying probability distribution.

The clinical relevance of these findings extends to several ongoing applications of LLMs in healthcare. In ambient documentation systems that generate clinical impressions and summarize patient encounters, low temperature settings may support accurate and consistent labeling of routine conditions—but could risk missing atypical or rare diagnoses. On the other hand, in decision support tools intended to aid diagnostic reasoning or hypothesis generation, especially during early diagnostic stages, higher temperatures might be used to increase the breadth of possibilities. However, when using higher temperatures, it may be necessary to either accept greater variability in outputs to capture a broader differential diagnosis or implement repeated sampling strategies to better characterize the full diagnostic distribution represented in the model's token space.

Furthermore, temperature settings are typically hidden or inaccessible in commercial products, making it difficult for clinicians to judge the reliability or reproducibility of LLM-generated suggestions[14]. This lack of transparency has major implications for trust, medicolegal accountability, and model governance. Even subtle variability in outputs can erode clinician confidence and create friction in environments like emergency medicine, where decisions must be made quickly and accurately, often with incomplete data.

Our findings offer guidance for developers and early adopters of LLM-based tools in clinical settings. Minimizing temperature can support focused, reproducible outputs in documentation and clinical decision support contexts. Alternatively, when broader exploration is desired, structured iterative approaches (e.g., consensus at 64 or ensemble prompting strategies) may mitigate risk by identifying stable diagnostic signals across multiple model outputs rather than relying on a single, potentially stochastic response.

The concept of diagnostic breadth versus diagnostic precision is central to this discussion. Clinicians often seek both: a focused working diagnosis for efficiency and a broad enough differential to avoid premature closure. Our study suggests that temperature acts as a tuning dial between these priorities, but tuning this parameter without understanding its downstream impact could lead to overconfident yet incorrect answers—or conversely, overwhelming and distracting lists of marginal possibilities.

Case-specific differences also emerged. Carbon monoxide poisoning maintained 100% accuracy across all temperatures, likely due to the presence of distinctive clinical cues. In contrast, ascending cholangitis—whose symptoms overlap with many intra-abdominal processes—proved highly sensitive to temperature changes. These observations suggest that model performance under varying temperatures may not be uniform across diseases, and temperature tuning strategies may need to be context- or syndrome-specific.

Finally, our findings reinforce the importance of methodological transparency in LLM research. The temperature parameter significantly influenced diagnostic performance, yet it remains inconsistently reported in published studies. Without clarity around settings like temperature, top-p, and context length, readers cannot interpret or reproduce findings—let alone assess real-world applicability. Standardized reporting frameworks (e.g., TRIPOD-LLM) should be adopted broadly to improve interpretability and comparability across studies and tools.

While the focus of this study is on diagnostic accuracy and breadth, the initial LLM prompts also requested diagnostic test recommendations. An analysis of how temperature affects the appropriateness and variability of test suggestions is ongoing, but preliminary observations suggest that test ordering patterns mirror diagnostic output behavior: broader and more variable at high temperatures, more targeted and repetitive at lower ones. These patterns may further influence clinical trust, cost-effectiveness, and patient safety in real-world deployments.

LIMITATIONS

This study has several limitations. The use of simulated cases—though complex—does not capture the full ambiguity and heterogeneity of real-world presentations. The gold-standard diagnoses were predefined, which simplifies scoring but does not reflect the probabilistic nature of actual clinical reasoning. We limited GPT-4o to a three-item differential to assess prioritization rather than exhaustive listing. Finally, these findings apply to a single autoregressive model version (GPT-4o) and may not generalize to other LLMs or future versions.

CONCLUSIONS

The temperature setting is a powerful modulator of GPT-4o's diagnostic outputs in the context of EM case evaluation. Lower temperatures promote diagnostic accuracy and consistency, making them preferable for applications requiring high reliability, such as ambient dictation tools that auto-populate diagnostic impressions or clinical decision support systems integrated into electronic health records. In contrast, higher temperatures increase diagnostic breadth and may be useful for hypothesis generation in uncertain or complex cases, such as through research tools or exploratory diagnostic chatbots. However, this breadth comes at the cost of decreased consistency and may necessitate iterative sampling strategies to identify stable diagnostic signals—introducing additional cost and complexity.

These findings have significant implications for the safe and effective integration of LLMs into clinical workflows. Temperature tuning must be informed by the intended clinical use case: whether the goal is confident diagnostic triage, broader differential exploration, or collaborative reasoning with human clinicians. Transparent reporting of temperature settings—by both researchers and LLM vendors—is essential for ensuring reproducibility, interpretability, and trust. Ultimately, while LLMs like GPT-4o show considerable promise in medical reasoning, human oversight remains indispensable, particularly when interpreting outputs generated with higher degrees of randomness or when applied in safety-critical environments like the emergency department.

REFERENCES

1. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, Cool JA, Kanjee Z, Parsons AS, Ahuja N, Horvitz E, Yang D, Milstein A, Olson APJ, Rodman A, Chen JH. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Netw Open. 2024 Oct 1;7(10):e2440969. doi: 10.1001/jamanetworkopen.2024.40969. PMID: 39466245; PMCID: PMC11519755.
2. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial [published correction appears in Nat Med. 2025 Apr;31(4):1370. doi: 10.1038/s41591-025-03586-x.]. *Nat Med*. 2025;31(4):1233-1238. doi:10.1038/s41591-024-03456-y
3. Buckley TA, Crowe B, Abdulnour RE, Rodman A, Manrai AK. Comparison of Frontier Open-Source and Proprietary Large Language Models for Complex Diagnoses. *JAMA Health Forum*. 2025;6(3):e250040. Published 2025 Mar 7. doi:10.1001/jamahealthforum.2025.0040
4. Bridges JM. Computerized diagnostic decision support systems - a comparative performance study of Isabel Pro vs. ChatGPT4. *Diagnosis (Berl)*. 2024;11(3):250-258. Published 2024 May 7. doi:10.1515/dx-2024-0033
5. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv. Published November 27, 2023. Accessed April 15, 2025. https://arxiv.org/abs/2311.16452

6. Liu X, Liu H, Yang G, et al. A generalist medical language model for disease diagnosis assistance. *Nat Med*. 2025;31(3):932-942. doi:10.1038/s41591-024-03416-6
7. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis. *J Med Internet Res*. 2024;26:e56110. Published 2024 Jul 8. doi:10.2196/56110
8. Kaboudi N, Firouzbakht S, Shahir Eftekhar M, et al. Diagnostic Accuracy of ChatGPT for Patients' Triage; a Systematic Review and Meta-Analysis. *Arch Acad Emerg Med*. 2024;12(1):e60. Published 2024 Jul 30. doi:10.22037/aaem.v12i1.2384
9. Kunze KN, Varady NH, Mazzucco M, et al. The Large Language Model ChatGPT-4 Exhibits Excellent Triage Capabilities and Diagnostic Performance for Patients Presenting With Various Causes of Knee Pain. *Arthroscopy*. Published online June 24, 2024. doi:10.1016/j.arthro.2024.06.021
10. Marcin T, Hautz SC, Singh H, et al. Effects of a computerised diagnostic decision support tool on diagnostic quality in emergency departments: study protocol of the DDx-BRO multicentre cluster randomised cross-over trial. *BMJ Open*. 2023;13(3):e072649. Published 2023 Mar 29. doi:10.1136/bmjopen-2023-072649
11. Hautz WE, Marcin T, Hautz SC, et al. Diagnoses supported by a computerised diagnostic decision support system versus conventional diagnoses in emergency patients (DDX-BRO): a multicentre, multiple-period, double-blind, cluster-randomised, crossover superiority trial. *Lancet Digit Health*. 2025;7(2):e136-e144. doi:10.1016/S2589-7500(24)00250-4
12. Hinson JS, Levin SR, Steinhart BD, et al. Enhancing Emergency Department Triage Equity With Artificial Intelligence: Outcomes From a Multisite Implementation. *Ann Emerg Med*. 2025;85(3):288-290. doi:10.1016/j.annemergmed.2024.10.014
13. Haberle T, Cleveland C, Snow GL, et al. The impact of nuance DAX ambient listening AI documentation: a cohort study. *J Am Med Inform Assoc*. 2024;31(4):975-979. doi:10.1093/jamia/ocae022
14. OpenAI. GPT-4o System Card. Published August 8, 2024. Accessed April 17, 2025. https://cdn.openai.com/gpt-4o-system-card.pdf
15. Suh CH, Yi J, Shim WH, Heo H. Insufficient Transparency in Stochasticity Reporting in Large Language Model Studies for Medical Applications in Leading Medical Journals. Korean J Radiol. 2024 Nov;25(11):1029-1031. doi: 10.3348/kjr.2024.0788. PMID: 39473094; PMCID: PMC11524687.
16. Davis J, Van Bulck L, Durieux BN, Lindvall C. The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research. *JMIR Hum Factors*. 2024;11:e53559. Published 2024 Mar 8. doi:10.2196/53559
17. Windisch P, Dennstädt F, Koechli C, et al. The Impact of Temperature on Extracting Information From Clinical Trial Publications Using Large Language Models. *Cureus*. 2024;16(12):e75748. Published 2024 Dec 15. doi:10.7759/cureus.75748
18. Park SH, Kim H, Yoon SH. Commendations and Concerns on the Analysis of Stochasticity in Large Language Models. *Radiology*. 2024;313(3):e241532. doi:10.1148/radiol.241532
19. Medical large language model for diagnostic reasoning across specialties. *Nat Med*. 2025;31(3):743-744. doi:10.1038/s41591-025-03520-1
20. Grabow Moore K, Ketterer A, Wheaton N, et al. Development, Implementation, and Evaluation of an Open Access, Level-Specific, Core Content Curriculum for Emergency Medicine Residents. *J Grad Med Educ*. 2021;13(5):699-710. doi:10.4300/JGME-D-21-00067.1
21. Jordan J, Wheaton N, Hartman ND, et al. Foundations of Emergency Medicine: Impact of a Standardized, Open-access, Core Content Curriculum on In-Training Exam Scores. *West J Emerg Med*. 2024;25(2):209-212. doi:10.5811/westjem.18387
22. Williams CYK, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nat Commun*. 2024;15(1):8236. Published 2024 Oct 8. doi:10.1038/s41467-024-52415-1

Table 1. Details of the four cases selected from the Foundations of Emergency Medicine curriculum to evaluate GPT-4o diagnostic accuracy. HPI: history of present illness. PMH: past medical history. PSH: past surgical history. HR: heart rate. BP: blood pressure. RR: respiratory rate. Sat: oxygen saturation. T: body temperature. HEENT: head, eyes, ears, nose, throat. ROS: review of systems.

Diagnosis	Complaint/ Vitals	HPI	PMH/PSH	Physical Exam
Myxedema Coma	Altered mental status - HR 58 BP 90/66 RR 9 Sat 96 T 35.5	A 62-year-old female who lives alone was found lying in bed by her neighbor after she hadn't been seen for the past 4 days. Neighbor reports that he is unsure of her medical history but knows that she sees multiple doctors. Further history and ROS is limited by the patient's altered mental status.	PMH: unknown PSH: unknown	General: wakes to voice but immediately falls back asleep; HEENT: normal; Chest: nontender; Lungs: normal; Heart: bradycardic, no murmurs; Abdomen: normal; Rectal: normal; Extremities: 2+ non-pitting edema of the legs; Back: normal; Neuro: oriented to self only, follows simple commands, falls asleep easily, otherwise non-focal; Skin: dry skin.
Ascending Cholangitis	Altered mental status - HR 110 BP 100/58 RR 20 Sat 93 T 35.8	This is an 81-year-old female who presents with fatigue, malaise, and vomiting overnight. The nurse at the patient's nursing facility noticed that she was more confused than usual and unable to get out of bed. She had an additional episode of non-bloody, nonbilious emesis right before EMS arrived. At baseline, she is a pleasant, ambulatory woman who is typically oriented to self but not time or place. ROS is otherwise negative.	PMH: hypertension PSH: hysterectomy	General: cachectic, fatigued; HEENT: dry mucous membranes, mild scleral icterus; Neck: normal; Heart: tachycardic, no murmurs, 1+ distal pulses; Lungs: normal breath sounds bilaterally; Abdomen: soft, mildly distended, patient grimaces with palpation of the right upper quadrant; Rectal: normal; Extremities: normal; Back: normal; Neuro: mumbling incoherently, oriented x 0, eyes open, moving all extremities and localizing to pain; Skin: jaundice, cool skin, 3 cm pressure ulcer (stage 1) to sacrum; Lymph: normal.
Carbon Monoxide Poisoning	Headache - HR 110 BP 90/50 RR 24 Sat 94 T 37.0	A 45-year-old male presents with headache, dizziness, and nausea. He reports he has been wearing extra layers to stay warm this winter. He reports that the headache is mild and diffuse, throbbing with slight nausea. He denies any neck stiffness, fever or head injuries. ROS is otherwise negative.	PMH: none PSH: none	General: awake and alert, malodorous, no acute distress; HEENT: normal; Neck: normal; Chest: nontender; Heart: tachycardic, otherwise normal; Lungs: tachypneic, otherwise normal; Abdomen: normal; Extremities: normal; Back: normal; Neuro: slightly ataxic gait, otherwise normal; Skin: normal.
Cryptococcal Meningitis	Headache - HR 106 BP 97/52 RR 16 Sat 98 T 37.7	A 27-year-old male presents with a frontal headache and intermittent chills for the past 3 days. This headache is unlike prior headaches, came on gradually and is now 8/10 in intensity. He also reports myalgias and vomiting that started after the headache. He also notes chronic diarrhea for several months with associated weight loss. ROS is otherwise negative.	PMH: none PSH: none	General: thin male, mildly confused on interview, appears uncomfortable and fatigued; HEENT: dry mucous membranes, no thrush, otherwise normal; Neck: supple; Heart: tachycardic, otherwise normal; Lungs: normal; Abdomen: normal; Genital: normal; Extremities: normal; Back: normal; Neuro: awake and alert, oriented to person and place, but disoriented to year, no focal neuro deficits; Skin: warm and clammy, otherwise normal.

Table 2. Diagnostic accuracy (95% confidence intervals [CI]) of GPT-4o on each case at five pre-specified temperatures including and excluding physical exam findings.

Case	Temp	Accuracy (+exams)		Accuracy (-exams)	
		Leading	Differential	Leading	Differential
Myxedema Coma	0.0	100% (98.5-100)	100% (98.5-100)	0% (0-1.5)	100% (98.5-100)
	0.25	100% (98.5-100)	100% (98.5-100)	0% (0-1.5)	96% (92.8-97.8)
	0.5	100% (98.5-100)	100% (98.5-100)	3.6% (1.9-6.7)	93.2% (89.4-95.7)
	0.75	99.6% (97.8-99.9)	99.6% (97.8-99.9)	3.6% (1.9-6.7)	89.2% (84.7-92.5)
	1.0	98.4% (96-99.4)	99.2% (97.1-99.8)	9.2% (6.2-13.4)	82.8% (77.6-87)
Ascending Cholangitis	0.0	100% (98.5-100)	100% (98.5-100)	0% (0-1.5)	0% (0-1.5)
	0.25	98.8% (96.5-99.6)	98.8% (96.5-99.6)	0% (0-1.5)	0% (0-1.5)
	0.5	94% (90.3-96.3)	94% (90.3-96.3)	0% (0-1.5)	0% (0-1.5)
	0.75	81.6% (76.3-85.9)	85.2% (80.3-89.1)	0% (0-1.5)	0% (0-1.5)
	1.0	70.4% (64.5-75.7)	78.8% (73.3-83.4)	0% (0-1.5)	0% (0-1.5)
Carbon Monoxide Toxicity	0.0	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)
	0.25	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)
	0.5	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)
	0.75	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)
	1.0	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)
Cryptococcal Meningitis	0.0	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)
	0.25	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)	100% (98.5-100)
	0.5	99.6% (97.8-99.9)	99.6% (97.8-99.9)	97.6% (94.9-98.9)	97.6% (94.9-98.9)
	0.75	93.2% (89.4-95.7)	97.2% (94.3-98.6)	94.4% (90.8-96.6)	95.6% (92.3-97.5)
	1.0	88.8% (84.3-92.1)	92.4% (88.4-95.1)	82.8% (77.6-87)	85.6% (80.7-89.4)
Overall	**All**	**96.2% (95.7-96.7)**	**97.2% (96.7-97.7)**	**49.6% (48.2-50.9)**	**72% (70.7-73.2)**

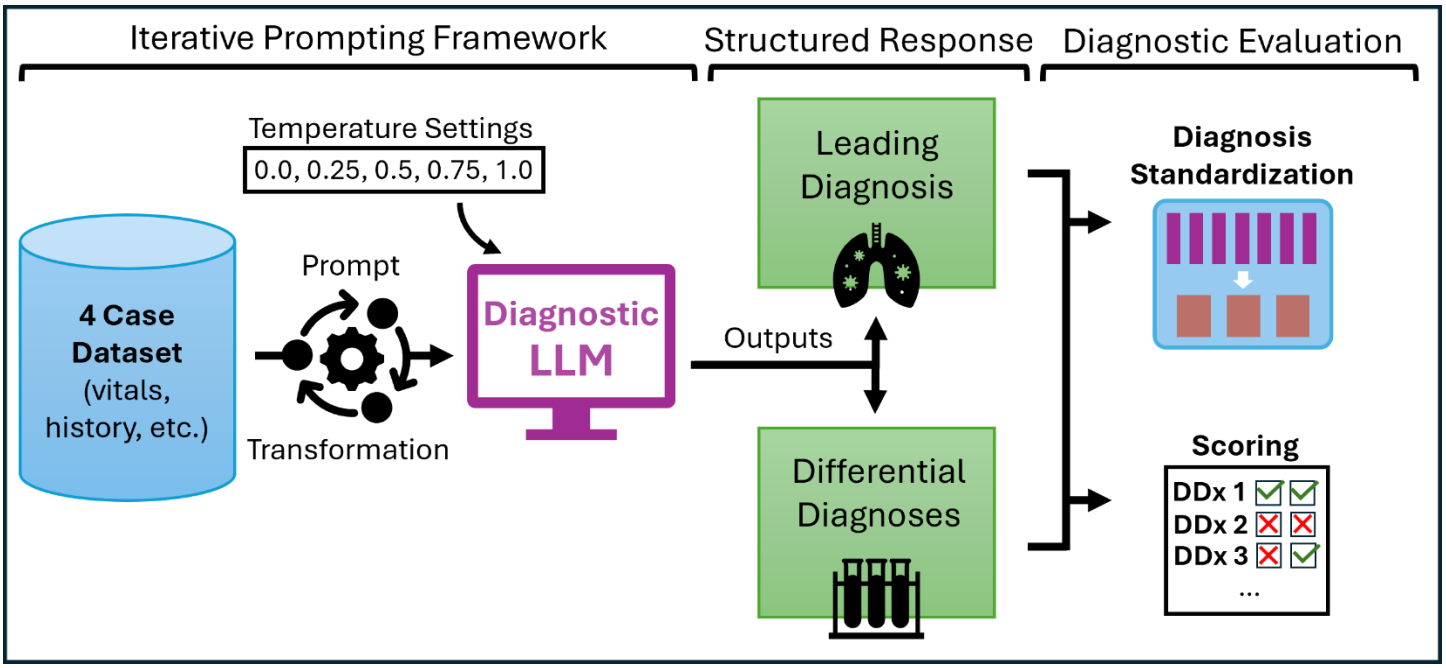

Figure 1. Study flow chart demonstrating the diagnostic evaluation pipeline using a large language model (LLM) and output scoring. Structured cases from Foundations of Emergency Medicine were adapted to represent early-encounter patient information. A standardized prompt specified the diagnostic reasoning task and output structure. The diagnostic LLM (GPT-4o) returned structured responses. Diagnostic lists (DDx) were scored for accuracy while unique diagnoses were standardized to assess diagnostic breadth.

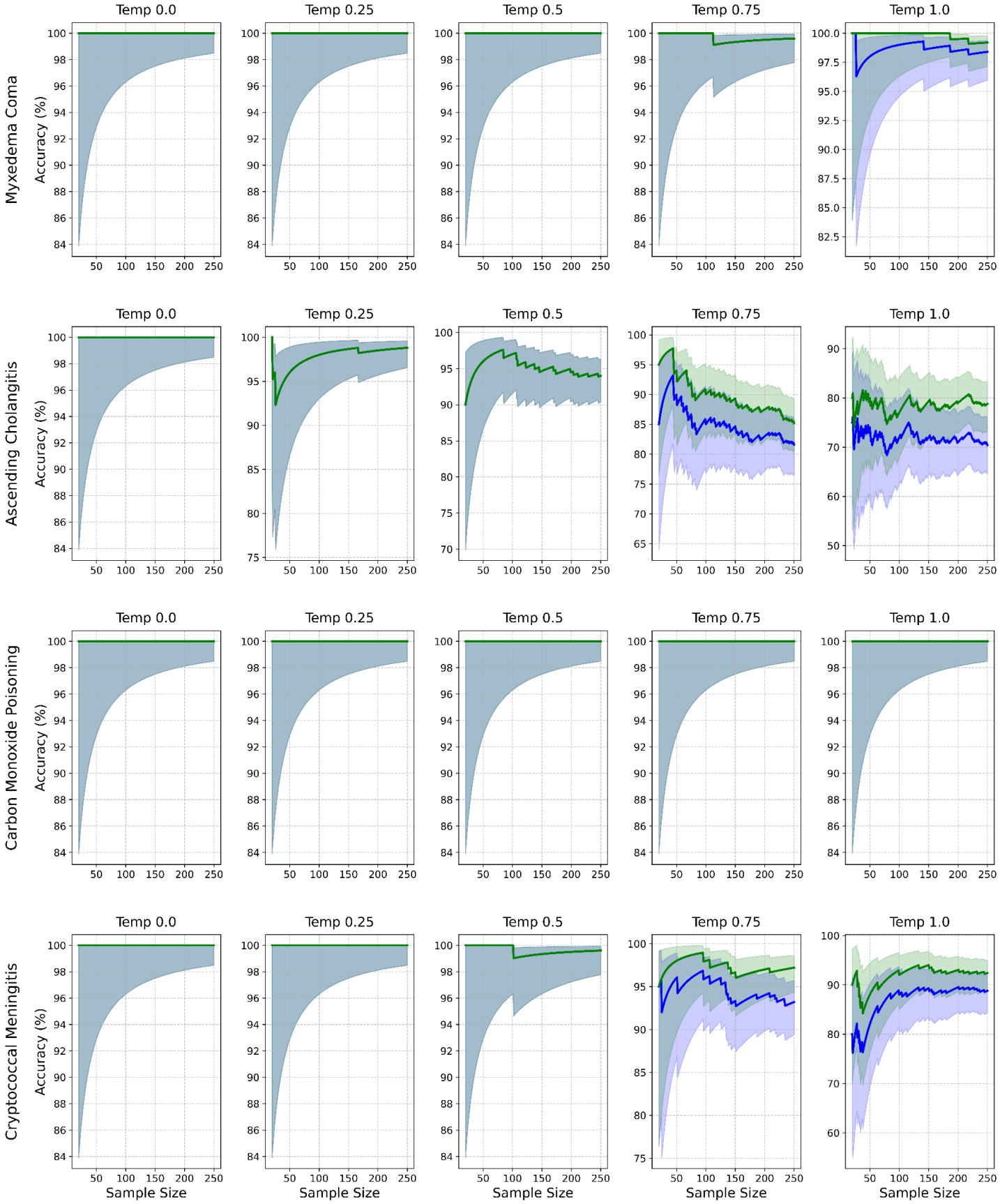

Figure 2. Diagnostic accuracy by case (rows) and temperature (columns) with exams. Lines show accuracy for leading (blue) and differential (green) diagnoses. Shaded pink and yellow areas represent the 95% confidence interval (CI) for the leading and differential diagnoses, respectively. Grey areas represent overlap in 95% CI.

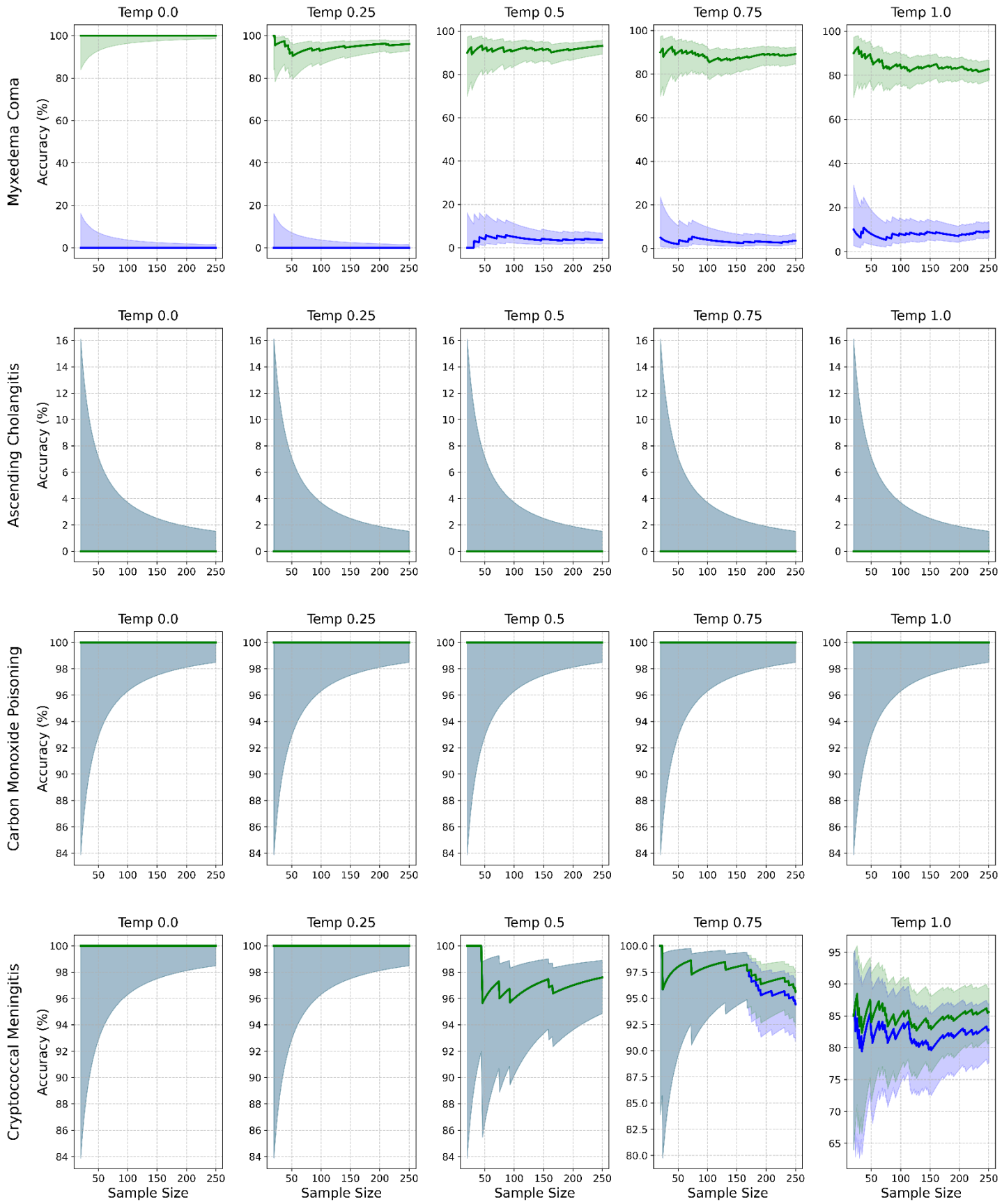

Figure 3. Diagnostic accuracy by case (rows) and temperature (columns) without exams. Lines show accuracy for leading (blue) and differential (green) diagnoses. Shaded pink and yellow areas represent the 95% confidence interval (CI) for the leading and differential diagnoses, respectively. Grey areas represent overlap in 95% CI.

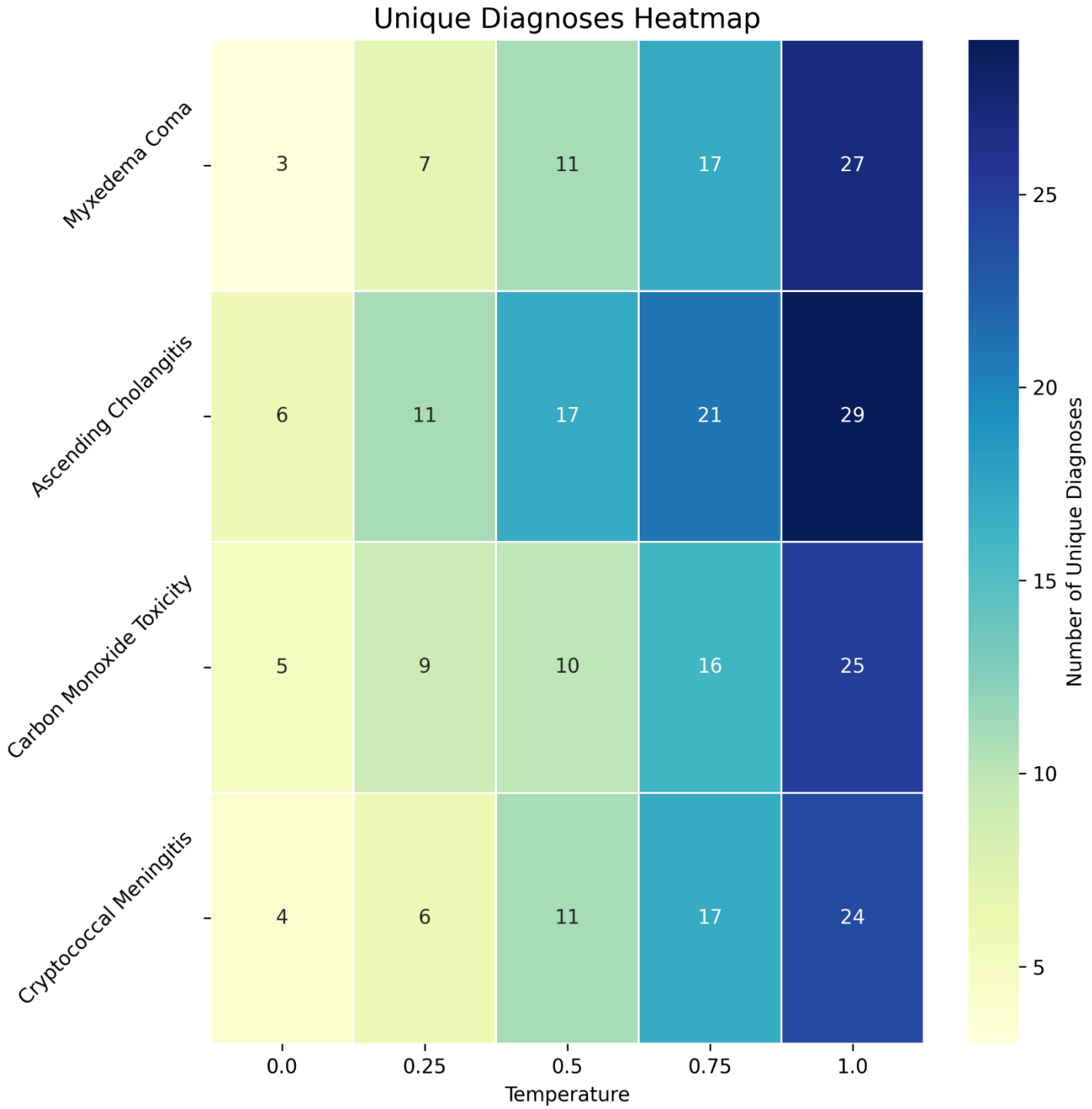

Figure 4. Heatmap depicting the number of standardized diagnoses considered in the three-item differential diagnoses across repetitive samples for each case-temperature pair. Diagnoses may overlap, such that items counted at low temperatures may also be counted at higher temperatures.

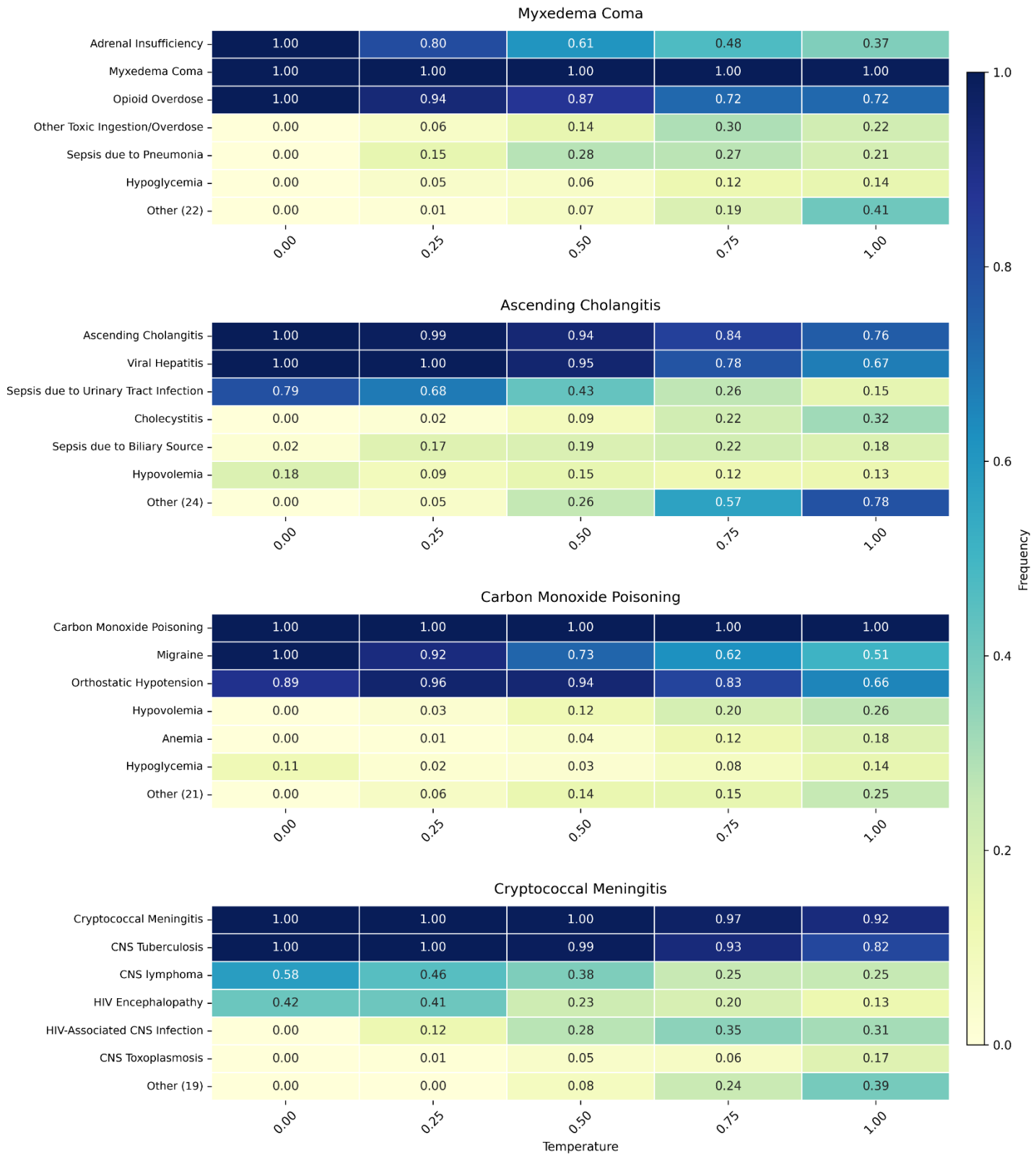

Figure 5. Heatmap depicting the top standardized diagnoses in the three-item differential with their corresponding frequency at each case-temperature pair. The “Other” diagnosis represents the cumulative frequency of all additional diagnoses considered for each case.

Supplement 1. Diagnostic assistant prompt header used to iteratively collect responses from GPT-4o.The individualized case details were adjusted for each of the four cases. When physical exam data was excluded during sensitivity analysis, the exam details were replaced by the statement: “unavailable at this time.”

You are a Large Language Model (LLM) assisting with emergency department diagnostic evaluations. You will be provided with patient data (chief complaint, vitals, history of present illness (HPI), past medical/surgical history (PMH/PSH), and physical exam) collected during triage evaluation of the patient.

Your task is to produce exactly three differential diagnoses for the most likely underlying condition, including a leading diagnosis and two additional differential diagnoses. Diagnoses should be as specific as possible. Do not use imprecise terms such as delirium, sepsis or shock - such conditions should have an underlying source. A diagnosis of "Delirium secondary to urinary tract infection" should just be "urinary tract infection", representing the underlying cause. Do not use terms such as "possible" or "probable" in the diagnoses. Each diagnosis should be a single condition without qualifiers.

Chief Complaint: [Case-specific complaint]

Vitals: [Case-specific vitals]

History of Present Illness: [Case-specific HPI]

Past Medical/Surgical History: [Case-specific PMH/PSH]

Physical Examination: [Case-specific exam findings]