ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

# Open-weight large language models for visceral leishmaniasis: Comparative diagnostic accuracy of seven locally-deployed models

**Aline Rafaela Soares da Silva[1,2,3,4†], Dino Schwingel[1,2†], Samuel Ricarte de Aquino[1,4], Adrillany da Costa Santos[1,2], Humberto Baudel Francisco[1,2], Luíza Azevedo Ferreira[1,2], John Alan Rodrigues Dantas[1,2], Márcio de Oliveira Silva[1,2], Paulo Gustavo Serafim de Carvalho[1,5], Rogério Fabiano Gonçalves[1,2,6], Fabiana Oliveira dos Santos Camatari[1,2], Maria Jacqueline Silva Ribeiro[1,7], Paula Andreatta Maduro[1,2,4,8], and Paulo Adriano Schwingel[1,2,3,8*]**

[1]AI-assisted Diagnostics Research Group, Universidade de Pernambuco, Petrolina, Pernambuco, Brazil

[2]Human Performance Research Laboratory, Universidade de Pernambuco, Petrolina, Pernambuco, Brazil

[3]Postgraduate Program in Rehabilitation and Functional Performance, Universidade de Pernambuco, Petrolina, Pernambuco, Brazil

[4]Hospital Universitário da Universidade Federal do Vale do São Francisco, Brazilian Hospital Services Company, Petrolina, Pernambuco, Brazil

[5]College of Agricultural and Environmental Sciences, Universidade Federal do Vale do São Francisco, Juazeiro, Bahia, Brazil

[6]Postgraduate Program in Public Health, Universidade de Pernambuco, Recife, Pernambuco, Brazil

[7]Health Sciences Center, Universidade Estadual do Maranhão, São Luís, Maranhão, Brazil

[8]Postgraduate Program in Health Sciences, Universidade de Pernambuco, Recife, Pernambuco, Brazil

[†]*These authors contributed equally to this work.*

***Corresponding author:**
Paulo Adriano Schwingel
(paulo.schwingel@upe.br)

**Citation:** Soares da Silva AR, Schwingel D, Ricarte de Aquino S, *et al.* Open-weight large language models for visceral leishmaniasis: Comparative diagnostic accuracy of seven locally-deployed models. *Artif Intell Health.*
doi: 10.36922/AIH026190041

**Received:** May 7, 2026

**Revised:** May 24, 2026

**Accepted:** May 28, 2026

**Published online:** June 16, 2026

**Copyright:** © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

**Publisher's Note:** AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Abstract

Large language models (LLMs) are increasingly considered as adjuncts for clinical reasoning support, yet most evaluations have focused on cloud-hosted proprietary models. Their open-weight counterparts, which can be deployed locally without transmitting patient data to external servers, remain underexplored—particularly for neglected tropical diseases (NTDs). This exploratory study evaluated the diagnostic performance of seven open-weight LLMs (Falcon 180B, Phi-3 mini 3.8B, LLaMA 2 70B, Meditron 70B, Mixtral 8x7B, DeepSeek-V2 16B, and DeepSeek-R1 70B) for visceral leishmaniasis, using eight clinical vignettes identical in content, prompt, and presentation order to those previously employed to evaluate ChatGPT/GPT-4 by our group. All models were served locally via the Ollama runtime under default inference settings. Top-five and top-one diagnostic accuracy were calculated, and a five-member specialist committee (three infectious-disease physicians and two clinical-diagnosis faculty) classified every generated hypothesis as plausible, non-plausible, or fabricated, after independent assessment and consensus adjudication. Top-five accuracy ranged from 100.0% (DeepSeek-R1 70B; 95% CI: 67.6–100.0) to 12.5% (LLaMA 2 70B and DeepSeek-V2 16B). DeepSeek-R1 70B exceeded the previously published ChatGPT/GPT-4 benchmark of 75.0%; Falcon 180B matched it. Three qualitatively distinct failure modes were identified: fabrication of nosological entities (Phi-3 mini),

generic vagueness (LLaMA 2 70B), and non-engagement with the diagnostic task (Meditron 70B). These findings demonstrate that locally-deployed open-weight LLMs can match or exceed proprietary models for VL differential diagnosis, while underscoring the need for expert-validated qualitative assessment beyond binary accuracy metrics. Institution-level evaluation should precede clinical adoption.

***Keywords:*** Artificial intelligence; Clinical decision support systems; Differential diagnosis; Large language models; Neglected tropical diseases; Open-weight models

## 1. Introduction

Visceral leishmaniasis (VL) is a severe systemic parasitic disease caused by protozoa of the *Leishmania donovani* complex and transmitted in the Americas by phlebotomine sandflies of the genus *Lutzomyia*.[1,2] The disease remains endemic in approximately one hundred countries, with substantial morbidity and mortality reaching 90% in untreated cases.[3] Updated Brazilian surveillance data confirm the persistence of VL as a public health priority in the country over the past two decades.[4] Its hallmark clinical presentation (prolonged fever, hepatosplenomegaly, and pancytopenia) overlaps with hematologic malignancies, hepatosplenic schistosomiasis, miliary tuberculosis, and other causes of fever of unknown origin, making timely diagnosis particularly challenging in endemic regions where specialized laboratory support is uneven.[1] The São Francisco Valley in northeastern Brazil exemplifies this tension: a hyperendemic area for both VL and Chagas disease, served by a network of public hospitals with variable access to specialist consultations and confirmatory testing.

Artificial intelligence (AI), and particularly large language models (LLMs) built on transformer architectures,[5] has emerged as a candidate adjunct for clinical reasoning support. By ingesting extensive medical text corpora, contemporary LLMs can generate human-like differential diagnoses and contextualized explanations,[6,7] extending the range of computational tools available to healthcare professionals.[8] Recent reviews have specifically highlighted potential applications in clinical microbiology and infectious diseases,[9,10] where pattern recognition across heterogeneous epidemiological, clinical, and laboratory data is essential.

A growing body of empirical work has tested LLM diagnostic capabilities. Earlier-generation ChatGPT models generated plausible top-five differential lists for clinical vignettes with common chief complaints;[11] ChatGPT-4 has shown agreement with physicians when judging whether a differential list contains the correct diagnosis;[12] and proprietary models have achieved competitive performance on standardized assessments such as the *New England Journal of Medicine* clinical problem-solving quiz[13] and emergency department triage benchmarks.[14] However, recent comparative work has cautioned that LLM accuracy in real clinical settings is often substantially lower than on examination-style questions, and highly sensitive to prompt construction.[15–17] In the specific context of neglected tropical diseases (NTDs), our group has previously reported that ChatGPT/GPT-4 included VL among the top-five differential diagnoses in 75% of clinical vignettes and identified it as the leading hypothesis in 50% of cases,[18] complementing related work by our group demonstrating the feasibility of LLM-based tools as supplementary aids in other clinical workflows such as nutritional prescription at hospital discharge.[19] A subsequent paired experimental study by our group has extended this line of research to real anonymized clinical cases of both Chagas disease and visceral leishmaniasis, providing a real-world complement to the present synthetic-vignette benchmark.[20]

A critical gap in this literature concerns open-weight LLMs. Most published evaluations have focused on proprietary models, with limited comparative data on open-weight alternatives that can be deployed on institutional hardware without transmitting patient data to external servers. This question is far from academic: privacy risks of cloud-hosted AI chatbots in healthcare are well documented,[21] and locally-deployed models offer potential advantages in cost, offline operation, and data sovereignty—concerns that are particularly salient for resource-constrained settings within Brazil's Unified Health System. These considerations have been systematically articulated in the broader literature on balancing control, collaboration, costs, and security in healthcare LLM deployment.[22] Lightweight runtimes such as Ollama[23] now allow general-purpose institutions to operate large models on dedicated workstations with minimal infrastructure overhead. However, the open-weight ecosystem is markedly heterogeneous, encompassing general-purpose

dense decoders (*e.g.*, LLaMA 2[24]), very large generalist models (*e.g.*, Falcon 180B[25]), sparse mixture-of-experts (MoE) designs (*e.g.*, Mixtral 8x7B,[26] DeepSeek-V2[27]), small efficient instruction-tuned models intended for on-device inference (*e.g.*, Phi-3[28]), reasoning-focused models trained via reinforcement learning (*e.g.*, DeepSeek-R1[29]), and biomedical-domain fine-tuned models (*e.g.*, Meditron-70B[30]). Whether these architectural choices translate into meaningful differences in NTD diagnostic performance is largely unknown, as is the question of whether medical fine-tuning provides a measurable advantage in this clinical context.[31]

The present study addresses this gap. Building directly on the same set of clinical vignettes used in our previous evaluation of ChatGPT/GPT-4,[18] we systematically compared the diagnostic performance of seven open-weight LLMs (Falcon 180B, Phi-3 mini, LLaMA 2 70B, Meditron-70B, Mixtral 8x7B, DeepSeek-V2 16B, and DeepSeek-R1 70B) deployed locally via Ollama on a dedicated workstation. The primary outcome was top-five diagnostic accuracy for VL; secondary outcomes included the rank position of the target diagnosis within the differential list and an independent specialist committee assessment of the biological plausibility of all generated hypotheses, addressing a limitation explicitly acknowledged in our previous work.[18] By keeping the case material, prompt structure, and presentation order identical to the earlier ChatGPT evaluation, this design allows direct, quantitative comparison between a state-of-the-art proprietary model and a diverse set of open-weight alternatives on a clinically relevant NTD benchmark.

## 2. Data and methods

### 2.1. Study design

This was an exploratory, comparative experimental study conducted in January 2025 at the Human Performance Research Laboratory (LAPEDH), Universidade de Pernambuco (UPE), Campus Petrolina, Pernambuco, Brazil. The study evaluated the diagnostic performance of seven open-weight LLMs deployed locally on dedicated institutional hardware, using clinical vignettes of VL identical to those previously employed in our group's evaluation of ChatGPT/GPT-4.[18] This design choice was deliberate: by holding case material, prompt, and presentation order constant, the present study enables direct quantitative comparison between the proprietary baseline already published[17] and a structurally diverse set of locally deployable open-weight alternatives. In contrast to that earlier evaluation, the present study extends the analysis in two specific dimensions: (i) broader model coverage, encompassing seven open-weight LLMs spanning multiple architectural and training paradigms rather than a single proprietary model; and (ii) the addition of an independent specialist-committee assessment of the biological plausibility of every generated diagnostic hypothesis.

### 2.2. Clinical vignettes

The eight clinical vignettes used in this study were the same set previously formulated in Brazilian Portuguese by an infectious-disease specialist with extensive clinical experience in VL diagnosis (S.R.deA., third investigator), based on the typical sociodemographic and clinical profile of outpatients diagnosed with VL at the University Hospital from the Universidade Federal do Vale do São Francisco (HU-UNIVASF), a Brazilian Hospital Services Company (EBSERH) teaching hospital located in Petrolina, Pernambuco, Brazil. The full content and structure of these vignettes have been described in detail elsewhere[18] and reflect the geographic and clinical context of the São Francisco Valley region. Each vignette included identifying information (age, biological sex, ethnicity, place of residence), social history (alcohol and tobacco use, occupation), history of present illness, past medical history with current medications, and a structured physical examination. No modifications were introduced to the clinical vignettes themselves for the present study, ensuring full comparability of the case material with the previous ChatGPT evaluation.[18]

### 2.3. Selected open-weight large language models

Seven open-weight LLMs were selected to span a deliberately heterogeneous set of architectural and training paradigms, reflecting the diversity of options currently available for local deployment in healthcare institutions. The models, identified by name, parameter count, architectural family, and Ollama image identifier in use at the time of the experiment, were: (i) Falcon 180B (dense decoder, image ID e2bc879d7cee),[25] included as a representative very large generalist model; (ii) Phi-3 mini 3.8B (instruction-tuned dense, 4-bit quantized image, ID 157684aa81ad),[28] representing efficient on-device-class models; (iii) LLaMA 2 70B (dense decoder, ID e7f6c06ffef4),[24] representing a previous-generation general-purpose foundation model; (iv) Meditron 70B (biomedical-domain fine-tuned, ID 315d466c452f),[30] included to assess whether medical-domain pretraining conferred advantages for NTD diagnostics; (v) Mixtral 8x7B (sparse MoE, ID 7708c059a8bb),[26] representing leading open-weight MoE designs; (vi) DeepSeek-V2 16B (sparse MoE, ID 7c8c332f2df7),[27] representing efficient MoE alternatives; and (vii) DeepSeek-R1 70B (reasoning-tuned dense, ID 0c1615a8ca32),[29] representing recent

reinforcement-learning-trained reasoning models. The decision to evaluate models with distinct architectural lineages (rather than only the latest available release) was motivated by the fact that healthcare institutions deploying LLMs locally typically operate with versions selected based on hardware availability, licensing, and deployment maturity at a given moment, rather than the absolute frontier.

### 2.4. Local deployment via Ollama

All seven models were served locally on a dedicated scientific workstation at LAPEDH using the Ollama runtime,[23] which provides a unified interface for downloading, quantizing, and serving open-weight LLMs without external API calls or transmission of input data to third-party servers. Default inference parameters as distributed by Ollama were used throughout (temperature 0.7–0.8 depending on model, no system prompt customization), without task-specific fine-tuning. This configuration was selected to reflect a "minimum-friction" deployment scenario consistent with how most clinical units would adopt such tools, where bespoke parameter tuning is impractical. This default-parameter, single-prompt configuration was a deliberate methodological choice rather than a technical limitation: it represents the operational scenario most likely to be encountered by institutions adopting locally-deployed open-weight LLMs in resource-constrained settings, where dedicated prompt-engineering expertise is typically unavailable.[22]

### 2.5. Prompt and case administration

A single prompt structure was used uniformly across all models and cases, identical in wording to the prompt employed in our previous ChatGPT evaluation:[18] "*Por favor, forneça-me os cinco diagnósticos mais prováveis para os seguintes sintomas:*" (English: "Please provide me with the five most likely diagnoses for the following symptoms:"), followed immediately by the full text of the clinical vignette in Brazilian Portuguese. Prompt and clinical case were delivered as a single concatenated message in each interaction. The order in which cases were presented to each model was the same randomized sequence (cases 02, 08, 04, 01, 06, 05, 03, and 07) generated by computer-generated random-number table (BioEstat 5.3, Instituto de Desenvolvimento Sustentável Mamirauá, Tefé, AM, Brazil) used in the previous study.[18] For each combination of model and case, a fresh chat session was opened, the model's conversational state was reset, and only the prompt-plus-case message was submitted, with no follow-up turns or clarifying exchanges. After receipt of the model's response, the session was terminated before the next case was administered. This procedure ensured full independence between trials and eliminated the possibility of carryover effects from earlier interactions. Responses were recorded verbatim, without editing or interpretation.

### 2.6. Outcome definitions

The primary outcome was top-five diagnostic accuracy, defined as the proportion of cases (out of eight) in which VL appeared anywhere within the model's list of five differential hypotheses. Secondary outcomes included: (i) the rank position of VL within the differential list, recorded on an ordinal scale from 1 (first/most-likely hypothesis) to 5 (fifth/least-likely hypothesis); (ii) top-one accuracy, defined as the proportion of cases in which VL was listed as the most likely diagnosis; (iii) two response-quality categories capturing departures from the requested output format—incomplete responses, in which the model produced fewer than five hypotheses, and non-engagement responses, in which the model returned generic text without any identifiable diagnostic hypothesis; and (iv) the biological plausibility of all generated hypotheses, as assessed by the specialist committee. For the primary analysis, incomplete and non-engagement responses were treated as failures to identify VL within the top-five list, preserving comparability with the binary scoring scheme used in the previous study.[18]

### 2.7. Specialist committee assessment

The specialist-committee plausibility assessment described in this section is a methodological component newly introduced in the present study; it was not part of our previous ChatGPT/GPT-4 evaluation and was added specifically to address the limitation, explicitly acknowledged in that earlier work, that binary scoring of differential-diagnosis lists does not capture qualitative variation in clinical-reasoning quality. A committee of five physicians (three infectious-disease specialists with documented clinical experience in NTDs and two clinical-diagnosis faculty members from Brazilian medical schools, both cardiologists with doctoral training and active teaching responsibilities in clinical diagnosis) assessed the biological plausibility of every diagnostic hypothesis generated by each of the seven models for each of the eight cases. The committee composition was chosen to combine domain expertise in the target disease with broader expertise in differential diagnostic reasoning, the latter contributed by faculty members responsible for clinical-diagnosis teaching at undergraduate medical level. The assessment followed a two-stage procedure designed to combine independent judgment with structured adjudication of discordances. In the first stage, each committee member independently classified every diagnostic hypothesis without access to the classifications of the other members. In the second stage,

the committee convened in a consensus meeting to discuss any discordant classifications and reach a final adjudicated decision. Of all individual classifications produced in the first stage, unanimous agreement was reached on every hypothesis except one, which was resolved by a 3-to-2 majority vote during the consensus meeting.

For each hypothesis, the final adjudicated classification placed it into one of four categories: (a) *VL correctly identified*, when the hypothesis was visceral leishmaniasis itself; (b) *plausible differential diagnosis*, when the hypothesis represented a clinically reasonable alternative consistent with the presenting features; (c) *non-plausible differential diagnosis*, when the hypothesis bore no clinical relationship to the case (*e.g.*, conditions whose typical presentation differed substantially from that described); and (d) *fabricated/no medical basis*, when the model produced what appeared to be a non-existent or otherwise unidentifiable nosological entity (*e.g.*, neologisms or distorted disease names with no recognized medical referent). Inclusion of this classification scheme was specifically intended to address a limitation explicitly acknowledged in our previous work,[18] in which binary scoring of differential-diagnosis lists did not capture the qualitative variation in clinical reasoning quality across model outputs.

### 2.8. Statistical analysis

All data were double-entered into a structured spreadsheet (Microsoft Excel, Microsoft Corporation, Redmond, WA, USA) and verified prior to analysis. Categorical variables were summarized as absolute (*n*) and relative (%) frequencies. For each model, top-five and top-one accuracy were calculated as proportions, with 95% confidence intervals (CIs) computed using the Wilson score method, which is appropriate for small samples and avoids zero-width intervals at extreme proportions. The Wilson score method is the canonical recommendation for small-sample binomial proportions and is preferred over Wald (normal-approximation) intervals (which can extend below 0 or above 1) and over percentile-bootstrap intervals, which, for proportions of 0/n or n/n, degenerate to intervals of zero width.[32] The mean rank position of VL among the differentials, when present, was calculated for each model. Committee-classified hypotheses were summarized per model as the mean number of plausible, non-plausible, and fabricated diagnoses per case (denominator: eight cases × five slots = 40 slots per model). Quantitative comparisons against the published ChatGPT/GPT-4 reference data[18] were descriptive, without inferential testing, given the exploratory nature of the design and the small sample size ($n$ = 8 cases per model). All analyses were conducted in IBM SPSS Statistics for Windows, release 22.0 (IBM Corp., Armonk, NY, USA, 2013).

### 2.9. Ethical considerations

This study was conducted in compliance with Brazilian National Health Council Resolution 466/2012 and used only anonymized synthetic clinical vignettes formulated by an infectious-disease specialist on the basis of typical patient profiles, without patient identifiers or direct patient involvement. Formal ethical review and approval for the broader research program of which this study forms a part were granted by the Research Ethics Committee of HU-UNIVASF under approval number 6967834. No patient consent was required, as no individual patient data were processed at any stage. As local deployment via Ollama ensures that no input or output data were transmitted to external servers, this design also addresses well-documented privacy concerns associated with cloud-hosted AI chatbots in healthcare.[21]

## 3. Results

The diagnostic performance of the seven open-weight LLMs is presented in three layers, mirroring the analytical structure of our previous evaluation:[18] top-5 and top-1 diagnostic accuracy (section 3.1), the rank position of VL within the differential list across cases (section 3.2), and the specialist committee's plausibility assessment of all generated hypotheses (section 3.3). A descriptive comparison with the previously published ChatGPT/GPT-4 reference data[18] is presented in section 3.4. An overview of the principal performance metrics is provided in Figure 1.

### 3.1. Top-five and top-one diagnostic accuracy

Table 1 presents the top-five and top-one diagnostic accuracy results for VL across the seven evaluated open-weight LLMs.

DeepSeek-R1 70B demonstrated the highest overall performance, correctly including VL among the top-five differential diagnoses in 8 of 8 cases (100%; 95% CI 67.6–100.0%), although it ranked VL as the most likely diagnosis in only three cases (37.5%). Falcon 180B was the second-best performer, achieving top-five accuracy in 6 of 8 cases (75.0%; 95% CI 40.9–92.9%), but failing to identify VL as the leading hypothesis in any case (0%; 95% CI 0.0–32.4%). Phi-3 mini 3.8B and Meditron 70B both achieved top-five accuracy in 3 of 8 cases (37.5%; 95% CI 13.7–69.4%); however, Phi-3 ranked VL as the leading hypothesis in 1 case (12.5%), whereas Meditron 70B never placed VL at the top of its list. Mixtral 8x7B identified VL among the top-five hypotheses in 2 of 8 cases (25.0%; 95% CI 7.1–59.1%) and never as the leading diagnosis.

**Table 1. Top-five and top-one diagnostic accuracy for visceral leishmaniasis across seven open-weight large language models (*N* = 8 cases)**

| Large language model (LLM) | Top-5 accuracy, *n*/*N* (%; 95% CI) | Top-1 accuracy, *n*/*N* (%) | Mean position when present | Incomplete responses, *n* | Non-engagement responses, *n* |
|---|---|---|---|---|---|
| DeepSeek-R1 70B | 8/8 (100.0; 67.6–100.0) | 3/8 (37.5) | 1.75 | 0 | 0 |
| Falcon 180B | 6/8 (75.0; 40.9–92.9) | 0/8 (0.0) | 3.33 | 0 | 0 |
| Phi-3 mini 3.8B | 3/8 (37.5; 13.7–69.4) | 1/8 (12.5) | 2.00 | 0 | 0 |
| Meditron 70B | 3/8 (37.5; 13.7–69.4) | 0/8 (0.0) | 3.00 | 1 | 2 |
| Mixtral 8x7B | 2/8 (25.0; 7.1–59.1) | 0/8 (0.0) | 3.00 | 0 | 0 |
| DeepSeek-V2 16B | 1/8 (12.5; 2.2–47.1) | 0/8 (0.0) | 5.00 | 0 | 0 |
| LLaMA 2 70B | 1/8 (12.5; 2.2–47.1) | 1/8 (12.5) | 1.00 | 0 | 0 |

Notes: Top-five accuracy: proportion of cases in which VL appeared within the model's five-hypothesis differential list. Top-one accuracy: proportion in which VL was listed as the most likely diagnosis. Mean position: rank of VL within the differential, calculated only on cases in which VL was identified (1 = most likely; 5 = least likely). Incomplete responses: model produced fewer than five hypotheses. Non-engagement responses: model returned generic text without identifiable diagnostic hypotheses. CI: 95% confidence interval (Wilson score method).

DeepSeek-V2 16B and LLaMA 2 70B exhibited the lowest top-five accuracy, each correctly listing VL in only 1 of 8 cases (12.5%; 95% CI 2.2–47.1%); LLaMA 2 ranked VL first on that single occasion (12.5%), while DeepSeek-V2 placed it at the fifth position.

A notable pattern emerged regarding output format compliance. Meditron 70B failed to engage with the diagnostic task in 2 of 8 cases (Cases 02 and 03), returning generic text without identifiable diagnostic hypotheses; in one of these cases (Case 02), the model produced what appeared to be a placeholder list (*e.g.*, "Doença A" through "Doença E") rather than substantive clinical reasoning, which the committee classified as non-engagement. In one additional case (Case 08), the model produced an incomplete differential containing only one entry. All other models produced complete five-item lists in every case. Two illustrative verbatim model outputs – DeepSeek-R1 70B for Case 04 (high-quality reasoning) and Phi-3 mini 3.8B for Case 06 (fabrication failure mode) – are provided as Supplementary File.

### 3.2. Position of the target diagnosis

When VL was correctly identified, its rank position within the differential list varied considerably across models. DeepSeek-R1 70B placed VL in the first position in three cases (Cases 02, 04, and 08), the second position in four cases (Cases 01, 03, 05, and 07), and the third position in one case (Case 06), yielding a mean position of 1.75. Falcon 180B placed VL at intermediate-to-low positions in all six cases in which it was identified, ranging from the second to the fifth hypothesis, with a mean position of 3.33. Phi-3 mini 3.8B identified VL once at each of the first, second, and third positions (mean position 2.00). Meditron 70B identified VL at intermediate positions (2nd, 3rd, and 4th) in three cases, but did not identify VL in five additional cases (two non-engagement responses, one incomplete response, and two complete differential lists without VL; section 3.1). Mixtral 8x7B identified VL at the third position in two cases (Cases 06 and 08); LLaMA 2 70B identified VL at the first position in a single case (Case 08); and DeepSeek-V2 16B identified VL at the fifth position in a single case (Case 05). Detailed positional data for each case and model are provided in Table 2.

### 3.3. Specialist committee plausibility assessment

Figure 1B and 1C summarize the committee's classification of the 280 diagnostic hypotheses generated across the 7 models × 8 cases × 5 slots, after the two-stage independent assessment and consensus adjudication described in section 2.7. Of all individual classifications produced in the first stage, the committee reached unanimous agreement on every hypothesis except one, indicating high inter-rater reliability prior to discussion.

DeepSeek-R1 70B produced differential lists of substantially higher overall plausibility than any other model, with a mean of 4.88 plausible hypotheses per case (corresponding to 8 correctly identified VL diagnoses plus 31 plausible alternative differentials, against only one non-plausible hypothesis across forty evaluated slots). Falcon 180B occupied a clear second position, with a mean of 3.88 plausible hypotheses per case (6 correctly identified VL diagnoses plus 25 plausible alternatives, against 9 non-plausible hypotheses). Mixtral 8x7B and DeepSeek-V2 16B occupied intermediate positions in plausibility (means of

**Table 2. Position of the target diagnosis (visceral leishmaniasis) within the top-five list, by case and large language model**

| Case | DeepSeek-R1 70B | Falcon 180B | Phi-3 mini 3.8B | Meditron 70B | Mixtral 8x7B | DeepSeek-V2 16B | LLaMA 2 70B |
|---|---|---|---|---|---|---|---|
| 01 | 2 | 3 | 2 | 4 | 0 | 0 | 0 |
| 02 | 1 | 3 | 0 | NE[1] | 0 | 0 | 0 |
| 03 | 2 | 4 | 0 | NE[1] | 0 | 0 | 0 |
| 04 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 05 | 2 | 0 | 0 | 3 | 0 | 5 | 0 |
| 06 | 3 | 5 | 0 | 2 | 3 | 0 | 0 |
| 07 | 2 | 3 | 3 | 0 | 0 | 0 | 0 |
| 08 | 1 | 2 | 1 | INC[2] | 3 | 0 | 1[3] |
| Hits | 8 | 6 | 3 | 3 | 2 | 1 | 1 |

Notes: Numeric values denote the rank position of VL within the model's differential list (1 = first/most-likely; 5 = fifth/least-likely). 0 = VL absent from a complete five-item list. [1]NE = non-engagement response: model returned generic text without identifiable diagnostic hypotheses. [2]INC = incomplete response: model produced fewer than five diagnostic hypotheses. [3]LLaMA 2 70B identified VL as the leading hypothesis in this case. The "Hits" row reports the number of cases (out of 8) in which VL appeared anywhere in the differential list.

3.50 and 3.12 plausible hypotheses per case, respectively), despite their low top-five accuracy for VL itself.

The remaining three models exhibited distinct patterns of clinically problematic output that the binary top-five accuracy metric did not capture. Phi-3 mini 3.8B produced three diagnostic hypotheses identified by the committee as fabricated entities with no recognizable medical basis ("infecção porvaciana humana," "linfangite proportional," and "linfangiose") – annotated as "(+3 fabr.)" in Figure 1C—in addition to a high number of non-plausible hypotheses (mean 2.38 per case). No other model in the evaluation produced fabricated diagnoses. Meditron 70B exhibited the previously noted non-engagement pattern (Cases 02 and 03 without diagnostic content and Case 08 with an incomplete one-item differential), resulting in only 26 evaluable hypothesis slots out of 40 and a mean of 2.00 plausible hypotheses per case. LLaMA 2 70B, in contrast, completed all slots but generated overwhelmingly generic or non-specific differentials ("doença renal," "doença gastrointestinal," "doença hepática" without further specification), yielding the lowest mean of plausible hypotheses per case (1.00) and the highest mean of non-plausible hypotheses (4.00) of any model.

### 3.4. Comparison with previously published ChatGPT/GPT-4 results

Table 3 places the present open-weight results in context against the ChatGPT/GPT-4 reference data previously published by our group[18] on the same eight clinical vignettes. Two open-weight models (DeepSeek-R1 70B [100% top-five accuracy] and Falcon 180B [75% top-five accuracy]) equaled or exceeded the proprietary baseline, although neither matched ChatGPT/GPT-4's 50% rate of placing VL as the most likely diagnosis. The remaining five open-weight models fell substantially below the ChatGPT/GPT-4 benchmark on both top-five and top-one accuracy.

## 4. Discussion

This study provides what is, to our knowledge, the first systematic evaluation of seven open-weight large language models—spanning a deliberately heterogeneous set of architectures, parameter scales, and training paradigms—for the differential diagnosis of visceral leishmaniasis using locally-deployed inference. By replicating the methodology of our previous evaluation of ChatGPT/GPT-4 on the same eight clinical vignettes,[18] the present results allow direct, quantitative comparison between proprietary and open-

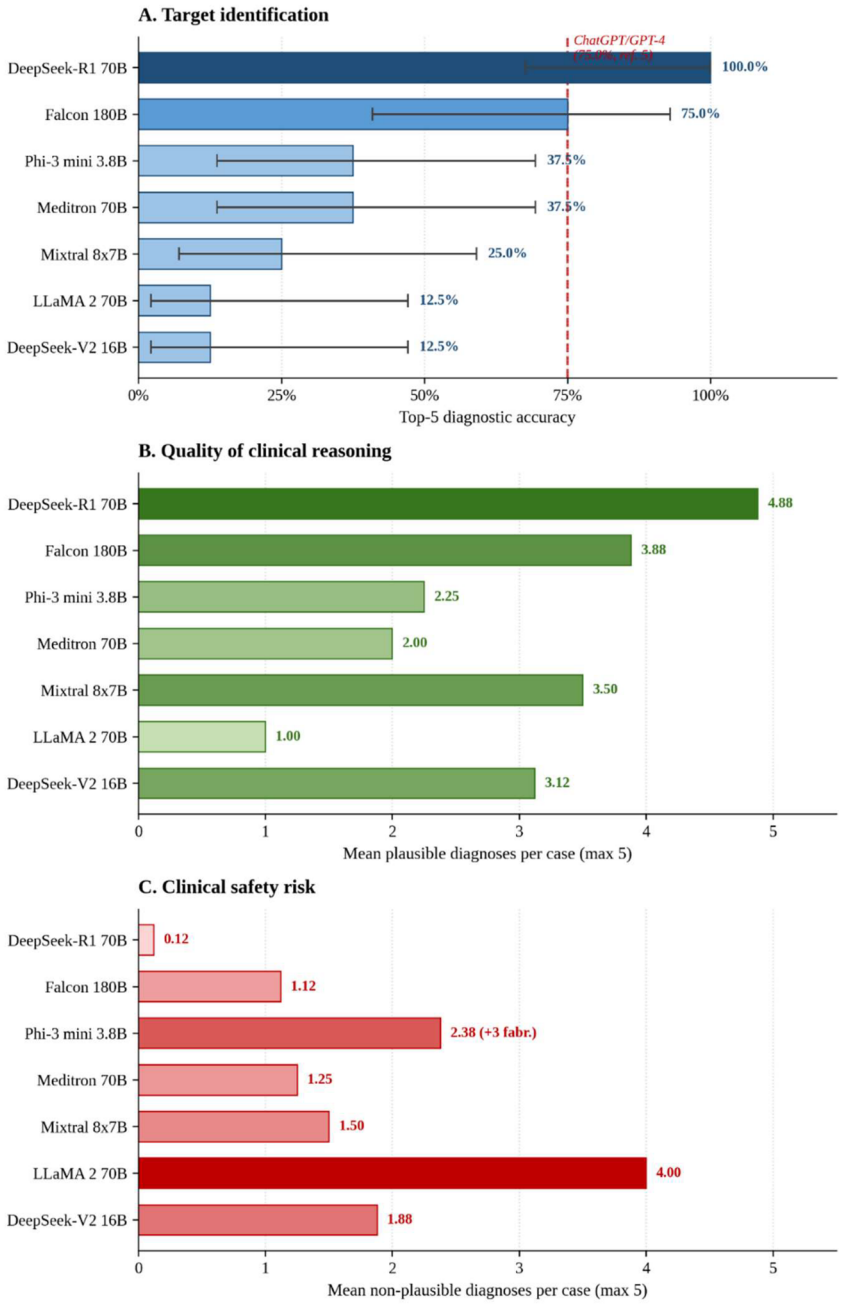



**Figure 1.** Comparative diagnostic performance of the seven open-weight large language models evaluated. (A) Top-5 diagnostic accuracy for visceral leishmaniasis ($n$ = 8 cases per model). Bars represent point estimates; error bars represent 95% Wilson score confidence intervals. The dashed red reference line at 75% indicates the previously published ChatGPT/GPT-4 benchmark on the same vignettes.[18] Bar shading distinguishes models with point estimates above (dark blue), equal to (medium blue), or below (light blue) the ChatGPT/GPT-4 reference value. (B) Mean number of plausible diagnoses generated per case by each model, as classified by the specialist committee (sum of correctly identified VL plus plausible alternative differentials, divided by 8 cases). Color intensity scales with the value plotted. (C) Mean number of non-plausible diagnoses per case (sum of non-plausible hypotheses divided by 8 cases). For Phi-3 mini 3.8B, the annotation "(+3 fabr.)" indicates three additional diagnostic hypotheses classified by the committee as fabricated nosological entities (*i.e.*, neologisms or distorted disease names with no recognized medical referent). All three panels share the same y-axis ordering—models are ranked top-to-bottom by descending top-5 accuracy (panel A)—to permit direct visual comparison across metrics.

**Table 3. Comparative top-five and top-one diagnostic accuracy: present open-weight LLM results versus previously published ChatGPT/GPT-4 reference data**

| LLM | Top-5% (95% CI) | Top-1% (95% CI) | Source |
|---|---|---|---|
| ChatGPT/GPT-4 | 75.0 (40.9–92.9) | 50.0 (21.5–78.5) | Schwingel *et al.*[18] |
| DeepSeek-R1 70B | 100.0 (67.6–100.0) | 37.5 (13.7–69.4) | Present study |
| Falcon 180B | 75.0 (40.9–92.9) | 0.0 (0.0–32.4) | Present study |
| Phi-3 mini 3.8B | 37.5 (13.7–69.4) | 12.5 (2.2–47.1) | Present study |
| Meditron 70B | 37.5 (13.7–69.4) | 0.0 (0.0–32.4) | Present study |
| Mixtral 8x7B | 25.0 (7.1–59.1) | 0.0 (0.0–32.4) | Present study |
| DeepSeek-V2 16B | 12.5 (2.2–47.1) | 0.0 (0.0–32.4) | Present study |
| LLaMA 2 70B | 12.5 (2.2–47.1) | 12.5 (2.2–47.1) | Present study |

Notes: ChatGPT/GPT-4 reference data are reproduced from our previously published evaluation on the same eight clinical vignettes[18]; the present study reports newly collected data from the seven open-weight models. All 95% confidence intervals (CIs) were recomputed using the Wilson score method to ensure consistency across rows. Top-five and top-one accuracy definitions are identical to those used in Table 1.

weight alternatives in a clinically relevant NTD context. Four principal findings emerge.

First, the diagnostic performance of open-weight LLMs varied dramatically by architecture and training paradigm, ranging from 100% top-five accuracy (DeepSeek-R1 70B) to 12.5% (LLaMA 2 70B and DeepSeek-V2 16B). This 87.5-percentage-point spread was not predicted by parameter count alone: the largest model evaluated (Falcon 180B, with approximately 2.6× the parameter budget of the next-largest models) achieved 75% accuracy, matching ChatGPT/GPT-4[18] but falling well below DeepSeek-R1 70B. Architectural lineage and training methodology (particularly the reinforcement-learning-based reasoning training applied to DeepSeek-R1[29]) appear to matter more than raw scale for this diagnostic task, consistent with recent reasoning-trained models matching or exceeding earlier proprietary models on complex clinical-reasoning benchmarks,[33] and extends that observation to neglected tropical diseases evaluated in Brazilian Portuguese, a setting underrepresented in existing diagnostic-LLM literature. These findings also complement recent reports that frontier proprietary models can achieve expert-level performance on standardized medical-QA benchmarks when combined with appropriate prompting and grounding strategies,[34] thereby framing the open-weight ecosystem as a rapidly closing (though still architecturally heterogeneous) alternative.

Second, the medically fine-tuned model evaluated, Meditron 70B,[30] exhibited paradoxically poor performance—37.5% top-five accuracy, two cases of complete non-engagement (Cases 02 and 03 returning generic text, Case 02 in particular producing what appeared to be a placeholder list of "Doença A" through "Doença E"), and a further case of incomplete output. A similar pattern of poor Meditron performance has been observed by other groups working on related tasks,[33] raising the question of whether biomedical pretraining on English-language corpora (as in Meditron's training data[30]) confers any benefit for diagnostic reasoning in Portuguese-language NTD scenarios. This pattern has been further corroborated by recent large-scale evaluations: in a benchmark study of 1933 Eurorad radiology cases, Meditron-7B achieved only 44.3% accuracy compared with 73.2% for the general-domain LLaMA-3-70B,[35] and an independent assessment of medical small language models similarly documented that Meditron-7B and BioMistral-7B underperformed their general-domain counterparts of equivalent scale.[36] The result echoes a broader concern in the LLM-in-medicine literature: domain-specific fine-tuning targeted at biomedical text does not automatically translate into clinical reasoning quality, particularly when the disease space (NTDs in Latin America) is underrepresented in

training corpora.[6,7] The question of whether language-specific medical instruction tuning[31] might address this gap warrants future investigation.

Third, the specialist committee assessment revealed three qualitatively distinct failure modes that the binary top-five accuracy metric did not capture, and which carry distinct implications for clinical safety. Phi-3 mini 3.8B exhibited a fabrication failure mode: in three of forty evaluable hypothesis slots, the model produced apparent neologisms or distorted disease names with no recognizable medical referent ("infecção porvaciana humana," "linfangite proportional," "linfangiose"). This is consistent with the broader phenomenon of LLM hallucination, in which models generate confident-sounding but fabricated content,[13,37] and is particularly concerning when the model in question (at 3.8 billion parameters) is precisely the kind that is most likely to be deployed in resource-constrained settings due to its modest hardware requirements. By contrast, LLaMA 2 70B exhibited a vagueness failure mode, completing all output slots but generating overwhelmingly generic differentials ("doença renal," "doença gastrointestinal," "doença hepática" without further specification), yielding the highest mean of non-plausible hypotheses (4.00 per case) and the lowest of plausible differentials (1.00 per case) of any model evaluated. Meditron 70B exhibited a third pattern (non-engagement) where the model returned generic text or placeholders rather than diagnostic content. From a safety standpoint, these three modes pose different risks: fabrication risks reinforcing false beliefs in the prescriber, vagueness risks failing to narrow the differential meaningfully, and non-engagement risks user disengagement from the tool. More concretely: a fabricated diagnostic entity such as those produced by Phi-3 mini 3.8B, if accepted at face value by a clinician unfamiliar with the term, could trigger inappropriate confirmatory testing, redirect clinical attention away from the genuine differential, and reinforce false beliefs that may propagate through the team; the risk is particularly acute for small models such as Phi-3 mini, which by virtue of their modest hardware requirements are precisely those most likely to be deployed in resource-constrained settings where expert verification may be less readily available. The vagueness failure mode exhibited by LLaMA 2 70B, although less likely to introduce factual error, dilutes the clinical utility of the differential and contributes to the broader phenomenon of alert fatigue when AI-generated outputs become routinely uninformative.[37] The non-engagement pattern observed in Meditron 70B poses a distinct organizational risk: when clinicians repeatedly encounter placeholder or refusal-style outputs from a deployed AI assistant, trust in the entire decision-support workflow erodes, undermining the adoption of properly-functioning models within the same clinical environment. Each warrants distinct mitigation strategies, and none is captured by accuracy metrics alone.[37-39]

Fourth, the present results have direct relevance for the clinical and policy debate around AI-assisted diagnostic deployment in resource-constrained settings, where data sovereignty, hardware limitations, and offline operation are critical constraints.[21] Among the seven open-weight models evaluated, two (DeepSeek-R1 70B and Falcon 180B) matched or exceeded the published top-five accuracy of ChatGPT/GPT-4 on the same vignettes,[18] while running entirely on institutional hardware via Ollama[23] without transmitting input data to external servers. This is a meaningful proof-of-concept for institutions in NTD-endemic regions of Brazil where the São Francisco Valley exemplifies the broader pattern: hyperendemic disease coexisting with limited specialist availability and active concerns about patient-data confidentiality on cloud platforms.[22] However, our findings also indicate that not all open-weight models are equivalent for this purpose. The choice of model matters: DeepSeek-R1 70B's combination of high accuracy, near-zero non-plausibility, and zero fabrications places it in a meaningfully different operating regime than the other six open-weight models tested, including Falcon 180B (which despite matching ChatGPT/GPT-4 on top-five accuracy never ranked VL as the leading hypothesis and produced 9 non-plausible differentials). Three operational considerations make the open-weight, locally-deployed scenario particularly attractive in Brazilian public-hospital settings. First, on the privacy axis, locally-served inference via Ollama keeps all input data (including potentially sensitive clinical narratives) within institutional infrastructure, eliminating the cross-border data-transfer concerns inherent to cloud-hosted proprietary APIs and aligning with the data-sovereignty principles increasingly required by Brazilian regulatory frameworks. Second, on the cost axis, the dedicated workstation used in the present study represents a one-time infrastructure investment rather than a recurring per-call API subscription, with implications for budget predictability in publicly-funded health systems. Third, on the operational-resilience axis, locally-deployed models continue to function during periods of restricted internet connectivity—a non-trivial consideration in remote NTD-endemic regions where connectivity may be intermittent. These trade-offs, which have been systematically articulated in the recent literature on enterprise LLM adoption in healthcare,[22] underscore that the open-weight pathway is not merely a technical alternative to proprietary cloud services but a qualitatively different operational regime. Practical deployment recommendations therefore cannot

be generic; institution-level evaluation against locally-relevant cases remains essential before clinical adoption.[40]

### 4.1. Strengths

This study has several strengths. First, the use of locally-deployed open-weight models on institutional hardware via Ollama represents a deployment scenario directly relevant to resource-constrained clinical settings, complementing the proprietary-cloud-API evaluations that dominate the existing literature.[38] Second, the deliberate selection of seven models spanning multiple architectural paradigms (dense decoder, sparse mixture-of-experts, reasoning-tuned, biomedical-fine-tuned, small efficient instruction-tuned) yields generalizable insights about the diversity of currently available open-weight options. Third, the strict methodological replication of our previous ChatGPT/GPT-4 evaluation[18] (same vignettes, same prompt, same presentation order) enables direct, quantitative comparison between proprietary and open-weight LLMs on identical clinical material, which is rare in the literature.[15] Fourth, the specialist committee plausibility assessment, based on independent classifications followed by structured consensus adjudication, captures dimensions of clinical reasoning quality that binary accuracy metrics miss, addressing a limitation explicitly acknowledged in our previous work.[18]

### 4.2. Limitations

Several limitations should be acknowledged. First, the sample size of eight clinical vignettes per model limits the statistical precision of accuracy estimates, as evidenced by the wide 95% Wilson confidence intervals reported in Table 1. The binary top-five accuracy and 8-case sample do not provide power to formally test inter-model differences, and all between-model comparisons should be interpreted as descriptive. A natural pathway to address this restriction is being pursued by our group: a recent paired experimental study by Soares da Silva *et al.*[20] evaluated four LLMs on 24 real anonymized clinical cases (12 Chagas disease and 12 visceral leishmaniasis) from the same teaching hospital, extending the research line to real-world clinical material with a larger paired sample. Second, the use of synthetic clinical vignettes (although formulated by an experienced infectious-disease specialist) is known to inflate diagnostic accuracy estimates relative to real clinical material,[15] a tension that future evaluations using real anonymized clinical data should address.[33] Third, the evaluation was restricted to a single prompt structure and single inference configuration (Ollama defaults); different prompt-engineering strategies, including chain-of-thought scaffolding and role-instructed examples, may yield different relative model rankings. Likewise, inference-parameter optimization (including reduction of the sampling temperature toward more deterministic outputs) was not investigated in the present study; a temperature-ablation experiment, applied uniformly across the seven models, constitutes a methodologically straightforward extension of the present design and is a priority for follow-up work.[16] Fourth, the open-weight model versions evaluated reflect the state of the Ollama ecosystem in January 2025; ongoing model updates may alter performance characteristics.[29] Fifth, the focus on a single disease (visceral leishmaniasis) and a single language (Brazilian Portuguese) limits external generalization; both broader disease coverage and additional language settings warrant evaluation. Sixth, although the specialist committee included expertise in both NTD diagnosis and clinical-diagnosis teaching, the four-category plausibility scheme is necessarily coarse and cannot capture the full nuance of clinical reasoning quality (*e.g.*, severity of error, ranking of plausible alternatives within a list); more granular frameworks should be developed for future evaluations.

## 5. Conclusion

In this exploratory study, seven open-weight large language models deployed locally on institutional hardware via Ollama were evaluated for differential diagnosis of visceral leishmaniasis using clinical vignettes identical to those previously used to evaluate ChatGPT/GPT-4.[18] Performance varied dramatically across models, with DeepSeek-R1 70B exceeding the previously published ChatGPT/GPT-4 top-five accuracy benchmark (100% vs. 75%), Falcon 180B matching it, and the remaining five models (including the biomedically fine-tuned Meditron 70B) falling substantially below. Beyond accuracy, the specialist committee assessment uncovered three qualitatively distinct failure modes (fabrication of nosological entities by Phi-3 mini, vagueness in LLaMA 2 70B, and non-engagement in Meditron 70B), each with distinct implications for clinical safety. These findings demonstrate the feasibility of deploying open-weight LLMs as adjuncts for NTD differential diagnosis in resource-constrained settings, while underscoring the importance of model selection, expert-validated qualitative assessment beyond binary accuracy, and institution-level evaluation prior to clinical adoption. Larger studies using real anonymized clinical material, broader disease coverage, and prompt-engineering optimization are warranted to translate these exploratory findings into actionable clinical guidance. Promising directions include the evaluation of recent reasoning-tuned small language models specifically trained for clinical reasoning on textbook-derived chain-of-thought trajectories,[36] systematic temperature-

ablation studies aligned with safety-critical deployment requirements, and continuity of the line of research already pursued by our group on real anonymized clinical material in NTD-endemic regions of Brazil.[20]

## Acknowledgments

The authors declare the use of generative artificial intelligence (AI) in the preparation of this manuscript. Claude Opus 4.7 (Anthropic, PBC, San Francisco, CA, USA) was employed for language refinement (including improvements in grammar, sentence structure, and readability) and technical writing assistance (such as suggesting more effective ways to structure complex technical descriptions). All AI-assisted processes were critically reviewed by the authors to ensure the integrity and reliability of the results. The final decisions and interpretations presented in this article were made solely by the authors.

## Funding

This study received financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under grant number 408003/2023-5 and from the Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) under grant number APQ-0238-4.01/24. Additionally, CNPq awarded Paulo Adriano Schwingel a Research Productivity Grant (PQ) under grant number 306628/2025-2, and FACEPE previously awarded a Research Productivity Grant (BPP) under grant number BPP-0003-4.01/24.

## Conflict of interest

Paulo Adriano Schwingel serves as an Editorial Board Member of this journal but was not in any way involved in the editorial and peer-review process conducted for this paper, directly or indirectly. Separately, other authors declared that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

## Author contributions

*Conceptualization*: Dino Schwingel, Samuel Ricarte de Aquino, Fabiana Oliveira dos Santos Camatari, Maria Jacqueline Silva Ribeiro, Paula Andreatta Maduro, Paulo Adriano Schwingel

*Formal analysis*: All authors

*Funding acquisition*: Paulo Adriano Schwingel

*Investigation*: All authors

*Methodology*: Dino Schwingel, Samuel Ricarte de Aquino, Paulo Gustavo Serafim de Carvalho, Rogério Fabiano Gonçalves, Maria Jacqueline Silva Ribeiro, Paulo Adriano Schwingel

*Resources*: All authors.

*Supervision*: Aline Rafaela Soares da Silva, Paula Andreatta Maduro.

*Validation*: Dino Schwingel, Samuel Ricarte de Aquino, Márcio Oliveira Silva, Paulo Serafim de Carvalho, Rogério Fabiano Gonçalves, Maria Jacqueline Silva Ribeiro

*Writing–original draft*: Aline Rafaela Soares da Silva, Adrillany da Costa Santos, Humberto Baudel Francisco, Luíza Azevedo Ferreira, John Alan Rodrigues Dantas, Paula Andreatta Maduro, Paulo Adriano Schwingel

*Writing–review & editing*: Dino Schwingel, Samuel Ricarte de Aquino, Márcio Oliveira Silva, Paulo Gustavo Serafim de Carvalho, Rogério Fabiano Gonçalves, Fabiana Oliveira dos Santos Camatari, Maria Jacqueline Silva Ribeiro

## Ethics approval and consent to participate

This study involved the use of clinical vignette cases and thus did not require individual informed consent. Research approval was obtained from the Ethics Committee of the HU-UNIVASF (approval number 6.967.834).

## Consent for publication

Not applicable.

## Availability of data

The data used to support the findings of this study are included within the article.

## References

1. Chappuis F, Sundar S, Hailu A, *et al.* Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? *Nat Rev Microbiol.* 2007;5(11):873-882.

   doi: 10.1038/nrmicro1748

2. Lainson R, Rangel EF. Lutzomyia longipalpis and the eco-epidemiology of American visceral leishmaniasis, with particular reference to Brazil: a review. *Mem Inst Oswaldo Cruz.* 2005;100(8):811-827.

   doi: 10.1590/s0074-02762005000800001

3. Alvar J, Vélez ID, Bern C, *et al.* Leishmaniasis worldwide and global estimates of its incidence. *PLoS ONE.* 2012;7(5):e35671.

   doi: 10.1371/journal.pone.0035671

4. de Almeida Soares FM, Rocha TS, Nascimento ER, *et al.* Human visceral leishmaniasis in Brazil in the past 20 years: an epidemiologic update. *Rev Soc Bras Med Trop.* 2025;58:e0019-2025.

doi: 10.1590/0037-8682-0019-2025

5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al.* Attention is all you need. *arXiv*. Preprint posted online 2017.

   doi: 10.48550/arXiv.1706.03762

6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29(8):1930-1940.

   doi: 10.1038/s41591-023-02448-8

7. Maity S, Saikia MJ. Large language models in healthcare and medical applications: a review. *Bioengineering.* 2025;12(6):631.

   doi: 10.3390/bioengineering12060631

8. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng.* 2018;2(10):719-731.

   doi: 10.1038/s41551-018-0305-z

9. Egli A. ChatGPT, GPT-4, and other large language models: the next revolution for clinical microbiology? *Clin Infect Dis.* 2023;77(9):1322-1328.

   doi: 10.1093/cid/ciad407

10. Cheng K, Li Z, He Y, *et al.* Potential use of artificial intelligence in infectious disease: take ChatGPT as an example. *Ann Biomed Eng.* 2023;51(6):1130-1135.

    doi: 10.1007/s10439-023-03203-3

11. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by Generative Pretrained Transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health.* 2023;20(4):3378.

    doi: 10.3390/ijerph20043378

12. Mizuta K, Hirosawa T, Harada Y, Shimizu T. Can ChatGPT-4 evaluate whether a differential diagnosis list contains the correct diagnosis as accurately as a physician? *Diagnosis.* 2024;11(3):321-324.

    doi: 10.1515/dx-2024-0027

13. Ueda D, Walston SL, Matsumoto T, Deguchi R, Tatekawa H, Miki Y. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. *BMC Digit Health.* 2024;2(1):4.

    doi: 10.1186/s44247-023-00058-5

14. Meral G, Ateş S, Günay S, Öztürk A, Kuşdoğan M. Comparative analysis of ChatGPT, Gemini and emergency medicine specialist in ESI triage assessment. *Am J Emerg Med.* 2024;81:146-150.

    doi: 10.1016/j.ajem.2024.05.001

15. Dinc MT, Kim S, Lee B, 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: 10.1093/jamiaopen/ooaf055

16. Gaebe K, van der Woerd B. Evaluation of large language models as a diagnostic tool for medical learners and clinicians using advanced prompting techniques. *PLoS ONE.* 2025;20(8):e0325803.

    doi: 10.1371/journal.pone.0325803

17. Su H, Sun Y, Li R, *et al.* Large language models in medical diagnostics: scoping review with bibliometric analysis. *J Med Internet Res.* 2025;27:e72062.

    doi: 10.2196/72062

18. Schwingel PA, Schwingel D, de Aquino SR, *et al.* An exploratory study on the potential of ChatGPT as an AI-assisted diagnostic tool for visceral leishmaniasis. *Artif Intell Health.* 2024;1(4):97-106.

    doi: 10.36922/aih.3930

19. da Cruz Pereira RA, Lima RR, Gomes ACA, *et al.* Exploring the potential of an AI chatbot as a supplementary tool for nutritional prescription at hospital discharge: a preliminary study. *Scientifica.* 2025;2025:2632410.

    doi: 10.1155/sci5/2632410

20. Soares da Silva ARS, Schwingel D, de Aquino SR, *et al.* Gender-attributed persona prompts and the diagnostic accuracy of proprietary and open-weight large language models in Chagas disease and visceral leishmaniasis: a paired experimental study. *Healthcare.* 2026;14(10):1385.

    doi: 10.3390/healthcare14101385

21. Kanter GP, Packel EA. Health care privacy risks of AI chatbots. *JAMA.* 2023;330(4):311-312.

    doi: 10.1001/jama.2023.9618

22. Cihoric N, Badra EV, Frei AL, *et al.* Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ Digit Med.* 2025;8(1):179.

    doi: 10.1038/s41746-025-01476-7

23. Ollama. Get up and running with large language models locally. Accessed January 15, 2025.https://github.com/ollama/ollama.

24. Touvron H, Martin L, Stone K, *et al.* Llama 2: open foundation and fine-tuned chat models. *arXiv*. Preprint posted online 2023.

    doi: 10.48550/arXiv.2307.09288

25. Almazrouei E, Alobeidli H, Alshamsi A, *et al.* The Falcon series of open language models. *arXiv*. Preprint posted online 2023.

    doi: 10.48550/arXiv.2311.16867

26. Jiang AQ, Sablayrolles A, Roux A, *et al.* Mixtral of experts. *arXiv*. Preprint posted online 2024.

doi: 10.48550/arXiv.2401.04088

27. DeepSeek-AI, Liu A, Feng B, *et al.* DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model. *arXiv*. Preprint posted online 2024.

doi: 10.48550/arXiv.2405.04434

28. Abdin M, Jacobs SA, Awan AA, *et al.* Phi-3 technical report: a highly capable language model locally on your phone. *arXiv*. Preprint posted online 2024.

doi: 10.48550/arXiv.2404.14219

29. Guo D, Yang D, Zhang H, *et al.* DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature.* 2025;645(8081):633-638.

doi: 10.1038/s41586-025-09422-z

30. Chen Z, Hernández Cano A, Romanou A, *et al.* MEDITRON-70B: scaling medical pretraining for large language models. *arXiv*. Preprint posted online 2023.

doi: 10.48550/arXiv.2311.16079

31. Sukeda I, Suzuki M, Sakaji H, Kodera S. Development and analysis of medical instruction-tuning for Japanese large language models. *Artif Intell Health.* 2024;1(2):107-116.

doi: 10.36922/aih.2695

32. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci.* 2001;16(2):101-133.

doi: 10.1214/ss/1009213286

33. Mumtaz U, Ahmed A, Mumtaz S. LLMs-healthcare: current applications and challenges of large language models in various medical specialties. *Artif Intell Health.* 2024;1(2):16-28.

doi: 10.36922/aih.2558

34. Singhal K, Tu T, Gottweis J, *et al.* Toward expert-level medical question answering with large language models. *Nat Med.* 2025;31(3):943-950.

doi: 10.1038/s41591-024-03423-7

35. Sorace L, Hoffmann F, Kottlors J, *et al.* Benchmarking the diagnostic performance of open source LLMs in 1933 Eurorad case reports. *NPJ Digit Med.* 2025;8(1):97.

doi: 10.1038/s41746-025-01488-3

36. Kim H, Hwang H, Lee J, *et al.* Small language models learn enhanced reasoning skills from medical textbooks. *NPJ Digit Med.* 2025;8(1):240.

doi: 10.1038/s41746-025-01653-8

37. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. *Interact J Med Res.* 2025;14:e59823.

doi: 10.2196/59823

38. Biswas S. ChatGPT and the future of medical writing. *Radiology.* 2023;307(2):e223312.

doi: 10.1148/radiol.223312

39. Aquino YS. Making decisions: bias in artificial intelligence and data-driven diagnostic tools. *Aust J Gen Pract.* 2023;52(7):439-442.

doi: 10.31128/AJGP-12-22-6630

40. Akhtar ZB. Artificial intelligence within medical diagnostics: a multi-disease perspective. *Artif Intell Health.* 2025;2(3):44-62.

doi: 10.36922/aih.5173