

Supplementary File

Illustrative verbatim outputs of two evaluated models

Companion to: Open-weight large language models for visceral leishmaniasis: comparative diagnostic accuracy of seven locally-deployed models

This supplementary file provides two verbatim model outputs collected during the experiment described in the main manuscript (Sections 2.5 and 3.1). The two examples were selected to illustrate the qualitative span of performance observed across the seven evaluated models. The prompt structure was identical across all model–case combinations: "*Por favor, forneça-me os cinco diagnósticos mais prováveis para os seguintes sintomas:*" followed immediately by the full text of the clinical vignette in Brazilian Portuguese.

The eight clinical vignettes are reproduced in full in our previously published work (Schwingel *et al.*, 2024; freely available at <https://doi.org/10.36922/aih.3930>; reference 18 of the main manuscript). All formatting choices made by each model — including bullet styles, emphasis, line breaks, and any whitespace artefacts produced during streaming inference — are preserved as produced, in accordance with the verbatim-recording protocol described in Section 2.5 of the main manuscript.

Example A — DeepSeek-R1 70B, Case 04 (high-quality reasoning)

Model: DeepSeek-R1 70B (Ollama image ID 0c1615a8ca32)

Case: Vignette #04 (top-five accuracy: VL identified at rank 1)

Specialist committee classification: 5/5 plausible hypotheses; VL correctly identified as the leading differential.

Verbatim output as produced by the model:

```
<think>
```

```
Okay, so I've got this query from a user asking for the top five possible diagnoses for a patient with specific symptoms. Let me try to unpack this step by step.
```

```
First, I need to look at the patient's history and symptoms. The patient is a 51-year-old Black woman living in Juazeiro, Bahia. She's unemployed, living in an urban area. Her main issues are daily fevers for 60 days, along with symptoms like asthenia, weight loss, diarrhea, abdominal discomfort, hyporexia, chills, and bleeding.
```

```
Looking at her past medical history, she has cirrhosis, hypertension, diabetes, and chronic kidney disease. That's a lot of comorbidities, which
```

might be relevant. She doesn't mention any current medications, and she doesn't use alcohol or tobacco.

In the physical exam, she's 81 kg and 1.57 m, which gives her a BMI of around 33, so she's obese. Her blood pressure is 100/80, which is on the lower side but could be due to dehydration or other issues. The abdominal exam shows hepatomegaly, splenomegaly of 5 cm, ascites, and abdominal pain. She also has jaundice, hypochromic skin without mucosal lesions, and her heart and lungs are normal except for a rapid heart rate.

Now, considering her symptoms and physical findings, I need to think about what could cause prolonged fever, weight loss, and abdominal issues. Let's start with the most likely possibilities.

1. Leishmaniasis visceral: Given the 60-day daily fever, marked weight loss, asthenia, diarrhea, and abdominal discomfort, visceral leishmaniasis is a strong candidate, particularly in endemic areas like Juazeiro, Bahia. Patients with visceral leishmaniasis are prone to severe systemic infections characterized by prolonged fever, hepatosplenomegaly, ascites, and hepatic involvement.

2. Spontaneous bacterial peritonitis: Given the fever, abdominal pain, ascites, and her history of cirrhosis, which can lead to spontaneous bacterial peritonitis, this seems like a strong candidate. Cirrhotic patients are prone to infections like SBP, which can cause fever, abdominal discomfort, and ascites.

3. Tuberculose abdominal: TB can present with fever, weight loss, diarrhea, and abdominal pain. The prolonged fever and systemic symptoms like asthenia and hyporexia support this. Plus, her splenomegaly and hepatomegaly could be due to TB affecting the organs.

4. Hepatitis (HBV, HCV or autoimmune): Given her documented cirrhosis, chronic hepatitis is clearly part of her medical history. The presentation of prolonged fever, jaundice, ascites, and hepatosplenomegaly could suggest acute hepatitis B reactivation, chronic hepatitis C with decompensation, or acute autoimmune hepatitis flare.

5. HIV infection: Given the prolonged fever lasting 60 days, weight loss, diarrhea, and hepatosplenomegaly, HIV infection should be strongly considered. Patients with AIDS are prone to opportunistic infections and fever of unknown origin, which can cause prolonged daily fever, abdominal symptoms, diarrhea, and organomegaly.

I also need to think about other possibilities. For example, her jaundice could be due to liver disease, but she already has cirrhosis, which might

explain that. However, other infections like malaria or parasites could also be a factor, especially considering her location in Bahia, where such infections are more common.

The user is likely a medical student or healthcare professional seeking a differential diagnosis to guide further testing or treatment.

In summary, the top five diagnoses seem to revolve around infections, given the fever and systemic symptoms. Leishmaniasis and tuberculosis are leading due to prolonged presentation, followed by opportunistic infections in HIV, complications of cirrhosis like SBP, and chronic viral hepatitis. It's important to consider her background and comorbidities when evaluating these options.

</think>

A partir dos sintomas, histórico e exame físico apresentados, seguem os **cinco diagnósticos mais prováveis**:

1. **Leishmaniose visceral**
2. **Peritonite bacteriana espontânea**
3. **Tuberculose abdominal**
4. **Hepatite crônica ou autoimune**
5. **Infecção por HIV**

Note: DeepSeek-R1 70B is a reasoning-tuned model that generates an internal reasoning trace, delimited by <think>...</think> tags, prior to producing its final response. In this example, the reasoning trace was produced predominantly in English even though the prompt was in Brazilian Portuguese, while the final five-hypothesis list was correctly returned in Portuguese as requested. This bilingual reasoning behaviour is a known characteristic of DeepSeek-R1 and does not affect the language of the final clinical output. The output is reproduced verbatim, including a tabulation artefact embedded within the word "hepa\ttosplenomegaly", which was produced during streaming inference and preserved per the recording protocol.

Example B — Phi-3 mini 3.8B, Case 06 (fabrication failure mode)

Model: Phi-3 mini 3.8B (Ollama image ID 157684aa81ad)

Case: Vignette #06 (VL not identified within top-five)

Specialist committee classification: This output contains one of the three diagnostic hypotheses classified by the committee as a fabricated nosological entity — "*infecção porvaciana humana*" — with no recognised medical referent. Two other hypotheses across the model's outputs ("*infangite porportional*" and "*infangiose*") exhibit the same pattern in other cases.

Verbatim output as produced by the model:

1. Infecção do trato gastrointestinal (TGI): os sintomas de astenia, perda de peso, desconforto abdominal e sangramentos podem indicar uma infecção no TGI, como a gastrite ou enterite.

2. Hepatite viral: devido à hepatomegalia e ao histórico de tabagismo, a hepatite viral pode estar envolvida, particularmente a hepatite C, que pode ser transmitida pelo contato direto com sangue ou fluidos corporais.

3. **Infecção porvaciana humana:** infecção por parvovírus humano causa febre prolongada, astenia, perda de peso e hepatoesplenomegalia em indivíduos imunocomprometidos.

4. Infecção bacteriana enterotoxigênica: infecção bacteriana por *Escherichia coli* ou *Vibrio cholerae* causa diarreia, desconforto abdominal, febre e perda peso, e translocação bacteriana piora em hepatoesplenomegalia.

5. Malária: embora rara no Brasil, a malária pode estar envolvida devido ao fato de o paciente viver em uma área onde a doença pode estar presente, juntamente com o histórico de tabagismo e etilismo.

É importante que o paciente seja avaliado por um médico para obter um diagnóstico preciso e um tratamento adequado.

Note: Hypothesis #3 ("*Infecção porvaciana humana*") was classified by the specialist committee as a fabricated nosological entity: the term "*porvaciana*" has no recognised medical referent in Portuguese or in any cognate language. The clinically concerning feature of this output is that the model not only produced an inexistent term but also generated a coherent-sounding etiological justification ("*infecção por parvovírus humano causa febre prolongada, astenia, perda de peso e hepatoesplenomegalia em indivíduos imunocomprometidos*"), which exemplifies the pattern of confidently-presented fabricated content discussed in Section 4 (third principal finding) of the

main manuscript. A clinician unfamiliar with the term might be misled by the surrounding clinical reasoning into accepting it as a legitimate diagnostic hypothesis, with potential downstream consequences for unnecessary confirmatory testing and reinforcement of false beliefs. This output illustrates why the fabrication failure mode is qualitatively distinct from — and arguably more concerning than — the vagueness or non-engagement failure modes also observed in the present study.

Notes on output integrity and data availability

All outputs were recorded verbatim, without editing or interpretation, at the time of model interaction. Each case was administered in an independent chat session with the model's conversational state explicitly reset, ensuring full independence between trials and eliminating carryover effects (Section 2.5 of the main manuscript).

The complete corpus of 56 model outputs (8 cases × 7 models) underlying the analyses reported in the main manuscript is retained by the authors and may be made available upon reasonable request, subject to compliance with applicable Brazilian data-protection regulations.