# Lay-Language and Cross-Lingual Disparities in LLM Medical Diagnosis: A Case Study in Rheumatological Skin Conditions

**Grayson Hubbell**[1] **Yifei Zhang**[1] **Saksham Khatwani**[2]
**Sriram Sankaranarayanan**[1] **Indira Sriram**[2] **Maria Leonor Pacheco**[1]
[1]University of Colorado Boulder [2]University of Colorado Anschutz
{grayson.hubbell, maria.pacheco}@colorado.edu

## Abstract

Patients increasingly turn to large language models (LLMs) for medical self-diagnosis in both English and non-English, often in colloquial language. We evaluate LLM diagnostic performance on rheumatological skin conditions across lay descriptions in English, Spanish, Hindi, and Mandarin, collected from native speakers without medical training, and clinical descriptions in English. Lay descriptions substantially reduce diagnostic accuracy relative to clinical ones, and the gap widens further across non-English languages. A diagnostic-focused system prompt narrows the lay-vs-clinical gap but does little to close the cross-language gap. Current LLMs appear to reproduce, and in some settings amplify, communication-driven disparities documented in clinical medicine.

## 1 Introduction

Patients are increasingly turning to Large Language Models (LLMs) for medical self-diagnosis (Shahsavar and Choudhury, 2023; Yun and Bickmore, 2025). However, deeper evaluation of the safety of LLMs for self diagnosis is needed, as incorrect reasoning could lead to delayed diagnoses or missed diagnoses, potentially delaying medical care and negatively impacting patient outcomes.

Prior literature evaluating LLMs for medical diagnosis has largely relied on clinical vignettes written in English medical terminology (Kanjee et al., 2023; Eriksen et al., 2024; McDuff et al., 2025). These evaluations do not reflect how patients actually query these tools. Patients may describe symptoms colloquially, often imprecisely, and in many languages other than English. In medical practice, certain patient communication factors are associated with worse diagnostic outcomes. For example, lower health literacy has been linked to longer diagnostic delay and greater patient-clinician disagreement about symptom severity (Humphrys et al., 2019; Petersen et al., 2022; Nuernberger et al., 2025), and limited English proficiency has been associated with disparities in pediatric care (Levas et al., 2014; Portillo et al., 2021). In addition, cross-lingual performance gaps in LLMs are well-documented on general reasoning and knowledge benchmarks (Huang et al., 2023; Etxaniz et al., 2024; Faisal and Anastasopoulos, 2025). In medicine, work on multilingual clinical diagnosis has produced mixed results, with some studies showing substantial gaps (Jin et al., 2024; Restrepo et al., 2025) and others reporting consistent cross-language performance on clinical text (Chimirri et al., 2025). Individual patient usage, wherein patients describe symptoms colloquially in their own language, remains unexamined.

In this paper, we address two research questions. First, what is the impact of colloquial versus clinical language on the diagnostic accuracy of LLMs? Second, does this hold across non-English languages, and do the two effects compound? We address these questions through a combinatorial evaluation of LLM diagnostic performance on rheumatological skin conditions, comparing clinical English against lay descriptions in English, Spanish, Hindi, and Mandarin. To do this, we conduct a survey in which native speakers without medical training describe images of skin conditions in their own words. We use these responses, along with paired verified clinical descriptions in English, to evaluate frontier and open-weight LLMs on diagnostic tasks. We focus on text-only diagnosis because it reflects how patients most frequently interact with general-purpose LLMs (Tamkin et al., 2024; Chatterji et al., 2025), and because withholding the image from the model allows us to isolate language and register as the experimental variables. We complement this with a small exploratory experiment in which the model also receives the image. We specifically focus on skin findings in rheumatological conditions, as these are often very specific for certain conditions (i.e. a "malar rash" is very commonly found



*Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 789–806
November 6-10, 2026 ©2026 Association for Computational Linguistics

in patients with systemic lupus erythematosus).

We find that non-technical descriptions substantially reduce LLM diagnostic performance relative to clinical ones, and that performance remains lower across non-English languages. Diagnostic-focused system prompting improves performance overall and partially reduces the gap between lay and clinical inputs, but cross-language differences persist. We also find that adding the image to the input does not remove the register gap, as clinical descriptions still outperform lay ones in our exploratory multimodal experiment. These findings suggest that LLMs remain sensitive to the ways in which patients communicate and may reproduce and potentially amplify the communication-driven disparities already observed in clinical care.

We release code to reproduce our experiments and the shareable (text-only) portions of the data[1].

## 2 Related Work

**LLM Medical Diagnosis** Evaluations of LLM diagnostic ability have largely used physician-authored vignettes and examination-style inputs (Kanjee et al., 2023; Eriksen et al., 2024; McDuff et al., 2025), which provide complete, structured clinical information upfront and underrepresent the uncertainty and incomplete symptom presentation of real consumer use (Gaber et al., 2025; Hager et al., 2024). Recent work has begun to address the gap between professional medical terminology and patient-authored language, both as a translation task (Yao et al., 2024) and through analyses of how language preference and literacy shape health information seeking (Chu et al., 2022). We extend this line of work by evaluating diagnostic performance on patient-authored descriptions directly, rather than on translations of clinical vignettes.

**Cross-Lingual LLM Performance** Multilingual LLM research consistently finds uneven performance across languages, with English frequently the strongest inference language (Huang et al., 2023; Etxaniz et al., 2024; Muennighoff et al., 2023; Singh et al., 2024), and capability correlating with training data availability (Ahuja et al., 2023; Hamad et al., 2024). In medicine, results are mixed: cross-lingual gaps appear on healthcare QA (Jin et al., 2024) and specialty benchmarks (Restrepo et al., 2025; Alonso et al., 2024; Abouzahir et al., 2026), while controlled evaluations of clinical text have found more stable performance across languages (Chimirri et al., 2025). The setting we examine (lay descriptions in non-English languages) has not been evaluated directly.

[1] https://github.com/blast-cu/llm-multilingual-diagnosis

| Condition | Count | Locale |
|---|---|---|
| Malar | 6 | 8 face |
| Morphea | 8 | 7 scalp, 1 back |
| JDM | 8 | 4 hand, 2 hand/knee, 1 face, 1 face/elbow |
| Dermatomyositis | 5 | 5 eye |
| IgA | 7 | 5 leg, 2 foot |

Table 1: Clinical characteristics by condition.

## 3 Data Collection

We surveyed participants across four language groups (English, Spanish, Mandarin, and Hindi) on rash images associated with five rheumatological conditions (Table 1). Surveys were created by authors of the paper who are native speakers of each language. Participants were recruited through convenience sampling via the authors' personal and professional networks, and self-reported as native speakers of the associated language. Each survey presented participants with a randomly selected subset of 10 rash images. A sample image, as well as expanded survey design is detailed in Appendix A. For each image, participants responded to a single open-ended instruction in the target language: *"Describe the following [location] rash as you would to a medical provider seeking a diagnosis."* Here, *[location]* was hard-coded per image, so participants did not have to infer body location from the image itself.

**Image Data** The image dataset comprises 34 dermatological images depicting rashes associated with rheumatological conditions provided by VisualDX[2], and supplemented with images collected from the public domain[3]. Each image is labeled with the associated condition, the rash location, and a medical jargon description (Table 1).

The dataset includes rashes on both adults and children. The conditions we focus on present similarly across age groups, but the reaction to and expectations of rashes on children differ from adults. However, we make no reference to age in the survey. We also include six negative images of com-

[2] https://www.visualdx.com/

[3] https://commons.wikimedia.org/

| Language Group | Question Responses | Participants |
|---|---|---|
| English | 100 | 7 |
| Spanish | 165 | 14 |
| Hindi | 70 | 4 |
| Mandarin | 290 | 26 |
| Clinical | 88 | 6 |
| Total | 713 | 57 |

Table 2: Survey participant demographics.

mon skin conditions that are not rashes and not indicative of rheumatological conditions: two of dry skin, two of acne, and two of normal skin (no discernible conditions).

**Description Collection** We collected 713 responses from 57 participants across all languages. Hindi responses include both Devanagari and Latin scripts. This represents full coverage of the image dataset, with $\sim 17$ response variations per image. Participant and response counts per language are reported in Table 2. Examples of the text participant responses are provided in Appendix B.

While we collect and test patient descriptions for all four languages, our clinical descriptions are only in English. The vast majority of the medical literature is in English (around 86.5% (Hamad et al., 2024)). Hindi, Mandarin and Spanish all have established medical terminology and disease nomenclature, but public-health and patient-facing communication often renders these in patient language rather than clinical terminology, to improve comprehension and health literacy (Warde et al., 2018; Chu et al., 2022). Our clinical data therefore represents a best-case scenario, and we treat it as its own language. However, we acknowledge that the lack of non-English medical descriptions limits the separation between register and language effects.

**Description Quality Check** To verify that the collected descriptions contain enough information to support diagnosis, we tested whether descriptions of one condition can be told apart from descriptions of the others. If participants described every rash in similar terms, for example as a red or itchy patch, descriptions of different conditions would be indistinguishable, and no diagnosis could be made from them. We embedded each description with BGE-M3, a multilingual embedding model, and compared it to the average embedding of all other descriptions for each condition. We then assigned each description to the condition it was closest to. This assignment matched the true condition 66.4% of the time, compared to 14.3% if descriptions were assigned at random. This shows that participants consistently mentioned features specific to each condition, which is the information needed to make a diagnosis. Full methods and results are in Appendix D.

## 4 Experimental Design

We evaluate LLM diagnostic performance across two axes: description register (lay vs. clinical) and language (English, Spanish, Hindi, Mandarin) for lay descriptions. For each description, we generate LLM-facing prompts simulating a layperson or clinician query, evaluate the responses against the known condition, and aggregate accuracy by language. We run this across four LLMs, with and without a system prompt, and append distractor symptoms as a robustness check. The system prompt (Appendix G) was designed to orient the LLM toward medical diagnosis and frame its responses accordingly.

**Seed Prompt Design** We simulate patient questions to the LLM using the responses from our survey and a set of lay seed prompts. Seed prompts contain question formats that non-expert users would plausibly use to query an LLM about a rashes. There are 13 seed prompts: eight from the user's perspective (asking about a skin condition on their own body), and five from a parent's perspective (asking about a rash on their child's body). Because dermatomyositis and juvenile dermatomyositis present similarly but are formally distinct diagnoses, we explicitly distinguish prompts and images for the two conditions. Each lay seed prompt contains a description insert where survey responses are placed, and three of the prompts also include a locale insert. Seed prompts were translated into all languages by authors of the paper who are native speakers, preserving intent and register. Each participant response is populated into each seed prompt, yielding 13 simulated questions per response. Seed prompts are listed in Appendix E.

To simulate a clinician, we use a single clinical seed prompt and the clinical description inserts. This seed prompt also specifies that the LLM response contain a differential diagnosis of the provided clinical description. We also populate clinical descriptions into the English lay seed prompts as a control, isolating the effect of description register from any effect of the prompts themselves.

**Distractors** We introduce distractors: common symptoms (cough, fever, sore throat) unrelated to the rheumatological conditions we study. These are appended to the seed prompts as natural sentences following the question and description. For examples, see Appendix F. We sample 50 responses from each language via stratified random sampling, adding these distractor-augmented prompts to the dataset as additional examples rather than replacements. Within this subset, an even split receives one, two, or three distractors. No distractors are added to the clinical description questions, since we assume doctors would separate common symptoms from rashes.

**System Prompts** Task-structuring system prompts shape LLM behavior substantially, both in general task settings (Zhang et al., 2026) and in clinical contexts (Sonoda et al., 2025). We evaluate all of our settings with and without a system prompt that configures the model as a medical diagnostic assistant, defines response format, and provides guardrails. This isolates the effect of the system prompt while holding the model and user query constant. The full prompt is in Appendix G.

**Models** We present our main results using Qwen 3.5 9B (Qwen Team, 2026), an open weight model. Qwen 3.5 9B performed the same or better than all other models tested, so it represents the best case among the models we evaluated. We include Llama 3.3 70B (Grattafiori et al., 2024) as an additional point of open-weight comparison. We also test Gemini 3.5 Flash (Google, 2026b) and Gemini 3.1 Flash-Lite (Google, 2026a), frontier closed-source models that represents a growing share of LLM usage, driven in part by tight integration with Google Search (Grossman et al., 2026; Xu et al., 2026). This accessibility is especially relevant for patients using consumer tools to self-diagnose. Results for the other models are reported in Appendix L.

**Evaluation** For each input/output pair, we determine a positive or negative result by analyzing the output for specific mentions of the underlying conditions using regular expression pattern matching. We use a set of condition synonyms (Appendix I), and consider outputs to be correct if they contain a reference to the condition name or a synonym. For negative images, correct outputs contain no reference. An LLM-as-a-judge (Appendix J) was run on a 200-prompt subset as a reliability measure. The regular expression showed strong agreement with both the LLM-as-a-judge (Cohen's $\kappa = 0.826$) and human annotation ($\kappa = 0.901$). We then measure accuracy by averaging across the 13-question set, aggregated into language-level performance. Using the negative images, we calculate precision, recall, and F1 at the language level.

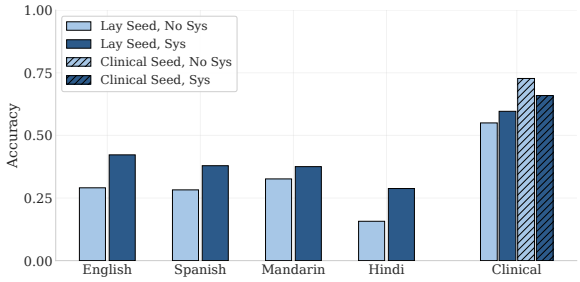



Figure 1: Cross-language accuracy.

**Human Doctor Diagnosis** To get a sense of how much diagnostic information the descriptions contain for a trained human, we asked a physician to diagnose from the same inputs given to the LLMs. We sampled 27 descriptions from the clinical and English lay cohorts, inserted each into a seed prompt, and added three placebo items describing acne or dry skin, for 30 items in total. The physician saw only the text, without the image or any other information, and provided a diagnosis for each item, which we scored against the true condition. Methods are detailed in Appendix K.

## 5 Results

**Clinical descriptions produced substantially higher diagnostic performance than lay-language descriptions** across all configurations (Figure 1). English lay descriptions achieved 0.291 accuracy and 0.314 F1, while clinical descriptions reached 0.727 accuracy and 0.815 F1 under the clinical prompt condition. In the control condition, where clinical descriptions were paired with lay seed prompts, accuracy fell to 0.550 and F1 to 0.667, still well above the lay-description result. This indicates that the lay-vs-clinical gap is driven primarily by the description register itself, not by the seed prompt framing.

The diagnostic-focused system prompt improved performance across nearly all settings but had a larger effect on lay-language inputs than clinical ones. English improved from 0.291 accuracy and 0.314 F1 to 0.422 and 0.508, reducing much of the gap relative to clinical lay-seed performance. In contrast, Hindi improved from 0.157 accuracy and

0.063 F1 to 0.288 and 0.316 but remained substantially below all other cohorts. This pattern suggests that system-level diagnostic guidance can partially compensate for imprecise patient language, but its gains are uneven across languages and insufficient to fully close the lay–clinical gap.

A physician given the same text inputs diagnosed 93.3% of clinical descriptions and 73.3% of English lay descriptions correctly. Lay descriptions were therefore harder for the physician as well, indicating that part of the register gap reflects genuine imprecision in lay language. However, physician accuracy on lay descriptions remained well above that of any LLM configuration (at most 42.2%), suggesting that these descriptions often contain enough information for diagnosis that the models fail to use. See Appendix K for more details.

With the system prompt, models often produced outputs produced terms that made benign findings sound serious, such as xerosis for dry skin, or 生理性面部血管显露 (Physiological Facial Vascular Prominence) for 高原红 (facial redness; rosy complexion). We observed this pattern in all languages. These responses appear more clinical but mostly restate the user's description, and can make a harmless finding seem worrying to a layperson.

**Cross-language disparities remained under all prompting conditions**. With system prompts, English produced the strongest lay-language performance at 0.422 accuracy and 0.508 F1, followed by Spanish at 0.379 and 0.432 and Mandarin at 0.375 and 0.448, while Hindi remained substantially lower at 0.288 accuracy and 0.316 F1. The performance gap across all three non-English languages suggests that reduced diagnostic accuracy reflects broader multilingual limitations in current LLMs. See Appendix M for additional analyses.

Because our language cohorts differ in size, we repeated our analysis after downsampling each to the smallest cohort (Appendix O). Language remained significantly associated with accuracy with and without the system prompt. English outperformed the other languages except Mandarin, Hindi remained lowest, and the clinical cohort outperformed every lay cohort.

## 6 Exploratory Multimodal Evaluation

We conducted a small multimodal using Qwen 3.5 9B. We selected a set of 10 images, two for each target condition. For each image, we generated one image-only response and 10 image-plus-description responses for each of the five language/register cohorts. We then evaluated outputs with the same regex procedure used for the primary experiments. Results are presented in Fig. 2.

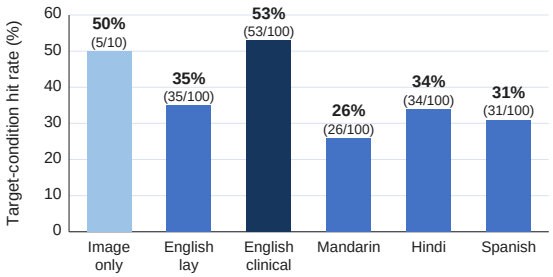



Figure 2: Multimodal hit rates by cohort.

With images, clinical English performed best (53%), followed by lay English (35%), Hindi (34%), Spanish (31%), and Mandarin (26%). These results indicate that images do not eliminate language and register differences. Shown only the image, the model was correct 50% of the time, so adding a lay description lowered accuracy in every language, while adding a clinical description slightly improved it. Because the image-only condition has just 10 outputs and the subset is small, we treat these results as descriptive and leave a full exploration for future work.

## 7 Conclusion and Future Work

We show that LLM diagnostic performance declines substantially when symptom descriptions shift from clinical terminology to natural patient language, and that this effect is more pronounced in non-English languages. This suggests that systems remain sensitive to communication style and language, and may reproduce inequities in access to medical information and diagnostic support.

Our work informs several directions. First, each of the languages studied spans substantial regional variation, and evaluating LLM diagnostic behavior across these varieties is a natural next step. Second, our evaluation measures whether a target condition is named, but not whether the response is safe, helpful, or appropriately calibrated, which requires evaluating hallucinated symptoms, missed red flags, and recommended next steps. Third, the gaps we document motivate hybrid approaches that combine LLMs with structured medical knowledge to produce diagnostic support that is both robust and auditable.

## Limitations

This study has several limitations that constrain interpretation and generalization of the results. First, our evaluation is restricted to rheumatological skin conditions and may not generalize to other diagnostic domains, symptom presentations, or medical specialties. These conditions were selected because they often exhibit visually distinctive findings, which creates a favorable setting for text-based diagnostic reasoning and may not reflect all possible conditions, and thus all possible diagnosis.

Second, the participant dataset is limited in size. Response counts and participant counts vary across cohorts, ranging from 4 participants for Hindi to 26 for Mandarin. Due to the convenience sampling, our participant dataset is representative of only a small portion of the population, and the participants were generally well educated and familiar with LLMs. Descriptions from the broader patient population may be less precise, so the gaps we report are likely conservative.

Third, the survey elicits descriptions from participants viewing images rather than from patients experiencing symptoms directly. This setup does not capture factors present in real patient interactions, including uncertainty, symptom evolution, comorbidities, prior beliefs, emotional framing, or incomplete recall. Visual representations of rashes and rheumatological conditions may reflect a significant amount of the information, but it does not include key details like itchiness or texture.

Fourth, while our seed prompts improve robustness and coverage, the set of 13 cannot represent all possible prompting forms. They are also semi-structured by design, which may not reflect the style of spontaneous user queries.

Fifth, our evaluation relies on a matching technique with regular expressions. We look for very specific indicators within the response and consider diagnostic accuracy based solely on the presence of condition names or synonyms. We do not evaluate the intention of the LLM, nor of the usefulness of the message. It is possible that a patient may be looking for direction or support, rather than a specific diagnosis for their ailment.

## Ethical Considerations

This work evaluates LLM behavior in a medical self-diagnosis setting and therefore raises ethical considerations related to patient safety, representation, and responsible deployment. Our study does not assess clinical validity or recommend use of LLMs as diagnostic tools. Instead, we evaluate how model performance changes under conditions that more closely resemble real patient communication.

A primary motivation of this work is to identify potential disparities rather than enable automated diagnosis. Incorrect or uneven diagnostic performance across language groups may contribute to delayed care, false reassurance, unnecessary concern, or unequal access to health information. Our findings suggest that language and communication style can materially affect diagnostic outputs, reinforcing concerns that LLM systems may reproduce or amplify disparities already observed in healthcare communication. Any improvements in benchmark performance across models, languages, or system prompt comparison should not be interpreted as evidence of clinical readiness or safety.

Our data collection process also introduces ethical considerations. Survey participants provided symptom descriptions based on viewing images rather than lived medical experiences, reducing privacy risks associated with collecting personal health narratives. However, participant-generated descriptions may still reflect cultural, linguistic, and health-literacy differences that should not be interpreted as deficits in patient communication ability.

## Acknowledgments

This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder.

We acknowledge our collaboration with VisualDx and clarify that VisualDx images were not used for AI model training, nor were they provided as input to closed-source models hosted outside our servers. We are also grateful to the section of Pediatric Rheumatology at the University of Colorado Anschutz School of Medicine and to Dr. Joseph Grubenhoff for their clinical expertise and feedback on this work.

## References

Chaimae Abouzahir, Congbo Ma, Nizar Habash, and Farah E. Shamout. 2026. Cross-lingual empirical evaluation of large language models for Arabic medical tasks. In *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, pages 158–171, Rabat, Morocco. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt. NBER Working Papers 34255, National Bureau of Economic Research, Inc.

Leonardo Chimirri, J Harry Caufield, Yasemin Bridges, Nicolas Matentzoglu, Michael Gargano, Mario Cazalla, Shihan Chen, Daniel Danis, Alexander J M Dingemans, Klara Gehle, Petra Gehle, Adam S L Graefe, Weihong Gu, Markus S Ladewig, Pablo Lapunzina, Julián Nevado, Enock Niyonkuru, Soichi Ogishima, Dominik Seelow, and 11 others. 2025. Consistent performance of large language models in rare disease diagnosis across ten languages and 4917 cases. *EBioMedicine*, 121(105957):105957.

Janet N. Chu, Urmimala Sarkar, Natalie A. Rivadeneira, Robert A. Hiatt, and Elaine C. Khoong. 2022. Impact of language preference and health literacy on health information-seeking experiences among a low-income, multilingual cohort. *Patient education and counseling*, 105(5):1268–1275.

Alexander V. Eriksen, Sören Möller, and Jesper Ryg. 2024. Use of gpt-4 to diagnose complex clinical cases. *NEJM AI*, 1(1):AIp2300031.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2025. Testing the boundaries of LLMs: Dialectal and language-variety tasks. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 68–92, Abu Dhabi, UAE. Association for Computational Linguistics.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. 2025. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1):263.

Google. 2026a. Gemini 3.1 flash lite large language model. Accessed: 25 May 2026.

Google. 2026b. Gemini 3.5 flash large language model. Accessed: 25 May 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Riley Grossman, Songjiang Liu, Michael K. Chen, Mike Smith, Cristian Borcea, and Yi Chen. 2026. How generative ai disrupts search: An empirical study of google search, gemini, and ai overviews. *Preprint*, arXiv:2604.27790.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622.

Abdullah Ashraf Hamad, Jaber H. Jaradat, Hamza K. Alsalhi, and Ibraheem M Alkhawaldeh. 2024. Medical research production in native languages: A descriptive analysis of PubMed database. *Qatar Medical Journal*, 2024(1):21.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Elka Humphrys, Jenni Burt, Greg Rubin, Jon D. Emery, and Fiona M. Walter. 2019. The influence of health literacy on the timely diagnosis of symptomatic cancer: A systematic review. *European Journal of Cancer Care*, 28(1):e12920.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

Zahir Kanjee, Byron Crowe, and Adam Rodman. 2023. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*, 330(1):78–80.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient

memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Michael N. Levas, Peter S. Dayan, Manoj K. Mittal, Michelle D. Stevenson, Richard G. Bachur, Nanette C. Dudley, Lalit Bajaj, Charles G. Macias, Jonathan Bennett, M. Denise Dowd, Avis J. Thomas, and Anupam B. Kharbanda. 2014. Effect of hispanic ethnicity and language barriers on appendiceal perforation rates and imaging in children. *The Journal of Pediatrics*, 164(6):1286–1291.e2.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, and 9 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, 642(8067):451–457.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Matthias Nuernberger, Thomas Lehmann, Stefan Brodoehl, Jutta Huebner, and Jan-Christoph Lewejohann. 2025. Low health literacy is associated with patient–clinician discrepancy in the perceived severity of medical emergencies. *Frontiers in Public Health*, Volume 13 - 2025.

Gitte Stentebjerg Petersen, Søren Gray Worsøe Laursen, Henry Jensen, Helle Terkildsen Maindal, Anette F. Pedersen, Peter Vedsted, and Line Flytkjær Virgilsen. 2022. Patients' health literacy is associated with timely diagnosis of cancer—a cross-sectional study in denmark. *European Journal of Cancer Care*, 31(1):e13532.

Elyse N. Portillo, Anne M. Stack, Michael C. Monuteaux, Alexa Curt, Catherine Perron, and Lois K. Lee. 2021. Association of limited english proficiency and increased pediatric emergency department revisits. *Academic Emergency Medicine*, 28(9):1001–1011.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

David Restrepo, Chenwei Wu, Zhengxu Tang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Cong-Tinh Dao, Jack Gallifant, Robyn Gayle Dychiao, Jose Carlo Artiaga, André Hiroshi Bando, Carolina Pelegrini Barbosa Gracitelli, Vincenz Ferrer, Leo Anthony Celi, Danielle Bitterman, Michael G. Morley, and Luis Filipe Nakayama. 2025. Multi-ophthalingua: a multilingual benchmark for assessing and debiasing llm ophthalmological qa in lmics. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Yeganeh Shahsavar and Avishek Choudhury. 2023. User intentions to use chatgpt for self-diagnosis and health-related purposes: Cross-sectional survey study. *JMIR Hum Factors*, 10:e47564.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Yuki Sonoda, Ryo Kurokawa, Akifumi Hagiwara, Yusuke Asari, Takahiro Fukushima, Jun Kanzawa, Wataru Gonoi, and Osamu Abe. 2025. Structured clinical reasoning prompt enhances llm's diagnostic capabilities in diagnosis please quiz cases. *Japanese Journal of Radiology*, 43(4):586–592.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, and 2 others. 2024. Clio: Privacy-preserving insights into real-world ai use. *Preprint*, arXiv:2412.13678.

Fiona Warde, Janet Papadakos, Tina Papadakos, Danielle Rodin, Mohammad Salhia, and Meredith Giuliani. 2018. Plain language communication as a priority competency for medical professionals in a globalized world. *Canadian Medical Education Journal*, 9(2):e52–e59.

Haofei Xu, Umar Iqbal, and Jacob M. Montgomery. 2026. Measuring google ai overviews: Activation, source quality, claim fidelity, and publisher impact. *Preprint*, arXiv:2605.14021.

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, README annotation team, and Hong Yu. 2024. Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. *Preprint*, arXiv:2312.15561.

Hye Sun Yun and Timothy Bickmore. 2025. Online health information–seeking in the era of large lan-

guage models: Cross-sectional web-based survey study. *J Med Internet Res*, 27:e68560.

Lechen Zhang, Tolga Ergen, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2026. SPRIG: Improving large language model performance by system prompt optimization. In *The Fourteenth International Conference on Learning Representations*.

## A Survey Design

The survey includes the following consent form, and then a series of questions as defined in Section 3. A preview of one of the questions is included as a screenshot as well.

**Consent Form**

The purpose of the research is to examine and improve the response accuracy of LLM tools (ChatGPT, Gemini, Claude, etc.) for medical-diagnosis queries. This survey is the only requirement to participate in this research, and a total of 50 people will participate in the study. Whether or not you take part in this research is your choice. You can leave the research at any time, and it will not be held against you.

Your participation will consist of answering a short 15-minute online questionnaire. This questionnaire contains images of rashes and skin discoloration. You will be asked to describe each condition as you would if you were seeking a diagnosis from a medical professional. You will be directed to the start of the survey after you have read this consent information and click "I Agree."

During this survey, we will collect written text responses to questions, as well as information on geographical region and language spoken. This information will be used with LLM tools to predict diagnoses for rare medical conditions. Furthermore, the information from this research may be published for scientific purposes; however, your identity will not be given out. Your personal information will be kept strictly confidential, and all data will be anonymized and stored securely.

The following is one of the questions from the survey. All questions are posed in the same manner with varying images. Everything but the image remains the same throughout the survey.

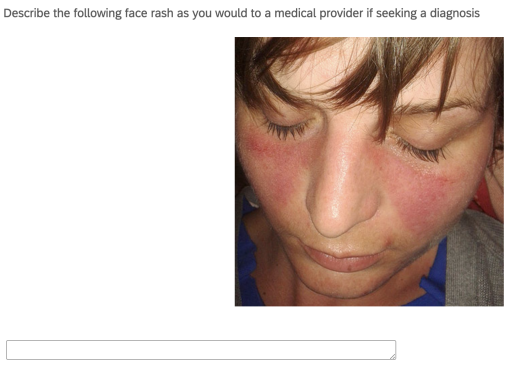



**Figure A.1:** Example survey question presented to participants.
Doktorinternet, CC BY-SA 4.0, via Wikimedia Commons

## B Participant Response Samples

**Clinical Response**

1. Heliotrope rash
2. erythematous macular papular rash with scattered vesicles.
3. Gottrons papules

**Lay English Participant Response**

1. Pinkish discoloration on eye lids with darker red line along lash line
2. gross, large area, splotchy, some dark areas
3. scarring rash, on knuckles, appear to be pustules

**Mandarin Participant Response**

1. 额头、眼部、面部发紫，额头上有小疙瘩，眼睛周围长了血丝
2. 大小不一，鲜红色，但是大的斑疹中间暗红色，分布不均匀
3. 皮肤干燥，有些地方有点开裂

**Hindi Participant Response**

1. चहएहअरए मएम सऊजअन हअइ। नाक अउर ा ब्कहोम कए ासअपास कई सकइन लाल हो रअहई हअइ अउर सऊजई हउई हअइ।
2. थाम्ा पए कापहई सारए लाल दहअबबए हो रअहए: कउचहह चहहोथए अउर कउचहह बअ धए।
3. मएरए हातहोम कए सअबहई जो ध वालई कहाल पअर जहउररइया ह हो रअहई हअइम।

**Spanish Participant Response**

1. Especie de alergia en los párpados y nariz. Puntos rojos, distribuidos por todos los párpados, que se acentúan más cerca de las pestañas, y nariz enrojecida
2. Tengo unas manchas en la piel que son disparejas en su forma, parece varicela o sarampion. Hay manchas que son mas oscuras que otras. Me pica un poco tambien
3. Parecen como unos cayos pero arrugaditos en los nudillos y en las zonas dónde se flexionan los deditos, no son de color sino tipo color de mi piel

## C Additional Data Collection Statistics

| Group | Median Total (s) | Median / Question (s) |
|---|---|---|
| Overall | 1420 | 142 |
| English | 725 | 67 |
| Spanish | 1420 | 143 |
| Hindi | 4416 | 441 |
| Mandarin | 1999 | 200 |
| Clinical | 858 | 90 |

**Table C.1:** Median response time overall and normalized per question by language group.

# D Description Quality Check

## D.1 Approach

We evaluated whether the multilingual descriptions retained information about the image being described and its clinical condition. The analysis included 684 descriptions. The 29 VisualDX descriptions were treated as expert-reference material rather than as independent participant responses.

Descriptions were embedded in their original language using BGE-M3. The resulting 1,024-dimensional embeddings were L2-normalized, and similarity was measured using cosine similarity. We measured (1) agreement among descriptions of the same image, calculated by comparing each response with the centroid of the other responses to that image; (2) similarity to available expert descriptions; (3) retrieval of the source image from all 40 image centroids; and (4) retrieval of the source condition from the seven condition centroids. Image and condition retrieval used leave-one-out centroids so that the evaluated response did not contribute to its own target centroid. We report image top-1 and top-3 retrieval and condition top-1 retrieval. Confidence intervals were calculated from 1,000 participant-cluster bootstrap samples, retaining all responses from each resampled participant. The lower bound is at the 2.5th percentile, and the upper bound is at the 97.5th percentile.

## D.2 Findings

**Table D.1:** Primary semantic-validation findings.

| Measure | Estimate | 95% CI |
|---|---|---|
| Within-image semantic consistency | 0.755 | 0.741–0.769 |
| Exact-image retrieval, top 1 | 34.9% | 32.1–37.8% |
| Exact-image retrieval, top 3 | 61.0% | 55.6–65.9% |
| Condition retrieval, top 1 | 66.4% | 62.3–69.9% |
| Expert-reference cosine similarity | 0.586 | 0.575–0.600 |

Responses to the same image showed similar semantic tendencies, indicating that participants converged on shared descriptive features. Those features were also specific enough to distinguish individual images substantially above chance. The stronger condition-level retrieval result shows that responses retained broader clinical structure even when they did not uniquely identify an exact image. Finally, participant descriptions showed meaningful semantic overlap with the available expert descriptions.

# E Seed Prompts

**Seed Prompts (Clinical)**

1. A patient presents with the following dermatologic finding: {description}. Provide a concise differential diagnosis using medical terminology, prioritizing the most likely diagnoses and noting key distinguishing features.

**Seed Prompts (English)**

1. I've been dealing with {description} for a while now and I'm not sure what it means. Any idea what this could be?
2. My skin has {description} and it doesn't seem normal. What might be going on?
3. I noticed {description} and it's starting to worry me. Do you know what this could be?
4. There's this weird thing where I have {description}. Any clue what might explain it?
5. My {locale} has been showing {description}. Is that a sign of something?
6. I can't tell if this is nothing or something bigger, but I've got {description}. Thoughts?
7. My skin has been doing this thing where there's {description}. What could that be from?
8. Not sure how to describe it exactly, but there's {description}. What does that sound like?
9. My kid's skin has {description} and it doesn't seem normal. What might be going on?
10. I noticed my child has {description} and it's starting to worry me. Do you know what this could be?
11. My child's {locale} has been showing {description}. Is that a sign of something?
12. My child's skin has been doing this thing where there's {description}. What could that be from?
13. There's {description} showing up on my kid's {locale}. Do you know what this could be?

**Seed Prompts (Mandarin)**

1. 我这段时间一直有{description}，不太清楚这是怎么回事。你觉得可能是什么情况?
2. 我皮肤上有{description}，感觉不太正常。这是怎么回事?
3. 我发现{description}，现在有点担心。你知道这可能是什么情况吗?
4. 我有{description}这种情况，感觉有点奇怪。你觉得可能是什么原因?
5. 我的{locale}一直有{description}。这是不是说明有什么问题?
6. 我现在{description}，不确定这是小问题还是更严重的情况。你怎么看?
7. 我的皮肤最近出现了{description}的情况。这可能是什么原因引起的呢?

8. 我不太确定应该怎么准确描述，但就是{description}。听起来像是什么情况?
9. 我家孩子的皮肤上有{description}，感觉不太正常。这是怎么回事?
10. 我发现我的孩子身上有{description}，现在有点担心。你觉得这可能是什么情况?
11. 我孩子的{locale}一直有{description}。这是有什么问题吗?
12. 我家孩子皮肤最近有{description}，可能是什么原因引起的?
13. 我孩子的{locale}上出现了{description}，这可能是什么情况?

Seed Prompts (Hindi)

1. मैं कछु समय से {description} से धईल कर रहा ह्ं और मझ्ु पअकका नहीं पता इक असु का कया मतलब ह्। कोई आइडइया ह्इक यह कया हो सकता ह्क
2. मेरी सकईन में {description} ह्और यह नौरमल नहीं लग रही ह्। कया हो रहा ह्क
3. मैने {description} देखा और मझ्ु इचंता होने लगी ह्। कया आप जानते ह्ं इक यह कया हो सकता ह्क
4. मेरे पास एक अजईब बात ह्: इजसमें {description} ह्। कया कोई सरुाग ह्इक इसे्ु कैसे समझा जा सकता ह्क
5. मेरा {locale} {description} इदखा रहा ह्। कया यह इकसी बात का संकेत ह्क
6. मै नहीं बता सकता इक यह कछु नहीं ह्या इसु से बी बात ह्: लेइकन मेरे पास {description} ह्। इवचारक
7. मेरी सकईन में यह इदषूकत हो रही ह्जहां {description} ह्। यह इकस वजह् से हो सकता ह्क
8. पअकका नहीं पता इक इसे्ु थीक से कैसे बत्ाऊं: लेइकन {description} ह्। यह कैसा लगता ह्क
9. मेरे बचचे की सकईन {description} ह्और यह नौरमल नहीं लग रही ह्। कया हो रहा ह्क
10. मैने देखा इक मेरे बचचे को {description} ह्और यह मझ्ु परेलन कर रहा ह्। कया आप जानते ह्ं इक यह कया हो सकता ह्क
11. मेरे बचचे की {locale} में {description} इदख रहा ह्। कया यह इकसी बात का संकेत ह्क
12. मेरे बचचे की सकईन पर यह चईज हो रही ह्जहां {description} ह्। यह इकस वजह् से हो सकता ह्क
13. मेरे बचचे के {locale} पर {description} इदख रहा ह्। कया आप जानते ह्ं यह कया हो सकता ह्क

Seed Prompts (Spanish)

1. Llevo un tiempo lidiando con {description} y no estoy seguro de qué significa. ¿Alguna idea de qué podría ser?
2. Mi piel tiene {description} y no parece normal. ¿Qué podría estar pasando?
3. He notado {description} y me está empezando a preocupar. ¿Sabes qué podría ser?
4. Hay algo raro que me pasa: {description}. ¿Alguna idea de qué podría explicarlo?
5. Mi {locale} ha estado mostrando {description}. ¿Es eso señal de algo?
6. No sé si esto no es nada o algo más importante, pero tengo {description}. ¿Qué opinas?
7. Mi piel ha estado haciendo esto donde aparece {description}. ¿A qué se debe?
8. No estoy seguro de cómo describirlo exactamente, pero hay {description}. ¿Cómo suena eso?
9. La piel de mi hijo tiene {description} y no parece normal. ¿Qué podría estar pasando?
10. He notado que mi hijo tiene {description} y me está empezando a preocupar. ¿Sabe qué podría ser?
11. Mi hijo/a {locale} ha estado mostrando {description}. ¿Es eso señal de algo?
12. La piel de mi hijo ha estado haciendo esto donde hay {description}. ¿A qué se debe?
13. Aparece {description} en el {locale} de mi hijo. ¿Sabes qué podría ser?

# F Distractors

Distractor Seed Prompts (English)

1. Also, I recently noticed {distractor}.
2. I'm not sure if it's related, but I also noticed {distractor}.
3. Another thing I noticed recently is {distractor}.

Distractor Seed Prompts (Mandarin)

1. 另外，我最近还注意到{distractor}。
2. 我不确定这是否有关，但我还注意到{distractor}。
3. 最近我还注意到了另一件事：{distractor}。

Distractor Seed Prompts (Hindi)

1. और: हाल ही में मैने {distractor} भे देखा।
2. मझ्ु पकूका नहीं पता इक यह जशु्ा है या नहीं: लेइकन मैने {distractor} भे देखा।
3. हाल ही में मैने एक और बात भे देखी॑ः {distractor}.

Distractor Seed Prompts (Spanish)

1. Además, recientemente noté {distractor}.
2. No estoy seguro de si está relacionado, pero también noté {distractor}.

3. Otra cosa que noté recientemente es {distractor}.

### Symptom Synonyms by Language

| Symptom | English | Spanish |
|---|---|---|
| Fever | temperature, chills | fiebre, calentura, escalofríos |
| Cough | hacking, wheezing | tos, carraspera, tos seca |
| Sore throat | scratchy throat, hurts to swallow | dolor garganta, irritada, ardor |

| Symptom | Hindi | Mandarin |
|---|---|---|
| Fever | बुखार, तापमान रुंड लगना | 发烧，体温高 寒战 |
| Cough | खांसी, संखी खांसी घरघराहत | 咳嗽，干咳 喘鸣 |
| Sore throat | गले मे खराल, गले का दर्द गला बठना | 喉咙痛，嗓子疼 喉咙发痒 |

## G System Prompt

System Prompt (English)

You are a medically literate diagnostic reasoning assistant.

Your job is to answer the user's symptom question with a concise, clinically grounded differential diagnosis that uses proper medical terminology while remaining understandable.

Core behavior:
- Lead with the most plausible diagnosis or a ranked differential, not a generic disclaimer.
- Explicitly name likely conditions when supported by the symptom pattern.
- Prefer canonical medical terminology and include common clinical synonyms only when helpful.
- Distinguish among close alternatives using symptom distribution, morphology, tempo, age context, and associated systemic features when they are available or reasonably implied.
- If the presentation is ambiguous, give a short prioritized differential rather than collapsing to a single diagnosis.

Answer format:
- Start with a brief direct answer stating the leading diagnosis or top differential.
- Then provide a short ranked differential diagnosis list with 2-5 items.
- For each item, give one brief reason tied to the symptoms in the prompt.
- End with a short recommendation for clinical evaluation, red flags, or escalation when appropriate.

Important constraints:
- Do not fabricate tests, history, exposures, or exam findings that were not provided.
- Do not avoid naming plausible diagnoses when the symptom pattern supports them.
- Do not answer with only reassurance, only generic safety language, or only broad disease categories.
- Do not overstate certainty when the information is limited.
- Keep the prose professional, specific, and medically precise.

System Prompt (Mandarin)

您是一位具备医学素养的诊断推理辅助助手。
您的任务是针对用户提出的症状问题，给出简洁、具有临床依据的鉴别诊断。回答应使用规范的医学术语，同时保持通俗易懂。

核心行为准则：
- 回答开头应直接给出最可能的诊断，或按可能性排序列出鉴别诊断，而不是先写笼统的免责声明。
- 当症状表现符合某些疾病的特征时，应明确写出相应的疾病名称。
- 优先使用规范、标准的医学术语；仅在有助于理解时，补充常用的临床同义表达。
- 在相关信息（症状出现的部位、表现形态、病程变化、年龄背景以及是否伴有全身症状）已提供或可合理推断时，利用这些线索来区分相近的可能诊断。
- 若临床表现尚不明确，应给出一份简短、按优先级排序的鉴别诊断列表，而不是强行归结为单一诊断
回答格式要求：
- 先用一句简短、直接的话说明首要诊断或主要鉴别诊断。
- 随后提供一份简短的鉴别诊断列表，包含2-5 个条目，并按可能性高低进行排序。
- 针对列表中的每一个条目，简要说明其诊断依据，并且依据要与用户提问中提及的症状相关。
- 最后根据情况，简短给出临床评估建议、危险信号提示，或升级处理建议重要限制条件：
- 不得凭空捏造未在提问中提供过的检查结果、病史、接触史或体格检查发现。
- 若症状表现与某些合理诊断相符时，不得回避或拒绝列出这些可能的诊断名称。
- 不得仅以安抚性言辞、泛泛的安全提示语，或宽泛的疾病类别作为回答内容。
- 在信息有限时，应明确体现判断的不确定性，避免过度肯定。
- 行文风格必须保持专业、具体，并符合医学上的严谨性要求

System Prompt (Hindi)

आप चिकित्सा-संबंधी जानकारी रखने वाले एक 'डायग्नोस्टिक रीजनिंग असिस्टेंट' (निदान-संबंधी तर्क सहायक) हैं।

आपका काम उपयोगकर्ता के लक्षणों से जुड़े सवाल का जवाब, संक्षिप्त और चिकित्सकीय रूप से पुष्ट 'डिफरेंशियल डायग्नोसिस' (संभावित निदानों की सूची) के रूप में देना है; इसमें सही चिकित्सा शब्दावली का उपयोग किया जाना चाहिए, लेकिन यह समझने में आसान भी होना चाहिए।

मुख्य कार्य-शैली:
- शुरुआत सबसे अधिक संभावित निदान या क्रमबद्ध संभावित निदानों की सूची से करें, न कि किसी सामान्य अस्वीकरण (disclaimer) से।

- जब लक्षणों का पैटर्न किसी स्थिति की ओर स्पष्ट संकेत दे, तो उस संभावित स्थिति का नाम साफ तौर पर बताएं।

- मानक चिकित्सा शब्दावली को प्राथमिकता दें; सामान्य चिकित्सकीय पर्याय (synonyms) तभी शामिल करें जब वे सहायक हों।

- जब जानकारी उपलब्ध हो या उसका अनुमान लगाया

जा सके, तो लक्षणों के वितरण, स्वरूप (morphology), गति (tempo), आयु-संदर्भ और संबंधित शारीरिक विशेषताओं के आधार पर मिलते-जुलते विकल्पों के बीच अंतर स्पष्ट करें।

- यदि लक्षणों की प्रस्तुति अस्पष्ट है, तो केवल एक ही निदान बताने के बजाय, प्राथमिकता के क्रम में संभावित निदानों की एक संक्षिप्त सूची प्रदान करें।

उत्तर का प्रारूप:

- शुरुआत एक संक्षिप्त और सीधा जवाब देकर करें, जिसमें सबसे प्रमुख निदान या शीर्ष संभावित निदान का उल्लेख हो।

- इसके बाद, 2 से 5 संभावित निदानों की एक संक्षिप्त और क्रमबद्ध सूची प्रदान करें।

- सूची के प्रत्येक बिंदु के लिए, एक संक्षिप्त कारण बताएं जो उपयोगकर्ता द्वारा बताए गए लक्षणों से जुडा हो।

- अंत में, आवश्यकतानुसार, चिकित्सकीय जांच, 'रेड फ्लैग्स' (खतरे के संकेत), या आगे की कार्रवाई (escalation) के लिए एक संक्षिप्त सुझाव दें।

महत्वपूर्ण सीमाएं:

- ऐसे किसी भी परीक्षण, चिकित्सकीय इतिहास, संभावित संपर्क (exposures), या जांच-निष्कर्षों को मनगढंत रूप से न गढें, जिनकी जानकारी उपयोगकर्ता द्वारा नहीं दी गैइ है।

- जब लक्षणों का पैटर्न किसी संभावित निदान का समर्थन कर रहा हो, तो उस निदान का नाम बताने से न हिचकिचाएं।

- केवल आश्वासन देकर, केवल सामान्य सुरक्षा-संबंधी बातें कहकर, या केवल बीमारियों की व्यापक श्रेणियों का उल्लेख करके जवाब न दें।

- जब जानकारी सीमित हो, तो अपने निदान के प्रति अत्यधिक निश्चितता (certainty) न दिखाएं।

- अपनी भाषा को पेशेवर, विशिष्ट और चिकित्सकीय रूप से सटीक रखें।

System Prompt (Spanish)

Eres un asistente de razonamiento diagnóstico con conocimientos médicos.

Tu función es responder a la consulta del usuario sobre sus síntomas ofreciendo un diagnóstico diferencial conciso y clínicamente fundamentado, utilizando terminología médica adecuada pero comprensible.

Pautas principales:
- Comienza con el diagnóstico más probable o con un diagnóstico diferencial jerarquizado, no con una advertencia genérica.
- Menciona explícitamente las afecciones probables cuando el patrón de síntomas las respalde.
- Prioriza la terminología médica estándar e incluye sinónimos clínicos comunes solo cuando sea útil.
- Diferencia entre opciones similares basándote en la distribución de los síntomas, la morfología, la evolución temporal, el contexto de edad y las manifestaciones sistémicas asociadas, siempre que esta información esté disponible o pueda inferirse razonablemente.
- Si la presentación clínica es ambigua, ofrece un diagnóstico diferencial breve y jerarquizado en lugar de limitarte a un único diagnóstico.

Formato de respuesta:
- Comienza con una respuesta breve y directa que indique el diagnóstico principal o la opción más probable.
- A continuación, presenta una lista breve y jerarquizada de diagnósticos diferenciales (de 2 a 5 elementos).
- Para cada elemento, aporta una razón breve vinculada a los síntomas descritos en la consulta.
- Finaliza con una recomendación breve sobre la evaluación clínica, los signos de alarma o la necesidad de derivación o atención especializada, según corresponda.

Restricciones importantes:
- No inventes pruebas, antecedentes, exposiciones o hallazgos de exploración física que no se hayan proporcionado.
- No evites mencionar diagnósticos plausibles cuando el patrón de síntomas los respalde.
- No respondas únicamente con mensajes tranquilizadores, advertencias de seguridad genéricas o categorías de enfermedades demasiado amplias.
- No afirmes tener una certeza excesiva cuando la información sea limitada.
- Mantén un estilo profesional, específico y médicamente preciso.

## H Compute Resources for Open-Weight Models

All experiments were conducted on Google Colab using a single NVIDIA RTX PRO 6000 Blackwell GPU. Inference was served with vLLM (Kwon et al., 2023) using single-GPU execution only; no distributed inference or multi-GPU tensor parallelism was used. Quantized model weights were used for local inference. Qwen 3.5 9B (Qwen Team, 2026) and Llama 3.3 70B (Grattafiori et al., 2024) were both run from a AWQ IN4 checkpoint.

| Model | GPU Count | Quantization | GPU Hours |
|---|---|---|---|
| Qwen 3.5 9B | 1 | AWQ INT4 | ∼7.5 |
| Llama 3.3 70B | 1 | AWQ INT4 | ∼27.5 |

**Table H.1:** Compute configuration and total GPU usage for local inference.

## I Condition Synonyms

Condition Synonyms (English)

**Lupus** systemic lupus erythematosus; systemic lupus; SLE; lupus erythematosus; cutaneous lupus; discoid lupus; subacute cutaneous lupus.

**Dermatomyositis** idiopathic inflammatory myopathy; inflammatory myopathy; myositis; heliotrope rash; heliotrope eruption; heliotrope; Gottron

papules; Gottron's papules; Gottron sign; periungual erythema; nailfold telangiectasia.

**IgA vasculitis**
immunoglobulin A vasculitis; IgA-mediated vasculitis; Henoch-Schonlein purpura; Henoch–Schönlein purpura; Henoch Schonlein purpura; Schönlein-Henoch purpura; HSP; vasculitis; small-vessel vasculitis; leukocytoclastic vasculitis; palpable purpura; purpuric rash; purpura; petechiae; petechial rash.

**Malar rash**
malar erythema; malar distribution; malar region; malar area; malar and nasal erythema; butterfly rash; butterfly-shaped rash; butterfly shaped rash; butterfly distribution; butterfly pattern; lupus; lupus erythematosus; systemic lupus erythematosus; systemic lupus; SLE.

**Morphea** localized scleroderma; linear scleroderma; linear morphea; scleroderma; en coup de sabre; coup de sabre; plaque morphea.

**Juv. dermatomyositis**
JDM; childhood dermatomyositis; pediatric dermatomyositis; dermatomyositis; idiopathic inflammatory myopathy; inflammatory myopathy; myositis; heliotrope rash; heliotrope eruption; heliotrope; Gottron papules; Gottron's papules; Gottron sign.

## Condition Synonyms (Mandarin)

**Lupus** 系统性红斑狼疮；SLE；SLE（系统性红斑狼疮）；红斑狼疮；皮肤型红斑狼疮；盘状红斑狼疮；亚急性皮肤型红斑狼疮。

**Dermatomyositis**
皮肌炎；特发性炎症性肌病；炎症性肌病；肌炎；向阳疹；眼睑紫红斑；紫红色眼睑疹；Gottron丘疹；戈特龙丘疹；Gottron征；戈特龙征；甲周红斑；甲襞毛细血管扩张。

**IgA vasculitis** 免疫球蛋白A血管炎；IgA介导性血管炎；过敏性紫癜；过敏性紫癜（Henoch–Schönlein紫癜）；Henoch-Schönlein紫癜；HSP；HSP（过敏性紫癜）；血管炎；小血管炎；白细胞破碎性血管炎；紫癜；可触性紫癜；紫癜性皮疹；瘀点；瘀点性皮疹；紫癜性血管炎。

**Malar rash** 蝶形红斑；蝴蝶疹；蝶形皮疹；颧部红斑；颧部皮疹；面颊红斑；面颊皮疹；蝶形分布；狼疮；红斑狼疮；系统性红斑狼疮；SLE。

**Morphea** 局限性硬皮病；线状硬皮病；线性硬皮病；线状硬斑病；硬皮病；剑伤样硬皮病；en coup de sabre；coup de sabre；斑块型硬斑病。

**Juv. dermatomyositis**
幼年型皮肌炎（JDM）；JDM；儿童皮肌炎；小儿皮肌炎；皮肌炎；炎症性肌病；肌炎；向阳疹；眼睑紫红斑；Gottron丘疹；戈特龙丘疹；Gottron征；戈特龙征。

## Condition Synonyms (Hindi)

**Lupus** परअबलईगत लयऊपस एरइतहएमएथोसस; SLE; SLE (परअबलईगत लयऊपस एरइतहएमएथोसस); लयऊपस एरइतहएमएथोसस; तवअचईय लयऊपस; दइसकोइद लयऊपस; सअबएकयऊथ कयऊथएनइयअस लयऊपस.

**Dermatomyositis** अजढातहएतउक सऊजअनअजअनय मांसपएसहई रोग; सऊजअनअजअनय मांसपएसहई रोग; मायोसाइथइस; हएलइयथिरोप रअइसह; हएलइयोथरोप चअकअतता; Gottron पअपयऊलस; Gottron चइहन; पएरइउनगउाल एरइतहएमअ; नएलपहोलद थएलअइजनजइएकतएसइया.

**IgA vasculitis** IgA वाहइकासहोतह; इमयऊनोगलोबउलइनर्स वाहइकासहोतह; IgA-मअदहयअसतहइत वअइसकउलयाइथइस; Henoch-Schonlein पअरपउरा; HSP; वअइसकउलयाइथइस; वाहइकासहोतह; चहोथई वाहइका वअइसकउलयाइथइस; लयऊकोसाइथोकलासथइक वअइसकउलयाइथइस; सपअरसहनईय पअरपउरा; पअरपउरा; पअरपउरइक रअइसह; पएथइकई; पएथईकइअल रअइसह.

**Malar rash** मअलार एरइतहएमा॥ मअलार वइतरअ७॥ तइतलई रअइसह॥ तइतलईसकङतइ चअकअतता॥ तइतलई जअइसा चअकअतता॥ गालों पअर तइतलई जअइसा दाना॥ लयऊपस॥ लयऊपस एरइतहएमएथोसस॥ परअबलईगत लयऊपस एरइतहएमएथोसस; SLE.

**Morphea** सूथानीयकत सूकूलेरोदर्मा॥ रैइखक सूकूलेरोदर्मा॥ रैइखक मौडइर्शया॥ सूकूलेरोदर्मा॥ नि कपु दी साबरे॥ कपु दी साबरे॥ पलूाक मौडइर्शया।

**Juv. dermatomyositis**
जउवएनाइल दअरमाथोमायोसाइथइस (JDM); JDM; बालयावअसतहा दअरमाथोमायोसाइथइस; बाल चइकइतसा दअरमाथोमायोसाइथइस; दअरमाथोमायोसाइथइस; सऊजअनअजअनय मांसपएसहई रोग; मायोसाइथइस; हएलइयोथरोप रअइसह; हएलइयोथरोप चअकअतता; Gottron पअपयऊलस; Gottron चइहन.

| Condition Synonyms (Spanish) | |
|---|---|
| **Lupus** | lupus eritematoso sistémico; LES; LES (lupus eritematoso sistémico); lupus eritematoso; lupus cutáneo; lupus discoide; lupus cutaneo; lupus subagudo cutáneo; lupus subagudo cutaneo. |
| **Dermatomyositis** | dermatomiositis; miopatía inflamatoria idiopática; miopatía inflamatoria; miositis; erupción heliotropa; rash heliotropo; heliotropo; pápulas de Gottron; papulas de Gottron; signo de Gottron; eritema periungueal; telangiectasias periungueales. |
| **IgA vasculitis** | vasculitis IgA; vasculitis por IgA; vasculitis por inmunoglobulina A; vasculitis mediada por IgA; púrpura de Henoch-Schönlein; púrpura de Henoch–Schönlein; purpura de Henoch-Schönlein; purpura de Henoch-Schonlein; PSH; PSH (púrpura de Henoch–Schönlein); HSP; vasculitis; vasculitis de vasos pequeños; vasculitis leucocitoclástica; púrpura palpable; purpura palpable; púrpura; purpura; erupción purpúrica; petequias; erupción petequial. |
| **Malar rash** | eritema malar; rash malar; erupción malar; eritema en alas de mariposa; eritema en mariposa; erupción en mariposa; distribución malar; distribución en alas de mariposa; patrón de mariposa; lupus; lupus eritematoso; lupus eritematoso sistémico; LES. |
| **Morphea** | morfea; esclerodermia localizada; esclerodermia lineal; morfea lineal; esclerodermia; en coup de sabre; coup de sabre; golpe de sable; morfea en placas. |
| **Juv. dermatomyositis** | dermatomiositis juvenil; DMJ; DMJ (dermatomiositis juvenil); dermatomiositis infantil; dermatomiositis pediátrica; dermatomiositis; miopatía inflamatoria; miositis; erupción heliotropa; rash heliotropo; pápulas de Gottron; papulas de Gottron; signo de Gottron. |

# J LLM-as-a-judge

As a robustness measure, we utilize an LLM-as-a-judge to verify our regex evaluation method, in addition to human annotators. The LLM-as-a-judge relies on the highly directed system prompt that follows.

```
You are a medical terminology evaluator.

Your task is NOT to judge quality, correctness, completeness, style, or usefulness.

Only determine whether the target medical concepts appear in the candidate response.

Rules:
1. Match medical meaning, not exact wording.
2. Accept standard synonyms, abbreviations, plural forms, and lexical variants.
3. Reject terms that appear only in negated contexts.
4. Ignore unrelated content.
5. If uncertain, return NOT_FOUND.
6. Output valid JSON only.

REQUIRED TERMS:
- {term_1}
- {term_2}
- ...

NEGATIVE MODE:
{true | false}

Return:
{    "term_results": [
       {
            "required_term": "...",
            "status": "FOUND | NOT_FOUND",
            "matched_text": "...",
            "reason": "..."
       }
   ],
   "final_verdict": "PASS | FAIL"
}
```

# K Human Doctor Diagnosis Performance

## K.1 Approach

One physician completed a 30-item diagnostic classification panel derived from sampled English clinical and English lay participant descriptions. The panel contained 27 target-condition items and three placebo items describing acne or dry skin. There were no images or further information provided outside of the sampled participant responses with seed prompts.

We report overall accuracy and one-versus-rest precision, recall, and F1 for each class. We also report macro, micro, and support-weighted averages. Accuracy uncertainty is summarized with a 95% Wilson binomial confidence interval. Because the data contain a single physician and 30 fixed items, these estimates describe this completed panel and are not treated as population-level estimates of physician performance.

### K.2 Results

**Table K.1:** Doctor diagnostic classification performance on the 30-item panel.

| Class or summary | Support | Precision | Recall | F1 |
|---|---|---|---|---|
| Lupus | 7 | 100.0% | 71.4% | 83.3% |
| JDM | 10 | 90.0% | 90.0% | 90.0% |
| Morphea | 6 | 100.0% | 100.0% | 100.0% |
| IgA vasculitis | 4 | 100.0% | 75.0% | 85.7% |
| N/A | 3 | 33.3% | 66.7% | 44.4% |
| Macro average | 30 | 84.7% | 80.6% | 80.7% |
| Micro average | 30 | 83.3% | 83.3% | 83.3% |
| Weighted average | 30 | 90.0% | 83.3% | 85.3% |

The physician classified 25 of 30 items correctly, for an accuracy of 83.3% (95% Wilson CI 66.4%–92.7%). Macro precision, recall, and F1 were 84.7%, 80.6%, and 80.7%, respectively.

**Table K.2:** Doctor diagnostic classification performance by participant-language cohort.

| Cohort | Items | Correct | Accuracy |
|---|---|---|---|
| Clinical | 15 | 14 | 93.3% |
| Lay | 15 | 11 | 73.3% |

| Cohort | Macro precision | Macro recall | Macro F1 |
|---|---|---|---|
| Clinical | 90.0% | 93.3% | 89.3% |
| Lay | 80.0% | 78.3% | 77.7% |

For clinical-language items, the physician classified 14 of 15 correctly, corresponding to 93.3% accuracy (95% Wilson CI 70.2%–98.8%) and a macro F1 of 89.3%. For lay-language items, 11 of 15 were classified correctly, corresponding to 73.3% accuracy (95% Wilson CI 48.0%–89.1%) and a macro F1 of 77.7%. These cohort estimates are descriptive; the panel was not sized for a formal clinical-versus-lay comparison.

## L Model Comparison

**Llama 3.3 70B:** Terminology retrieval performance by prompt condition and language.

| Prompt | Lang. | Group | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|---|
| sys | EN | lay | .295 | .808 | .224 | .351 |
| sys | ES | lay | .218 | .745 | .099 | .175 |
| sys | ZH | lay | .251 | .922 | .140 | .243 |
| sys | HI | lay | .155 | .474 | .023 | .045 |
| sys | CL | lay | .675 | .937 | .673 | .784 |
| sys | CL | clinical | .761 | 1.000 | .727 | .842 |
| nosys | EN | lay | .242 | .769 | .154 | .256 |
| nosys | ES | lay | .202 | .697 | .082 | .147 |
| nosys | ZH | lay | .233 | .902 | .120 | .212 |
| nosys | HI | lay | .132 | .129 | .005 | .010 |
| nosys | CL | lay | .640 | .932 | .634 | .755 |
| nosys | CL | clinical | .750 | 1.000 | .714 | .833 |

**Gemini 3.1 Flash Lite:** Terminology retrieval performance by prompt condition and language.

| Prompt | Lang. | Group | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|---|
| sys | EN | lay | .218 | 1.000 | .130 | .230 |
| sys | ES | lay | .267 | 1.000 | .146 | .256 |
| sys | ZH | lay | .185 | 1.000 | .132 | .234 |
| sys | HI | lay | .278 | 1.000 | .142 | .249 |
| sys | CL | lay | .264 | 1.000 | .163 | .281 |
| sys | CL | clinical | .292 | 1.000 | .190 | .320 |
| nosys | EN | lay | .200 | .984 | .119 | .212 |
| nosys | ES | lay | .274 | 1.000 | .152 | .264 |
| nosys | ZH | lay | .172 | 1.000 | .120 | .214 |
| nosys | HI | lay | .233 | .958 | .092 | .168 |
| nosys | CL | lay | .221 | 1.000 | .116 | .208 |
| nosys | CL | clinical | .196 | 1.000 | .075 | .140 |

**Gemini 3.5 Flash:** Terminology retrieval performance by prompt condition and language.

| Prompt | Lang. | Group | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|---|
| sys | EN | lay | .234 | 1.000 | .149 | .259 |
| sys | ES | lay | .255 | .987 | .136 | .239 |
| sys | ZH | lay | .178 | 1.000 | .126 | .224 |
| sys | HI | lay | .232 | .912 | .095 | .172 |
| sys | CL | lay | .271 | 1.000 | .171 | .293 |
| sys | CL | clinical | .320 | 1.000 | .227 | .370 |
| nosys | EN | lay | .189 | 1.000 | .104 | .189 |
| nosys | ES | lay | .256 | 1.000 | .140 | .246 |
| nosys | ZH | lay | .200 | 1.000 | .138 | .243 |
| nosys | HI | lay | .280 | 1.000 | .136 | .239 |
| nosys | CL | lay | .213 | 1.000 | .106 | .191 |
| nosys | CL | clinical | .262 | 1.000 | .139 | .244 |

**Qwen 3.5 9B:** Terminology retrieval performance by prompt condition and language.

| Prompt | Lang. | Group | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|---|
| sys | EN | lay | .422 | .921 | .350 | .508 |
| sys | ES | lay | .379 | .920 | .282 | .432 |
| sys | ZH | lay | .375 | .928 | .295 | .448 |
| sys | HI | lay | .288 | .829 | .196 | .316 |
| sys | CL | lay | .596 | .970 | .555 | .706 |
| sys | CL | clinical | .659 | 1.000 | .610 | .758 |
| nosys | EN | lay | .291 | .883 | .191 | .314 |
| nosys | ES | lay | .283 | .849 | .173 | .287 |
| nosys | ZH | lay | .326 | .893 | .245 | .384 |
| nosys | HI | lay | .157 | .500 | .034 | .063 |
| nosys | CL | lay | .550 | .947 | .514 | .667 |
| nosys | CL | clinical | .727 | 1.000 | .688 | .815 |

## M Condition Level Results

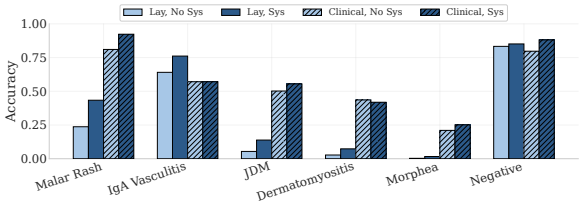



**Figure M.1:** Condition Level Hit Rates.

## N Distractor Results

**Llama 3.3 70B (Distractor):** Terminology retrieval performance under distractor prompt conditions.

| Prompt | Lang. | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|
| sys | EN | .192 | 1.000 | .192 | .323 |
| sys | ES | .124 | 1.000 | .124 | .221 |
| sys | ZH | .112 | 1.000 | .112 | .202 |
| sys | HI | .013 | 1.000 | .013 | .025 |
| nosys | EN | .192 | 1.000 | .192 | .323 |
| nosys | ES | .059 | 1.000 | .059 | .112 |
| nosys | ZH | .136 | 1.000 | .136 | .240 |
| nosys | HI | .013 | 1.000 | .013 | .025 |

**Qwen 3.5 (Distractor):** Terminology retrieval performance under distractor prompt conditions.

| Prompt | Lang. | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|
| sys | EN | .321 | 1.000 | .321 | .485 |
| sys | ES | .219 | 1.000 | .219 | .359 |
| sys | ZH | .249 | 1.000 | .249 | .398 |
| sys | HI | .147 | 1.000 | .147 | .257 |
| nosys | EN | .212 | 1.000 | .212 | .349 |
| nosys | ES | .148 | 1.000 | .148 | .258 |
| nosys | ZH | .284 | 1.000 | .284 | .442 |
| nosys | HI | .019 | 1.000 | .019 | .038 |

## O Downsampled Analysis

### O.1 Approach

Because the language cohorts contained unequal numbers of evaluated prompts, we performed a balanced-sample sensitivity analysis for Qwen 3.5. For the four-language omnibus analysis, the eligible target-positive rows were balanced to the smallest language cohort, yielding 767 outputs per language (3,068 outputs in each system-prompt setting). For comparisons based on clinical seed prompts, each of the five cohorts was balanced to 88 outputs (440 per system-prompt setting). For comparisons based on lay seed prompts, each cohort was balanced to 910 outputs (4,550 per system-prompt setting). Larger cohorts were sampled without replacement; no output was duplicated to reach these totals.

The omnibus analysis tested equality across English, Spanish, Mandarin, and Hindi separately for outputs generated with and without the system prompt. We then examined all six pairwise language contrasts and applied Holm's step-down correction within each system-prompt group. We separately compared the clinical cohort with each of the four lay-language cohorts under clinical and lay seed prompts.

### O.2 Language Comparisons

**Table O.1:** Balanced-sample Qwen 3.5 accuracy by language.

| Language | System prompt | No system prompt |
|---|---|---|
| English | 36.8% | 19.6% |
| Spanish | 24.3% | 14.3% |
| Mandarin | 27.1% | 21.6% |
| Hindi | 19.6% | 3.4% |
| Omnibus $p$ | $2.51 \times 10^{-13}$ | $2.85 \times 10^{-26}$ |

Language remained associated with target-condition accuracy in both prompting settings (Table O1). With the system prompt, English had the highest balanced accuracy (36.8%), followed by Mandarin (27.1%), Spanish (24.3%), and Hindi (19.6%). Without the system prompt, Mandarin had the highest accuracy (21.6%), narrowly exceeding English (19.6%); Hindi was lowest at 3.4%.

**Table O.2:** Pairwise language comparisons in the balanced sample.

| Setting | Comparison | Holm-adjusted $p$ |
|---|---|---|
| System | English–Spanish | 0.00018 |
| System | English–Mandarin | 0.0048 |
| System | English–Hindi | $1.80 \times 10^{-10}$ |
| System | Spanish–Mandarin | 0.352 |
| System | Spanish–Hindi | 0.0268 |
| System | Mandarin–Hindi | 0.0022 |
| No system | English–Spanish | 0.0309 |
| No system | English–Mandarin | 0.576 |
| No system | English–Hindi | $3.08 \times 10^{-12}$ |
| No system | Spanish–Mandarin | 0.0073 |
| No system | Spanish–Hindi | $8.45 \times 10^{-6}$ |
| No system | Mandarin–Hindi | $3.71 \times 10^{-14}$ |

After Holm adjustment, all system-prompt language contrasts were significant at $\alpha = 0.05$ except Spanish versus Mandarin. Without the system prompt, English versus Mandarin was not significant; the remaining contrasts were significant. Thus, balancing reduced neither the overall evidence of language heterogeneity nor the especially large separation involving Hindi.

### O.3 Register Comparisons

**Table O.3:** Clinical-register comparisons after balanced downsampling using clinical seed prompt.

| Language | System | No system |
|---|---|---|
| Clinical | 65.9% | 72.7% |
| English | 44.3% (0.0062) | 33.0% ($1.99 \times 10^{-7}$) |
| Spanish | 35.2% (0.00016) | 21.6% ($3.22 \times 10^{-11}$) |
| Mandarin | 31.8% ($3.07 \times 10^{-5}$) | 28.4% ($1.16 \times 10^{-8}$) |
| Hindi | 29.5% ($9.11 \times 10^{-6}$) | 9.1% ($4.01 \times 10^{-18}$) |

**Table O.4:** Clinical-register comparisons after balanced downsampling using lay seed prompt.

| Language | System | No system |
|---|---|---|
| Clinical | 60.5% | 55.2% |
| English | 43.3% ($7.93 \times 10^{-13}$) | 29.7% ($8.80 \times 10^{-27}$) |
| Spanish | 36.1% ($5.05 \times 10^{-24}$) | 27.0% ($7.85 \times 10^{-33}$) |
| Mandarin | 37.0% ($1.78 \times 10^{-22}$) | 33.6% ($1.38 \times 10^{-19}$) |
| Hindi | 29.7% ($7.76 \times 10^{-38}$) | 16.4% ($1.45 \times 10^{-65}$) |

The clinical cohort outperformed each lay-language cohort in every displayed clinical comparison after Holm adjustment. Under clinical seed prompts, clinical accuracy was 65.9% with the system prompt and 72.7% without it. Under lay seed prompts, the corresponding clinical accuracies were 60.5% and 55.2%. Although the magnitude varied by system-prompt and seed-prompt setting, the register contrast was therefore robust to balanced downsampling.

# P Multimodal Reference

## P.1 Approach

The multimodal experiment, with the subset of 10 images (two each of malar rash, morphea, dermatomyositis, juvenile dermatomyositis, and IgA vasculitis) produced 510 outputs in total: 10 image-only outputs and 100 outputs per image-plus-description cohort. The same English diagnostic system prompt was used in every condition. Inference used a temperature of 0, disabled thinking, and a 384-token output limit.

Sampling maximized unique participant-response identifiers within each image and cohort. When a cohort contained fewer than 10 unique responses for an image, the remaining cells used distinct existing seed-prompt variants paired with available participant descriptions. There are no normal controls or distractor cases, thus we report target-condition hit rate (recall for this subset).

## P.2 System Prompt

System Prompt (English)

You are a medically literate diagnostic reasoning assistant analyzing dermatologic images and, when supplied, a participant's description of the same image.

Your job is to identify the most plausible diagnosis and give a concise, clinically grounded differential diagnosis. Use the image as primary evidence and the participant description as additional evidence; do not assume that every detail in the description is correct.

Core behavior:
- Lead with the most plausible diagnosis or a ranked differential, not a generic disclaimer.
- Explicitly name likely conditions when supported by the image and description.
- Prefer canonical medical terminology and include common clinical synonyms only when helpful.
- Distinguish close alternatives using visible distribution and morphology plus any supplied tempo, age context, or systemic features.
- If the presentation is ambiguous, give a short prioritized differential rather than forcing a single diagnosis.

Answer format:
- Start with a brief direct answer stating the leading diagnosis.
- Give a ranked differential of 2-5 items, each with one brief reason.
- End with a short recommendation for clinical evaluation or urgent escalation when appropriate.

Important constraints:
- Do not fabricate tests, history, exposures, symptoms, demographics, or examination findings.
- Clearly separate what is visible in the image from what the participant reports.
- Do not avoid naming plausible diagnoses when the evidence supports them.
- Do not answer with only reassurance, only generic safety language, or only broad disease categories.
- Do not overstate certainty from an image alone.
- Keep the prose professional, specific, medically precise, and understandable.

## P.3 Extended Results

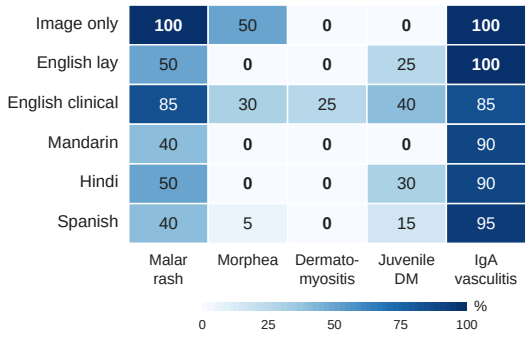



**Figure P.1:** Target-condition hit rates by diagnosis.

The small image set, absence of negative examples, repeated use of some participant descriptions across prompt variants, and shared images within cells preclude inferential comparisons or claims of population-level multimodal accuracy. Figure 2 contains the cohort level hit rates.