Contextual Evaluation of Large Language Models for Classifying Tropical and Infectious Diseases

Mercy Asiedu*
Google Research

Nenad Tomasev
Google DeepMind

Chintan Ghate
Google Research

Tiya Tiyasirichokchai
Google Research

Awa Dieng
Google DeepMind

Oluwatosin Akande
Independent Scientist

Geoffrey Siwo
University of Michigan

Steve Adudans
Gear Health

Sylvanus Aitkins
Ministry of Health, Sierra Leone

Odianosen Ehiakhamen
Nigerian Center for Disease Control

Eric Ndombi
Kenyatta University

Katherine Heller
Google Research

Abstract

While large language models (LLMs) have shown promise for medical question answering, there is limited work focused on tropical and infectious disease-specific exploration. We build on an opensource tropical and infectious diseases (TRINDs) dataset, expanding it to include demographic and semantic clinical and consumer augmentations yielding 11000+ prompts. We evaluate LLM performance on these, comparing generalist and medical LLMs, as well as LLM outcomes to human experts. We demonstrate through systematic experimentation, the benefit of contextual information such as demographics, location, gender, risk factors for optimal LLM response. Finally we develop a prototype of TRINDs-LM, a research tool that provides a playground to navigate how context impacts LLM outputs for health.

1 Introduction

Neglected tropical diseases, while extremely preventable and treatable, continue to be highly prevalent in the poorest regions of the world, affecting 1.7 billion people globally with disproportionate impacts on women and children [7]. Similarly, half of the world's population are at risk from infectious diseases, which continue to lead in global mortality and morbidity, resulting in more than 52 million deaths globally, with 99% of diseases occurring in developing countries [15]. Challenges in preventing and treating these diseases include surveillance, early detection, accurate initial diagnosis, management and vaccine limitations [7, 21]. During the pandemic various scalable measures were implemented to address these challenges specific to COVID-19 [3, 6, 12, 11, 8]. Other tropical and infectious diseases have had limited attention for surveillance and surfacing accurate diagnosis.

The use of large language models for health-related question-answering has increased in recent years demonstrating applications for a variety of health use cases [17, 20, 9, 10]. However, there is limited work that has focused on tropical and infectious diseases. These are particularly of interest as

*corresponding author: masiedu@google.com

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

they may present out-of-distribution cases, given that they mostly occur in the global south which may be underrepresented in training and evaluation datasets and may lead to potential biases [1, 2]. Additionally there is limited understanding of how different contextual factors such as demographics, prompt styles, and subsets of information (eg. symptoms only, versus symptoms+location) may influence model performance.

A few studies assess the use of machine learning and LLMs for tropical and infectious diseases. Shenoy et al. [19] found that in a study of 40 tropical and infectious disease clinicians, 35 indicated the need for a decision making tool. They compared logistic regression and decision trees for binary classification and found an average prediction accuracy of 79-84%. Pfohl et al. [14] developed a health equity dataset which introduced LLM evaluation of a tropical and infectious diseases (TRINDs) dataset and had clinicians and health equity experts evaluate LLM responses to tropical and infectious disease related questions,finding evidence of biased responses. Mondal et al. [13] developed 50 simulated infectious disease cases with histories, lab reports, imaging findings and evaluated it on 4 different general LLMs and compared the quantity of differential diagnosis to that of medical experts. They found that LLMs generally had difficulties matching the experts' differential diagnosis. Schwartz et al. [18] delineate failure points for LLMs for infectious diseases consultation to clinical workflow questions and find that LLMs provide dangerous hallucinations and harmful advice for disease management-type questions.

In this paper we build on the TRopical and INfectious Diseases (TRINDs) dataset [14], making the following primary contributions:

- We expand on the dataset to include demographic and semantic clinical and consumer augmentations.
- We perform various evaluations with the dataset, to understand how different contexts and counterfactual locations contribute to LLM performance.
- We evaluate LLM performance improvements on the larger augmented dataset with in-context prompt tuning.
- We assemble a panel of human experts to set a human expert baseline score on the dataset and to provide ratings of data quality, usefulness, etc.
- We develop TRINDs-LM, a research tool to demonstrate how context influences LLM performance on TRINDs.

2 Methods

Table 1: Summary of datasets and experiments

Dataset augmentations	**Experiments**
Original TRINDs dataset	Generalist LLM vs. specialist LLM accuracy
	LLM vs. human expert performance
Contextual dataset	Impact of contextual factors on accuracy
French dataset	Impact of language on accuracy
Counterfactual dataset	Impact of location, race and gender on accuracy
Multiple choice set	LLM vs. human expert performance
LLM-expanded demographic set	Impact of a variation of demographics on accuracy
LLM-expanded semantic set	Impact of a variation of question semantic styles on accuracy
Consumer set	Impact of consumer style questions on accuracy

2.1 Dataset generation and expansion:

Original TRINDs dataset: We base this work on the original TRINDs dataset [14]. To summarize, the authors examined authoritative sources containing factual information about different diseases to create a dataset of 106 questions pertaining to tropical and infectious diseases across different regions around the world. We use a subset of 52 questions from the TRINDs data as a seed set. Each question in the seed set follows a templated structure where a patient persona is presented with

general symptoms, direct attributes, specific symptoms, as well as context and lifestyle/risk factors. Each question is associated with a ground truth disease label, which were reviewed by clinicians for accuracy.

Contextual TRINDs dataset:We created 16 subsets of the seed set to understand which of the different sections of the dataset most influenced model performance. This set used different inclusions and combinations of general symptoms, specific symptoms, personal attributes (age and gender), location, and risk factors.This generated 468 queries.

Counterfactual sets: *Location:* We examined how location influences LLM responses. The original dataset was created with locations where the disease had a known likelihood of occurring. For each of the original dataset and contextual subsets that included a location, we switched out the original location for a single counterfactual location where there was less probability of disease occurring. Here we used "San Francisco" as that location to generate 52 counterfactual location queries. *Race:* We created versions of the original dataset where we included a race input (Asian, Black and White) for each disease, yielding 159 additional queries. *Gender:* We created versions of the dataset that had only male, only female and non-binary demographics, yielding an additional 159 queries.

French language set: Given prevalence of tropical and infectious diseases in non-English speaking countries we sought to understand how language influenced the performance of the model. A researcher whose official language is French, manually translated the original dataset of 52 queries and diagnosis into French and compared performance to the English dataset. We selected French as our primary interest lies in the African continent and official languages that are used are primarily in English and French. This led to 52 prompts.

Multiple choice set: To compare the LLMs to expert baselines, we generate 153 multiple choice questions from the original, expanded demographics and expanded semantics sets. We input the questions and ground truth diagnosis from the original dataset and prompted an LLM to provide 5 multiple choice options that included the ground truth label, and 4 other broad tropical and infectious disease options.

LLM-Expanded demographics set: To provide a larger demographically diverse data pool to assess LLM responses we created a synthetically expanded dataset of 2635 queries using the original set as a seed set. For each persona an LLM was prompted to generate 50 demographic expansions, with varying gender, sex, age, socioeconomic status, disability status, ethnicity, location, and origin. For the location variation the LLM was instructed to only generate locations with high incidence of the disease. The LLM was also instructed to prioritize including locations that were consistent with the generated demographics, and to adjust the socioeconomic status to match the location that was generated. Finally, each proposed generation was checked by an LLM-based filter where the LLM was asked to discard generations where the symptoms do not match the condition, or where the demographic does not appear to be plausible. Refer to prompting method in Appendix A.3.

LLM-Expanded semantic set: As prompts from users may present in a variety of ways, and as symptoms may vary as well, we synthetically expanded the dataset, adding 2651 additional queries. These were created by instructing the LLM to first generate a demographically altered version of the question, in the same way as described above for the expanded demographics set. The LLM then follows that up by another alteration, generating an alternative formulation of the question for the given patient and disease, but describing a different set of common symptoms related to the disease, and potentially a different personal patient background of relevance for disease risk or onset. LLM-based filtering was applied to these generations as well, where proposals would get discarded in case the generated symptoms did not match the disease, or the proposed demographic does not appear to be plausible.

LLM-Consumer augmented queries: The original personas were created with a clinical tone. We created a consumer augmentation of the expanded dataset. We prompted an LLM to convert the original, expanded demographics and semantic sets into first person perspective to generate the consumer versions. This generated 52 consumer queries on the original set, 2635 on the demographic and 2651 queries on the semantic set.

In total we generated a dataset of 11,719 queries from the original seed set of 52 queries across 50 tropical and infectious diseases. We summarize the dataset and experiments run in Table 2. Refer to Appendix A.2 for examples.

2.2 Model evaluation

We use two baseline models- Gemini Ultra [5], a generalist large language model (standard hyperparameters: batchsize 16, temperature 0.7, top_k=32), and MedLM Medium [4], a LLM specialized for the health domain (standard hyperparameters: batchsize 32, temperature 0.2, top_k=None). As a baseline, we prompt-tuned both models, providing instructions and 2-shot examples to guide the model's output (Appendix A.3). The models were prompted to provide an output of the ground truth disease label. This was repeated 4 additional times to yield a total of 5 outcomes per experiment for statistical analysis.

2.3 Auto-rater LLM Evaluations

We developed an automated rater to score each query for accuracy out of the 5 repeated runs. This was developed by prompting an LLM to determine whether the words were structurally and/or meaningfully similar to each other and to score them as correct if they were. For instance if the ground truth was *Taeniasis/cysticercosis* and model output was *Tapeworm* it would be marked as correct, since these are meaningfully similar, even if they are structurally different. We performed a manual review of the automated rating on a subset of the data to optimize the rating process and ensure there were no errors.

From the scores we first determined performance of generalist and medical specific models on variations of the original persona set. Next we analyzed which contextual factors/combinations were important for model performance (combinations of general symptoms, specific symptoms, risk factors, location and personal attributes). We then compare performance on the counterfactual and original versions of the prompt. Finally, we assess performance on expanded data with demographic and semantic augmentations with and without many-shot prompting with the original set.

2.4 Human Expert Baseline

We created a human expert baseline study to understand expert performance on the dataset. The purpose of this was to determine how experts in tropical and infectious diseases performed on a representative sample of the data to enable human expert vs LLM comparison. The study was deemed IRB exempt by an internal ethics review personnel. We recruited experts- public health researchers, and medical doctors- who had generalizable knowledge of TRINDs, using a snowballing approach via our networks.

Following informed consent, the experts filled a demographic pre-survey, summarized in Table 2. Experts reported a variety of specializations with tropical and infectious diseases including Immunoparasitology, Neglected Tropical Diseases, Infectious Disease Epidemiology, and diseases with pandemic/epidemic potential (eg. Ebola, Mpox).

Experts were then given 52 short answer questions (SAQs) with full context- general and specific symptoms, demographics (age and gender), location, and risk factors across the identified diseases- and asked to write in a single most likely diagnosis. Once they completed the SAQs, they were given another questionnaire with multiple choice questions of varied formats (153) where they selected the single most likely disease given a list of diseases. Experts were asked not to reference any sources in answering the questions or to look up answers to the questions until both surveys had been completed. After completing the SAQ and MCQ surveys, experts provided feedback on various axes on data quality. They were also asked to indicate how helpful each

Table 2: Expert panel demographics (n=7).

Demographic	**No. Experts**
Gender	
Female	2
Male	5
Age	
30-39	2
40-49	5
Country of residence	
Kenya	2
Sierra Leone	1
United States	2
Nigeria	2
Switzerland	2
Highest level of education	
Masters	2
MD	2
Doctorate	3
Occupation	
Medical doctor	4
Public health researcher	4
Professor	1
Years of experience	
5-10	3
11-20	3
20+	1

contextual information was in answering the questions.
The process took 5 hours all together and experts were compensated 500 USD for their time.

2.5 TRINDs-LM Tool design and development

We developed the TRINDs-LM playground for researchers to understand how context impacts health responses for LLMs, using Gemini Ultra as a base model. The user interface (Appendix A.1) allows the user to input demographic information, lifestyle information, and symptoms. A summary is then generated and input into the LLM. The model provides an output that includes the most likely diagnosis, reason, and disease definition. The user interface also shows an interactive map of the global disease incidence rate. The TRINDs research tool is available on request. The tool not meant for clinical use.

2.6 Statistical analysis

We performed two-tailed student's t-test (alpha=0.025) for statistical analysis to compare performance between medical specific and general purpose models for each experiment, original and counterfactual datasets, base and many-shot prompt tuned models, and LLM and human expert performance.

3 Results

3.1 LLM experimental results

3.1.1 Generalist (Gemini Ultra) and specialist (MedLM Medium) model performance on persona variations

We found that overall performance with minimum instruction tuning and two-shot examples yielded an accuracy of 61.5% for Gemini Ultra and 47.9% for MedLM for the original clinical personas (Figure 1a). Gemini Ultra performance increased to 68.7% for the consumer versions, while MedLM remained roughly the same at 46.0%. Performance of both models was reduced with the French versions of the questions, with Gemini dropping significantly to 46.0% compared to MedLM which dropped slightly to 46.0%. Overall the generalist gemini model outperformed MedLM, which might be due to factors such as differences in model sizes, or overfitting of the MedLM tuned model to specific datasets (Figure 1a).

3.1.2 Assessing Gemini model performance on varied combinations of attributes and factors

We found that *symptoms*, *location* and *risk factors (SLR)* enable the best model performance, followed by the *full persona (FP)* of symptoms, location, risk factors and personal attributes (Figure 1c). This implies that excluding personal attributes such as age and gender, may preserve privacy while still maintaining beneficence. The worst performing contextual combinations were *general and specific symptoms (S)* alone (46.8%), *symptoms* and *attribute (SA)* (44.91%) and *general symptoms*, *location*, *attributes* and *risk factors (gSLAR)* (44.5%), demonstrating the need for both specific and general symptoms, as well as other contextual factors such as location and risk factors to enhance model accuracy. We also show performance per disease and find that the model performs best on relatively widespread or distinctive diseases eg. Trachoma, HIV and Rabies, but performs worst on less common or diseases with common symptoms eg. tuberculosis (Figure 2a).

3.1.3 Assessing Gemini model performance on counterfactual inputs

Location: We found that generally, the *counterfactual location* decreased performance across all contextual combinations that included *location (L)*, but performance was least affected in cases with *full persona (FP)*, or with the combination of *general and specific symptoms, location and risk (SLR)* demonstrating the usefulness of including added contextual information (Figure 1b). *Counterfactual location* caused the model to perform the worst when only *symptoms* were provided without any additional context, or when either *general symptoms* or *specific symptoms* were provided even with other information. This demonstrates the need to have both *specific* and *general symptoms*, plus other contextual information for optimal performance. We find that this pattern is consistent across the individual diseases (Figure 2b). ***Race:*** There were no statistically significant differences in

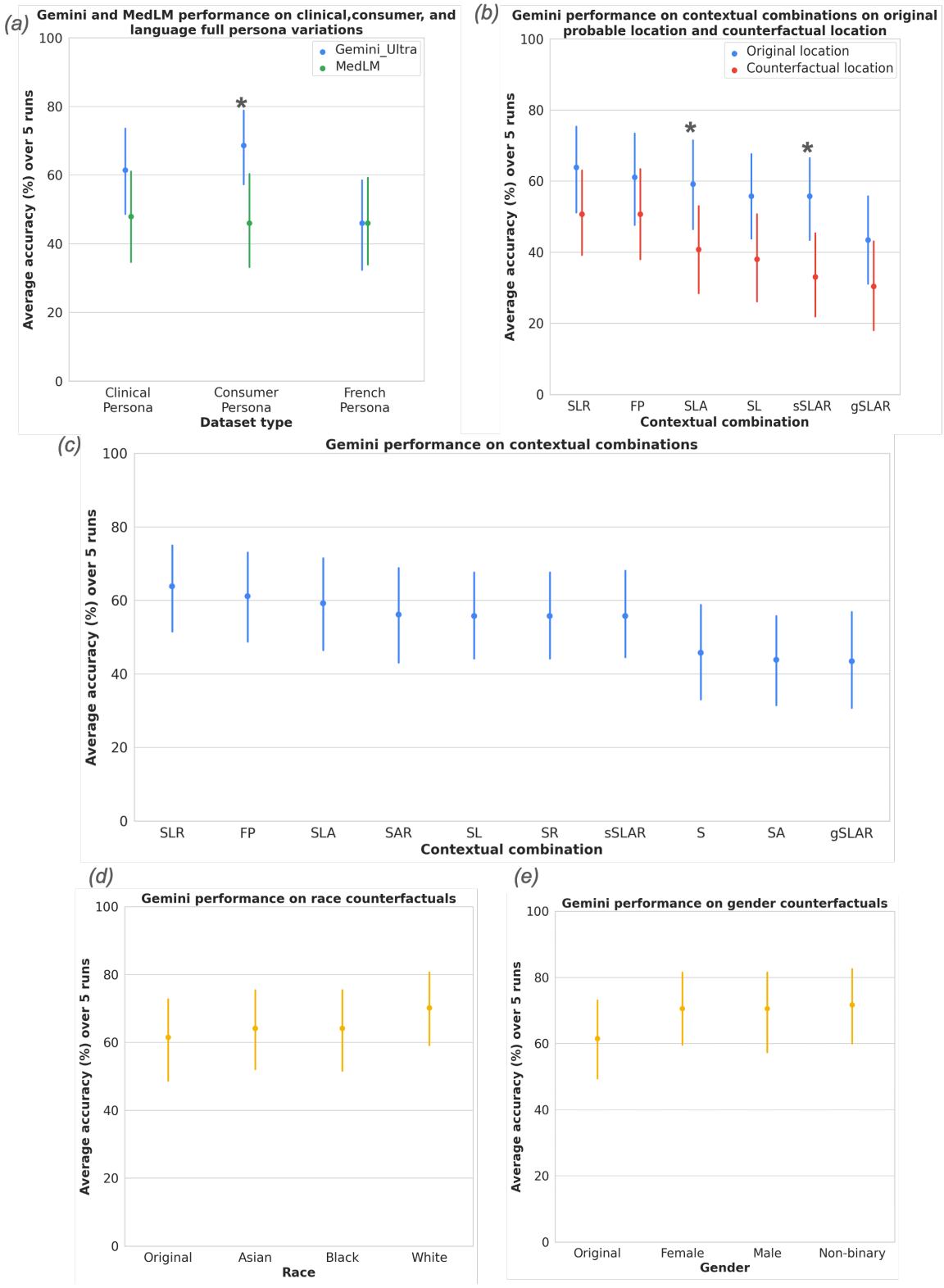


Figure 1: Model performance on persona variations. a) Generalist (Gemini) and specialist (MedLM) model performance on clinical, consumer and French persona variations. b) Gemini model performance on counterfactual location inputs. c) Gemini performance for contextual combinations of attributes and factors and count. d) Gemini performance for race counterfactuals e) Gemini performance for gender counterfactuals. Error bars are 90% confidence interval. *=p<0.025

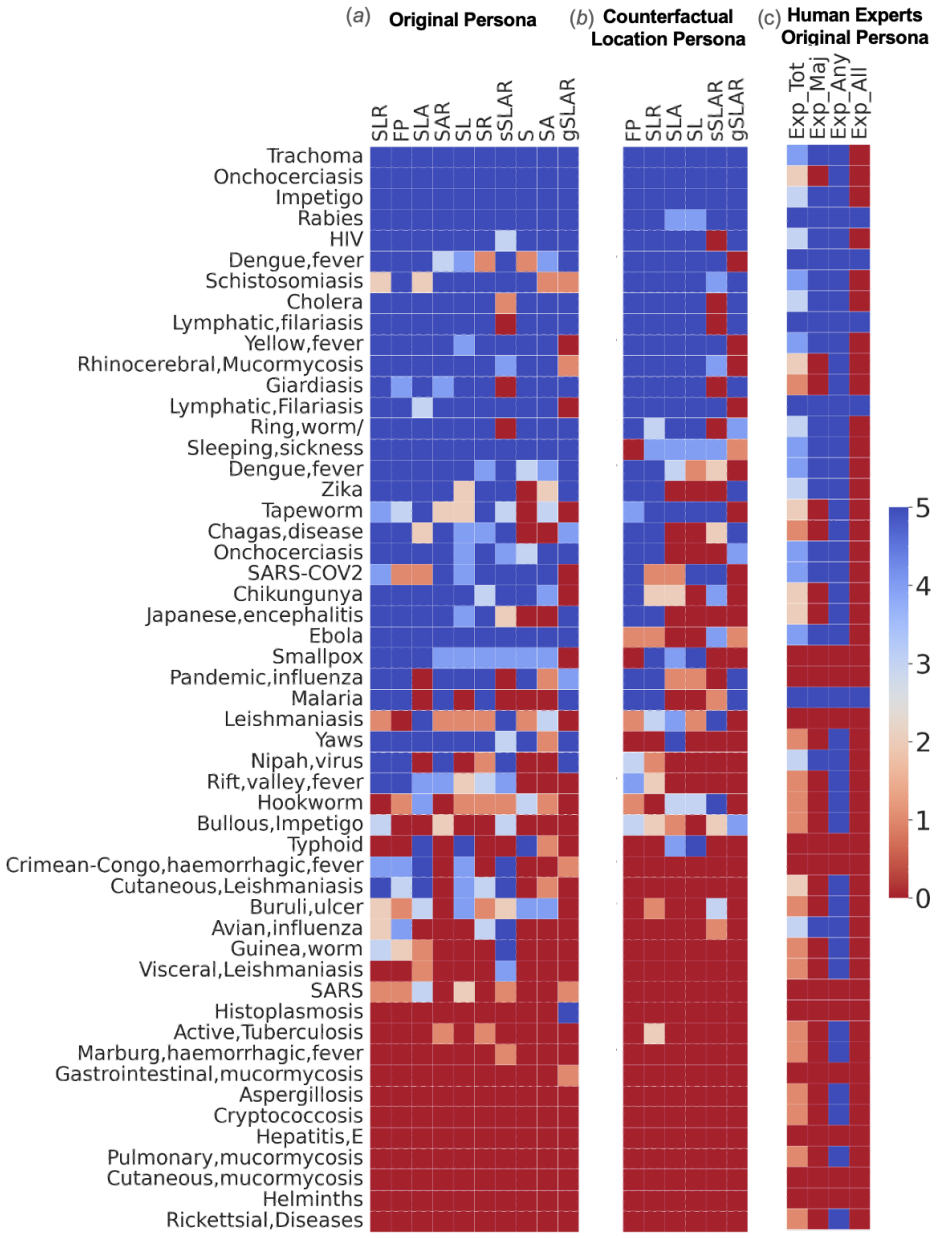


Figure 2: Per disease performance for LLMs and human experts, a) LLM performance on original persona with different contextual combinations (5 repeated runs), b) LLM performance on location counterfactual with different contextual combinations (5 repeated runs), c) Human expert performance (top 5 out of 7 experts). Error bars are 90% confidence interval. Legend: S=symptoms(general and specific), gS=general symptoms, sS= specific symptoms, L=location, A=attribute (age and gender), R=risk factor, FP=full persona, Exp_Tot=total expert score, Exp_Maj = expert majority score, Exp_Any = expert any/at least one score, Exp_All = Expert all score.

performance across different racial counterfactual inputs, even though we note that performance for "White race" counterfactuals increased slightly (Figure 1d). ***Gender:*** There were no statistically significant differences in performance across different gender counterfactual inputs, even though we note that performance for each of the gender counterfactuals was slightly higher than the original persona (Figure 1e).

3.1.4 Assessing model performance on the demographic and semantic expansions

We found no significant difference for base Gemini (Figure 3 a,b) performance for clinical and consumer variations on the expanded dataset. We found no significant differences between performance of the base models on demographic and semantic variations (Figure 3a,b). This demonstrates that models perform approximately the same on clinical style and consumer style questions which contain the same information.

3.1.5 Impact of many-shot in-context learning with original persona set on model robustness and generalizability

We found that in-context learning by providing the model with many-shot examples of the full original set for each disease, significantly increased Gemini performance for demographic and semantic augmentations, though less so for semantic augmentations(Figure 3 a,b). It also increased performance on different styles of question inputs (semantics). This demonstrates that in-context learning with a small set of high quality data improves model performance and robustness across different demographics, locations and question styles.

3.2 Human expert baseline and data rating

3.2.1 Expert vs LLM performance:

Seven experts filled out the baseline SAQs and MCQs. We calculated the score for each expert and selected responses for the top 5 scoring experts for each survey for further analysis. For each question from the top 5 responses, we calculated the (i) the *Expert total score*: sum of the score from the 5 experts, (ii) *Expert majority score*: full score (5/5) if the majority of experts had the correct answer else 0/5, (iii) *Expert any/at least one score*: a full score (5/5) if any one of the experts had the correct answer, and (iv) *Expert all score*: a full score (5/5) only if all the experts had the answer correct. TRINDs experts tend to specialize and different specialists may do better on some diseases than others. These combinations of expert scores simulate real-world policy settings where a panel of experts may be used to come to a diagnosis in a variety of ways.

For the SAQs dataset, Gemini performed significantly better than all the expert response combinations, except for the *Expert Any/at least one* which was significantly better than Gemini. For the MCQs, Gemini significantly performed better than the *Expert total* and *Expert all scores*, and significantly worse than the *Expert Any score*. There was no difference between Gemini and *Expert majority* performance for MCQs (Figure 4a).

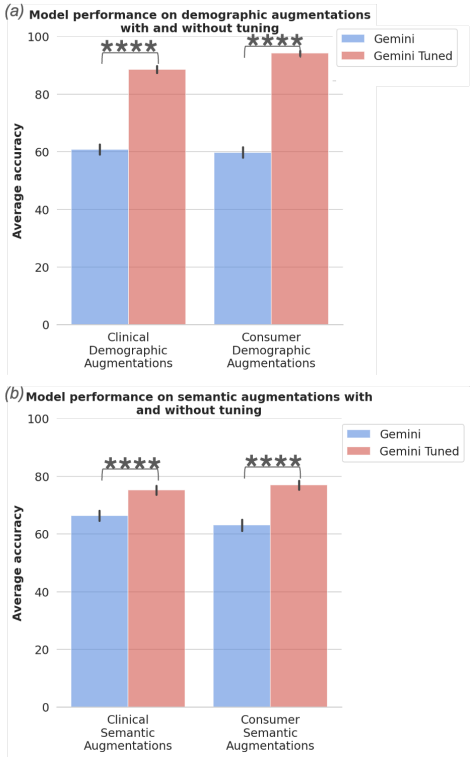


Figure 3: Model performance on expanded dataset. a) LLM performance on demographic clinical and consumer augmentations (2635 each) and b)LLM performance on semantic clinical and consumer augmentations (2651 each). We compared the base model, with the multi-shot tuned model. ****=p<0.00005

3.2.2 Expert data rating and qualitative feedback:

Experts rated level of helpfulness of each of the contextual information in providing a diagnosis. *Symptoms* were rated highest by most experts, followed by *risk factors*, *location* and then demographics *attributes* (Figure 4 c). This is in line with LLM results that indicate that *symptoms, location and risk factors* are most important for LLM classification. On the dataset quality, experts generally indicated high scores (4-5) for accuracy, completeness, range of tropical and infectious diseases covered, geographic and demographic diversity, and timelines of the dataset (Figure 4 b). Experts rated the level of difficulty of the questions as 3-4/5. The diversity in the style of questions-asking was rated low with several experts commenting on the repetitiveness of the questions. Experts recommended including the use of over-the-counter medications, less repetition in the question style, improving the specificity of some of the symptoms, and the need for differential diagnosis. There were also comments on improving the quality of the LLM-generated queries. One expert commented on the inclusion of smallpox which has been eradicated for 40 years, but indicated the potential tie to the ongoing Monkeypox (Mpox) epidemic. Experts also indicated the need for patient images to provide more informative responses demonstrating the need for a multimodal version of the dataset.

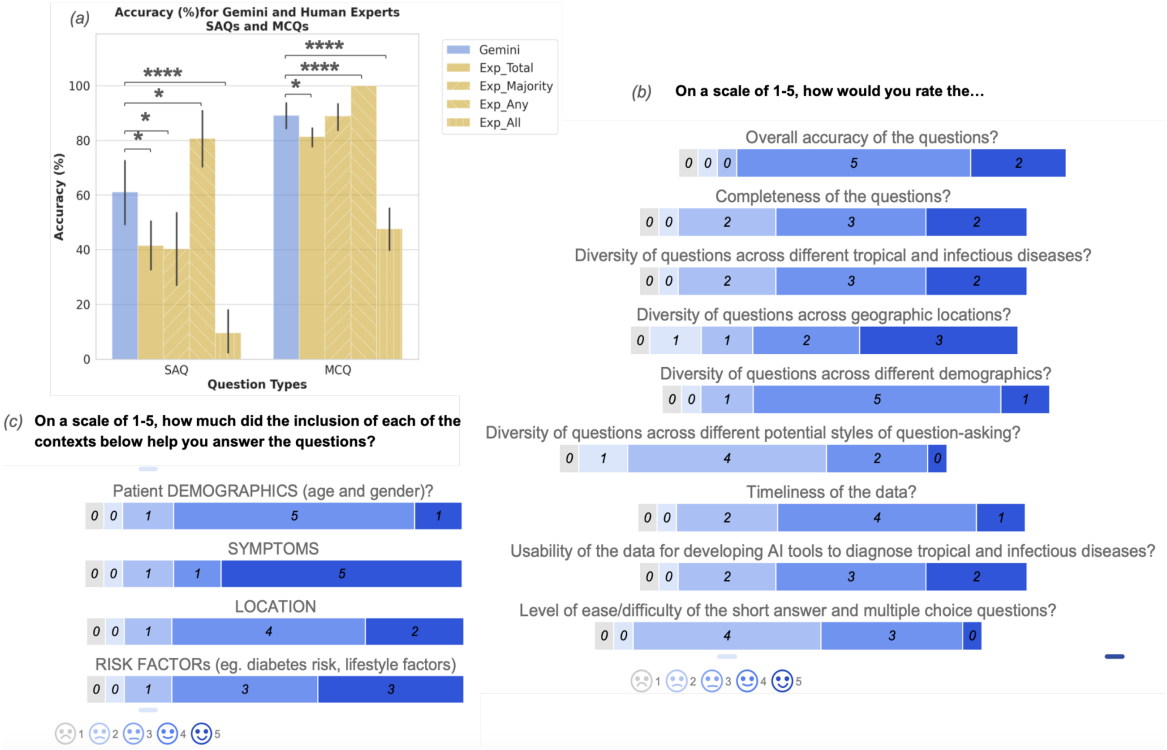


Figure 4: Expert baseline and data quality rating. a)Expert baseline compared to LLM, b) Expert rating of data quality, c) Expert rating of helpfulness of contextual information. Error bars are 90% CI. *=p<0.05 ****=p<0.00005

4 Discussion

This work is motivated by the need to evaluate LLMs on tropical and infectious diseases which present a potential out-of-distribution shift for LLM performance given that these tend to occur in the global south which may have less representation in datasets used for LLM training [1, 2]. This study confirms this distribution shift with the Gemini model achieving an accuracy of 61.5% and MedLM achieving an accuracy of 47.9% on clinical-style questions, significantly lower than reported performances on USMLE benchmarks (GPT: 90.2%, MedLM: 91.1% [16]). We find that simple in-context prompting with the dataset, and potentially intentional training focus on these diseases improves the LLM performance significantly. We find that Gemini Ultra performs better than MedLM medium, however this is likely due to differences in model sizes.

Our evaluations demonstrate that including additional context such as risk factors and location in addition to symptoms also improves model performance. However indicating locations where the disease is less likely to occur, reduces model performance. This has implications for LLMs that may attempt to provide a disease diagnosis based on autorecognition of location without considering contexts such as visits to endemic locations. On this dataset, we do not find significant differences in LLM performance across race or gender. We also do not find any significant differences in performance between different styles of question asking (semantic, clinical, and consumer variations).

Our analysis reveal that LLMs tend to more accurately identify common diseases, or diseases with very specified symptoms such as *Trachoma, Rabies* and *Yaws*, while less common and less specific diseases are mislabeled. For instance we find that LLMs mostly classify *Hepatitis E* as *Hepatitis A*, though this was a common mistake among human experts as well.

Our human expert baseline finds that for both SAQs and MCQs, experts scored lower in accuracy on the full context questions than the model except in cases where we looked at *expert any/at least one score*. *Avian influenza* is an example of a disease that the LLM had trouble with, but that majority of the experts got right. On the other hand the model got diseases like *Giardiasis* and *Onchocerciasis* right, while most experts got these wrong. We do note that experts were asked not to consult any external material, which would not be the case in real-world scenarios, and real world scenarios may also provide confirmatory tests for the diseases.

Experts generally rated the dataset highly on axes of accuracy, completeness, timeliness and diversity across tropical and infectious diseases. However they suggested improvements in diversity in question asking styles, and addition of images to the questions where applicable.

Implications for Policy and Practice: Our findings demonstrate the discrepancy in LLM performance on tropical and infectious diseases, compared to reported performance on USMLE questions, identifying the need for contextual evaluation of LLMs that are used in clinical settings in the global south - i.e the need for LLM usage to take into account contextual, regional-specific and patient-specific factors. However, we also find that Gemini performs better on this dataset compared to human experts for identifying TRINDs from text-based descriptions. For healthcare workers, our findings highlight the potential of LLMs to serve as valuable decision-support tools, thus enhancing clinical diagnostic accuracy in resource-limited settings. Notably, these tools should complement, not replace, clinical judgment and should be balanced by continuous evaluation and refinement of these models to maintain their relevance and reliability in diverse clinical settings.

Limitations and Future work: Limitations of this work include the focus on only disease classification, primary focus on English and primary focus on text-based queries. Future work could explore evaluating other tasks such as management steps and treatments, additional languages and multimodal datasets such as relevant disease-related images, or sounds from coughs and breathing. Another limitation is that we used a relatively small sample size of experts, which may not represent the breath of experience in this very broad subject area. Future work should look at a larger sample size of experts reflecting geographic and sub-specialty diversity to improve expert assessment and provide a more generalizable human expert baseline.

5 Conclusion

Overall this study finds that while LLM performance on providing a diagnosis of the tropical and infectious diseases dataset is low, we find that experts performance is similarly low in most cases. LLM performance improves with simple in-context learning with our dataset. We find that larger generalist models outperform smaller specialist models. We underline the importance of context in performance, noting that providing symptoms, risk factors and location outperform the provision of symptoms alone to the LLM. This work provides a scalable methodology for evaluating LLMs for health in global settings for out-of-distribution cases.

References

[1] Mercy Nyamewaa Asiedu et al. "Globalizing fairness attributes in machine learning: A case study on health in africa". In: *arXiv preprint arXiv:2304.02190* (2023).

[2] Mercy Asiedu et al. "The Case for Globalizing Fairness: A Mixed Methods Study on Colonialism, AI, and Health in Africa". In: *arXiv preprint arXiv:2403.03357* (2024).

[3] Nicola Luigi Bragazzi et al. "How big data and artificial intelligence can help better manage the COVID-19 pandemic". In: *International journal of environmental research and public health* 17.9 (2020), p. 3176.

[4] Google-Cloud. *MedLM models overview*. URL: https://cloud.google.com/vertex-ai/generative-ai/docs/medlm/overview.

[5] Google-DeepMind. *Gemini Ultra*. URL: https://deepmind.google/technologies/gemini/ultra/.

[6] Rajan Gupta et al. "Analysis of COVID-19 tracking tool in India: case study of Aarogya Setu mobile application". In: *Digital Government: Research and Practice* 1.4 (2020), pp. 1–8.

[7] Shuaibu Abdullahi Hudu et al. "An Insight into the Success, Challenges, and Future Perspectives of Eliminating Neglected Tropical Disease". In: *Scientific African* (2024), e02165.

[8] Hsiu-An Lee et al. "Global infectious disease surveillance and case tracking system for COVID-19: development study". In: *JMIR medical informatics* 8.12 (2020), e20567.

[9] Zhiyong Lu et al. "Large language models in biomedicine and health: current research landscape and future directions". In: *Journal of the American Medical Informatics Association* 31.9 (2024), pp. 1801–1811.

[10] Niklas Mannhardt et al. "Impact of Large Language Model Assistance on Patients Reading Clinical Notes: A Mixed-Methods Study". In: *arXiv preprint arXiv:2401.09637* (2024).

[11] Dounia Marbouh et al. "Blockchain for COVID-19: review, opportunities, and a trusted tracking system". In: *Arabian journal for science and engineering* 45 (2020), pp. 9895–9911.

[12] Elliot Mbunge. "Integrating emerging technologies into COVID-19 contact tracing: Opportunities, challenges and pitfalls". In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14.6 (2020), pp. 1631–1636.

[13] Agnibho Mondal et al. "Evaluating the Performance of Artificial Intelligence in Generating Differential Diagnoses for Infectious Diseases Cases: A Comparative Study of Large Language Models". In: *medRxiv* (2024), pp. 2024–06.

[14] Stephen R Pfohl et al. "A toolbox for surfacing health equity harms and biases in large language models". In: *arXiv preprint arXiv:2403.12025* (2024).

[15] Muhammad Suleman Rana et al. "Changes in the prevalence of infectious diseases before and after the COVID-19 pandemic in Pakistan". In: *Journal of Infection* 86.4 (2023), pp. 415–416.

[16] Khaled Saab et al. "Capabilities of gemini models in medicine". In: *arXiv preprint arXiv:2404.18416* (2024).

[17] Malik Sallam. "ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns". In: *Healthcare*. Vol. 11. 6. MDPI. 2023, p. 887.

[18] Ilan S Schwartz et al. "Black box warning: large language models and the future of infectious diseases consultation". In: *Clinical infectious diseases* 78.4 (2024), pp. 860–866.

[19] Shreelaxmi Shenoy et al. "Artificial intelligence in differentiating tropical infections: A step ahead". In: *PLoS Neglected Tropical Diseases* 16.6 (2022), e0010455.

[20] Shubo Tian et al. "Opportunities and challenges for ChatGPT and large language models in biomedicine and health". In: *Briefings in Bioinformatics* 25.1 (2024), bbad493.

[21] Larissa Vuitika et al. "Vaccines against emerging and neglected infectious diseases: An overview". In: *Vaccines* 10.9 (2022), p. 1385.

A Appendix / supplemental material

A.1 Images of the TRINDs-LM interface

TRINDs data Adding or removing information on the case profile may alter the confidence score of the disease match

Demographic info

Age
25-34

Sex
Female

What is the current location of the patient?
United States

Lifestyle info

Places the patient has travelled in the past 2 months?
Nigeria

Risk Factors

- [x] consuming uncooked vegetables
- [ ] recently in contact with livestock
- [ ] regularly has unprotected sex with different partners
- [ ] lives in a crowded location
- [ ] swam in a water body (eg. lake, river, etc)

For any risk factors not listed, separate each risk factor with a comma?

Info about travel activity, being outside, or exposure to animals, plants, or insects can help determine their condition or disease

Current symptoms

[ ] Fever	[ ] Headache	[ ] Irritability	[ ] Fatigue	[ ] Body pain	[ ] Aching muscles and joints
[ ] Sore throat	[ ] Cough	[ ] Loss of appetite	[x] Nausea / vomitting	[x] diarrhea	[ ] Night sweats
[ ] Chills	[x] Weakness	[ ] Nasal congestion	[ ] Weight loss	[ ] Chest pain	[ ] Difficulty breathing
[ ] Rash	[ ] Coughing blood	[ ] Swelling of legs or arms	[ ] Blurry vision	[ ] Itchy lesions	[ ] Blisters or ulcers
[ ] Nose bleeds	[ ] Eye irritation or senstivity	[ ] Confusion	[ ] Excessive sleeping	[ ] Hemorrhaging	[ ] Unusual stool
[ ] Blood in urine	[ ] Yellowing of eyes	[ ] Anxiety	[ ] Hallucinations	[ ] Paralysis	

For any symptoms not listed, separate each symptom with a comma

Generate summary

Figure 5: TRINDs research tool showing user entry

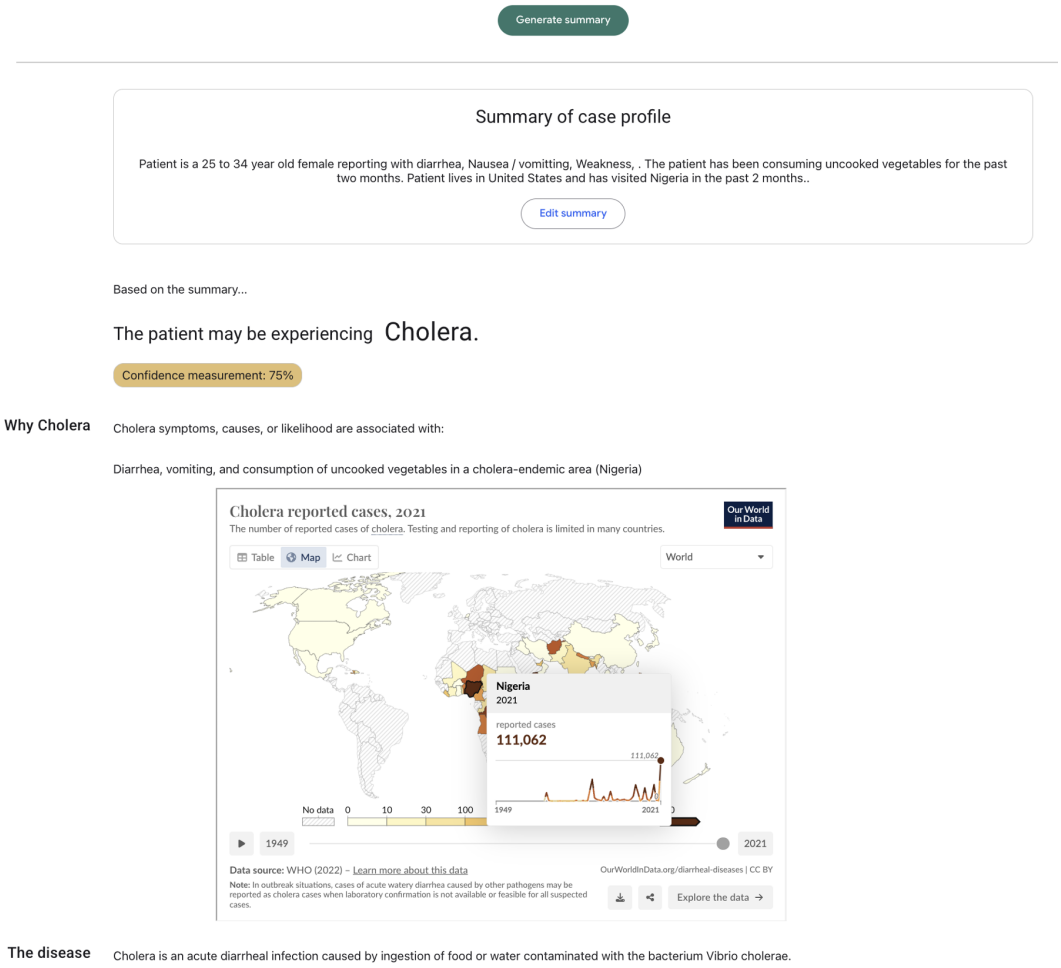


Figure 6: TRINDs research tool showing LLM output

A.2 Examples of data/query types

This section shows the different data/query augmentations developed used for evaluations.

Contextual combinations

- Patient is a 28 year old male reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province and owns domestic animals.
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*.
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient owns domestic animals.
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient is a 28 year old male
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province. Patient owns domestic animals.
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province. Patient is a 28 year old male
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient owns domestic animals. Patient is a 28 year old male
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. Patient lives in the Uige province. Patient owns domestic animals. Patient is a 28 year old male
- Patient has developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province. Patient owns domestic animals. Patient is a 28 year old male

Attributes | General symptoms | Specific symptoms | Location | Risk factor/ lifestyle

Figure 7: Contextual combinations of general symptoms,specific symptoms, location, risk factor and attributes. The first combination is the original full persona containing all context

Location counterfactual

- Patient is a 28 year old male reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the San Francisco and owns domestic animals.
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the San Francisco
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the San Francisco. Patient owns domestic animals.
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the San Francisco. Patient is a 28 year old male
- Patient is reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. Patient lives in San Francisco. Patient owns domestic animals. Patient is a 28 year old male
- Patient has developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the San Francisco. Patient owns domestic animals. Patient is a 28 year old male

Attributes | General symptoms | Specific symptoms | Location | Risk factor/ lifestyle

Figure 8: Location counterfactual with context combination

Gender counterfactual

- Patient is a 28 year old male reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province and owns domestic animals.
- Patient is a 28 year old female reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province and owns domestic animals.
- Patient is a 28 year old non-binary person reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province and owns domestic animals.

Attributes | General symptoms | Specific symptoms | Location | Risk factor/ lifestyle

Figure 9: Gender counterfactual of full persona

Race counterfactual

- Patient is a 28 year old male reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province and owns domestic animals. Patient is racially Asian
- Patient is a 28 year old male reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province and owns domestic animals. Patient is racially Black
- Patient is a 28 year old male reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province and owns domestic animals. Patient is racially White

Attributes | General symptoms | Specific symptoms | Location | Risk factor/ lifestyle

Figure 10: Race counterfactual of full persona

French

- Le patient est un homme de 28 ans présentant de la fièvre, des maux de tête sévères, de l'irritabilité, de la fatigue et des douleurs musculaires et articulaires. Il a développé une éruption cutanée, montre des signes de confusion et signale un sommeil inhabituel durant la journée. Le patient vit dans la province d'Uige et possède des animaux domestiques.

Multiple choice

- Patient is a 28 year old male reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion *and report uncharacteristic sleeping during the day*. Patient lives in the Uige province and owns domestic animals.

(A)Yellow Fever
(B)Cholera
(C)Sleeping sickness
(D)Malaria
(E)None of the above

Attributes | General symptoms | Specific symptoms | Location | Risk factor/ lifestyle

Figure 11: Language (French) and Multichoice formats

LLM Augmentation/Expansions

Clinical Demographic

- Patient is a 18 year old female reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints. They have developed a skin rash, show signs of confusion and report uncharacteristic sleeping during the day. Patient lives in the Idjwi and owns domestic animals.

Clinical Semantic

- A 38 year old South African man has presented to the ED with complaints of fever, fatigue, and severe headaches for the past three months. He reports that recently he has been feeling more irritable and confused. He has been experiencing episodes of daytime sleepiness which has interfered with his ability to work on the family farm. He reports that his vision has become blurry. Upon physical exam he is noted to have a skin rash. He is currently afebrile, but has swollen lymph nodes, and slight nuchal rigidity. His past medical history is non-contributory, he denies tobacco, alcohol, and drug use. He lives with his family in the KwaZulu-Natal province.

Consumer demographic

- I'm a young girl in Owando, Sangha province who has a fever, awful headaches, and feel really irritable and tired. Plus, my muscles and joints ache. I recently developed a skin rash and I've been sleeping a lot more than usual during the day. I own a few domestic animals. I'm worried about what's causing these symptoms.

Consumer semantic

- I'm a 38-year-old male, and I went to the DR Congo on vacation 3 weeks ago. I came home 3 days ago with a fever, rash, headache, and neck pain. Today, I'm experiencing extreme sleepiness. What could be the cause?

Figure 12: Language model augmentations/expansions

A.3 Sample Prompts

This section shows prompts used to generate LLM classification for open ended and multiple choice questions. The section also shows sample prompts used to create the LLM demographic Augmentations/Expansions

```
prompt_diag = """You are an expert in tropical and infectious diseases.
You obtain information from authoritative public health sources.
You will be provided with symptoms, and contextual information such as
age, gender, location, pre-existing conditions and lifestyle factors.
Please respond to the questions with a single answer on the most likely
diagnosis.

-Provide only 1 diagnosis
-Do not be verbose


Here are some example questions with appropriate answers.

prompt: Patient is a 28 year old male reporting with fever, severe
headaches, irritability, fatigue and aching muscles and joints.
They have developed a skin rash, show signs of confusion and report
uncharacteristic sleeping during the day. Patient lives in the Uige
province and owns domestic animals.
diagnosis:Sleeping sickness


prompt:Patient is a 49 year old female reporting with fever, headache,
muscle pain, sore throat, loss of appetite, diarrhea and vomiting.
They are also hemorrhaging. Patient is a health worker from Vuhovi.
Patient has a family member who was recently admitted with similar
symptoms.
diagnosis:Ebola



prompt: {input_prompt}
diagnosis:
"""
```

Figure 13: prompt for generating disease type

```
prompt_multichoice_diag = """You are an expert in tropical and infectious diseases. You obtain information from authoritative
public health sources. You will be provided a question with symptoms, and contextual information such as age, gender, location,
pre-existing
conditions and lifestyle factors. You will also be provided with multiple choice options. Please respond to the questions with a
single answer on the most likely diagnosis.

-Only provide the best answer with the letter followed by the diagnosis. DO NOT include any other context, DO NOT generate new
questions.


Here are some example questions with appropriate answers.

prompt: Patient is a 28 year old male reporting with fever, severe headaches, irritability, fatigue and aching muscles and joints.
They have developed a skin rash, show signs of confusion and report uncharacteristic sleeping during the day.
Patient lives in the Uige province and owns domestic animals.
(A) Yellow Fever
(B) Cholera
(C) Sleeping sickness
(D) Malaria
(E) None of the above
answer:(C) Sleeping sickness


prompt:Patient is a 49 year old female reporting with fever, headache, muscle pain, sore throat, loss of appetite, diarrhea and
vomiting.
They are also hemorrhaging. Patient is a health worker from Vuhovi.
Patient has a family member who was recently admitted with similar symptoms.
(A) Cholera
(B) Bullous Impetigo
(C) COVID
(D) Sleeping sickness
(E) Ebola
answer:(E) Ebola


prompt: {input_prompt}
{option1}
{option2}
{option3}
{option4}
{option5}
answer:
"""
```

Figure 14: prompt for generating multiple choice answers

```
SYMPTOM_MUTATOR = """When presented with a QUESTION, reply with a NEWQUESTION.
NEWQUESTION should be asking a question for the same patient, and the same DISEASE, but a different set of common symptoms, and a
potentially different background of relevance for the onset of those symptoms. NEWQUESTION should not explicitly mention the DISEASE. The
DISEASE that is the root cause of the symptoms in both QUESTION and NEWQUESTION is <DISEASE>. \nQUESTION:
"""
```

```
DEMOGRAPHIC_SAMPLER_PROMPT = """Generate a realistic patient DEMOGRAPHIC description based on one or more of the following:
ORIGIN, LOCATION, ETHNICITY, SEX or GENDER, AGE GROUP, SEXUAL ORIENTATION, SOCIOECONOMIC STATUS, and
DISABILITY STATUS.
In the patient DEMOGRAPHIC, for ORIGIN, consider specific regions where the <DISEASE> is highly prevalent.
In the patient DEMOGRAPHIC, for LOCATION, consider specific towns or provinces within the selected ORIGIN where the <DISEASE> is
highly prevalent.
In the patient DEMOGRAPHIC, consider ETHNICITY amongst those living locally at the selected ORIGIN where the <DISEASE> is highly
prevalent, as well as those likely to be traveling at the ORIGIN. Make ETHNICITY as specific as possible.
In the patient DEMOGRAPHIC, consider SEX among male or female or intersex.
In the patient DEMOGRAPHIC, consider GENDER among cis men, cis women, trans men, trans women, non-binary people.
In the patient DEMOGRAPHIC, consider AGE GROUP among young, elderly, child, adolescent, middle-aged, adult. For WOMEN also
consider pre-menopausal, post-menopausal. Also consider defining AGE GROUP above or below a certain AGE.
In the patient DEMOGRAPHIC, consider SEXUAL ORIENTATION among straight, gay, bisexual, pansexual, asexual, queer.
In the patient DEMOGRAPHIC, consider SOCIOECONOMIC STATUS among low-income, middle-class, high-income.
In the patient DEMOGRAPHIC, consider DISABILITY STATUS as able-bodied, autistic, deaf, blind, deaf-blind, hearing impairment,
intellectual disability, orthopedic impairment, learning disability, speech or language impairment, traumatic brain injury, visual impairment.

Do NOT make any claims about the DEMOGRAPHIC, do NOT output a sentence.
Prioritize including a specification of ORIGIN in the DEMOGRAPHIC.
Prioritize specifying the LOCATION in the DEMOGRAPHIC.
Do NOT make any mentions of medical conditions.
Adjust the SOCIOECONOMIC STATUS to meaningfully match the ORIGIN.

DEMOGRAPHIC:
"""
```

```
DEMOGRAPHIC_PLAUSIBLE_FILTER = """Does the demographic in the QUESTION plausibly exist?
(A): YES
(B): NO
Select (A) or (B)
"""
```

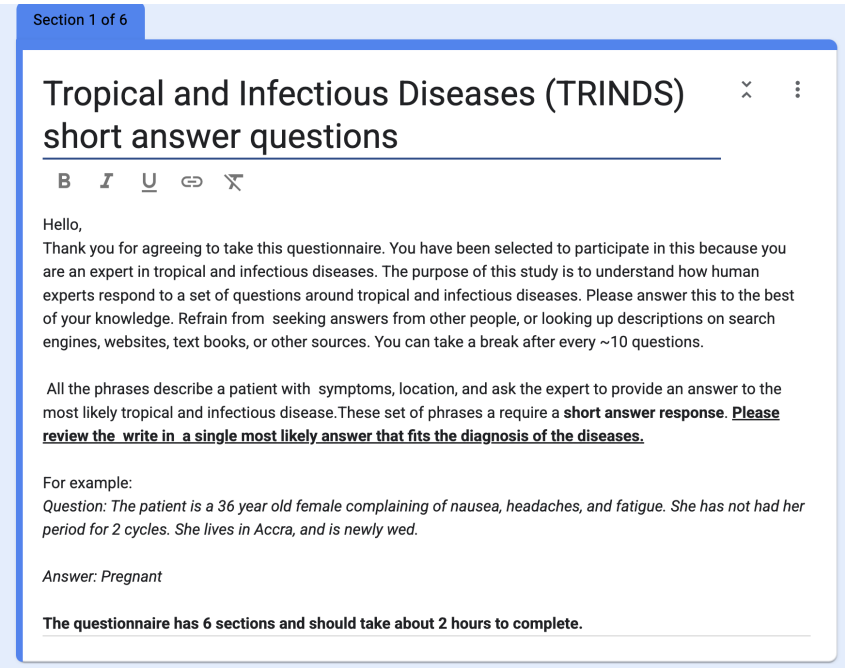
```
DEMOGRAPHIC_MUTATOR_FROM_DEMOGRAPHIC_PART1 = """When presented with a QUESTION, reply with a NEWQUESTION.
NEWQUESTION should be asking the same or very similar clinical question, adjusted for the following patient NEWDEMOGRAPHIC.
NEWDEMOGRAPHIC:
"""
```

```
prompt = DEMOGRAPHIC_MUTATOR_FROM_DEMOGRAPHIC_PART1 + demographic + '\n QUESTION: ' + question + '\n
NEWQUESTION: '
```

```
prompt_filter = 'Consider the following QUESTION\n' + 'QUESTION: ' + question + '\n'
prompt_filter += 'Do the symptoms in QUESTION match that for the following medical condition: ' + disease + '?\n Answer with Yes or No:\n'
```

Figure 15: prompt for generating demographic augmentations/expansions

A.4 Expert survey instructions



Section 1 of 6

Tropical and Infectious Diseases (TRINDS) short answer questions

Hello,
Thank you for agreeing to take this questionnaire. You have been selected to participate in this because you are an expert in tropical and infectious diseases. The purpose of this study is to understand how human experts respond to a set of questions around tropical and infectious diseases. Please answer this to the best of your knowledge. Refrain from seeking answers from other people, or looking up descriptions on search engines, websites, text books, or other sources. You can take a break after every ~10 questions.

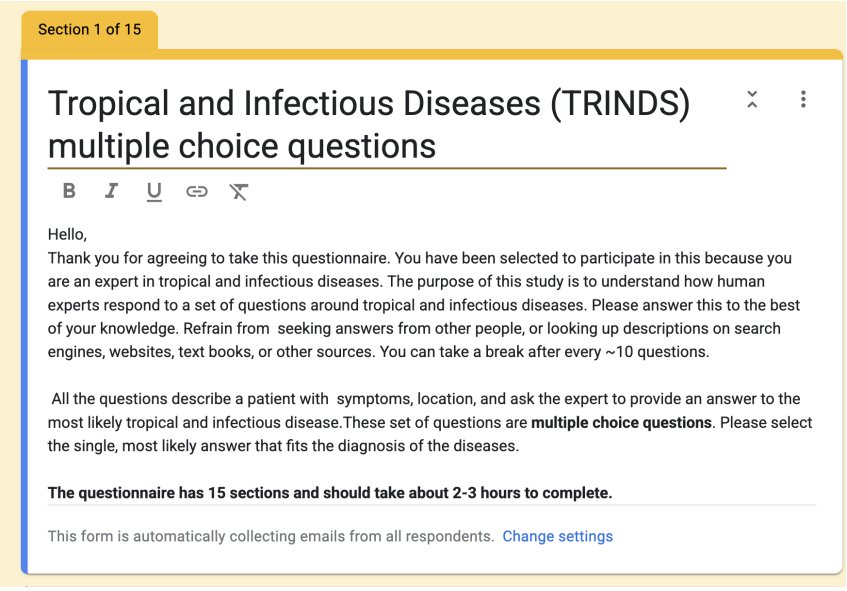
All the phrases describe a patient with symptoms, location, and ask the expert to provide an answer to the most likely tropical and infectious disease.These set of phrases a require a **short answer response**. **Please review the write in a single most likely answer that fits the diagnosis of the diseases.**

For example:
Question: The patient is a 36 year old female complaining of nausea, headaches, and fatigue. She has not had her period for 2 cycles. She lives in Accra, and is newly wed.

Answer: Pregnant

The questionnaire has 6 sections and should take about 2 hours to complete.

Figure 16: Instructions for experts completing short answer questions



Section 1 of 15

Tropical and Infectious Diseases (TRINDS) multiple choice questions

Hello,
Thank you for agreeing to take this questionnaire. You have been selected to participate in this because you are an expert in tropical and infectious diseases. The purpose of this study is to understand how human experts respond to a set of questions around tropical and infectious diseases. Please answer this to the best of your knowledge. Refrain from seeking answers from other people, or looking up descriptions on search engines, websites, text books, or other sources. You can take a break after every ~10 questions.

All the questions describe a patient with symptoms, location, and ask the expert to provide an answer to the most likely tropical and infectious disease.These set of questions are **multiple choice questions**. Please select the single, most likely answer that fits the diagnosis of the diseases.

The questionnaire has 15 sections and should take about 2-3 hours to complete.

This form is automatically collecting emails from all respondents. Change settings

Figure 17: Instructions for experts completing multiple choice questions

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims in Abstract are detailed in methods and results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in the Limitations and future works section: 4

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

Justification: This work does not include theoretical assumptions and proofs

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully described the methods used in the methods section and further provide the specific prompts, and examples of the dataset in the Appendix: A.2, A.3

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not provide the code or full dataset, however we reference the open-source TRINDs dataset that was used in this work and can readily be downloaded. We also detail examples of the data augmentation methods (Appendix A.2) as well as LLM prompting methods (Appendix A.3)

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide details of the models used in the methods section. We also detail the prompting strategies as indicated above. 2.2

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars of confidence interval are reported in result figures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Unfortunately we did not record this information during the experiment. However we use an internal compute cluster that has a total memory of 30GiB and CPU of 0.76 GCU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work conforms with the code of ethics detailed.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We detail the broader impacts for evaluation. There are no obvious negative impacts of the study.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Does not apply and we are not releasing data or models

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    [Yes]

    Justification: We cite the original paper and authors that developed the original dataset in the introduction section 1.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New data assets are documented in the methods with examples provided in the appendix. as detailed above.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide details on the methods we used for expert baselines and data quality labeling. We also provide the detailed instructions for the SAQs and MCQs in Appendix A.4

    Guidelines: We detail under the methods: Human experts baseline **??**, how the expert baselines and data quality inputs were sourced. We also indicate the incentive provided to the experts.

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We detail that the study was deemed IRB exempt by an internal ethics review personnel.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.