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

Mercy Asiedu[1], Nenad Tomasev[2], Chintan Ghate[1], Tiya Tiyasirichokchai[1], Awa Dieng[2], Oluwatosin Akande[3], Geoffrey Siwo[4], Steve Adudans[5], Sylvanus Aitkins[6], Odianosen Ehiakhamen[7], Eric Ndombi[8], Katherine Heller[1]

1 Google Research, 2 Google Deepmind, 3 World Health Organization, 4 University of Michigan, 5 Gear Health Kenya, 6 University of Sierra Leone, 7 Nigerian Center for Disease Control, 8 Kenyatta University. **Corresponding author contact:** masiedu@google.com

Introduction

- Tropical and infectious diseases (TRINDs) continues to be highly prevalent in the poorest regions of the world, affecting 1.7 billion people globally with disproportionate impacts on women and children[1].
- Challenges in preventing and treating these diseases include limitations in **surveillance, early detection, accurate initial diagnosis**, management and vaccines[2].
- The use of large language models(LLMs) for health-related question-answering has demonstrated promise however, there is limited work focused on TRINDs.
- There is also 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.
- We develop the **TRINDs dataset** for evaluating LLMs and demonstrate through systematic experimentation, **the effect of contextual information** on LLM outputs for health.

Methods

- We manually create the TRINDs dataset of synthetic seed personas (n=50) across 50 diseases, using authoritative sources.
- We utilize LLM prompting to expand the dataset to include demographic and semantic clinical and consumer augmentations 11000+).
- We perform evaluations with the dataset, to understand how different contexts, types (clinical vs consumer), demographics, semantic styles and counterfactuals contribute to LLM performance on disease classification.
- We evaluate LLM performance improvements on expanded demographic and semantic datasets after simple in-context prompt tuning with the seed set.
- 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.

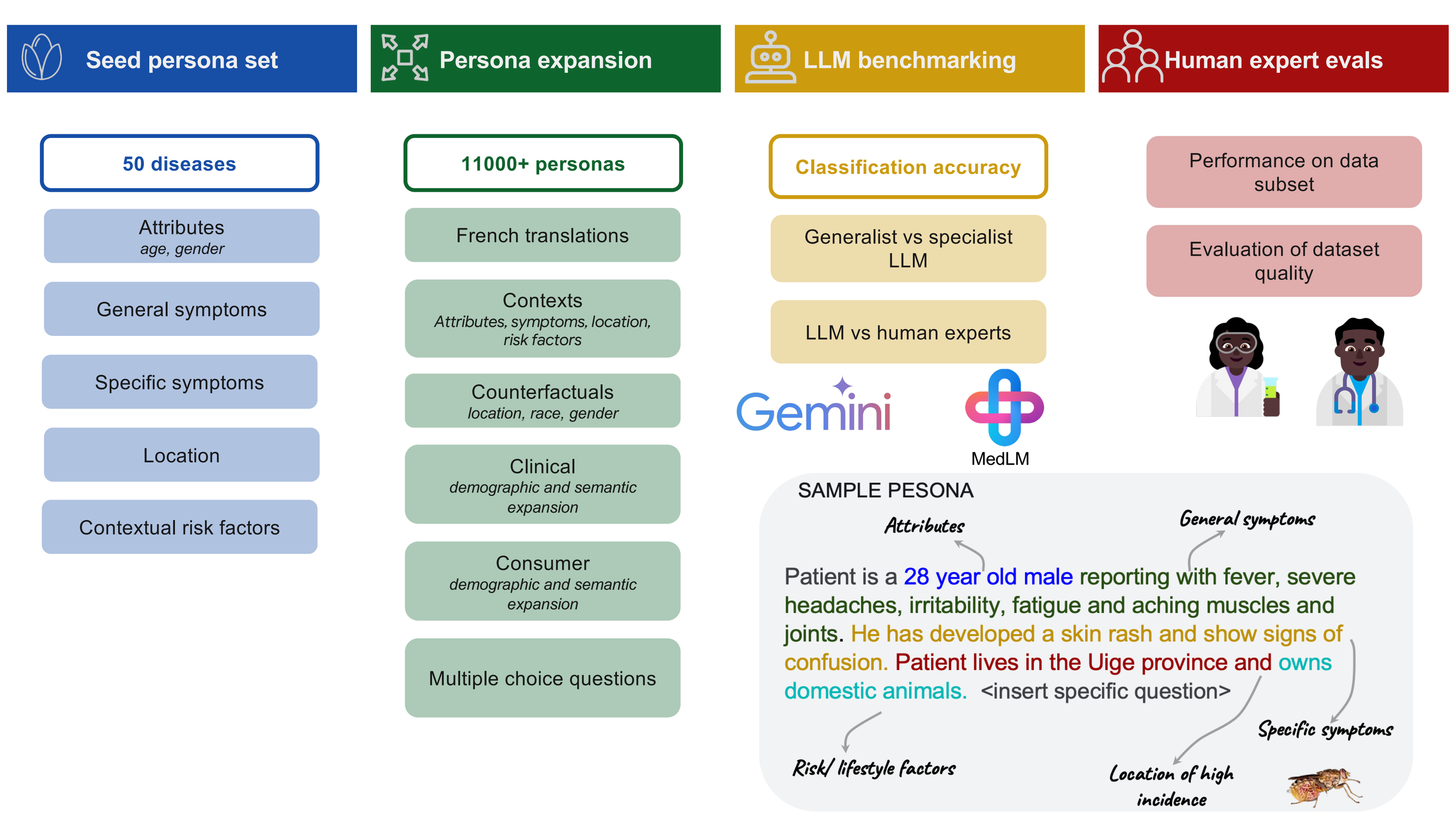

Figure 1: Methodology overview with sample persona

Table 1: Summary of datasets and experiments

Dataset Augmentation	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

Results

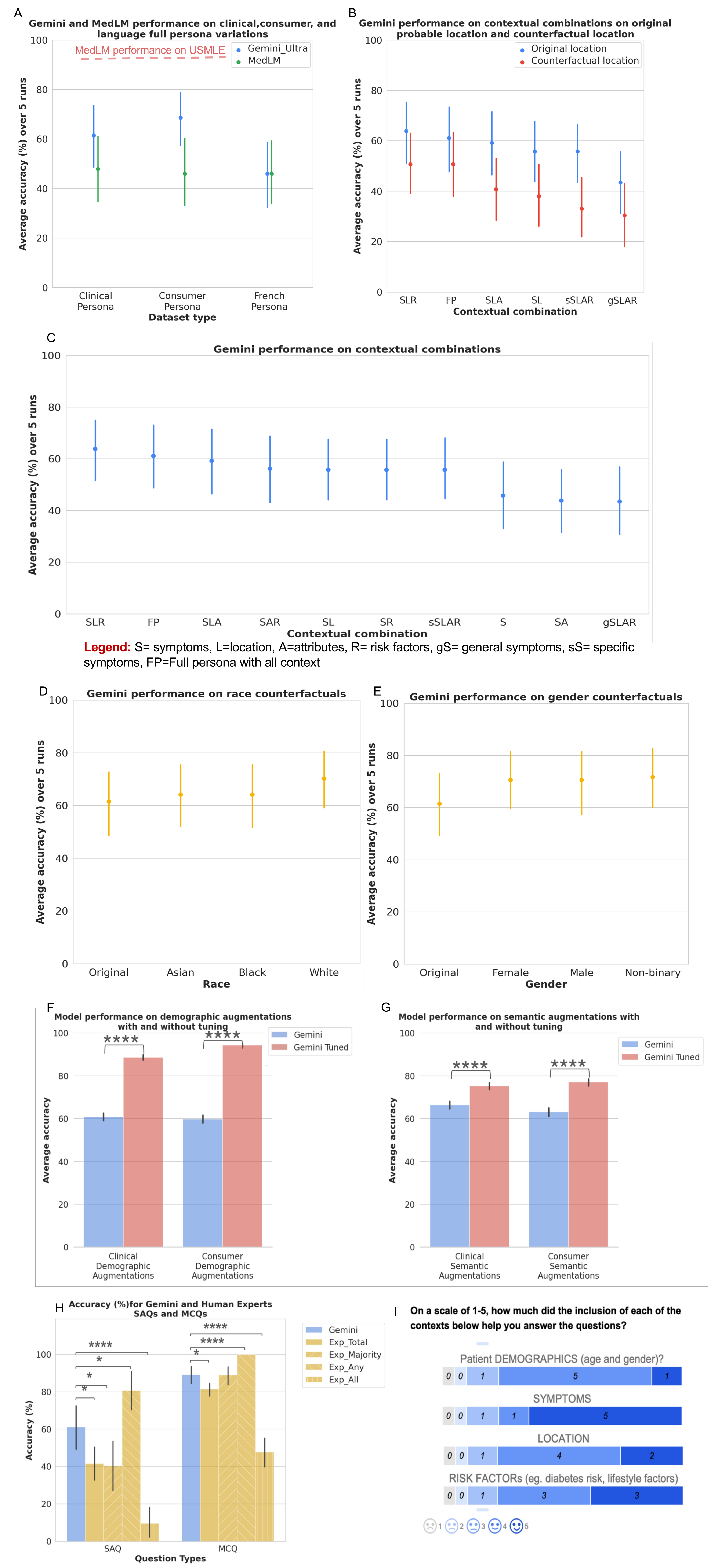

Figure 2: Summary of results. A) Generalist vs specialist model. B) Contextual location counterfactuals. C) Contextual combinations. D-E) Race and gender counterfactuals. F-G) Clinical and consumer demographic and semantic augmentations with and without in-context tuning. H) LLM vs expert performance. I) Perceived assistance of context on expert decisions.

Results

- Results demonstrate a distribution shift with Gemini 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[3] (GPT: 90.2%, MedLM: 91.1%) (Fig.2A). Simple in-context prompting with the dataset improves the LLM performance (Fig.2D,E).
- We find that generalist model-Gemini Ultra performs better than specialist model-MedLM, however this is likely due to differences in model sizes (Fig.2A).
- We find that LLMs tend to more accurately identify common diseases, or diseases with very specified symptoms (Figure in paper).
- Evaluations demonstrate that including additional context such as risk factors and location in addition to symptoms also improves model performance (Fig.2B,C).
- Our human expert baseline finds that for both short answer response questions (SAQs) and multiple choice questions (MCQs), experts scored lower in accuracy on the full context questions than the model except in cases where we looked at scenarios where any/at least one expert was correct (Fig. 2H).
- Experts found symptoms and risk factors to be most helpful in decision making (Fig.2I). They 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.

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.

References

Shuaibu Abdullahi Hudu et al. "An Insight into the Success, Challenges, and Future Perspectives of Eliminating Neglected Tropical Disease". In: Scientific African (2024)
Larissa Vuitika et al. "Vaccines against emerging and neglected infectious diseases: An overview". In: Vaccines 10.9 (2022), p. 1385.
Khaled Saab et al. "Capabilities of gemini models in medicine". In: arXiv preprint arXiv:2404.18416 (2024).