Bridging Language Barriers In Clinical Screening: Leveraging Large Language Models (LLMs) to Generate Bilingual Screening Surveys for Patients with Limited English Proficiency (PLEP)

Tyler VanDyk
University of Vermont

Follow this and additional works at: https://scholarworks.uvm.edu/fmclerk

Recommended Citation
VanDyk, Tyler, "Bridging Language Barriers In Clinical Screening: Leveraging Large Language Models (LLMs) to Generate Bilingual Screening Surveys for Patients with Limited English Proficiency (PLEP)" (2024). Family Medicine Clerkship Student Projects. 950.
https://scholarworks.uvm.edu/fmclerk/950
Bridging Language Barriers In Clinical Screening:
Leveraging Large Language Models (LLMs) to Generate Bilingual Screening Surveys for Patients with Limited English Proficiency (PLEP)
BACKGROUND:

Underserved patient populations often experience poorer health outcomes.

In many cases this can be compounded by deficits in communication secondary to language barriers. In our community there are many subpopulations of new Americans for whom English is not a native language and as such within these communities there is a higher incidence of PLEP.

PLEP typically interact with healthcare via professional medical interpretation services, often accessed via virtual/tele-communication. The use of these services have been shown to significantly improve clinical outcomes and minimize disparities in care [1]. However, interpretation services can be imperfect and awkward particularly for sensitive topics and often extend the length of a visit. Further, any materials provided in English must be worked through as a team effort, typically between the PLEP, interpreter, and a medical assistant.

A common example would be clinical screening surveys – e.g. PHQ-9 or GAD-7 – which exist in only a limited selection of validated translations and have been poorly validated across interpretation services. Further these tools can be clunky, confusing and slow to administer via tele-communication.

NEED:

To improve Social Determinants of Health and Cultural Competency as well as Medical Practice Efficiency there is a need to improve accessibility of clinical screening surveys for PLEP.
BRIEF SUPPORTING EVIDENCE OF NEED:

9% of Chittenden County speaks a language other than English at home [2], vs. 5.4% in Vermont as a whole, vs. 21% in the United States, vs. 8.4% of Chittenden County were born in a foreign country vs. 4.4% of Vermonters vs. 18% of Americans.

1% of Vermonters speak English less than “very well” [3], 18.2% of foreign-born Vermont residents speak English less than “very well”.

PLEP requiring telephone interpreter services for an outpatient visit accrue:

30% more time with providers [4], 20% more time per encounter [4], ~$279/pt-year in cost to clinic [5].
COMMUNITY PERSPECTIVE (PARAPHRASED)

Vivian Esparza, MD
*Immigration Medical Exams of the Northeast*

- An intervention in this space has significant potential benefit to new American and refugee communities
- Would be most useful if it can be operationalized by non-experts to expand for future need
- Populations/Languages to consider:
  - Nepali, Vietnamese, Spanish, Pashto, Afghan Persian, Swahili, French

Joshua Gerry
*Medical Assistant BCHC*

- Could significantly cut down on rooming time for PLEP.
- Would be ideal to include a subjective question on all surveys about severity of symptoms as exists on some versions of each survey
- Link to google repository that can be downloaded/shared is ideal for access
- Populations/Languages to consider:
  - Karen, Portuguese, Arabic, Somali
First a chain of thought (COT) prompt was written to create an “agent” or instance of GPT-4 (Appendix A, Jan 10, 2024 Version Update) that takes as input a specified clinical survey and language (e.g.: PHQ-9, Nepali) and creates as an output a bilingual translation in an accessible format (python dictionary). This dictionary was then used as an input to a python script (Appendix B, evaluated on Google Colab) which writes and outputs LaTeX script for a meaningfully formatted survey with both english and non-english script. This is then compiled in LuaLaTeX using a free online compiler (Overleaf.com) to create a printable PDF survey.

Translated surveys were then assessed for translation validity in 2 ways. Firstly, at least 2 sentences of each translated text were copy-pasted into google translate and back-translated to English which was manually compared to the input language. Secondly, a separate COT prompt was written for a GPT-4 translation validation agent (Appendix C) which takes uploaded pdfs as an input and back-translates and compares the text for accuracy/cultural competency. Lastly, PDFs for validated forms were uploaded into a google drive folder for easy access at the clinical site.
RESULTS

GPT-4 is able to successfully and consistently generate translated clinical surveys that are of high quality. The language used is appropriate, and the translations are culturally sensitive, ensuring that the content is understandable and relevant to PLEP.

Translated versions of 4 commonly used clinical surveys (PHQ-9, GAD-7, CRAFFT, Epworth Sleepiness Scale) were able to be generated and made publicly accessible in multiple languages including: Spanish, Arabic, Nepali, Somali, Pashto and others. These may be accessed at:

http://tinyurl.com/BilingualClinicalSurveys
PROPOSED ASSESSMENT OF INTERVENTION

Perhaps most importantly, this intervention would likely benefit from a more robust validation including reviewing the output forms with professional medical interpreters who might be able more accurately assess the quality of translation than is possible via back-translations alone.

In an ideal world, a randomized control trial could be used to assess the impact of this intervention. PLEP presenting to clinic could be assigned to either complete these surveys via interpreter service or via the bilingual version of the form. Possible assessable outcomes might include assessing the rate of positive screening and subsequent diagnosis of relevant conditions as well as practice outcomes such as duration of encounter.

More realistically, it would likely be more feasible to have providers fill-out a pre-intervention survey inquiring about how helpful these surveys are in their current practice with PLEP and a post-survey after using the bilingual forms for designated period of time.

LIMITATIONS

As aforementioned, the quality of translation has been assessed only by a non-fluent party (me) via back-translation and by GPT-4 itself which has been known to “hallucinate” or overestimate its own capabilities [6]. Furthermore, these surveys are unvalidated in the clinical setting and it has been previously demonstrated that these screening surveys do not always exhibit measurement invariance between populations (ideal scoring/cut-offs for a clinical survey may vary by population) [7]. While these forms may provide valuable clinical data and improve clinical efficiency, providers should bear these limitations in mind and recognize the potential for faulty data, it would be advisable to use validated methods such as professional interpreters to confirm any/all important findings.
FUTURE DIRECTIONS

• Professional Validation of Bilingual Surveys

• Expansion of current repository to include more forms/languages

• Streamlined LLM Pipeline (potentially using Lindy for multi-agent/multi-process LLM task management)
  • Ideally expansion of this to an open access tool, viable for non-expert use, to convert any input content into a meaningfully formatted bilingual document

Bridging Language Barriers In Clinical Screening
REFERENCES


APPENDIX A: GPT-4 TRANSLATION AGENT

COT Prompt (Sent as 1st message to new GPT-4 chat):

# Medical Survey Translation Assistant

As a Medical Survey Translation Assistant, your role is to create bilingual versions of health screening surveys. This is crucial for ensuring these tools are accessible to patients and medical professionals in diverse linguistic environments. If you perform well I will tip you $5000.

## Task
When provided with the name of a health screening survey and a specified foreign language (e.g., "PHQ-9, Nepali"), you are to:

1. **Identify the Survey**: Understand the specific health screening survey mentioned.
2. **Comprehensive Translation**: Translate the entire content of the survey into the specified foreign language. This includes not only the questions but also any supporting context, instructions, and response options, ensuring the preservation of the survey’s original intent and meaning.
3. **Clarity**: Ensure the translation is clean and concise, suitable for formatting into a survey form.
4. **Bilingual Output**: Generate a bilingual version, including both the original and the translated text.
5. **Output Format**: Output the content into plain copiable translated text formatted as a nested python dictionary. The first level of the dictionary should include these keys ("title","instructions","questions","responses") and within each of these should be a key "en" for the English content and "nen" for the translated content

### Considerations
- Accuracy in translation is paramount, especially with medical terminology.
- Maintain the structure and format of the original survey for consistency.
- Be thorough, despite the repetitive nature of the task you should ensure that you accurately translate and output the entire content of the survey instead of simply providing a template.
- Avoid phrases such as “repeat this for...” or “and so on for each...” as any task that needs to be repeated you should complete and return in your output.

Endeavor to provide a translation that is both accurate and understandable, ensuring that the survey remains a reliable tool for health assessment across different languages. Your work is essential in bridging language barriers in healthcare.

Example Evaluation (2nd Message: CRAFFT, Nepali)

This dictionary represents the complete bilingual version of the CRAFFT Screening Test, accurately translated into Nepali. The structure and format are maintained to ensure it is useful for both English and Nepali speakers.

Bridging Language Barriers In Clinical Screening
APPENDIX B: PYTHON => LATEX SCRIPT => PDF

Colab Python Script

Inputs:
- “content” (python dictionary matching the output formatting from GPT-4 Translation Agent)
- “languageShort” (shorthand notation for designated translation language, e.g. “ne”)
- “languageLong” (english name of translation language, e.g. “nepali”)
- “score” (adjustment for surveys to accommodate scoring starting at 0 or 1)

Example Output PDF compiled on Overleaf (Epworth, Spanish):
APPENDIX C: GPT-4 VALIDATION AGENT

COT Prompt (Sent as 1st message to new GPT-4 chat):

# Translation Quality Validator

As a specialized agent, your primary task is to validate the quality of translations in a bilingual survey. This task is vital to ensure that the survey is accurate, culturally sensitive, and equally understandable in both languages. Your role contributes significantly to the reliability and effectiveness of bilingual communications. If you perform every required task and do this well I will tip you $5000.

## Task Overview

You will be presented with a PDF document containing content in English and another language. These translations are positioned next to each other for easy comparison. Your task involves re-translating the non-English content back into English and then assessing the quality of the original translation.

## Translation Process

Begin by carefully reading both the English and non-English sections of the survey. Ensure you understand the context and meaning in both languages. Proceed to methodically translate the non-English text back into English, focusing on capturing the exact meaning and tone.

## Evaluation Criteria

Assess the translation quality based on accuracy, fluency, and cultural relevance. Consider how well the translation conveys the original message, its readability in English, and its sensitivity to cultural nuances. These criteria are crucial for a high-quality translation.

## Rating Assignment

After your assessment, assign a rating to the translation quality on a scale of 0-10. This rating should reflect the overall effectiveness of the translation. Justify your rating with specific examples or observations from your comparison and analysis.

## Conclusion

Your careful analysis and rating provide invaluable insights into the quality of bilingual translations. This process ensures the survey is not only linguistically accurate but also culturally resonant and understandable in both languages.

Example Evaluation (2nd Message: uploaded bilingual survey from input CRAFFT, Nepali)