Assessing the Feasibility of Processing a Paper-based Multilingual Social Needs Screening Questionnaire Using Artificial Intelligence

Obinna I. Ekekezie, M.D.[a,b]

[a] *Cambridge Health Alliance, Cambridge, MA, United States of America*
[b] *Harvard Medical School, Boston, MA, United States of America*

Description

This study explores the feasibility of using artificial intelligence (AI) to transform a response to a paper-based Social Determinants of Health (SDoH) questionnaire into a structured representation that could theoretically be persisted to an electronic health record (EHR). It suggests there is potential for AI, particularly document understanding and large language models, to offer a cost-effective, accurate alternative to manual data entry of responses to paper-based questionnaires.

Abstract

Recent initiatives by healthcare payers to mandate the collection of Social Determinants of Health (SDoH) data underscore the importance of efficient and accurate data capture methods in improving patient care. Traditional paper-based questionnaires, while widely used, present challenges in terms of cost, accuracy, and completeness of data entry into Electronic Health Records (EHRs). This study investigates the application of artificial intelligence (AI), specifically document understanding models and large language models (LLMs), to automate the transformation of paper-based SDoH questionnaires into structured, machine-readable formats suitable for EHR integration. Utilizing a test dataset derived from the Cambridge Health Alliance SDoH questionnaire, available in eight languages, this study explored the feasibility of using Microsoft Azure Document Intelligence and OpenAI's GPT-4 and GPT-3.5 Turbo, which were selected for their advanced capabilities in document understanding and language processing, respectively, for this purpose. The findings show that GPT-4 outperforms GPT-3.5 Turbo across various metrics, including accuracy and consistency, albeit at a higher cost. This feasibility study highlights the potential of using AI as a relatively accurate and potentially cost-effective alternative to manual data entry of SDoH data collected using paper-based questionnaires. It also suggests there will be challenges such as data privacy and security considerations as well as the integration of AI-generated data into EHR systems that merit further research.

Introduction

In recent years, both federal and state healthcare payers have been incentivizing or even mandating the collection of information about the social determinants of health (SDoH), which refer to the social and economic factors, such as social isolation, interpersonal violence, or lack of access to housing, transportation, or food, that can impact the health and well-being of individuals[1,2].

Like many other forms of personal or demographic information asked of patients, this data is often collected using paper-based questionnaires that are later manually entered into an electronic health record (EHR) by clinicians or non-clinicians[2]. In addition to manual data entry being expensive, it is worth noting that previous studies have found that manually entered data is often less accurate and less complete than information that is extracted using automations[3]. Some of these issues can be mitigated by administering the questionnaires in an electronic format using tablets and other devices; however, developing such capabilities can require deep technical expertise as well as substantial investments of time and resources that many healthcare providers may not have[4].

With recent advancements in artificial intelligence (AI), particularly with respect to document understanding models and using large language models (LLMs) for natural language processing (NLP) tasks, it is now more feasible than ever to extract structured data from documents[5]. Being able to reliably and efficiently extract structured data from paper-based questionnaires, could help more healthcare providers meet these SDoH-related mandates. This study sought to explore the feasibility of using AI to generate a structured representation of a paper-based SDoH questionnaire.

Methods

Dataset

The SDoH questionnaire used in this feasibility study was created by Cambridge Health Alliance (CHA) and is based on existing research[6,7,8,9]. To meet the needs of its diverse patient population, the CHA SDoH questionnaire is available in eight languages: Arabic, Chinese, English, Haitian Creole, Hindi, Nepali, Portuguese, and Spanish. Regardless of the language, each questionnaire contains 10 questions and a total of 26 different answer choices each represented by a checkbox.

To avoid exposing any personal health information (PHI), a test dataset was developed. For each language there were three examples and each example (i.e., a questionnaire that was filled out by hand and then scanned as a PDF using an iPhone 12 Mini smartphone camera). The questionnaires were represented using the JavaScript Object Notation (JSON) format, and these representations were considered the gold standard labeled dataset **(see "Structured Representation of Blank Questionnaire" in the Supplementary Materials)**. This dataset was

labeled by hand twice, and a script written in the Python programming language was used to automate ensuring the labels were consistent.

Document Understanding Model

A document understanding model uses AI to comprehend, classify, and extract meaningful information from unstructured or semi-structured documents[5]. For this study, both Microsoft Azure Document Intelligence and Google Document AI were considered as potential document understanding models. Ultimately, Document Intelligence was chosen due to Document AI being less reliable in detecting checkboxes in documents as per the Google Document AI July 18, 2023 release notes.

The Document Intelligence Layout Model application programming interface (API) was leveraged for representing the scanned questionnaires as text with special selection mark tokens denoting whether a checkbox was “selected” or "unselected.” At the time of writing this article, the Layout API was priced at $10 per 1,000 pages (i.e., $0.01 per page) processed. The Document Intelligence Layout API was also used to extract text and checkboxes from blank questionnaires in each of the eight languages **(see “Extracting and Cleaning Text from the Blank Questionnaires” in the Supplementary Materials)**.

LLM

OpenAI’s GPT-4 (“gpt-4-0125-preview”), which has remained amongst the best LLMs available, was one of the LLMs used. Because GPT-4 is still quite expensive, the other LLM that was used was OpenAI’s GPT-3.5 Turbo (“gpt-3.5-turbo-0125”) as it costs substantially less than GPT-4 does. At the time of writing this article, GPT-3.5 Turbo was priced at $0.0005 per 1,000 request tokens used and $0.0015 per 1,000 response tokens generated. GPT-4 was 20 times more expensive at $0.01 per 1,000 request tokens used and $0.03 per 1,000 response tokens generated. In order to make the LLMs’ outputs more deterministic, the temperature was set to zero and a seed was used. JSON mode was also utilized to ensure the LLMs’ responses were formatted in valid JSON.

Experiments

The LLM’s task was to create a structured JSON representation of a scanned, filled out questionnaire. Three different prompt templates were tested: two of the templates (“Example Only” and “Example Only - Multilingual”) used an example-based approach in which the LLM was given an example of the expected input and desired output while the third employed an instruction-based approach that included detailed instructions for the LLM to follow **(see “Prompt Templates” in the Supplementary Materials)**.

In addition, Rivet, an open source visual AI programming environment, was used to construct a computational graph for running these experiments i.e., prompting the LLM with the correct set of parameters for each of the experiments. The results of running these computational graphs were persisted and underwent additional post-processing to extract the relevant information

(e.g., the LLM's response, the number of tokens in the request, the number of tokens in the response, the latency) for assessing the model's performance.

Evaluation

The model's predictions were programmatically compared against the gold standard. If the model correctly predicted the state of a checkbox (i.e., "selected" or "unselected") on the scanned, filled out questionnaire, the prediction was deemed to be "consistent" with the gold standard; otherwise, the prediction was deemed "inconsistent" if the states differed or "missing" if the model failed to make a prediction. The resulting comparisons were then used to generate heatmaps visualizing the areas in which the predictions and the gold standard were in alignment or differed.

These resulting comparisons were also used to assess the model's performance using traditional machine learning metrics, such as accuracy (the percentage of all predictions the model gets right), precision (the percentage of positive identifications that were actually correct), recall (the percentage of actual positives the model correctly identifies), and F1 score (a balance between precision and recall, measuring the test's accuracy).

In addition, Cohen's Kappa, which is a statistical measure used to quantify the level of agreement between two raters on a classification task, was also included in the analysis. In this experiment, the classification task involved correctly labeling whether a "selection mark" (i.e., a checkbox) was "selected" or "unselected. The gold standard classification ratings were done by a human and compared against the model's predictions. In general, a Cohen's Kappa of less than or equal to zero indicates no agreement while a value of 0.01 to 0.20 indicates no to slight agreement. A value between 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80, and 0.81 to 1.00 could be considered fair, moderate, substantial, and perfect agreement, respectively.

In addition, the costs of model inferencing were calculated based on the the number of tokens used in the prompt, or the instructions for the LLM, and on the number of tokens generated in response. The analysis was performed using Python in Jupyter Notebooks and was conducted between February 3 and February 12, 2024.

Results

As expected, GPT-4 generally outperformed GPT-3.5 Turbo across all metrics, including accuracy, precision, recall, F1 score, and Cohen's Kappa as detailed in **(Table 1)**. For instance, when considering the overall performance across all languages, GPT-4 achieved an accuracy of 0.95, precision of 0.94, recall of 0.93, F1 score of 0.94, and Cohen's Kappa of 0.90. In contrast, GPT-3.5 Turbo exhibited lower overall scores, with an accuracy of 0.79, precision of 0.73, recall of 0.71, F1 score of 0.72, and Cohen's Kappa of 0.56.

It is worth noting though that this greater performance is associated with markedly higher costs for GPT-4. The average cost for using GPT-4 Turbo to complete the task was $0.0932 overall,

which is significantly higher than the $0.0044 associated with using GPT-3.5 Turbo. This trend of higher costs for superior performance is consistent across different languages. For instance, for the Arabic language questionnaires, the cost for GPT-4 was $0.0987, compared to $0.0047 for GPT-3.5. Similarly, in the Hindi language category, the cost for GPT-4 was $0.1129, markedly higher than the $0.0055 for GPT-3.5 Turbo. The differences in inferencing costs between the languages is possibly due to differences in how the languages are tokenized resulting in the input of some languages being split up into more tokens[10].

The instruction-based prompt ("Hand Crafted") noticeably outperformed the two example-based prompts in most cases. The improvement in performance was most noticeable in experiments that were run using GPT-3.5 Turbo as the LLM. That said, in a few experiments in which GPT-4 was used, the multilingual example-based prompt slightly outperformed the instruction-based prompt; however, this was by a much smaller margin in comparison to the margin by which GPT-3.5 Turbo's performance improved upon switching prompting approaches. Finally, it is worth recalling that the third prompt included a final instruction to output a specific stop token, which seemed to significantly increase the likelihood of the model completing the task in its entirety **(Figure 1)**.

Discussion

This study suggests AI could be used to transform paper-based questionnaires into structured, machine-readable formats for inclusion in EHRs. Previous research indicates that administering these questionnaires electronically using tablets in primary care clinics can improve screening quality, and that requiring manual entry of paper-based questionnaire responses into EHRs can potentially compromise screening effectiveness[11]. Employing AI to automate data extraction from paper-based questionnaires could enhance the efficiency and quality of SDoH screening in clinics lacking the necessary devices and technical expertise.

The National Bureau of Labor Statistics indicates the average hourly wage for a medical assistant was approximately $20 in 2022[12]. Assuming a minute is required for a medical assistant to manually input a questionnaire response into the EHR, this translates to a cost of roughly $0.67 per entry. Conversely, employing GPT-4 to create structured representations of questionnaire responses incurs an average cost of about $0.10 per questionnaire when including the costs for processing a page using the Document Intelligence Layout API and for LLM inferencing, suggesting that utilizing AI for this task could be more economically efficient for clinics than using manual data entry. There is potential for further savings with GPT-3.5 Turbo, although its accuracy would need to improve, which could be achieved through additional prompt engineering or model fine-tuning[13,14].

Although using AI for generating structured representations of paper-based questionnaires is promising, there could be some concerns with sharing sensitive information like a patient's social needs with the third parties that own these AI models. Additionally, it is crucial to evaluate potential security threats healthcare organizations employing LLMs could face, such as the risk of prompt injection attacks where attackers hijack LLMs with malicious instructions[15].

A limitation of this study is the single execution of each experiment, which, despite attempts to increase determinism by setting the LLM's temperature to zero and using a seed, restricts insights into the consistency of the LLM's outputs. Furthermore, integrating the LLM-generated structured representation into the EHR would necessitate post-processing to convert it into an EHR-compatible format, e.g. the HL7 Fast Healthcare Interoperability Resources (FHIR) standard[16], and additional validation to ensure the output's integrity. It would also be worthwhile to explore the degree to which the cost differential between using manual data entry versus using AI could be impacted by running such an application on HIPAA-compliant infrastructure as would be required if real patient data were used. That said, the results thus far are quite promising and suggest that it would be worthwhile to address these considerations in future research.

Acknowledgements

Lara Jirmanus, MD, MPH and the rest of the team involved in spearheading the efforts to implement screening for social needs in primary care at Cambridge Health Alliance.

Hannah Galvin, MD, FAAP, FAMIA, CHCIO and Hsiang Huang, MD, MPH for providing constructive feedback regarding the final draft of this article.

References

1. Daniel H, Bornstein SS, Kane GC, et al. Addressing Social Determinants to Improve Patient Care and Promote Health Equity: An American College of Physicians Position Paper. Ann Intern Med. 2018;168(8):577-578. doi:10.7326/M17-2441
2. Gruß I, Bunce A, Davis J, Dambrun K, Cottrell E, Gold R. Initiating and Implementing Social Determinants of Health Data Collection in Community Health Centers. Popul Health Manag. 2021;24(1):52-58. doi:10.1089/pop.2019.0205
3. Reza F, Jones C, Reed JH. CIC2022: An Informatics-Driven Strategy for Improving Immunization Healthcare Data Quality using 2D Barcoding and Barcode Scanning Practices. Appl Clin Inform. Published online January 29, 2024. doi:10.1055/a-2255-9749
4. Rogers CK, Parulekar M, Malik F, Torres CA. A Local Perspective into Electronic Health Record Design, Integration, and Implementation of Screening and Referral for Social

Determinants of Health. Perspect Health Inf Manag. 2022;19(Spring):1g. Published 2022 Mar 15.

5. Wang D., Raman N., Sibue M., et al. (2023). DocLLM: A layout-aware generative language model for multimodal document understanding. ArXiv, abs/2401.00908.
6. Hager E. R., Quigg A. M., Black M. M., et al. Development and Validity of a 2-Item Screen to Identify Families at Risk for Food Insecurity. Pediatrics 2010;126(1):26-32. DOI: 10.1542/peds.2009-3146.
7. Cook J. T., Frank D. A., Casey P. H., et al. A Brief Indicator of Household Energy Security: Associations with Food Security, Child Health, and Child Development in US Infants and Toddlers. Pediatrics 2008;122(4):867-875. DOI: 10.1542/peds.2008-0286.
8. National Association of Community Health Centers and Partners, National Association of Community Health Centers, Association of Asian Pacific Community Health Organizations, Association OPC, Institute for Alternative Futures. PRAPARE. 2017. Available at: http://www.nachc.org/research-and-data/prapare/.
9. Sherin K. M., Sinacore J. M., Li X. Q., Zitter R. E., Shakil A. HITS: A Short Domestic Violence Screening Tool for Use in a Family Practice Setting. Family Medicine 1998;30(7):508-512.
10. Maskey U., Bhatta M., Bhatt S., Dhungel S., Bal B.K. (2022). Nepali Encoder Transformers: An Analysis of Auto Encoding Transformer Language Models for Nepali Text Classification. SIGUL.
11. Buitron de la Vega P, Losi S, Sprague Martinez L, et al. Implementing an EHR-based Screening and Referral System to Address Social Determinants of Health in Primary Care. Med Care. 2019;57 Suppl 6 Suppl 2:S133-S139. doi:10.1097/MLR.0000000000001029
12. Government Report: Occupational Employment and Wages, May 2022: 31-9092 Medical Assistants. Bureau of Labor Statistics. (https://www.bls.gov/oes/current/oes319092.htm)

13. Zhang X., Talukdar N., Vemulapalli S., et al. (2024). Comparison of Prompt Engineering and Fine-Tuning Strategies in Large Language Models in the Classification of Clinical Notes. medRxiv.
14. Trad F., Chehab A. (2024). Prompt Engineering or Fine-Tuning? A Case Study on Phishing Detection with Large Language Models. Machine Learning and Knowledge Extraction.
15. Liu Y., Jia Y., Geng R., Jia J., Gong N.Z. (2023). Prompt Injection Attacks and Defenses in LLM-Integrated Applications. ArXiv, abs/2310.12815.
16. Watkins M, Viernes B, Nguyen V, Rojas Mezarina L, Silva Valencia J, Borbolla D. Translating Social Determinants of Health into Standardized Clinical Entities. Stud Health Technol Inform. 2020;270:474-478. doi:10.3233/SHTI200205

Table 1: Performance and Cost Overall by Language and Prompt

Language	Model	Prompt	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa	Average Cost ($)
Overall	*GPT-3.5 Turbo 0125*	*Example Only*	0.72	0.64	0.63	0.64	0.42	$0.0051
Overall	*GPT-3.5 Turbo 0125*	*Example Only - Multilingual*	0.78	0.71	0.71	0.71	0.54	$0.0044
Overall	*GPT-3.5 Turbo 0125*	*Hand Crafted*	0.87	0.83	0.80	0.82	0.71	$0.0037
Overall	*GPT-4 Turbo 0125*	*Example Only*	0.95	0.94	0.93	0.93	0.89	$0.1043
Overall	*GPT-4 Turbo 0125*	*Example Only - Multilingual*	0.96	0.95	0.93	0.94	0.91	$0.0889
Overall	*GPT-4 Turbo 0125*	*Hand Crafted*	0.95	0.94	0.93	0.94	0.90	$0.0865
Overall	***GPT-3.5 Turbo 0125***	**-**	**0.79**	**0.73**	**0.71**	**0.72**	**0.56**	**$0.0044**
Overall	***GPT-4 Turbo 0125***	**-**	**0.95**	**0.94**	**0.93**	**0.94**	**0.90**	**$0.0932**
Arabic	*GPT-3.5 Turbo 0125*	*Example Only*	0.68	0.58	0.54	0.56	0.31	$0.0058
Arabic	*GPT-3.5 Turbo 0125*	*Example Only - Multilingual*	0.63	0.50	0.45	0.47	0.19	$0.0046
Arabic	*GPT-3.5 Turbo 0125*	*Hand Crafted*	0.71	0.62	0.55	0.58	0.36	$0.0037
Arabic	*GPT-4 Turbo 0125*	*Example Only*	0.86	0.82	0.79	0.81	0.70	$0.1165
Arabic	*GPT-4 Turbo 0125*	*Example Only - Multilingual*	0.86	0.85	0.76	0.80	0.69	$0.0920
Arabic	*GPT-4 Turbo 0125*	*Hand Crafted*	0.86	0.82	0.79	0.81	0.70	$0.0876
Arabic	***GPT-3.5 Turbo 0125***	**-**	**0.67**	**0.56**	**0.51**	**0.54**	**0.28**	**$0.0047**
Arabic	***GPT-4 Turbo 0125***	**-**	**0.86**	**0.83**	**0.78**	**0.80**	**0.69**	**$0.0987**
Chinese	*GPT-3.5 Turbo 0125*	*Example Only*	0.87	0.88	0.75	0.81	0.71	$0.0044
Chinese	*GPT-3.5 Turbo 0125*	*Example Only - Multilingual*	0.88	0.83	0.86	0.84	0.75	$0.0038
Chinese	*GPT-3.5 Turbo 0125*	*Hand Crafted*	0.95	0.93	0.93	0.93	0.89	$0.0035
Chinese	*GPT-4 Turbo 0125*	*Example Only*	0.99	1.00	0.96	0.98	0.97	$0.0871
Chinese	*GPT-4 Turbo 0125*	*Example Only - Multilingual*	1.00	1.00	1.00	1.00	1.00	$0.0835
Chinese	*GPT-4 Turbo 0125*	*Hand Crafted*	1.00	1.00	1.00	1.00	1.00	$0.0833
Chinese	***GPT-3.5 Turbo 0125***	**-**	**0.90**	**0.88**	**0.85**	**0.86**	**0.78**	**$0.0039**
Chinese	***GPT-4 Turbo 0125***	**-**	**1.00**	**1.00**	**0.99**	**0.99**	**0.99**	**$0.0847**
English	*GPT-3.5 Turbo 0125*	*Example Only*	0.47	0.37	0.40	0.38	0.01	$0.0036
English	*GPT-3.5 Turbo 0125*	*Example Only - Multilingual*	0.49	0.38	0.40	0.39	0.03	$0.0036
English	*GPT-3.5 Turbo 0125*	*Hand Crafted*	0.88	0.88	0.79	0.84	0.75	$0.0034
English	*GPT-4 Turbo 0125*	*Example Only*	0.96	0.96	0.93	0.95	0.92	$0.0797
English	*GPT-4 Turbo 0125*	*Example Only - Multilingual*	0.96	0.96	0.93	0.95	0.92	$0.0797
English	*GPT-4 Turbo 0125*	*Hand Crafted*	0.95	0.96	0.90	0.93	0.89	$0.0821

English	***GPT-3.5 Turbo 0125***	**-**	**0.62**	**0.55**	**0.53**	**0.54**	**0.26**	**$0.0035**
English	***GPT-4 Turbo 0125***	**-**	**0.96**	**0.96**	**0.92**	**0.94**	**0.91**	**$0.0805**
Haitian Creole	*GPT-3.5 Turbo 0125*	*Example Only*	0.68	0.55	0.62	0.58	0.33	$0.0046
Haitian Creole	*GPT-3.5 Turbo 0125*	*Example Only - Multilingual*	0.79	0.69	0.74	0.71	0.55	$0.0045
Haitian Creole	*GPT-3.5 Turbo 0125*	*Hand Crafted*	0.81	0.71	0.74	0.73	0.58	$0.0036
Haitian Creole	*GPT-4 Turbo 0125*	*Example Only*	0.88	0.88	0.88	0.88	0.76	$0.0916
Haitian Creole	*GPT-4 Turbo 0125*	*Example Only - Multilingual*	0.96	0.93	0.96	0.95	0.92	$0.0846
Haitian Creole	*GPT-4 Turbo 0125*	*Hand Crafted*	0.96	0.93	0.96	0.95	0.92	$0.0842
Haitian Creole	***GPT-3.5 Turbo 0125***	**-**	**0.76**	**0.65**	**0.70**	**0.67**	**0.49**	**$0.0042**
Haitian Creole	***GPT-4 Turbo 0125***	**-**	**0.94**	**0.91**	**0.94**	**0.93**	**0.86**	**$0.0868**
Hindi	*GPT-3.5 Turbo 0125*	*Example Only*	0.72	0.64	0.60	0.62	0.40	$0.0071
Hindi	*GPT-3.5 Turbo 0125*	*Example Only - Multilingual*	0.85	0.80	0.80	0.80	0.68	$0.0054
Hindi	*GPT-3.5 Turbo 0125*	*Hand Crafted*	0.86	0.83	0.80	0.81	0.70	$0.0040
Hindi	*GPT-4 Turbo 0125*	*Example Only*	0.99	1.00	0.97	0.98	0.97	$0.1418
Hindi	*GPT-4 Turbo 0125*	*Example Only - Multilingual*	0.99	1.00	0.97	0.98	0.97	$0.1035
Hindi	*GPT-4 Turbo 0125*	*Hand Crafted*	1.00	1.00	1.00	1.00	1.00	$0.0935
Hindi	***GPT-3.5 Turbo 0125***	**-**	**0.81**	**0.76**	**0.73**	**0.74**	**0.59**	**$0.0055**
Hindi	***GPT-4 Turbo 0125***	**-**	**0.99**	**1.00**	**0.98**	**0.99**	**0.98**	**$0.1129**
Nepali	*GPT-3.5 Turbo 0125*	*Example Only*	0.71	0.60	0.54	0.57	0.34	$0.0075
Nepali	*GPT-3.5 Turbo 0125*	*Example Only - Multilingual*	0.87	0.85	0.79	0.81	0.72	$0.0049
Nepali	*GPT-3.5 Turbo 0125*	*Hand Crafted*	0.86	0.84	0.75	0.79	0.69	$0.0041
Nepali	*GPT-4 Turbo 0125*	*Example Only*	0.99	1.00	0.96	0.98	0.97	$0.1490
Nepali	*GPT-4 Turbo 0125*	*Example Only - Multilingual*	0.91	0.89	0.86	0.87	0.80	$0.1059
Nepali	*GPT-4 Turbo 0125*	*Hand Crafted*	0.91	0.89	0.86	0.87	0.80	$0.0950
Nepali	***GPT-3.5 Turbo 0125***	**-**	**0.81**	**0.76**	**0.69**	**0.72**	**0.58**	**$0.0055**
Nepali	***GPT-4 Turbo 0125***	**-**	**0.94**	**0.93**	**0.89**	**0.91**	**0.86**	**$0.1166**
Portuguese	*GPT-3.5 Turbo 0125*	*Example Only*	0.73	0.65	0.74	0.69	0.46	$0.0041
Portuguese	*GPT-3.5 Turbo 0125*	*Example Only - Multilingual*	0.85	0.77	0.83	0.80	0.68	$0.0040
Portuguese	*GPT-3.5 Turbo 0125*	*Hand Crafted*	0.90	0.84	0.90	0.87	0.78	$0.0035
Portuguese	*GPT-4 Turbo 0125*	*Example Only*	0.91	0.87	0.90	0.88	0.81	$0.0838
Portuguese	*GPT-4 Turbo 0125*	*Example Only - Multilingual*	0.99	0.97	1.00	0.98	0.97	$0.0816
Portuguese	*GPT-4 Turbo 0125*	*Hand Crafted*	0.96	0.93	0.97	0.95	0.92	$0.0832
Portuguese	***GPT-3.5 Turbo 0125***	**-**	**0.82**	**0.75**	**0.82**	**0.79**	**0.64**	**$0.0039**
Portuguese	***GPT-4 Turbo 0125***	**-**	**0.95**	**0.92**	**0.95**	**0.94**	**0.90**	**$0.0828**

Spanish	*GPT-3.5 Turbo 0125*	*Example Only*	0.91	0.89	0.86	0.88	0.81	$0.0039
Spanish	*GPT-3.5 Turbo 0125*	*Example Only - Multilingual*	0.87	0.83	0.83	0.83	0.73	$0.0040
Spanish	*GPT-3.5 Turbo 0125*	*Hand Crafted*	0.99	1.00	0.97	0.98	0.97	$0.0035
Spanish	*GPT-4 Turbo 0125*	*Example Only*	1.00	1.00	1.00	1.00	1.00	$0.0850
Spanish	*GPT-4 Turbo 0125*	*Example Only - Multilingual*	1.00	1.00	1.00	1.00	1.00	$0.0807
Spanish	*GPT-4 Turbo 0125*	*Hand Crafted*	0.99	1.00	0.97	0.98	0.97	$0.0830
Spanish	***GPT-3.5 Turbo 0125***	**-**	**0.92**	**0.91**	**0.89**	**0.90**	**0.83**	**$0.0038**
Spanish	***GPT-4 Turbo 0125***	**-**	**1.00**	**1.00**	**0.99**	**0.99**	**0.99**	**$0.0829**