

Just Don't Make It Worse: A custom GPT to develop Generative AI literacy in specialized translation

Jean-Ezra Yeung¹, Michelle Rufrano², Katrin Herget³

¹Columbia Startup Lab, Columbia University, U.S.A., ²Department of Sociology & Anthropology, Fordham University, U.S.A. ³Department of Languages and Cultures / CLLC, University of Aveiro, Portugal

How to cite: Yeung, J-E.; Rufrano, M.; Herget, K. (2026). Just Don't Make It Worse: A custom GPT to develop Generative AI literacy in specialized translation. In: 12th International Conference on Higher Education Advances (HEAd'26). Valencia, 15-18 June 2026. <https://doi.org/10.4995/HEAd26.2026.21626>

Abstract

The landscape of higher education requires both students and educators to develop Generative AI (GenAI) literacy. With an increasing number of citizens accessing health information on digital platforms, specialized translation of health information using GenAI can improve care outcomes if translators can account for bias embedded in data. Customizing a GPT is an accessible and quick way to adapt LLMs, especially in specialized translation where there is scarce high-quality data for pre-training or fine-tuning. We customize ChatGPT with role and task instructions, supply it with knowledge graph analysis data files for retrieval-augmented generation (RAG), and develop special commands to address domain knowledge gaps of LLMs and technical accessibility. Preliminary testing shows that the custom GPT supports translators' diagnosis of machine translation when comparing to a human translation, and enables them to see how the machine may alter the biomedical relationship structure.

Keywords: Custom GPT; Specialized Translation; Higher Education; AI Literacy; Generative AI; Large Language Models (LLMs).

1. Introduction

Large Language Models (LLMs) present a new opportunity for Generative AI (GenAI) to automate complex knowledge-intensive tasks. As researchers from different disciplines interested in the health sciences and climate policies, we sought to apply LLMs in higher educational settings to prepare students to use, understand, and assess them in school and the workforce. The academic landscape requires educators and students to develop skills for using GenAI with integrity (Sejdiu & Sejdiu, 2025). Stereotyped biases are embedded in textual data (Caliskan et al., 2017). If educators and students translate these biases without scrutiny,

misinformation seeded by algorithmic bias is magnified, which can exacerbate health inequity by widening gaps in knowledge between people with high and low health literacy (Franklin et al., 2024; Joseph, 2025). This project provides a unique example of the challenges involved in fitting technological solutions into the pedagogy of higher education (Zawacki-Richter et al., 2019). In the case of specialized translation, we wish to teach students how to diagnose and improve machine translations for large scale corpora in the age of misinformation, through understanding and assessing how human translations may differ due to the domain and audience.

1.1. Motivation for customizing a GPT

General purpose GenAI, such as ChatGPT, do not serve our translation and pedagogical needs because they do not have the required domain training. Due to the novelty of the domain data, there is also insufficient high-quality data for pre-training or fine-tuning a model considering that the data is limited to what is readily available, and these data can be missing nonrandomly (Cross et al., 2024). Therefore, we focus on adapting an existing LLM. Previously, we found that GenAI could not identify misinformation as a topic until we prompted it with key words using our existing expertise on misinformation (Rufrano et al., [Forthcoming]; Herget et al., 2024). Consequently, we thought we could configure GenAI with effective prompt engineering.

We propose customizing ChatGPT with role and task instructions, retrieval-augmented generation (RAG), and special commands to address domain knowledge gaps and technical accessibility issues of general purpose LLMs. Assigning a role and providing a description can enhance translation quality (Yamada, 2024; He, 2023). RAG can incorporate in interactions retrieved information that are out-of-domain (Lewis et al., 2020). Special commands can develop input semantics and prompt patterns to increase the accessibility of a successful interaction and control the output template, respectively (White et al., 2023). For RAG, we create linguistic data files that segment and summarize the knowledge graph of the source text and machine translation. We also provide users human translations to submit in their interaction.

1.2. A custom GPT to develop Generative AI literacy

The project goal is to show students how to interact with LLMs, or learn prompt engineering. Since the release of LLMs, anyone could access them, like ChatGPT, for free without undergoing training, and may encounter risks, including spreading misinformation (Stevic et al., 2025). We design a custom GPT to develop GenAI literacy, and organize learning objectives into navigation, appraisal and understanding (Stevic et al., 2025). In this paper, we detail the learning objectives; the design and implementation of the custom GPT; and the user workflow and reflections. We end with a discussion on how our design could apply to other domains and improved in higher education. The main contribution of this paper is to showcase how to

develop and improve GenAI literacy learning objectives for specialized translation through co-designing with instructors and testing with specialized users.

2. Generative AI Literacy and Technical Requirements

We came up with six learning objectives, organizing them by GenAI literacy category, and mapping them to custom GPT technical specifications (Table 1).

Table 1. Learning objectives for the usage of the custom GPT

GenAI literacy	Learning Objectives	Technical Specification
Navigation	Develop the ability to interact critically with an LLM by analyzing both its translation output and its handling of source text entities.	Instructions for LLM and RAG
Comprehension	- Use RAG to enrich the LLM with additional domain-specific information, improving translation accuracy and contextualization. - Experience how LLMs can be customized without fine-tuning to support specialized tasks such as medical or patient-oriented translation. - Understand the role of AI as an augmented intelligence tool, where the human translator actively evaluates and improves machine output rather than passively accepting it.	Data, natural language processing, and knowledge graph
Appraisal	- Practice identifying and comparing entities across source text, AI draft, and human reference translation. - Reflect on translation decision-making by justifying improvements and evaluating which version best serves the communicative purpose and audience.	Special commands

3. Custom GPT Design

The custom GPT design guides users through a specific workflow that eliminates inconsistent results and serves as guardrails against hallucinations (Jaźwińska & Chandrasekar, 2025; Burford et al., 2024). We preprocess the translation and data analysis, and customize the output templates into special commands.

3.1. Data

Our source text corpus is compiled from a Portuguese medical information website (<https://www.cuf.pt/saude-a-z/>), where we select ten conditions. We scrape the data from the urls and format them into a table with the title, header, content, url, and a paragraph-sentence index. We use the HuggingFace Facebook NLLB 600M model to translate the Portuguese source text into German and English; and the NLTK library for parts-of-speech tagging to extract triplets of subject-verb-object. These triplets construct the knowledge graph, which enables the LLM to reason about the linguistic analysis via RAG. For each language, we generate two files, one to view content in chunked segments with extracted relational structures; and the other to summarize and compare these structures across texts. We also generate a master file combining all three languages side-by-side.

3.2. LLM instructions

We provide the LLM with general instructions that define its role (Table 2); data instructions to enable the special commands to use the GPT's "Code Interpreter & Data Analysis" capability (Table 3); and define special commands (Table 4).

Table 2. General instructions: role

Your task is to help the user produce or examine a translation draft into German intended for patients. For shorter texts, you may be asked to create a human reference translation yourself; for longer texts, a reference translation will be provided for benchmarking.

Help the user focus on the meaning of the text by identifying entities that are altered, omitted, or inaccurately translated, and suggest revisions to ensure the source text's intended meaning is preserved. Do this by executing the special commands to help the user see how entities are related in the source text and the translation, independently and comparatively. Note the differences to the user.

Table 3. First paragraph of data instructions: data files, primary keys, and column headers

You are provided with a source text in Portuguese containing patient information or general public guidance about illnesses in the file "translation_medical_090825.csv", where the "link" column provides the url; the "content_source_portuguese" column provides relevant chunked source text in Portuguese by sentence from the url; and the "id" column is the unique key for the dataset that is a period-separated concatenation of the columns "title", "header_source_portuguese" and "content_source_portuguese". This dataset is ordered by "id", and in addition, the "paragraph_sentence_index" provides the paragraph number and sentence number separated by a period. The "header_source_portuguese" column provides the header name.

Table 4. Special command to summarize one language using the knowledge graph

/summarize: Ask the user whether they would like to summarize the source text in Portuguese, or the translation text in German or English. Based on the user's answer, use the file "triplets_medical<user's answer on what they would like to summarize>_090825.csv", where <user's answer on what they would like to summarize> is "por" for Portuguese, "ger" for German, and "eng" for English. Visualize a histogram of the "value" column and then list the first 5 values from the file with the highest value in the "degrees" column, showing columns "nodes", "status", "in_degree", "out_degree", "degrees", and "triplet_connected_concept".

3.3. Workflow

Users were asked to follow a four-step workflow, using the special commands /view, /summarize, /compare, and then to sketch simplified diagrams of a human translation and document their reflections in comparing them to the machine translations. Users were supplied the human translation, which they submitted to the custom GPT in the beginning of their interaction. The reflection exercise was about translation quality with a target audience being a patient, and this provided qualitative data on how to improve the custom GPT and GenAI learning objectives for the future.

4. Preliminary Findings

In the preliminary test with five experienced translators, users found that they could see the knowledge graph structure of the translations and how the source text, machine and human translations differ in accuracy and depth: the machine tended to flatten the relationship structure, whereas the human retained the relationships and sometimes elaborated upon them. The users assessed the translation in a non-linear way because they were reflecting upon the topology of the key relationships.

5. Discussion

Based on the documented workflow and reflections, the custom GPT was missing functionality to draw a truncated knowledge graph with only biomedical concepts. The instructor and users had to develop their own prompts to draw the graph because the preprocessed linguistic files only aggregated network analysis statistics. In addition, the custom GPT did not have sufficient context surrounding biomedical discourse conventions and patient information quality standards to improve machine translations. As such, some users noted this in their assessment of the machine translation; or they focused on understanding and assessing the reconstruction of the source text and translations' structure. These two gaps could be future GenAI literacy learning objectives on how to manage missing functionality.

5.1. Responsible AI

Translation of medical information without consideration of inherent biases can undermine inclusive practices and magnify inequity (Francis et al., 2025; Franklin et al., 2024). Therefore, empowering students to understand the strengths and weaknesses of translation using GenAI is tantamount to generating the best care options for people who directly access health information websites—about half of citizens in the European Union (Eurostat, 2022; Francis et al., 2025). Overall, this project highlights the importance of demonstrating the limitations of GenAI to students in higher education. They must learn that GenAI is a tool humans can use to help organize large amounts of text data but cannot be used alone as an end-to-end tool without the threat of perpetuating misinformation.

Especially in multilingual settings, LLMs are known to perpetuate social bias from the training data used (Talat et al. 2022; Francis et al., 2025). While we chose Facebook's NLLB multilingual model that represents many languages for machine translation, we did not address the scientific biomedical model bias present in the data source (Francis et al., 2025). Incorporating the Social Determinants of Health (SDoHs) is often missing from health information resources and is crucial to producing better medical interpretations (Cross et al., 2024).

The resource consumption of training LLMs and using GenAI has significant environmental and ethical concerns, as impacted marginalized communities are not always the primary beneficiary (Bender et al., 2021). Even if GenAI is used to mitigate these concerns, a future GenAI literacy learning objective would be to consider costs and emissions in comparison to the benefit of machine(-assisted) translation (Gailhofer et al., 2021; Nordgren, 2023).

5.2. Future Research

For low-resource languages, it has been shown that a combination of parameter efficient fine-tuning, and custom prompt engineering in the form of a dictionary has improved benchmark scores for multilingual translation (Khoboko et al., 2025). Future applications must account for the resource load of the language and domain (e.g. health sciences, climate policies). This distinction will allow researchers to focus more on fine-tuning versus prompt engineering. In the case of both language and domain, a custom GPT can help identify challenges in multilingual out-of-domain machine translation, such as domain mismatch and rare words (Pang et al., 2025). This can help users understand the need to build dictionaries to address these challenges for custom prompt engineering, including correcting knowledge graphs. Customizing prompt engineering is also a natural pathway in higher education as many of us have compiled dictionaries and indices in relation to our expertise as disciplines evolve.

References

- Althubaiti, A. (2016). Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare*, 211-217. <https://doi.org/10.2147/JMDH.S104807>.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, (pp. 610–623).. <https://doi.org/10.1145/3442188.3445922>.
- Burford, K. G., Itzkowitz, N. G., Ortega, A. G., Teitler, J. O., & Rundle, A. G. (2024). Use of generative AI to identify helmet status among patients with micromobility-related injuries from unstructured clinical notes. *JAMA Network Open*, 7(8), e2425981-e2425981.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186. <https://doi.org/10.1126/science.aal4230>. PMID: 28408601.
- Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, 3(11), 1-19. <https://doi.org/10.1371/journal.pdig.0000651>.
- Eurostat (6 April 2022). EU citizens: over half seek health information online. <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220406-1>
- Francis, N. J., Jones, S., & Smith, D. P. (2025). Generative AI in higher education: Balancing innovation and integrity. *British Journal of Biomedical Science*, 81, 1-9. <https://doi.org/10.3389/bjbs.2024.14048>.
- Franklin, G., Stephens, R., Piracha, M., Tiosano, S., Lehouillier, F., Koppel, R., & Elkin, P. L. (2024). The sociodemographic biases in machine learning algorithms: a biomedical informatics perspective. *Life*, 14(6), 652. <https://doi.org/10.3390/life14060652>.
- Gailhofer, P., Herold, A., Schemmel, J.P., Scherf, C.-S., Urrutia, C., Köhler, A.R., Braungardt, S. (2021). The role of artificial intelligence in the european green deal., *Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament*, EU, available at [www.europarl.europa.eu/RegData/etudes/STUD/2021/662906/IPOL_STU\(2021\)662906_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/STUD/2021/662906/IPOL_STU(2021)662906_EN.pdf) (accessed 25 January 2026).
- He, S. (2024). Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. *arXiv preprint arXiv:2403.00127*.
- Herget, K., Yeung, J. E., Rufrano, M., & Alegre, T. (2025). Exploring Semantic-Thematic Fields and Lexical Patterns in Parliamentary Debates—Topic Modeling Across Comparable Corpora. In S. García Fernández, F. Gómez-Cáñeba, B. Guerrero García, F. Placidi, M. Savchenkova, & S,choer-Granado, S. (Eds). *Traducción y sostenibilidad cultural II: retos y nuevos escenarios* (pp. 559-569). Ediciones Universidad de Salamanca. <https://doi.org/10.14201/0AQ0373559569>.
- Jaźwińska, K., & Chandrasekar, A. (2025). AI search has a citation problem. *Columbia Journalism Review*, www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyreall-bad-at-citing-news.php.

- Joseph, J. (2025). Algorithmic bias in public health AI: a silent threat to equity in low-resource settings. *Frontiers in Public Health*, 13, 1-5. <https://doi.org/10.3389/fpubh.2025>.
- Khoboko, P. W., Marivate, V., & Sefara, J. (2025). Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, 20, 100649.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.
- Nordgren, A. (2023). Artificial intelligence and climate change: ethical issues. *Journal of Information, Communication and Ethics in Society*, 21(1), 1-15. <https://doi.org/10.1108/jices-11-2021-0106> 19.
- Pang, J., Ye, F., Wong, D. F., Yu, D., Shi, S., Tu, Z., & Wang, L. (2025). Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13, 73-95.
- Rufrano, M., Yeung, J. E., Herget, K., & Alegre, T. (24024) [Forthcoming]. "Misinformation" Everywhere but Nowhere: the Use of Topic Modeling and Generative AI in Linguistics. DDHUM24, the 2nd International Conference on Digital Humanities and Data.
- Sejdiu, N. P., & Sejdiu, S. (2025). The quiet transformation of higher education in the AI era. *Open Research Europe*, 5,1-9. <https://doi.org/10.12688/openreseurope.20715.1>.
- Stevic, A., Hodzic, S., Matthes, J., Nanz, A., Binder, A., & Chan, M. (2025). Generative Ai Literacy: Key Antecedents and Outcomes Among Youth in Four Countries. *Available at SSRN 5277061*.
- Talat, Z., Névéal, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., ... & Van Der Wal, O. (2022, May). You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5--Workshop on Challenges & Perspectives in Creating Large Language Models* (pp. 26-41).
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- Yamada, M. (2023). Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. *arXiv preprint arXiv:2308.01391*.
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators?. *International journal of educational technology in higher education*, 16(1), 1-27. <https://doi.org/10.1186/s41239-019-0171-0> 73.