Evaluating the Feasibility of Using ChatGPT for Cross-cultural Survey Translation

Danielly Sorato
Universitat Pompeu Fabra
Barcelona, Spain
danielly.sorato@upf.edu

Diana Zavala-Rojas
European Social Survey ERIC
Universitat Pompeu Fabra
Barcelona, Spain
diana.zavala@upf.edu

Abstract

Cross-national survey projects such as the European Social Survey measure attitudes in European countries by applying standardized questionnaires to population samples. In this context, low-quality translations can affect data collected through questionnaires, hampering data comparability and increasing measurement errors across countries. For these reasons, cross-national survey projects employ rigorous methods in their translation process and refrain from using Machine Translation (MT). However, new advances in language models and their wide-ranging successful application across several Natural Language Processing tasks, including MT, have shown promising results. In the present study, we explore using `GPT-4o mini` for the questionnaire translation domain. We analyze `GPT-4o mini` translations for the English-German language pair by comparing them with reference translations produced by humans in an experimental setting. To assess the quality of automatic translations, we apply quantitative MT evaluation metrics and complement our results with qualitative analysis. Our findings show that the `GPT` model achieved overall good quality, although it failed to produce adequate translations for survey-specific terminology.

1 Introduction

The European Social Survey (ESS) and the European Values Study (EVS) are large-scale international survey projects that produce cross-national and cross-cultural social science data. These projects aim to measure respondents' attitudes, beliefs, and behavior regarding socially relevant topics (e.g., immigration, climate change, social trust) by administering standardized and structured questionnaires to representative population samples across European countries. The data collected through the questionnaires is publicly available and can be utilized for social and political analysis, serving as a tool for policy-making, for instance.

The survey questionnaires are written in a source language, which in Europe is usually English from Great Britain, and then translated into other target languages. The translated survey questions and their response options must capture the same opinions and attitudes across linguistic contexts to allow for cross-national statistical comparisons (Zavala-Rojas et al., 2022; Harkness et al., 2010; Mohler and Johnson, 2010; Zavala-Rojas et al., 2018a). Within this framework, the formulation of questions and the lack of proper cultural adaptation in the translation process can affect the quality of the data collected through questionnaires, as low-quality translations can hinder data comparability and increase measurement errors between countries (Davidov and De Beuckelaer, 2010; Oberski et al., 2007; Zavala-Rojas et al., 2022; Bound et al., 2001; Zavala-Rojas et al., 2018b).

To avoid translation issues, the Translation, Review, Adjudication, Pretesting, and Documentation (TRAPD) method (Harkness et al., 2003) has been increasingly adopted by large-scale survey projects, becoming a gold standard in the field of questionnaire translation. In this method, given a questionnaire written in the source language, two translators either produce independent translations into the respective target language or split the questionnaire so that each translator works separately on one of the parts. Then, in a review meeting, the translation drafts are compared to foster discussion, and the reviewer and the translators reconcile the different translation options. Finally, an adjudicator makes the final decisions. During the meeting, the team can opt for (i) one of the translation options; (ii) create a new translation by combining the strengths of each translation; or (iii) create a new translation from scratch. Subsequently, the translated questionnaire is pretested before being administered in fieldwork, and the process is documented.

TRAPD is a scientific methodology tailored to survey texts that circumvents problematic transla-

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 177–191, 2025

tion practices. Furthermore, leveraging the contributions of several translators facilitates dealing with regional variances and individual preferences that can influence a translator's decision (Harkness, 2011; Walde and Völlm, 2023). However, TRAPD is time- and work-intensive, as it requires many interactions with a multidisciplinary team.

To date, Machine Translation (MT) has not been incorporated into the TRAPD, as the quality of automatic translations was considered below acceptable in most cases until recently. However, new advances in Large Language Models (LLMs) and their wide-ranging successful application across many Natural Language Processing (NLP) tasks, including MT, show promising results (He et al., 2024; Agrawal et al., 2023; Vilar et al., 2023; Moslem et al., 2023; Pilault et al., 2023; Hendy et al., 2023; Jiao et al., 2023; Wang et al., 2023; Peng et al., 2023; Bawden and Yvon, 2023). Furthermore, the ease of integrating relevant information in the prompts sent to LLMs represents a good opportunity to leverage expert knowledge and translations produced in previous questionnaires.

In this paper, we investigate the quality of the translations produced by `GPT-4o mini` for the English-German language pair in the survey translation domain, exploring different prompts and model temperature settings. Namely, we are interested in analyzing (i) the overall quality of the translated questionnaire; (ii) if the `GPT` translations of the response scales accurately keep the intensity attached to verbal labels (e.g., qualifiers, intensifiers); (iii) if survey-specific terminology and elements (e.g., instructions to the respondent or the interviewer) are correctly translated, and; (iv) to what extent the usage of examples and linguistic resources extracted from past questionnaires can improve the quality of the translations.

We analyze the quality of `GPT` translations by comparing them with human reference translations produced through the TRAPD methodology. To achieve this, we employ quantitative metrics for MT evaluation, namely `ChrF++` (Popović, 2017) and `COMET-22` (Rei et al., 2022). To deepen our study, we manually analyze certain aspects of the `GPT` translations that are key to answering our research questions. Finally, we create a translation assessment questionnaire containing the source text, reference, and `GPT` translations and ask five domain specialists to select the translation options that are most adequate for the survey domain, considering accuracy, fluency, and cultural adaptation criteria. The source code and data used to produce the results and analyses presented in this study are openly available on a `Github` repository[1].

Our findings show that `GPT-4o mini` produced translations largely similar to human references, but struggled with gender adaptation, country-localized terms (e.g., job occupations), and survey-specific terminology, particularly instructions for interviewers. Our analysis also indicates that some translations are influenced by the English language, thus sounding unnatural and/or non-idiomatic.

This paper is organized as follows. In Section 2 we discuss related work. Then, we present our data and explain the process through which the reference translations used in this work were created in Section 3. Subsequently, in Section 4, we introduce our prompting strategy and MT evaluation methods. In 5, we discuss the findings derived from this study. Finally, we present our conclusions and limitations in Section 6.

2 Related Work

Translation quality is a complex and context-dependent topic (Nord, 2006; Koby et al., 2014). Using different translation theories, the academic literature provides various definitions of translation quality and how to evaluate it (González-Jover, 2002; Han et al., 2021).

An extensive body of research employs methods to evaluate and estimate the quality of MT (Papineni et al., 2002; Doddington, 2002; Specia et al., 2010, 2018; Bawden et al., 2018; Popović, 2018; Kepler et al., 2019; Mathur et al., 2020; Fomicheva et al., 2020; Thompson and Post, 2020; nll, 2024; Leiter et al., 2024). Given the varied motivations behind such analyses, we focus on those most aligned with ours, namely, studies that assess MT quality to evaluate the viability of integrating automatic translation in the process of producing cross-linguistic and cultural materials.

Turner et al., 2015 studied the feasibility of using MT and post-editing (PE) to produce Chinese translations of public health materials. The authors collected 60 health promotion documents from public websites for which validated human reference translations from English to Chinese were available. The documents were translated using `Google Translate` and then manually analyzed for translation errors. Additionally, a subset of the auto-

[1] https://github.com/dsorato/KONVENS_2025_MT_paper

matic translations was post-edited by native Chinese speakers with health sciences backgrounds. Their results showed that the evaluators consistently selected the human translation over MT in combination with human PE. Their evaluation indicated poor MT quality, where word meaning, word order, and missing words were the most common annotated translation errors.

Ugas et al., 2025 investigated the viability of using MT to translate and simplify patient education materials to make them more accessible for populations with limited English proficiency. The authors selected 5 patient education pamphlets, which were translated by domain specialists and `Google Translate` into Spanish, Portuguese, Punjabi, simplified Chinese, and Vietnamese. The translations were manually evaluated by domain specialists considering fluency, adequacy, meaning, and risk severity criteria. Their results show significant quality differences between human and automatic translations, although the authors report that the translated engine performed relatively well.

Kunst and Bierwiaczonek, 2023 analyzed automatic translations of the `HEXACO` personality inventory (Lee and Ashton, 2004), which is one of the most used cross-cultural psychometric instruments with validated translations in several languages. The authors assessed the quality of the translations of the English inventory produced by `Google Translate` and `GPT-3.5` in 33 languages for which validated human reference translations exist. To measure the similarity between machine and human translations, they applied metrics such as Levenshtein and Jaccard similarity, also providing human judgments for Norwegian translations.

Their findings show that the similarity between human and machine translations varied across target languages, being highest for languages from the same language family as the source language. They also found that `GPT`'s temperature had little influence on similarity estimates, although high temperatures tended to result in lower similarity. Manual evaluation showed that, although human translations achieved higher quality, the evaluators did not rate them significantly better than `GPT 3.5` translations with a low temperature setting.

Brewster et al., 2024 assessed the performance of `Google Translate` and `ChatGPT` for the translation of English pediatric discharge instructions into Spanish, Brazilian Portuguese, and Haitian Creole. The authors analyzed both automatic and human translations of 20 standardized discharge instructions for pediatric conditions in the specified languages, evaluating adequacy, fluency, meaning, and severity, along with an assessment of overall preference. According to the authors, `Google Translate` and `ChatGPT` achieved performances comparable to professional translations, at least for Spanish and Portuguese languages.

3 Data

A social survey, or questionnaire, comprises different items composed of questions and their respective response options, and may also include instructions for the interviewer or the respondent. We will refer to these elements as "survey items" throughout this paper. Figure 1 illustrates a survey item that contains a request for an answer, i.e., a question, two instructions, and response options.

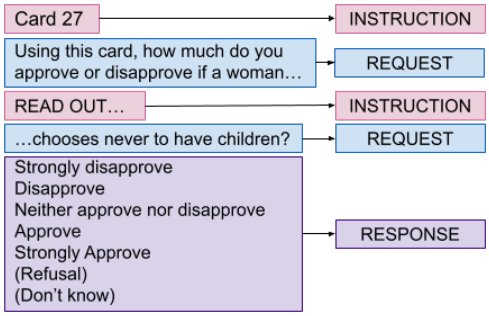

Figure 1: Source survey item from ESS round 3 (2006).

For our controlled translation experiment, we created a laboratory questionnaire using a sample of ESS and EVS questions, that is, survey items that were used in past officially published ESS and EVS questionnaires. First, we filtered a set of approximately 3,500 survey items written in English from the EVS and ESS questionnaires, favoring the variability of topics and response scales, as well as ensuring that the selected items had at least two sentences. From this process, we sampled 300 survey items. Since the translation of the laboratory questionnaire should simulate the translation of a real questionnaire, it was necessary to ensure that the survey items had enough context so that the translators could conduct their assignments properly. Therefore, when a survey item belonging to a grid of items (i.e., a battery of questions on the same topic) was selected but was not the first item in the grid, then the first survey item of that grid was also added to the sample, which guarantees that the laboratory questionnaire was read in a "natural" way; that is, it did not contain questions without context. The combination of random sampling and

qualitative criteria helped ensure satisfactory coverage of the characteristics of a survey.

Finally, 26 survey items comprising 268 sentences were selected according to a qualitative scheme provided by questionnaire translation experts. This scheme was defined to select survey items that are considered challenging for both human and machine translation. For instance, survey items that contain survey-specific terminology, complex syntax/grammar, terms related to one's feelings, state of mind, well-being (e.g., feeling lonely), and technical terms that may or may not be adapted to a specific country context (e.g., unemployment benefits). Few modifications were introduced to the source items to create the laboratory questionnaire, such as harmonizing special answer categories such as "Don't know" and "Refusal" across ESS and EVS survey items.

Following the TRAPD guidelines, the laboratory questionnaire was independently translated from English to German by two human translators. Afterward, meetings led by a reviewer in conjunction with the translators took place to discuss the options and decide on the final translations. The outputs derived from the review meeting were then used as reference translations in this work.

We specified the following criteria for the study participants. First, the two human translators should be (i) one professional translator with at least 5 years of translation practice and previous experience translating survey questionnaires, and; (ii) one social scientist with at least 2 years of work experience, with a background in questionnaire design and translation. The reviewer should also be a social scientist who has experience in questionnaire translation and questionnaire design and is willing to lead a review discussion. The combination of having both professional translators and social scientists collaborating in teams goes back to how the interdisciplinary TRAPD method is set up.

Other than the reference translations produced through the aforementioned process, we leverage the translations present in the Multilingual Corpus of Survey Questionnaires (MCSQ) (Zavala-Rojas et al., 2022), which is a corpus of questionnaires from the ESS, EVS, Survey of Health Ageing and Retirement in Europe (SHARE), and Wage Indicator survey (WIS). The MCSQ comprises 306 different questionnaires and over 4 million tokens in the source language (English) and their translations into Catalan, Czech, French, German, Norwegian, Portuguese, Spanish, and Russian, as well as 29 language varieties (e.g., Swiss-French, Austrian-German). Aside from the Wage Indicator, these survey projects employ the TRAPD as the recommended translation method, and thus serve the purpose of gathering other valid translation options to compare against the GPT translations.

4 Method

One way to evaluate the quality of MT outputs is to compare them with reference translations. In this context, reference translations serve as a quality benchmark, i.e., the more similar a given MT output is to a reference translation, the better. To properly assess MT similarity, lexical, syntactic, and semantic aspects should be considered.

While lexical similarity can be assessed straightforwardly, measuring semantic similarity imposes greater challenges, including handling polysemy and paraphrasing. Although some classic MT evaluation metrics, e.g., BLEU, mainly assess lexical similarity, recent metrics such as BERTScore and COMET integrate semantic similarity through the use of neural network-based models that leverage semantic distributional hypothesis, bidirectionality[2], and self-attention mechanisms to improve meaning quantification. Recent research recommends the application of neural network-based metrics for MT quality assessment, as they have shown high correlations with human evaluation and are considered resilient to domain shift (Freitag et al., 2022; Hendy et al., 2023).

In this paper, we study the performance of GPT-4o mini in the survey translation domain. We specifically chose GPT-4o mini because it is a cost-efficient model that surpasses the performance of past GPT models, such as GPT-3.5 Turbo.

4.1 Metrics

To evaluate the quality of automatic translations produced by GPT-4o mini at the sentence level, we employ COMET-22 and ChrF++. COMET-22 is a neural network-based metric that has been shown to correlate well with human judgments. In addition to automatic and reference translations, this metric takes into account the source text and a combination of direct assessments (DA), sentence-level scores, and word-level tags from Multidimensional Quality Metrics (MQM) error annotations to produce the scoring. ChrF++ is not neural network-

[2]Taking into account both left and right contexts simultaneously to understand and represent word meaning.

based; however, it has also been shown to correlate well with human judgments, especially for morphologically rich target languages like German (Popović, 2017). It uses `F-score` statistic for character and word n-gram matches, the latter being more strongly correlated with DA. We include `ChrF++` in this work because it provides direct information about the lexical similarity of translations, as the specific wording of the survey items matters in the context of survey translation.

Translation in the context of cross-national and cultural research can be challenging for MT since it is imperative for the text to be functionally equivalent across languages to accurately capture respondents' emotions and attitudes across different languages, cultures, and countries. For example, response scales containing qualifiers, e.g., "extremely satisfied", should ideally keep the same intensity when translated (Zavala-Rojas et al., 2022). Furthermore, survey questionnaires frequently contain survey-specific terms (e.g., instructions for the interviewer or the respondent), cultural and country-specific terminology (e.g., job titles, education levels), and short sentences in the case of some response scales and interviewer instructions.

Therefore, to supplement the insights derived from our qualitative analysis, we manually analyze some translation errors concerning survey-specific terminology, job descriptions, and qualifiers present in certain response options.

Finally, we created a questionnaire containing the source, reference, and `GPT`-translated items and asked five native German speakers acquainted with the survey translation domain to indicate, for each item, which translation option was more appropriate. Participants were blinded to the notion of which translation option was automatically generated and were asked to leverage accuracy, fluency, and cultural adaptation aspects. The order of the `GPT` and reference translations was randomized for each question. This questionnaire is available for consultation in the code repository.

4.2 Model prompting and temperature

To investigate how the integration of domain information in the prompts and the changes in the model temperature affect the quality of translations produced by `GPT-4o mini`, we translated our source laboratory questionnaire, testing 5 prompts and 6 temperature parameter values. Since the output of generative models may vary, we tested each prompt/temperature combination 5 times. A separate API call was used each time the questionnaire was translated with a given prompt/temperature combination. In all cases, the following persona description was given to the system: *"You are a professional translator specialized in translating survey questionnaires from English (from Great Britain) to German (from Germany) who works for social survey projects like the European Social Survey and the European Values Study."*.

We added domain knowledge to our prompts in a cumulative manner. `Prompt 0` is zero-shot, and therefore instructs the model to perform the translation task without any additional examples to steer it, containing only a few guidelines concerning the formatting of the output and the tone of the text.

Subsequently, in `prompt 1`, we added a list of automatically extracted terms in English and their German translations. We employ the MCSQ to this end. Using only the sentence-aligned English-German (from Germany) data from the corpus and the `tm2tb` library[3], we extracted a list of 477 biterms and integrated this list into `prompt 1` as well as the following prompts.

In `prompts` 2 and 3, other than the biterms, we also included guidelines and examples concerning using country-specific terminology for cultural adaptation and translating interviewer/respondent instructions and response scales, respectively. Finally, in `prompt 4`, we added four examples of English survey items and their translations to German, which were selected from EVS and ESS questionnaires, ensuring that the chosen survey items were not present in the laboratory questionnaire.

Concerning the model temperature, through preliminary tests, we observed that temperatures equal to or higher than 0.8 generated many instances of missing sentences and/or questionnaire formatting problems; therefore, we limit the temperature to 0.75 (the highest possible value would be 1). Namely, we tested the following temperatures: 0, 0.15, 0.3, 0.45, 0.6, 0.75.

The model prompts are available in the Appendix A. Our code, laboratory questionnaire, MT and reference translations, and scores of evaluation metrics are available on our Github repository.

5 Results

We start this Section by discussing the assessment of the overall questionnaire quality. Then, we focus

[3] https://github.com/luismond/tm2tb

on the translation quality of response options and respondent/interviewer instructions.

Figures 2 and 3 illustrate the similarity between GPT and reference translations according to COMET-22 and ChrF++, respectively, where the black dots and bars represent the mean and a 95% Confidence Interval. The similarity values range from 0 to 1 (or 0 to 100 in the case of ChrF++), where higher values indicate more similarity between automatic and reference translation. The X-axis indicates the prompt/temperature pairs, for instance p0_0.0 stands for prompt 0 with model temperature equal to 0, p4_0.75 stands for prompt 4 with model temperature equal to 0.75, etc.

For each sentence in the laboratory questionnaire, we took the average of the similarity values achieved in the five batches[4], taking into account the prompt and model temperature.

As can be observed in Figures 2 and 3, similarity values are overall high, and the prompt that achieved results most similar to reference translations is prompt 4, which was the prompt that contained examples of past survey items. Although prompt 1 also includes examples of bilingual terms extracted from MCSQ sentence-aligned data, the quality improvement is not as significant compared to prompt 4. The metrics also indicate that semantic similarity values are fairly high; however, there are differences concerning the wording of automatic and reference translations, as depicted in Figure 3. The model temperature does not appear to significantly impact the translation quality of the analyzed GPT translations.

We now turn to the translation quality of responses and instructions within the survey items analyzed in this study. Unlike Figures 2 and 2, which evaluated complete survey items without distinguishing sentence types, the remainder of this section focuses specifically on sentences that refer to responses or instructions[5].

We start by analyzing response translations, highlighting the response categories that include qualifiers and/or intensifiers in their text. Figures 4 and 5 illustrate the heatmap of similarity values between GPT and human translations according to ChrF++ and COMET-22, respectively. As in Figures 2 and 3, the X-axis of the heatmaps also indicates the prompt/temperature pairs.

As shown in the heatmaps, the GPT translations are overall similar to the references, especially when integrating examples in the prompt, i.e., translations produced with prompt 4. Figure 5, which depicts the ChrF++ values, indicates that using the prompt 4, the translations are also similar at the word level, since the metric computes similarity through character and word n-gram matches.

When manually analyzing responses that achieved low similarity values, we find that the sentence IDs 42, 126, 153, 228, and 246 correspond to the translations of the response option "Neither agree nor disagree", commonly present in response scales similar to the one depicted in Figure 1. The reference translation for this response option was *"Stimme weder zu noch nicht zu"*, while the most frequent GPT translations were *"Weder noch"* and *"Weder zustimmen noch ablehnen"*. Both translation options were present in past ESS and EVS questionnaires, as we verified using the MCSQ. The list of questionnaires in which these translation options are present is available in the Appendix B. The first option was more frequent, while the latter has been used only in a few EVS questionnaires localized for Luxembourgish-German.

The IDs 135 and 136 refer to a 0 to 10 response scale, where the extremes meant "Extremely important/unimportant". Reference translations for these response options were *"Äußerst unwichtig/wichtig"*. Although the GPT model succeeded in translating *"unwichtig/wichtig"*, in all cases, it failed to translate "Extremely", repeatedly translating the term as *"Extrem"*, which is often used in German to describe something radical or excessive, instead of indicating a high degree of something without implying excessiveness.

Beyond responses containing qualifiers, by revising the GPT translations and their metric scores, we found that one of the most challenging survey items was the translation of the response categories regarding the question *"Welchen Beruf üben Sie aus?"* ("What is your job?").

The first aspect that the GPT translations failed to achieve was gender adaptation (Savoldi et al., 2021). Since German is a gendered language, response options such as "Skilled manual worker" also provided gender inflections in the reference translations, e.g., *"Facharbeiter**/in**"*. Then, the GPT translations showed inconsistent terminology and mistranslations of job categories. Although

[4]We measured the variance of metric values and found that for most cases $variance < 0.01$. The mean and variance per sentence ID, prompt, and temperature are available in the Github repository.

[5]As illustrated in Figure 1 survey items may contain requests, responses, and instructions

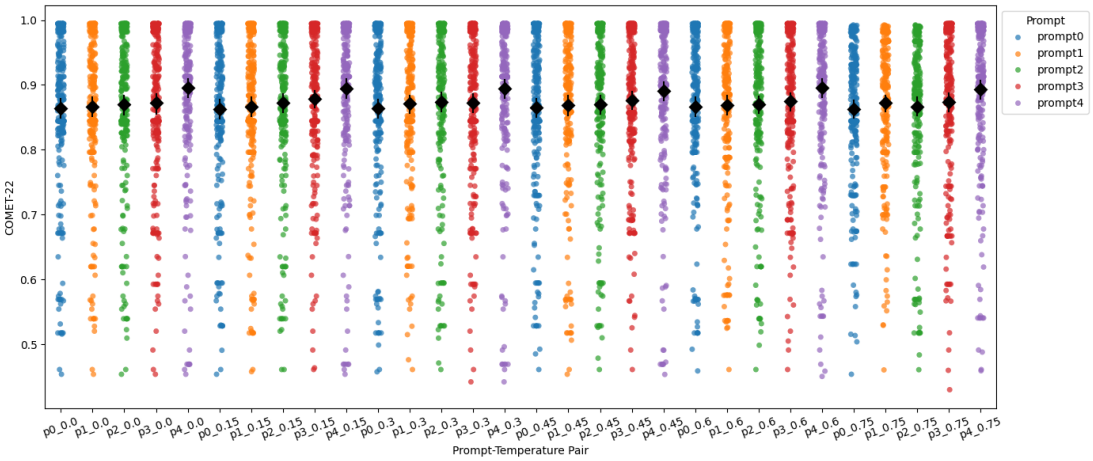

Figure 2: Similarity between `GPT-4o mini` and reference translations according to `COMET-22`.

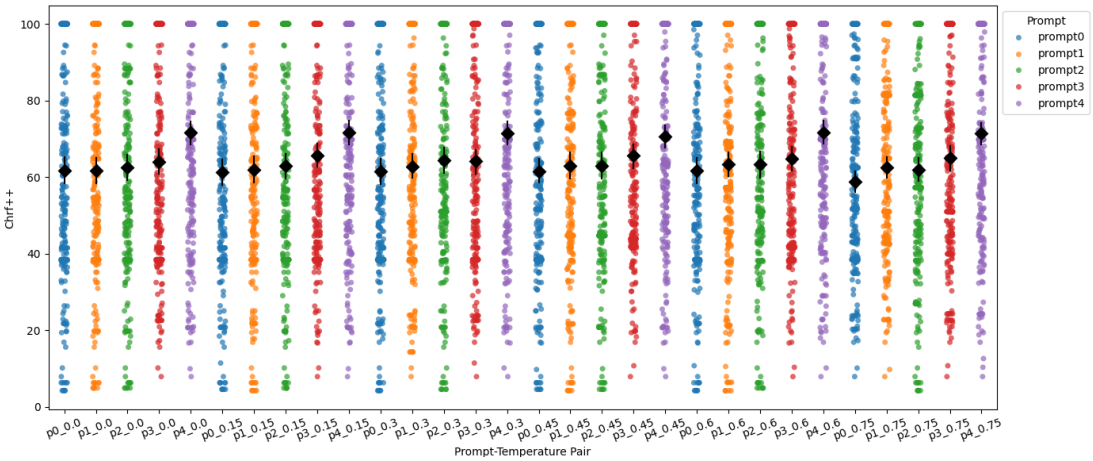

Figure 3: Similarity between `GPT-4o mini` and reference translations according to `ChrF++`.

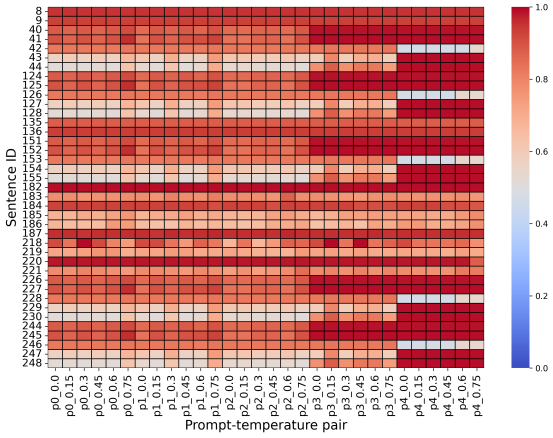

Figure 4: Heatmap of `COMET-22` for responses.

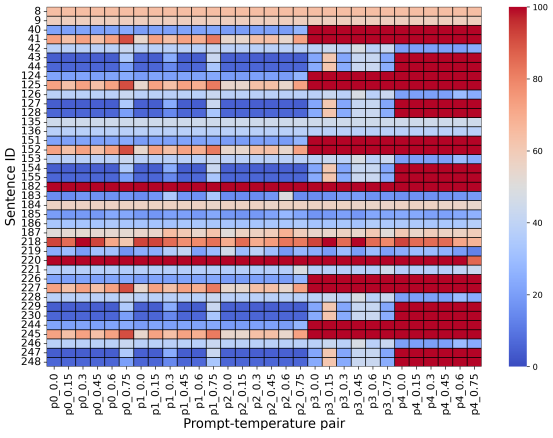

Figure 5: Heatmap of `ChrF++` for responses.

the model managed to correctly translate categories such as *Vorarbeiter und Aufsichtsperson* ("Foreman and supervisor"), *"Facharbeiter"*, *"Ungelernter Arbeiter"* ("Unskilled manual worker"), it struggled to produce idiomatic translations for terms like "middle/junior level non-manual", often resulting in English-influenced translations, e.g., *"Junior-Stufe/Mittelstufe nicht-manual"* or *"Mitarbeiter auf mittlerer/unterer Ebene"*. We also notice the influence of English on some partially correct translations, for instance, in *"Landwirt: Arbeitgeber, Manager oder selbstständig"* ("Farmer: employer,

manager or own account"), the term *"Manager"* is used instead of *"Leiter"*.

Now we comment on instruction translations. Interviewer instructions can be challenging elements for MT, and in many cases, the automatic translations are dissimilar to the references, as observed in Figures 6 and 7. Interviewer instructions often contain survey-specific terminology such as "Write in and code below" (ID 252), "Read out the statement and code in grid" (ID 122), and variations, which refer to the action of an interviewer registering the number of a response category provided by the respondent.

Although the GPT model was able to achieve translations equal or similar to those employed in past questionnaires for some complex instructions, e.g., *"Vorlesen und eine Antwort für jede ankreuzen"*("Read out and code one answer for each", ID 80) is similar to *"Vorlesen und eine Antwort ankreuzen"*, which was used in past EVS questionnaires, we observe that several translation options are literal and/or influenced by English syntax. For instance, the GPT translation options *"**Code** alle genannten"/"**Kodieren** Sie alle genannten"* and *"**Code** nur eine"/"**Kodieren** Sie nur eine"* were generated for the instructions "Code all mentioned" (ID 60) and "Code one only" (ID 115). *"**Erfassen**"*, i.e., to capture/record, would be a more idiomatic German equivalent term than "Code", however, terms like *"**Code**"* and *"**Kodieren**"*, are present in the MCSQ. *"Nur eine Antwort kodieren"*, for example, is a frequent translation in past ESS, EVS, and SHARE questionnaires, thus *"Kodieren Sie nur eine"* would be a good translation option if the syntax was not English-influenced.

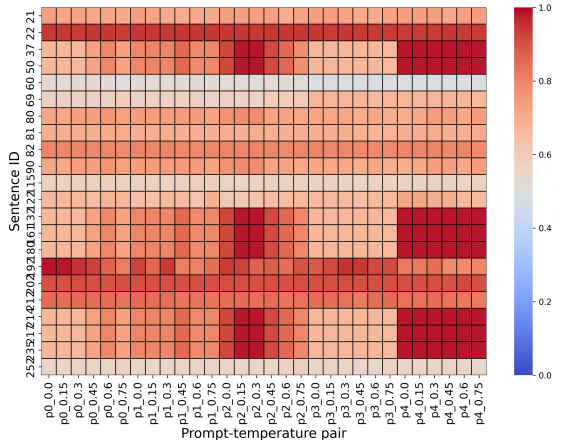

Figure 6: Heatmap of COMET-22 for instructions.

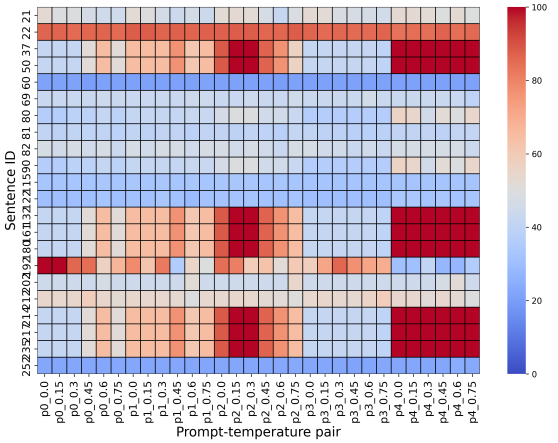

Figure 7: Heatmap of ChrF++ for instructions.

Finally, our experiment with five native German speakers showed that out of the 26 items included in our questionnaire, five of the survey items generated by the GPT model were deemed to have better quality than their human translation counterparts, taking into account accuracy, fluency, and cultural adaptation criteria.

Figure 8 shows the distribution of the responses for each of the questions. As can be observed in the graph, the participants generally preferred human translations, however, there were only five items for which all participants unanimously chose the human translation option over the GPT translation. An examination of the answer distribution also reveals instances of disagreement among the participants.

When verifying the inter-annotator agreement using Fleiss' Kappa, we obtained a 0.18 coefficient, confirming the low agreement between the participants. A certain degree of disagreement was anticipated for this task, as regional linguistic variations and individual preferences may influence participants' decisions. This variability underscores the importance of review meetings in the TRAPD method, where translation options are collaboratively discussed and evaluated.

On the one hand, the low inter-annotator agreement indicates that the GPT model achieved human-like translations in most cases, otherwise, the participants would have unanimously chosen the human translation options. On the other hand, it is not possible to draw conclusions concerning the perceived quality and adequacy of the GPT translations since the answers are mixed.

We cite the following limitations of our findings. This analysis considers only the English-German language pair, therefore, our conclusions cannot be

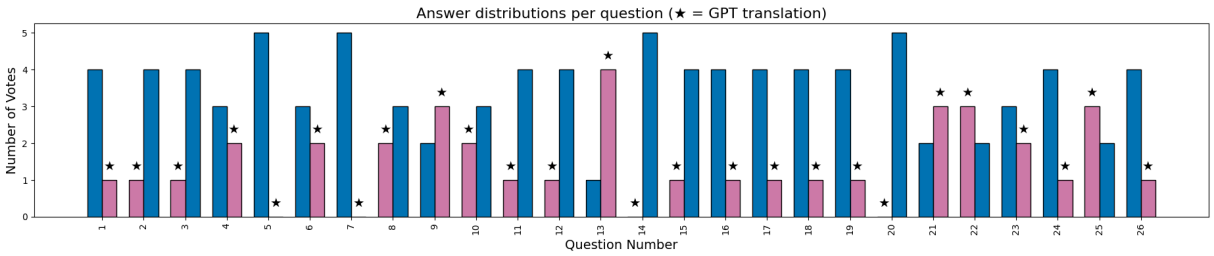

Figure 8: Answer distributions per question in the questionnaire. Stars indicate which answer option was the GPT translation.

generalized to other languages. Our quantitative evaluation was conducted at the sentence level; a word-level assessment could better identify problematic terms and expressions, offering deeper insights into the adequacy of GPT translations in the survey domain.

Concerning the experiment with native German speakers, the evaluation group could be increased, as increasing the number of annotators may improve the inter-annotator agreement. Furthermore, adding examples and more definitions to the task guidelines could help reduce annotator subjectivity.

6 Conclusion

In this study, we analyzed the quality of English to German questionnaire translations produced by `GPT-4o mini`, aiming to evaluate the feasibility of incorporating MT in the translation step of TRAPD method. We conducted a controlled translation experiment with human translators, and the automatic translation task included different prompts and model temperatures. Then, we analyzed the overall quality of the translations, focusing on survey-specific elements that can be challenging for MT, such as response scales and instructions.

Our findings indicate that `GPT-4o mini` achieved acceptable performance in the translation task, as depicted in our quantitative analysis and our experiment with five native German speakers acquainted with the survey translation domain. We found that providing examples of survey items fielded in past questionnaires was the best stimulus to improve translation quality. `GPT-4o mini` was able to keep the intensity of quantifiers in the response options in most cases, although it struggled to perform some crucial cultural adaptations, for instance, when referring to job descriptions. Moreover, the `ChrF++` scores suggest that, in some cases, the vocabulary used by the GPT model deviates from that employed by human translators, which can be problematic in the context of cross-national surveys.

Although LLM-based MT can be used as a tool to speed up the translation process potentially, human translators are essential for verifying and rectifying machine-translated text, as well as for mitigating the lack of appropriate gender, cultural, and localization adaptations by understanding the cultural context in which the survey is conducted (Benlidayi and Gupta, 2024). This is of particular importance when formulating and translating questions for gathering socio-demographic data, e.g., personal details related to household, gender identity, or health issues, which is a task that can be complex even for humans. Employing culturally adequate and high-quality translations in the questionnaires is fundamental to minimizing the survey error measurements, or in other words, it is necessary to ensure that the survey items indeed capture what they intend to measure and that possible differences found across countries cannot be attributed to translation errors.

In future research, it would be beneficial to explore the use of translation tables (e.g., for job titles, response scales), MQM and/or DA-annotated examples in the translation prompts, and a Retrieval-Augmented Generation (RAG) system to improve linguistic and cultural accuracy. Further experimentation would be required to evaluate if integrating MT or MT plus PE at the translation step of the TRAPD improves the overall efficiency of the process. For a more comprehensible study of the quality of `GPT-4o mini` translations, it is necessary to take into account other language pairs, since the performance of LLMs is reported to vary in function of the target language.

Acknowledgments

This work was supported by the Social Sciences and Humanities Open Cloud (SSHOC) project, and received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018, grant agreement #823782.

References

2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1304–1313.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170.

Ilke Coskun Benlidayi and Latika Gupta. 2024. Translation and cross-cultural adaptation: A critical step in multi-national survey studies. *Journal of Korean Medical Science*, 39(49).

John Bound, Charles Brown, and Nancy Mathiowetz. 2001. Measurement error in survey data. In *Handbook of econometrics*, volume 5, pages 3705–3843. Elsevier.

Ryan CL Brewster, Priscilla Gonzalez, Rohan Khazanchi, Alex Butler, Raquel Selcer, Derrick Chu, Barbara Pontes Aires, Marcella Luercio, and Jonathan D Hron. 2024. Performance of chatgpt and google translate for pediatric discharge instruction translation. *Pediatrics*.

Eldad Davidov and Alain De Beuckelaer. 2010. How harmful are survey translations? a test with schwartz's human values instrument. *International Journal of Public Opinion Research*, 22(4):485–510.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Adelina Gómez González-Jover. 2002. La equivalencia como cuestión central de la traducción en las instituciones de la unión europea. In *on line]. In: Actas del I Congreso Internacional sobre El español, lengua de traducción. Almagro*.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33.

J Harkness. 2011. Guidelines for best practice in cross-cultural surveys. *Survey Research Center, Institute for Social Research, University of Michigan*.

Janet A Harkness, Fons JR van de Vijver, Peter Ph Mohler, and John Wiley. 2003. *Cross-cultural survey methods*, volume 325. Wiley-Interscience Hoboken, NJ.

Janet A Harkness, Ana Villar, and Brad Edwards. 2010. Translation, adaptation, and design. *Survey methods in multinational, multiregional, and multicultural contexts*, pages 115–140.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019. Openkiwi: An open source framework for quality estimation. *ACL 2019*, page 117.

Geoffrey S Koby, Paul Fields, Daryl R Hague, Arle Lommel, and Alan Melby. 2014. Defining translation quality. *Tradumàtica*, 12:0413–420.

Jonas R Kunst and Kinga Bierwiaczonek. 2023. Utilizing ai questionnaire translations in cross-cultural and intercultural research: Insights and recommendations. *International Journal of Intercultural Relations*, 97:101888.

Kibeom Lee and Michael C Ashton. 2004. Psychometric properties of the hexaco personality inventory. *Multivariate behavioral research*, 39(2):329–358.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, 25(75):1–49.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997.

Peter Ph Mohler and Timothy P Johnson. 2010. Equivalence, comparability, and methodological progress. *Survey methods in multinational, multiregional, and multicultural contexts*, pages 17–29.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237.

Christiane Nord. 2006. Translating as a purposeful activity: a prospective approach. *Teflin Journal*, 17(2):131–143.

Daniel Oberski, Willem E Saris, and Jacques Hagenaars. 2007. Why are there differences in measurement quality across countries. *Measuring Meaningful Data in Social Research. Acco, Leuven*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633.

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. 2023. Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–483.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Maja Popović. 2018. Error classification and analysis for machine translation quality assessment. In *Translation quality assessment: From principles to practice*, pages 129–158. Springer.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Lucia Specia, Carolina Scarton, Gustavo Henrique Paetzold, and Graeme Hirst. 2018. *Quality estimation for machine translation*, volume 11. Springer.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121.

Anne M Turner, Kristin N Dew, Loma Desai, Nathalie Martin, and Katrin Kirchhoff. 2015. Machine translation of public health materials from english to chinese: a feasibility study. *JMIR public health and surveillance*, 1(2):e4779.

Mohamed Ugas, Maria Anna Calamia, Jessica Tan, Ben Umakanthan, Christine Hill, Karen Tse, Angela Cashell, Zaynab Muraj, Meredith Giuliani, and Janet Papadakos. 2025. Evaluating the feasibility and utility of machine translation for patient education materials written in plain language to increase accessibility for populations with limited english proficiency. *Patient Education and Counseling*, 131:108560.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.

Peggy Walde and Birgit Angela Völlm. 2023. The trapd approach as a method for questionnaire translation. *Frontiers in Psychiatry*, 14:1199989.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661.

Diana Zavala-Rojas, Willem E Saris, and Irmtraud N Gallhofer. 2018a. Preventing differences in translated survey items using the survey quality predictor. *Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts (3MC). New York: Wiley Series in Survey Methodology*.

Diana Zavala-Rojas, Willem E. Saris, and Irmtraud N. Gallhofer. 2018b. Preventing Differences in Translated Survey Items using the Survey Quality Predictor. In Timothy P. Johnson, Beth-Ellen Pennell, Ineke A. L. Stoop, and Brita Dorer, editors, *Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts (3MC)*, chapter 17, pages 357–384. Wiley Series in Survey Methodology, New York.

Diana Zavala-Rojas, Danielly Sorato, Lidun Hareide, and Knut Hofland. 2022. The multilingual corpus of survey questionnaires: a tool for refining survey translation. *Meta*, 67(1):71–93.

A Model Prompts

Besides the texts included in this section, the prompts 1, 2, 3, and 4 included a list of 477 biterms automatically extracted from the MCSQ. The complete list of biterms can be consulted in the repository that contains the code and data related to this article[6].

Prompt 0: Please translate the following questionnaire written in English (from Great Britain) to German (from Germany). Your translated text should read as if it were written by a native German speaker from Germany. Use a direct and clear communication style. Use a professional, formal, and neutral tone suitable for survey questions. Replace any instances of the token "[country]" with the appropriate country in your output. Your translation output must have the same number of lines as the source (English) questionnaire. Include only the translated questionnaire in your output.

Prompt 1: Please translate the following questionnaire written in English (from Great Britain) to German (from Germany). After the questionnaire, you can find a list of English terms and their translations to German extracted from the European Social Survey and European Values Study questionnaires. This list may contain useful survey-specific terminology translation examples for you. Your translated text should read as if it were written by a native German speaker from Germany. Use a direct and clear communication style. Use a professional, formal, and neutral tone suitable for survey questions. Replace any instances of the token "[country]" with the appropriate country in your output. Your translation output must have the same number of lines as the source (English) questionnaire. Include only the translated questionnaire in your output.

Prompt 2: Please translate the following questionnaire written in English (from Great Britain) to German (from Germany). After the questionnaire, you can find a list of English terms and their translations to German extracted from the European Social Survey and European Values Study questionnaires. This list may contain useful survey-specific terminology translation examples for you. Use expressions and terminology that are specific from Germany to ensure proper cultural adaptation when applicable, especially when translating terms related to education levels and job descriptions/titles. For instance, the "Qualification from vocational ISCED 2C programmes of 2 years or longer duration, no access to ISCED 3" education level would be equivalent to "Volks-/Hauptschulabschluss bzw. Polytechnische Oberschule mit Abschluss 8. oder 9. Klasse" in Germany; the job description "in education, (not paid for by employer) even if on vacation" would be equivalent to "Schule/Ausbildung (nicht vom Arbeitgeber bezahlt; auch während der Ferien oder im Urlaub)" in Germany. Your translated text should read as if it were written by a native German speaker from Germany. Use a direct and clear communication style. Use a professional, formal, and neutral tone suitable for survey questions. Replace any instances of the token "[country]" with the appropriate country in your output. Your translation output must have the same number of lines as the source (English) questionnaire. Include only the translated questionnaire in your output.

Prompt 3: Please translate the following questionnaire written in English (from Great Britain) to German (from Germany). After the questionnaire, you can find a list of English terms and their translations to German extracted from the European Social Survey and European Values Study questionnaires. This list may contain useful survey-specific terminology translation examples for you. Use expressions and terminology that are specific from Germany to ensure proper cultural adaptation when applicable, especially when translating terms related to education levels and job descriptions/titles. For instance, the "Qualification from vocational ISCED 2C programmes of 2 years or longer duration, no access to ISCED 3" education level would be equivalent to "Volks-/Hauptschulabschluss bzw. Polytechnische

[6] https://github.com/dsorato/KONVENS_2025_MT_paper

Oberschule mit Abschluss 8. oder 9. Klasse" in Germany; the job description "in education, (not paid for by employer) even if on vacation" would be equivalent to "Schule/Ausbildung (nicht vom Arbeitgeber bezahlt; auch während der Ferien oder im Urlaub)" in Germany. When translating response scales, focus on keeping the concepts of interest and intensity of qualitifies/intensifiers the same across languages (e.g., "Agree strongly" -> "Stimme voll und ganz zu", "Very dissatisfied" -> "Sehr unzufrieden"). Do not forget to correctly translate and add to your output important survey-specific elements such as instructions to the respondent/interviewer like "READ OUT" and "Code all mentioned". Your translated text should read as if it were written by a native German speaker from Germany. Use a direct and clear communication style. Use a professional, formal, and neutral tone suitable for survey questions. Replace any instances of the token "[country]" with the appropriate country in your output. Your translation output must have the same number of lines as the source (English) questionnaire. Include only the translated questionnaire in your output.

Prompt 4: Please translate the following questionnaire written in English (from Great Britain) to German (from Germany). After the questionnaire, you can find a list of English terms and their translations to German extracted from the European Social Survey and European Values Study questionnaires. This list may contain useful survey-specific terminology translation examples for you. Use expressions and terminology that are specific from Germany to ensure proper cultural adaptation when applicable, especially when translating terms related to education levels and job descriptions/titles. For instance, the "Qualification from vocational ISCED 2C programmes of 2 years or longer duration, no access to ISCED 3" education level would be equivalent to "Volks-/Hauptschulabschluss bzw. Polytechnische Oberschule mit Abschluss 8. oder 9. Klasse" in Germany; the job description "in education, (not paid for by employer) even if on vacation" would be equivalent to "Schule/Ausbildung (nicht vom Arbeitgeber bezahlt; auch während der Ferien oder im Urlaub)" in Germany. When translating response scales, focus on keeping the concepts of interest and intensity of qualitifies/intensifiers the same across languages (e.g., "Agree strongly" -> "Stimme voll und ganz zu", "Very dissatisfied" -> "Sehr unzufrieden"). Do not forget to correctly translate and add to your output important survey-specific elements such as instructions to the respondent/interviewer like "READ OUT" and "Code all mentioned". Your translated text should read as if it were written by a native German speaker from Germany. Use a direct and clear communication style. Use a professional, formal, and neutral tone suitable for survey questions. Replace any instances of the token "[country]" with the appropriate country in your output. Your translation output must have the same number of lines as the source (English) questionnaire. Include only the translated questionnaire in your output. To further help you in your task, below you can find some examples of survey items extracted from past European Values Study and European Social Survey questionnaires translated from English (United Kingdom) to German (Germany), observe these examples and apply the same text style and terminology to your translations:

Example 1)
Source:
Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?
Most people can be trusted
Can't be too careful
Don't know
No answer

Translation:
Würden Sie ganz allgemein sagen, dass man den meisten Menschen vertrauen kann, oder dass man da nicht vorsichtig genug sein kann?
Man kann den meisten vertrauen
Man kann nicht vorsichtig genug sein
Weiß nicht
Verweigert

Example 2)
Source:
SHOW CARD 41 – READ OUT AND CODE ONE ANSWER PER LINE
How much do you agree or disagree with each of the following:
Religious leaders should not influence government decisions
agree strongly
agree
neither agree nor disagree

disagree
disagree strongly
Don't know
No answer

Translation:
LISTE 41 VORLEGEN - VORGABEN VORLESEN UND EINE ANTWORT PRO ZEILE ANKREUZEN
Sagen Sie mir bitte zu jeder der folgenden Aussagen, ob Sie ihr voll und ganz zustimmen, zustimmen, nicht zustimmen oder überhaupt nicht zustimmen.
Die Kirchenoberhäupter sollten nicht versuchen, Entscheidungen der Regierung zu beeinflussen
Stimme voll und ganz zu
Stimme zu
Weder noch
Stimme nicht zu
Stimme überhaupt nicht zu
Weiß nicht
Verweigert

Example 3)
Source:
In the past 2 years, did the police in [country] approach you, stop you or make contact with you for any reason?
Yes
No
Don't know

Translation:
In den letzten 2 Jahren, hat sich die Polizei in Deutschland aus irgendeinem Grund an Sie gewendet, Sie angehalten oder kontaktiert?
Ja
Nein
Weiß nicht

Example 4)
Source:
Now some questions about whether or not the police in [country] treat victims of crime equally.
Please answer based on what you have heard or your own experience.
When victims report crimes, do you think the police treat rich people worse, poor people worse, or are rich and poor treated equally?
Choose your answer from this card.
Rich people treated worse
Poor people treated worse
Rich and poor treated equally
Don't know

Translation:
Nun einige Fragen dazu, ob die Polizei in Deutschland alle Opfer von Straftaten gleich behandelt oder nicht.
Bitte denken Sie bei Ihrer Antwort an das, was Sie gehört oder selbsterlebt haben.
Wenn Opfer zur Polizei gehen, um eine Straftat zu melden, glauben Sie, dass die Polizei reiche Leute schlechter behandelt, arme Leute schlechter behandelt oder dass beide gleich behandelt werden?
Wählen Sie Ihre Antwort aus Liste 29 aus.
Reiche Leute werden schlechter behandelt
Arme Leute werden schlechter behandelt
Reiche und arme Leute werden gleich behandelt
Weiß nicht.

B Translations retrieved from the MCSQ

Tables 1 and 2 show the list of published questionnaires in which the German translation options for the response option "Neither agree nor disagree" appear. The MCSQ does not include all published questionnaires from the European Social Survey (ESS), European Values Study (EVS), Survey of Health, Ageing and Retirement in Europe (SHARE), and Wage Indicator Survey (WIS), although it covers most of them, especially in the case of the ESS.

Weder noch		
Study	Round/Year	Country
ESS	1 (2002), 2 (2004), 3 (2006). 4 (2008), 5 (2010), 6 (2012), 8 (2016), 9 (2018)	Austria
EVS	3 (1999), 4 (2008), 5 (2017)	Austria
SHARE	7 (2017), 8 (2019)	Austria
ESS	1 (2002), 2 (2004), 3 (2006). 4 (2008), 5 (2010), 6 (2012), 7 (2014), 8 (2016), 9 (2018)	Switzerland
EVS	4 (2008), 5 (2017)	Switzerland
SHARE	7 (2017)	Switzerland
ESS	1 (2002), 2 (2004), 3 (2006). 4 (2008), 5 (2010), 6 (2012), 7 (2014), 8 (2016), 9 (2018)	Germany
EVS	3 (1999), 4 (2008), 5 (2017)	Germany
WIS	1 (2000)	Germany
ESS	2 (2004)	Luxembourg
EVS	4 (2008)	Luxembourg

Table 1: Questionnaires included in the MCSQ in which the translation *"Weder noch"* appears.

Weder zustimmen noch ablehnen		
Study	Round/Year	Country
EVS	4 (2008), 5 (2017)	Luxembourg

Table 2: Questionnaires included in the MCSQ in which the translation *"Weder zustimmen noch ablehnen"* appears.