When Short Responses Are Hard to Distinguish in Translation Audits

João Victor M. Correa[1,2] **Pedro Shiozawa**[3]
Victor Hierro Moraes Ferreira[4] **Emanoel Cristhian Leão Marinho**[4]
[1]Universidade de São Paulo
[2]Great People Mental Health
[3]Faculdade de Ciências Médicas da Santa Casa de São Paulo
[4]Universidade Federal do Pará
victormcorrea@usp.br, pedro.shiozawa@fcmsantacasasp.edu.br
victor.hierro@icen.ufpa.br, emanoel.marinho@itec.ufpa.br

Abstract

Reference-free translation audits have little to work with when a survey answer is only one or two words long. A translation may be reasonable and still be nearly impossible to reconnect to its source among thousands of similar answers. We examined this problem in 44,118 Portuguese–English response pairs from 14,706 survey-specific respondent records across four workplace surveys. One third of the pairs addressed mental health or well-being. Our analysis kept duplicate-aware retrieval, COMETKiwi quality estimates, language identification, and affective-lexicon coverage separate because they answer different questions. The retrieval encoder found the paired target or an exact normalized duplicate at rank one for 71.80% of English queries and 73.37% of Portuguese queries overall. Response length changed the picture. Rank-one recovery ranged from 8.39–12.96% for one- or two-word answers to 98.45–99.08% for answers longer than 20 words. COMETKiwi scores did not rise steadily with length, and about one quarter of responses in either language contained no emotion term covered by the lexicons. A failed match therefore does not, by itself, identify a poor translation. For brief and repetitive survey text, reference-free audits should report how distinguishable the response is and how much of it the diagnostic resources cover.

1 Introduction

Open-ended surveys capture what fixed-choice questions miss. In workplace and mental-health research, they allow respondents to describe psychological safety, stressors, and well-being in their own words (Hobbs and Ong, 2023). Translation can make those answers accessible to English-centered tools, but it may also alter the distinctions that made the open response informative. This tension matters because the resources used to organize and analyze free text remain unevenly distributed across languages, while both language and culture can alter what a feature means (Kučera and Mehl, 2022). Multilingual mental-health NLP faces the same imbalance at a larger scale, with non-English evidence remaining sparse relative to the populations that these systems may eventually affect (Bucur et al., 2026).

Machine translation offers a practical route through this imbalance. Instead of rebuilding every downstream instrument for Brazilian Portuguese, researchers can translate responses into English and use existing models. Studies of automated content analysis suggest that this strategy can preserve useful aggregate patterns in some languages and domains (Windsor et al., 2019; Boot, 2021; Chew et al., 2023). That convenience does not settle whether a particular corpus retains the distinctions that matter. Translation can obfuscate author-linked stylistic signals and alter affective information (Rabinovich et al., 2017; Troiano et al., 2020). A recent preprint evaluating multilingual mental-health datasets reported substantial variation in downstream performance on translated data (Raihan et al., 2026). Machine-translation-based continued training also did not improve the cross-lingual language-understanding tasks tested by Ji et al. (2024).

This uncertainty is consequential for mental-health and well-being research. Text embeddings can represent components of the empirical association structure among psychiatric questionnaire items (Kambeitz et al., 2025), but a model's ability to represent a construct is separate from whether translation preserves distinctions expressed in a respondent's phrasing. A high score from one automated metric cannot establish clinical or cultural equivalence. Modern quality estimators can operate without a reference translation (Rei et al., 2022),

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 15–25
November 9, 2026 ©2026 Association for Computational Linguistics

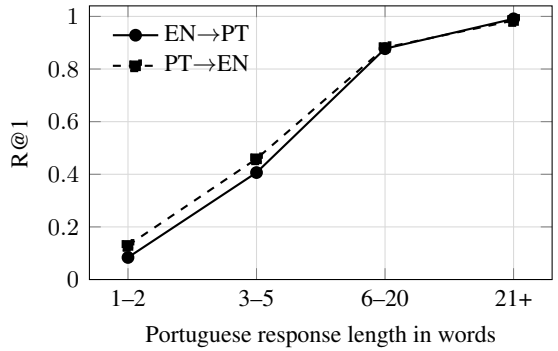


Figure 1: Duplicate-aware rank-one recovery by Portuguese response length. Error bars show 95% survey-specific respondent-record-cluster bootstrap intervals. Every query is ranked against the full corpus.

although their scores depend on model configuration, data, and the error phenomena under study (Zouhar et al., 2024; Moghe et al., 2025). Evaluation therefore needs several views that answer different questions.

Against this background, we ask what complementary evidence reference-free tools can provide for Portuguese-to-English translations of short workplace survey responses and where that evidence becomes unreliable. The corpus contains three balanced prompt families covering positive workplace qualities, areas for improvement, and mental health or emotional well-being. Rather than collapse the evidence into a single score, we examine the translations through three complementary views. Duplicate-aware retrieval tests whether paired texts remain recoverable among 44,118 candidates, while a second model estimates translation quality without a human reference. A diagnostic branch combines language identification with bilingual affective-lexicon coverage. Together, the analyses show a strong association between response length and retrieval, while repetition and lexical coverage impose additional interpretive constraints. The contribution is a joint account of retrieval, COMETKiwi, and lexical coverage in operational survey text, including the response conditions under which their outputs diverge.

2 Related Work

Prior work has directly compared translating the text with translating the analysis instrument. Machine-translated text can sometimes support content analysis more effectively than a translated dictionary, particularly when the target-language tool is mature (Windsor et al., 2019; Boot, 2021). Yet this advantage varies across linguistic categories, and aggregate agreement can obscure language-specific failures (Dudău and Sava, 2021; Kučera and Mehl, 2022). Our study examines this problem in very short, naturally occurring survey responses.

When human references are unavailable, reference-free estimators offer a second way to examine translated text. COMETKiwi learns to estimate translation quality from a source and candidate translation without a reference (Rei et al., 2022). Its output remains a model-based criterion rather than ground truth. COMET-family scores can change with implementation choices (Zouhar et al., 2024), and favorable intrinsic scores do not guarantee that distinctions needed by a downstream task survive (Moghe et al., 2023). Challenge-set studies also show that metrics differ in the errors they detect (Moghe et al., 2025). For this reason, we preserve disagreement among metrics rather than combine them into one score.

Response length adds another constraint because it changes what can be inferred from an open answer. Longer well-being responses have produced more stable psycholinguistic associations than shorter answers (Hobbs and Ong, 2023). In Brazilian Portuguese emotion classification, transformer models have outperformed lexicon-based methods, while very short texts are particularly difficult for EmoAtlas (Domingues Aparecido et al., 2025). Human-calibrated work in Chinese–English literary translation has shown the value of comparing source and target affective vectors, but we do not assume that its thresholds transfer to workplace survey text (Zhou et al., 2026). These findings motivate an audit that reports affective coverage explicitly and treats lexical profiles as exploratory diagnostics rather than direct measures of a respondent's emotional state.

3 Data and Methods

3.1 Survey Cohorts and Response Selection

The source material consists of four workplace survey cohorts administered in Brazil by Great Place to Work (GPTW) on behalf of client companies in 2023, 2024, and 2025. The companies commissioned the surveys for internal organizational assessment, and study-team governance records indicate that they authorized secondary research use of pseudonymized responses. Two open-ended GPTW questions recurred across the surveys. They

Cohort	Year	Records	Pairs
A	2023	4,858	14,574
B	2024	4,022	12,066
C	2025	5,049	15,147
D	2025	777	2,331
Total		14,706	44,118

Table 1: Survey-specific respondent records and paired responses. Each record contributes one response to each of three prompt families.

asked what makes the respondent's company a great place to work and what could be improved. Each of the four selected cohorts also included one question about mental health or emotional well-being, although its wording varied by cohort. Every retained survey-specific respondent record therefore contains three open-ended answers. The analysis preserves the twelve cohort–prompt combinations while mapping them to three common families. No response text is reproduced in this paper.

Identifiers were created within each survey cohort. The unit count is therefore 14,706 survey-specific respondent records rather than 14,706 known unique people. A person who answered in more than one year cannot be linked across cohorts. Cohort sizes are shown in Table 1. Each retained survey-specific record contributes a complete triad, producing 44,118 Portuguese source responses and 44,118 English translations.

After applying source- and target-language metadata, prompt, and nonempty-text filters, the frozen source snapshot contained 67,809 eligible stored rows from 17,187 candidate survey-specific respondent records. We excluded 2,481 survey-specific respondent records without all three questions, removing 6,213 stored rows. Because an English output was required before this count, the selection cannot estimate translation-service failure rates. The remaining 61,596 rows included exact duplicate ingestions in two 2025 cohorts. Deterministic de-duplication removed 17,478 storage duplicates and retained 44,118 logical pairs. No retained Portuguese response had competing English translation variants in the archived data.

Unicode NFKC normalization and whitespace collapse were used only to identify duplicates and quality-control flags. Original text, including case and punctuation, was retained for model inference. Very short, repeated, and source-equal-to-translation responses were not removed from the primary analysis because they are part of the deployment conditions the audit is intended to characterize. There were 836 normalized source–translation copies (1.90%). These were flagged rather than classified as failures because names, acronyms, numbers, and shared forms can remain unchanged legitimately. Figure 2 summarizes cohort construction and the three complementary audit views applied to every logical pair.

3.2 Translation Workflow and Provenance

Available study records and the source-code history of the service indicate that the English texts were generated before this study through an operational large-language-model translation workflow. The service submitted keyed batches of Portuguese text with English as the requested target language and used translation-only output instructions. It included automated retry and model-fallback mechanisms. The analytic database retained the paired source and translated texts but did not preserve linkable row-level records of the model, model version, prompt version, processing time, retry path, fallback path, or post-editing history. We therefore evaluated the archived translation corpus and did not attribute its outputs to a specific model or configuration.

3.3 Language Quality-Control Flags

We ran Lingua 2.2.0 offline with Portuguese, English, and Spanish as the only candidate languages and a minimum relative distance of 0.05 (Stahl, 2026). The expected labels were Portuguese for the source and English for the translation. A low-confidence result was marked ambiguous. These labels were retained as descriptive flags because language identification is unstable for brief or name-like strings. The primary analysis includes every pair.

3.4 Instruction-Conditioned Paired-Text Retrieval

The primary encoder was Qwen3-Embedding-4B with frozen weights, normalized embeddings, bfloat16 inference, and the model's default pooling (Zhang et al., 2025). Portuguese-to-English queries were instructed to retrieve the English translation that retained the meaning of the workplace response. The reverse direction used the corresponding English-to-Portuguese instruction. Candidate documents were encoded without an instruction. Qwen3-Embedding-8B repeated the analysis as a

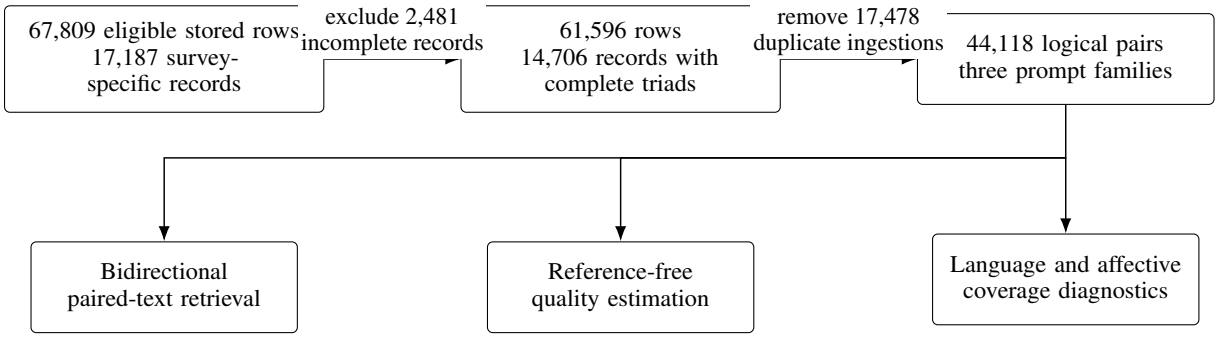


Figure 2: Cohort construction and the three audit views. The views are interpreted separately because none provides human ground truth.

prespecified model-scale sensitivity. No model was fine-tuned on the survey data. Appendix B.1 gives the exact instructions and query format.

For every query, all 44,118 texts in the other language formed the candidate set. We report recall at rank one (R@1), recall at rank five (R@5), and mean reciprocal rank (MRR). Retrieval uses a multi-positive definition. When several candidate rows have identical normalized target text, any member of that target-language duplicate group is counted as correct. This prevents repeated generic answers from being mislabeled as distinct negative examples. Each match identifies a target-text group shared by one or more records. Semantically equivalent targets with different normalized wording count as competing candidates. Since only queries receive an instruction, directional results are reported separately rather than averaged.

A sensitivity analysis retained every query but collapsed the target candidates into unique normalized-text groups separately for each retrieval direction. A target group's score was the maximum similarity among its member rows. This analysis tests whether repeated target rows occupy several rank positions even after the primary metric correctly treats them as multiple valid matches.

3.5 Reference-Free Quality Estimation

We applied the frozen `Unbabel/wmt23-cometkiwi-da-xl` model[1] to the Portuguese source and English translation (Rei et al., 2022, 2023). The model produces a reference-free automated quality estimate without requiring a human translation. Scores are summarized as a continuous distribution. We do not use a pass threshold because no cutoff has been calibrated for short Brazilian Portuguese workplace responses. COMETKiwi and retrieval are treated as different views. The former estimates learned translation quality, while the latter measures whether a pair remains distinguishable within this corpus.

[1] Immutable revision `33858b2239a139d497d9c74952c88b89a8c06213`.

3.6 Exploratory Affective Profiles

EmoAtlas was run independently in Portuguese and English with language-specific spaCy pipelines and Plutchik emotion and valence resources (Semeraro et al., 2025). For each pair, we calculated the square root of base-2 Jensen–Shannon divergence between the language-specific emotion and valence distributions. Emotion-lexicon coverage was the number of unique semantic-network nodes matched by an emotion lexicon divided by the number of semantic nodes. Under this implementation, two all-zero profiles receive a distance of 0, while a zero profile compared with a nonzero profile receives approximately 0.7071. An all-zero distribution indicates that no emotion-bearing terms were found, not that the response contained no emotion. We therefore report lexicon coverage and the frequency of zero distributions before interpreting distances. These measures describe lexical profiles rather than respondent states or clinical constructs.

3.7 Outcome Hierarchy and Statistical Analysis

The primary outcomes are duplicate-aware bidirectional R@1, R@5, and MRR from the 4B encoder. COMETKiwi is a separate automated criterion. Prespecified length analyses stratify by Portuguese response length using bands of 1–2, 3–5, 6–20, and more than 20 whitespace-delimited words. Prompt-family and 8B-encoder results are secondary. Exact normalized copies, repeated responses, and language flags are treated as corpus diagnostics rather than inferential outcomes.

Uncertainty intervals use a percentile bootstrap with 2,000 repetitions and seed 20260820. Sampling clusters are survey-specific respondent records, so all eligible responses from a record receive the same bootstrap multiplicity. The resulting estimates are row-weighted means across the resampled records. For filtered estimates, each replicate samples the full universe of 14,706 records before applying the filter, including records with no eligible row. Repeat people cannot be linked across cohorts and therefore cannot be clustered together. The confidence level is 95%. The retrieval candidate corpus remains fixed, so the intervals quantify variation across sampled query records rather than the construction of a new candidate corpus.

This fixed-corpus design also governs the length-stratified and sensitivity analyses. Prompt-family and response-length estimates describe selected queries against all 44,118 targets rather than smaller within-group retrieval tasks. The expected-language sensitivity likewise filters evaluated query pairs while preserving their original full-corpus ranks. The remaining analyses compare paired row-level outcomes from the 8B and 4B encoders, collapse repeated targets into normalized-text groups, and restrict affective distances to pairs with nonzero profiles in both languages.

4 Results

4.1 Overall Automated Evidence

The primary encoder placed the paired target-language text or an exact normalized duplicate at rank one for 71.80% of English-to-Portuguese queries and 73.37% of Portuguese-to-English queries. R@5 reached 80.12% and 82.46%, while MRR was 75.82% and 77.63% (Table 2). The mean COMETKiwi score was 0.7344 (95% CI 0.7330–0.7357), with a median of 0.7686 and an interquartile range of 0.6953–0.8140. These values describe model outputs over the corpus and do not classify individual translations as correct or incorrect.

One third of the corpus was elicited by mental-health or well-being prompts. In this family, R@1 was 71.85% in the English-to-Portuguese direction and 74.90% in the reverse direction, with mean COMETKiwi 0.7270. The improvement family had the highest R@1 values, 75.35% and 76.35%, but the lowest COMETKiwi mean at 0.7143. Positive-workplace responses had lower R@1 values, 68.21% and 68.85%, yet the highest COMETKiwi mean at 0.7619. Retrieval distinctiveness and quality estimation therefore ranked the three prompt families differently.

View	Measure	Estimate	95% CI
EN→PT	R@1	.7180	[.7130, .7234]
	R@5	.8012	[.7968, .8056]
	MRR	.7582	[.7531, .7626]
PT→EN	R@1	.7337	[.7287, .7386]
	R@5	.8246	[.8205, .8288]
	MRR	.7763	[.7717, .7808]
COMETKiwi	Mean score	.7344	[.7330, .7357]

Table 2: Primary automated outcomes over 44,118 pairs. Intervals use a 2,000-repetition bootstrap clustered by survey-specific respondent record.

4.2 Retrieval Varies Sharply with Response Length

The aggregate retrieval estimates concealed a large difference by response length (Figure 1). Among 5,569 responses containing one or two Portuguese words, R@1 was 8.39% for English-to-Portuguese retrieval and 12.96% for Portuguese-to-English retrieval. The rates rose to 40.65% and 45.89% among 8,157 responses with three to five words. For the 19,543 responses containing 6–20 words, R@1 reached 87.73% and 88.12%. The 10,849 responses longer than 20 words reached 99.08% and 98.45%.

Among the 30,392 queries whose Portuguese source exceeded five words, R@1 was 91.78% and 91.81%. These estimates do not show that short translations were worse. Short answers offer fewer distinguishing cues and are more likely to be shared by many respondents. COMETKiwi followed a different pattern, with means of 0.6414, 0.7798, 0.7526, and 0.7152 across the four ascending length bands. The disagreement supports the interpretation of retrieval as corpus-level identifiability rather than a direct translation-quality score.

4.3 Repetition, Model Scale, and Language Flags

Repeated content was common. The Portuguese responses formed 982 repeated-text groups covering 5,684 rows, and the English responses formed 1,104 groups covering 6,598 rows. The largest group in either language contained 405 rows. The primary multi-positive metric prevents identical target strings from becoming false negatives. When target candidates were instead collapsed into 39,416 Portuguese or 38,624 English normalized-

text groups, R@1 was unchanged at 71.80% and 73.37%. R@5 rose from 80.12% to 80.67% and from 82.46% to 83.49%, while MRR rose from 75.82% to 76.07% and from 77.63% to 78.02%. Maximum-similarity grouping preserves the top-ranked target group and therefore R@1. The changes in R@5 and MRR describe how repeated target rows occupy rank positions.

The prespecified encoder-scale sensitivity produced a directional trade-off. Relative to the 4B model, the 8B model reduced English-to-Portuguese R@1 by 2.37 percentage points (95% CI -2.68 to -2.08) and increased Portuguese-to-English R@1 by 1.47 points (95% CI 1.19 to 1.76). The larger encoder therefore did not provide a general improvement.

A separate quality-control sensitivity examined the automated language flags. Both texts received their expected label for 41,925 pairs (95.03%). The Portuguese source was labeled as expected in 95.73% of pairs, while 2.60% were ambiguous and 1.67% received another candidate label. English translations were labeled as expected in 97.46%, with 2.04% ambiguous and 0.50% mismatched. Both labels matched for only 71.92% of one- or two-word responses, compared with 99.43% at 6–20 words and 99.99% above 20 words. In the both-labels-matched subset, R@1 was 74.70% (95% CI 74.20–75.19) and 76.26% (95% CI 75.78–76.72), while mean COMETKiwi was 0.7431 (95% CI 0.7420–0.7443). The subset's higher scores do not verify language or translation quality because the omitted pairs are concentrated among short strings, which also have lower retrieval rates.

4.4 Affective Coverage Limits Interpretation

Where language identification was limited mainly by short strings, the affective branch faced a broader coverage problem. Mean emotion-profile Jensen–Shannon distance was 0.3314 (95% CI 0.3285–0.3344), and mean valence distance was 0.2045 (95% CI 0.2022–0.2067). Both languages had nonzero emotion distributions in 68.81% of pairs. The Portuguese and English profiles were both all zero in 16.18% of pairs, which mechanically produced a distance of zero. Exactly one profile was zero in 15.02%, producing a large distance fixed by the implementation. Emotion-lexicon coverage was zero in 24.67% of Portuguese responses and 23.48% of English responses. Among the 30,356 pairs with nonzero emotion profiles in both languages, mean distance was 0.3274 (95% CI 0.3242–0.3305). The corresponding valence analysis retained 41,926 pairs with nonzero language-specific profiles and produced a mean distance of 0.1943 (95% CI 0.1923–0.1964). These conditional estimates avoid mechanical all-zero distances, but they remain dependent on the associated semantic networks and language-specific lexicons.

5 Discussion

The length and repetition analyses describe the distinguishing information available to retrieval. Retrieval rates rose with response length, and target grouping changed R@5 and MRR. The opposing directional shifts under the larger encoder show that recovery also depends on encoder choice. Interpreting individual failures would require bilingual assessment of translation quality alongside the retrieval results.

The disagreement between retrieval and COMETKiwi clarifies that the metrics answer different questions. Retrieval asks whether a translation points back to a matching source-text group within this corpus, whereas COMETKiwi estimates quality from learned source–translation relations. Positive-workplace responses received the highest mean COMETKiwi score but the lowest retrieval rate among the three prompt families, and the divergence was even stronger across response-length bands. This pattern is consistent with evidence that COMET-family behavior depends on implementation choices and that intrinsic metric scores need not preserve downstream distinctions (Zouhar et al., 2024; Moghe et al., 2023). The views should therefore remain separate, and neither should serve as a pass criterion without target-domain calibration.

The prompt-family comparison also places the mental-health findings in context. Responses elicited by mental-health and well-being prompts produced retrieval and COMETKiwi estimates between those of the other two families. The findings concern workplace surveys with a mental-health component. Prompt-family comparisons reflect differences in wording and response length. Prompt wording also varies across cohorts. Attributing differences to mental-health content would require comparisons that account for these factors. The methodological problem is nevertheless relevant to multilingual mental-health research because translated free text may later be used to infer themes, affect, or risk. Analysts must then distinguish

whether an audit tool measures translation behavior, response distinctiveness, lexicon coverage, or some combination of the three.

The affective results likewise show why coverage must be separated from preservation. EmoAtlas offers interpretable language-specific profiles, but roughly one quarter of responses in each language contained no covered emotion term. Restricting the analysis to two nonzero profiles removed the mechanically determined cases without resolving the dependence on language-specific lexicons and semantic networks. Very short text is already known to challenge Brazilian Portuguese emotion classification (Domingues Aparecido et al., 2025). Treating an all-zero profile as evidence of neutral affect would therefore be incorrect. The affective analysis is most defensible as a coverage diagnostic that identifies where lexical comparison is possible and where another method or human review is needed.

For datasets with similar constraints, these findings motivate a staged audit. Analysts can begin by preserving paired source and translated text alongside translation provenance, then report response length and repetition before interpreting retrieval. Reference-free quality estimates add a separate view, but their distributions should remain visible and threshold-free unless calibrated in the target domain. Language and affective analyses likewise require explicit coverage diagnostics. Human bilingual review would still be required when the intended claim concerns clinical meaning, cultural equivalence, or construct validity.

6 Data and Code Availability

The workplace comments remain confidential because free text may contain identifying details or confidential organizational information. The public artifact provides the analysis code, frozen configurations, and synthetic tests through the project repository. Readers can inspect the implementation and run the synthetic tests. Reproducing the empirical estimates requires authorized access to the survey data. Sharing additional aggregate outputs requires authorization from the data controller. The appendices record the prompt wording, immutable model revisions, software versions, and selected sensitivity estimates.

7 Conclusion

Reference-free tools characterize complementary aspects of Portuguese-to-English translations in this survey corpus. Retrieval is informative for distinctive longer responses but weak for brief or repeated answers, while COMETKiwi and lexical coverage expose different constraints. Reporting metric disagreement and coverage gaps keeps these interpretations explicit. Connecting the automated outputs to translation accuracy and construct equivalence requires human assessment and validation in the target domain.

8 Limitations

The evaluation relies on pretrained systems and automated criteria. Bilingual quality ratings or reference translations would be needed to calibrate their scores against human judgments. Claims about clinical, emotional, or cultural equivalence would also require construct-specific validation. Agreement among the automated measures may reflect shared language and domain biases.

The complete-case cohort excludes survey-specific respondent records without all three prompts, and the candidate count already required a nonempty archived English output. The study cannot assess whether incomplete records differ systematically from the retained records or estimate translation-service failure rates. The analyzed sample should therefore not be treated as representative of every survey respondent or every submitted Portuguese response.

Translation provenance is incomplete. Although the available implementation records identify an operational language-model workflow, the analytic database does not retain row-level provenance for models, prompts, timestamps, retries, or fallback routes. Results characterize the archived corpus rather than a named translation model. The source surveys also cannot link people across cohorts, so respondent counts are survey-specific instances.

The 4B and 8B encoders share the Qwen3 family, so their comparison assesses scale sensitivity within that family. Assessment across independently trained encoder families would require additional comparisons.

Retrieval is sensitive to corpus size, response repetition, and task construction. Queries received direction-specific instructions while documents did not, so cosine values and directional ranks are not symmetric equivalence measures. Subgroup results

use the full corpus as the candidate set and should not be read as within-group retrieval benchmarks. The language detector was restricted to three candidate languages, and its flags are unreliable for very short strings. Finally, EmoAtlas relies on language-specific lexicons and parsers without demonstrated measurement invariance for these survey responses.

9 Ethical Considerations

The study used previously collected workplace survey data administered by GPTW for internal assessment on behalf of the participating companies. The companies provided the data and authorized their secondary use for research after removal of direct identifiers. No new data were collected for this study, and the research team had no contact with respondents. Data handling followed study procedures established for compliance with Brazil's Lei Geral de Proteção de Dados Pessoais (LGPD).

The analyzed records use pseudonymous keys in place of direct respondent identifiers, and publication materials omit organization names. Free-text comments may still contain identifying details, so they were treated as confidential and are neither quoted nor released. Access was restricted to the research workflow, and the paper reports only aggregate, text-free outputs. No claim about individual mental health is produced, and the automated measures are not intended for individual-level decisions.

Acknowledgments

Generative AI tools were used to suggest alternatives for rewriting some sentences in a concise way and to correct spelling. All suggestions were manually verified by the authors.

References

Peter Boot. 2021. Machine-translated texts as an alternative to translated dictionaries for LIWC. Preprint.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2026. A survey on multilingual mental disorders detection from social media data. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pages 893–918. Association for Computational Linguistics.

Edward W. Chew, William D. Weisman, Jingying Huang, and Seth Frey. 2023. Machine translation for accessible multi-language text analysis. *Preprint*, arXiv:2301.08416.

Thales David Domingues Aparecido, Alexis Carrillo, Chico Q. Camargo, and Massimo Stella. 2025. Benchmarking psychological lexicons and large language models for emotion detection in brazilian portuguese. *AI*, 6(10):249.

Diana Paula Dudău and Florin Alin Sava. 2021. Performing multilingual analysis with linguistic inquiry and word count 2015 (LIWC2015): An equivalence study of four languages. *Frontiers in Psychology*, 12:570568.

William R. Hobbs and Anthony D. Ong. 2023. For living well, behaviors and circumstances matter just as much as psychological traits. *Proceedings of the National Academy of Sciences*, 120(12):e2212867120.

Shaoxiong Ji, Timothee Mickus, Vincent Segonne, and Jörg Tiedemann. 2024. Can machine translation bridge multilingual pretraining and cross-lingual transfer learning? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 2809–2818. ELRA and ICCL.

Joseph Kambeitz, Jason Schiffman, Lana Kambeitz-Ilankovic, Vijay A. Mittal, Ulrich Ettinger, and Kai Vogeley. 2025. The empirical structure of psychopathology is represented in large language models. *Nature Mental Health*, 3(12):1482–1492.

Dalibor Kučera and Matthias R. Mehl. 2022. Beyond english: Considering language and culture in psychological text analysis. *Frontiers in Psychology*, 13:819543.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta-evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13060–13078. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1074–1084. Association for Computational Linguistics.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Ana-Maria Bucur, Stevie Chancellor, and Marcos Zampieri. 2026. Large language models for mental health: A multilingual evaluation. *Preprint*, arXiv:2602.02440.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up

CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645. Association for Computational Linguistics.

Alfonso Semeraro, Salvatore Vilella, Riccardo Improta, Edoardo Sebastiano De Duro, Saif M. Mohammad, Giancarlo Ruffo, and Massimo Stella. 2025. EmoAtlas: An emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science. *Behavior Research Methods*, 57(2):77.

Peter M. Stahl. 2026. Lingua language detector. Python package. Accessed 21 August 2026.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354. International Committee on Computational Linguistics.

Leah Cathryn Windsor, James Grayson Cupit, and Alistair James Windsor. 2019. Automated content analysis across six languages. *PLOS ONE*, 14(11):e0224425.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, Yang An, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Jingshi Zhou, Fei Cheng, and Xiaojun Zhang. 2026. EMOS: A multidimensional emotional fidelity evaluation framework for machine translation quality assessment. *Cognitive Computation*, 18(1):65.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288. Association for Computational Linguistics.

A Cohort-Specific Prompt Wording

Table 3 reports every Portuguese cohort–prompt combination used to define the three prompt families in the GPTW survey cohorts. The placeholder {organização} replaces the organization name. Whitespace was normalized for presentation, but wording and punctuation were otherwise retained. The English renderings help readers understand the prompts and were not used as analytical inputs. Cohort labels remain generic because the survey source is confidential.

B Computational Provenance

Table 4 records the model and software versions that generated the reported automated outputs. Model weights were frozen and no component was fine-tuned on the survey corpus. The translation system is omitted because the archived data do not retain row-level model and version provenance.

B.1 Retrieval instructions

Portuguese-to-English queries used this instruction.

> Retrieve the English translation that preserves the meaning of this Brazilian Portuguese workplace response.

English-to-Portuguese queries used this instruction.

> Retrieve the Brazilian Portuguese source that has the same meaning as this English workplace response.

Both encoders used these instructions. Each query starts with `Instruct`, a colon and a space, then the instruction. A newline follows. The next line starts with `Query`, a colon and a space, then the response text. Candidate documents contain the response text alone.

C Sensitivity Analyses

Table 5 reports the principal sensitivity estimates. Target de-duplication retains all 44,118 queries while replacing repeated target rows with direction-specific normalized-text groups. Each target group receives the maximum similarity among its member rows. Encoder differences are paired within response pair and use the same survey-specific respondent-record bootstrap draws. The expected-language subset is based on an automated detector

Cohort	Family	Portuguese prompt	English rendering
A	Positive workplace	O que torna a sua {organização} um excelente lugar para trabalhar?	What makes your {organization} a great place to work?
A	Improvement	O que precisa ser melhorado na sua {organização}?	What needs to be improved at your {organization}?
A	Mental health and well-being	No dia-a-dia, que fatores interferem na construção de um ambiente psicologicamente e emocionalmente saudável para trabalhar?	In day-to-day work, which factors affect the creation of a psychologically and emotionally healthy work environment?
B	Positive workplace	O que torna a sua {organização} um excelente lugar para trabalhar?	What makes your {organization} a great place to work?
B	Improvement	O que precisa ser melhorado na sua {organização}?	What needs to be improved at your {organization}?
B	Mental health and well-being	Para você, o que é saúde mental?	What does mental health mean to you?
C	Positive workplace	O que torna a sua {organização} um excelente lugar para trabalhar?	What makes your {organization} a great place to work?
C	Improvement	O que precisa ser melhorado na sua {organização}?	What needs to be improved at your {organization}?
C	Mental health and well-being	O que te faz sentir acolhido(a) e seguro(a) para falar sobre saúde mental no ambiente de trabalho?	What makes you feel welcomed and safe when talking about mental health in the workplace?
D	Positive workplace	O que torna a sua {organização} um excelente lugar para trabalhar?	What makes your {organization} a great place to work?
D	Improvement	O que precisa ser melhorado na sua {organização}?	What needs to be improved at your {organization}?
D	Mental health and well-being	Quais fatores no ambiente de trabalho mais impactam seu bem-estar emocional e psicológico?	Which workplace factors have the greatest effect on your emotional and psychological well-being?

Table 3: The twelve Portuguese cohort–prompt combinations and explanatory English renderings.

Component	Identifier	Version or immutable revision
Primary encoder	`Qwen/Qwen3-Embedding-4B`	`5cf2132abc99cad020ac570b19d031efec650f2b`; bfloat16; normalized embeddings; model-default pooling
Scale sensitivity	`Qwen/Qwen3-Embedding-8B`	`1d8ad4ca9b3dd8059ad90a75d4983776a23d44af`; bfloat16; normalized embeddings; model-default pooling
Embedding runtime	Sentence Transformers; Transformers	5.7.0; 4.57.6
Quality estimator	`Unbabel/wmt23-cometkiwi-da-xl`	`33858b2239a139d497d9c74952c88b89a8c06213`; float16; Unbabel COMET 2.2.7
Affective profiles	EmoAtlas	`17b16d736e6935a63ccbbddeff739cd7baddb66c`
Language pipelines	`pt_core_news_lg`; `en_core_web_lg`	Both 3.8.0; spaCy 3.8.15
Language flags	Lingua	2.2.0; Portuguese, English, and Spanish candidates; minimum relative distance 0.05
Bootstrap	Percentile cluster bootstrap	2,000 repetitions; seed 20260820; survey-specific record clusters

Table 4: Model and software provenance.

Sensitivity	Direction or profile	n	Estimate	95% CI
Target-group R@1	EN→PT	44,118	.7180	[.7130, .7234]
Target-group R@1	PT→EN	44,118	.7337	[.7287, .7386]
Target-group R@5	EN→PT	44,118	.8067	[.8022, .8111]
Target-group R@5	PT→EN	44,118	.8349	[.8307, .8392]
Target-group MRR	EN→PT	44,118	.7607	[.7559, .7653]
Target-group MRR	PT→EN	44,118	.7802	[.7759, .7848]
8B minus 4B R@1	EN→PT	44,118	−.0237	[−.0268, −.0208]
8B minus 4B R@1	PT→EN	44,118	.0147	[.0119, .0176]
Expected-language R@1	EN→PT	41,925	.7470	[.7420, .7519]
Expected-language R@1	PT→EN	41,925	.7626	[.7578, .7672]
Expected-language COMETKiwi	Not applicable	41,925	.7431	[.7420, .7443]
Two nonzero profiles, Jensen–Shannon distance	Emotion	30,356	.3274	[.3242, .3305]
Two nonzero profiles, Jensen–Shannon distance	Valence	41,926	.1943	[.1923, .1964]

Table 5: Selected sensitivity estimates. All intervals use 2,000 bootstrap repetitions over the complete 14,706-record sampling universe. Target grouping left every per-query R@1 outcome unchanged, so those rows repeat the primary intervals rather than an independent bootstrap draw.

and is not language ground truth. Affective estimates conditioned on two nonzero profiles compare only the pairs for which both language-specific resources returned a distribution.