

EVALUATING LARGE LANGUAGE MODELS FOR CLINICAL TERMINOLOGY TRANSLATION IN LOW-RESOURCE LANGUAGES: EVIDENCE FROM ENGLISH-TO-SINDHI NEURODEVELOPMENTAL DISORDER TERMS

Aoun Ali^{*1}, Barkat Ali Dahri², Syed Hussain Ali Shah Jillani³

^{*1}Government College of Education Federal B Area, Block 15, Karachi, Pakistan

²Government College of Education Federal B Area, Block 15, Karachi, Pakistan

³Institute of Professional Psychology, Bahria University, Karachi, Pakistan

^{*1}aounaliwadho@gmail.com

DOI: <https://doi.org/10.5281/zenodo.22684171>

Keywords

large language models; technology assessment; Sindhi language; neurodevelopmental disorders; clinical terminology; low-resource languages; machine translation; healthcare technology management

Article History

Received: 01 June 2026

Accepted: 23 August 2026

Published: 10 September 2026

Copyright @Author

Corresponding Author: *

Aoun Ali

Abstract

Diagnostic labels for neurodevelopmental disorders exist in English but have no agreed Sindhi equivalents, yet large language models are increasingly used to fill that gap without systematic evaluation of their handling of clinical vocabulary. This study evaluated four widely available systems (ChatGPT, Gemini, Microsoft Copilot and DeepSeek) on eighteen DSM-5-TR neurodevelopmental disorder terms. Model output was compared with a benchmark constructed by three independent Sindhi-language experts, who translated each term separately before reconciling discrepancies through consensus. Eight metrics were computed: Exact Match, Normalized Levenshtein Similarity, chrF2, BLEU, Translation Edit Rate, Jaccard Similarity, Cosine Similarity and ROUGE-L. Differences across systems were assessed using Friedman's test with Holm-Bonferroni correction. Leadership was distributed rather than concentrated: Copilot led on chrF2, Jaccard Similarity and Cosine Similarity; ChatGPT on BLEU, Translation Edit Rate and ROUGE-L; Gemini on Normalized Levenshtein Similarity; DeepSeek on none. No system reproduced an expert translation exactly for any term. Of the seven testable metrics, only chrF2 differed significantly after correction ($\chi^2 = 14.33$, $df = 3$, $p = .002$, adjusted $p = .017$, Kendall's $W = 0.265$); post-hoc comparisons resolved no specific pair, and the composite ranking spanned only 0.35 of a rank, from ChatGPT (2.34) to DeepSeek (2.69). For healthcare organizations considering LLM-based translation, these systems can generate a usable first draft of specialized terminology in an under-resourced language, but none currently produces output that a clinical service could deploy without bilingual expert review.

1. INTRODUCTION

Conversational AI has quietly become a default translation tool in many healthcare and education settings without anyone formally approving it for that purpose. Healthcare technology managers normally vet tools before they reach clinical workflows, through usability testing, safety review and sometimes formal health technology assessment; translation by large language models has largely skipped that step, used because it is available rather than because anyone confirmed it works for the vocabulary in question.

The stakes depend on content. A mistranslated diagnostic term can misdirect a referral, confuse a care plan or leave a family misunderstanding a child's diagnosis, and communication failures linked to language barriers remain a documented driver of unequal access to care, with reviewers repeatedly urging caution before treating machine-translated clinical content as reliable (Dew et al., 2018; Khoong and Rodriguez, 2022; Zappatore and Ruggieri, 2024). Fluency does not settle the question: output can read smoothly in dozens of languages (Han et al., 2024) while distorting the underlying clinical concept, and the errors that matter are meaning slips a non-expert reader would not notice (Mehandru, Robertson and Salehi, 2022).

The risk compounds for languages outside the small set well served by translation research, where the data needed to build and evaluate systems is scarce (Haddow et al., 2022; Ranathunga et al., 2023); a 2025 review of 69 low-resource studies still found a persistent deficit relative to high-resource pairs (Tafa et al., 2025). Sindhi sits squarely in that category: spoken by tens of millions, but with none of the clinical glossaries, parallel corpora or validated terminology banks available for English or Mandarin, and identified as needing substantial further resource development before reliable machine translation becomes realistic (Mujadia and Sharma, 2023, 2024). Clinical terminology compounds the problem again, because terms like Autism Spectrum Disorder or Specific Learning Disorder are diagnostic categories with defined criteria

rather than ordinary vocabulary, and machine-translated medical content has been found to need substantial human correction before use (Kasperė, Mikelionienė and Venckienė, 2023).

Two distinct failure modes make this hard. Existing lexicographic resources rarely encode etymology or etiology: the Sindhi Language Authority's online dictionary renders Autism Spectrum Disorder as ڏهنی مونجه ڪميونڪيشن نه ٿيڻ ڪري, literally "due to a lack of mental communication," which asserts a supposed cause rather than naming the diagnostic construct (Online Sindhi Dictionaries, 2025). AI systems, meanwhile, disagree with one another on the same term: DeepSeek returns ڏهنی سڀيڪٽرم ڊسارڊر, Gemini returns ڏهنی اسپيڪٽرم خرابي, and ChatGPT and Copilot both return ڏهنی اسپيڪٽرم بيماري, while the expert consensus rendering used in this study is ڀاڻ پڻي جي تسلسل ۾ بي ترتيبي. The variants differ in whether the condition is framed as a transliterated borrowing, a defect or an illness – distinctions that carry clinical weight – and general-purpose engines tend toward simplified output that loses contextual nuance.

The consequences reach past lexicography. Without standardized clinical language, diagnosis is harder to deliver consistently and therapy harder to conduct, because clinicians cannot reliably explain constructs to patients and families cannot articulate what they observe. Training compounds the gap: clinical psychology programs in Pakistan include little formal terminology work, which risks inconsistent application of diagnostic standards (American Psychiatric Association, 2022). Nor is the problem confined to Sindhi – Urdu, Punjabi, Pashto and Balochi all lack adequate neurodevelopmental terminology, so the gap examined here is likely representative rather than exceptional.

No study has systematically tested this specific combination: clinical terminology translated into Sindhi and evaluated against an expert benchmark rather than one translator's opinion. That is the gap addressed here, framed practically – not only how good the translation is, but what a healthcare

organization would need to know before relying on it.

2. Background

2.1 Large language models in healthcare translation

Health systems have long used machine translation to stretch limited interpreter capacity, particularly out of hours (Dew et al., 2018; Khoong and Rodriguez, 2022), and large language models extend that promise by covering more languages with less dedicated engineering than statistical or rule-based systems required (Han et al., 2024). The recurring catch is that fluent output is not accurate output: clinicians interviewed by Mehandru et al. (2022) reported that a translation can sound entirely natural while carrying the wrong clinical meaning, which busy staff cannot detect without knowing the target language. A multi-criteria review of healthcare adoption reached the same conclusion from the technology-management side – useful for low-stakes content, but specialized clinical material still carries real safety risk even with modern neural systems (Zappatore and Ruggieri, 2024).

2.2 Low-resource languages in machine translation research

Translation quality has improved unevenly. Haddow et al. (2022) and Ranathunga et al. (2023) identify the same root cause: building or properly evaluating a translation system requires data that does not exist in quantity for most of the world's roughly seven thousand languages. Transfer learning, multilingual pretraining and synthetic data generation help without closing the gap, and Tafa et al. (2025) concluded from 69 studies published between 2020 and 2024 that low-resource quality still trails high-resource pairs. South Asian work tells the same story: Das et al. (2023) report gains from multilingual neural architectures for Indic-to-Indic translation, a 36-language subcontinental translation ecosystem identifies Sindhi as still requiring dedicated corpus development (Mujadia and Sharma, 2024), and a WMT24 submission achieved solid BLEU

gains for Assamese and Manipuri only by fine-tuning existing open-source models – a starting point Sindhi clinical translation does not have (Wei et al., 2024).

2.3 The distinct demands of clinical terminology

General-purpose translation tolerates flexibility, since several phrasings can convey roughly the same idea. Clinical terminology is designed to do the opposite – minimize ambiguity, so a diagnosis means the same thing to every clinician reading it – which makes errors costlier rather than merely more frequent. Kasperé et al. (2023) found machine-translated medical vocabulary in a low-resource setting regularly needed substantial revision rather than occasional polishing, and a scoping review of natural language processing applied to standardized medical terminology found that automated translation of controlled clinical vocabularies consistently required expert validation (Noll et al., 2023). The consequences are concrete: a diagnosis such as Intellectual Developmental Disorder shapes what services a child is entitled to and how a school plans support, so translating the label is closer to translating a legal definition than a vocabulary item.

2.4 Measuring translation quality: the case for multiple metrics

No single scoring method is accepted, particularly where more than one wording is reasonably correct. BLEU is the most commonly reported metric but relies on exact word overlap, penalizing a translation that is right in meaning and differently phrased; character-level metrics such as chrF2 reward partial overlap, which matters more in morphologically rich languages where one root takes many surface forms. Translation Edit Rate, Jaccard Similarity and Cosine Similarity add edit distance, vocabulary overlap and vector-space proximity respectively. Han et al. (2024) and Tafa et al. (2025) argue for using several together, since a system can look strong on one metric and unremarkable on another. Even a full slate has limits in a specialized domain: expert judgment

remains necessary for translations intended for clinical use, because no automated score captures whether practitioners would accept a translated term (Merx, Phillips and Suominen, 2024; Noll et al., 2025).

2.5 Gaps in the existing literature

Three strands of research – large language models, clinical terminology translation, and Sindhi computational resources – rarely intersect. Neither the healthcare translation literature nor the low-resource machine translation literature has examined neurodevelopmental terminology moving from English into Sindhi, and no validated benchmark exists for that task, so there is no way to test whether claims about multilingual AI capability hold for a language and domain this specific. This study benchmarks four widely deployed models against an expert-built Sindhi reference set using eight complementary metrics, read through a technology-assessment lens relevant to organizations weighing these tools for multilingual clinical or educational service delivery.

3. Method

3.1 Design

A quantitative comparative design was used: eighteen English source terms were translated by four large language models, scored against an expert-built Sindhi benchmark on eight metrics, and compared using Friedman's test. Multiple metrics were necessary because a translation can track the reference lexically while drifting semantically, or the reverse.

3.2 Term set construction

The eighteen terms were drawn from the Neurodevelopmental Disorders chapter of the DSM-5-TR (American Psychiatric Association, 2022), covering Intellectual Developmental Disorders, Communication Disorders, Autism Spectrum Disorder, Attention-Deficit/Hyperactivity Disorder, Specific Learning Disorder, Motor Disorders and Tic Disorders. Isolated terms were tested rather than sentences or

case narratives, because the diagnostic label is the foundation on which assessment reports, school plans and family conversations are built; an unreliable label propagates into every longer text containing it.

3.3 Reference benchmark construction

No validated Sindhi clinical glossary exists for this terminology, so the benchmark was built for the study. Three Sindhi-language experts in education, psychology and translation each translated all eighteen terms independently, then met to resolve disagreements and agree one consensus rendering per term, weighing conceptual accuracy, clarity and professional acceptability. Each expert's independent pre-consensus translation set and the verification record from the reconciliation stage were retained. The consensus reference for each of the eighteen terms, alongside the raw output of all four systems, is reproduced in full in Appendix A. Expert consensus is the most defensible available substitute for an institutional glossary and is the approach used in comparable terminology-translation work (Noll et al., 2025).

3.4 Systems evaluated

Four systems were tested: ChatGPT (OpenAI), Gemini (Google), Copilot (Microsoft) and DeepSeek (DeepSeek AI). All four were accessed through their publicly available interfaces. They were selected because they are the systems organizations are already using or considering – widely accessible, free or low-cost, and commonly used by staff without specialist translation training rather than experimental research systems.

3.5 Procedure

Every model received an identical instruction: translate the given term from English into Sindhi, with no additional context. Holding the prompt constant isolated differences in model capability from differences in how each system was asked. All four sets of translations were collected within the same short window so that no model's output reflected a version update the others had not received.

3.6 Metrics

Sindhi uses an extended Perso-Arabic alphabet, so all strings were normalized before scoring: NFC, removal of Arabic combining marks, tatweel and zero-width characters, and folding of the yeh and he allographs onto one code point each (U+064A/U+06CC, U+0647/U+06C1), because the same phoneme reaches the page as different code points depending on a model's keyboard locale, and short vowels are optional in running Sindhi. The phonemically contrastive pairs U+06A9/U+06AA and do-chashmi he U+06BE/U+0647 were deliberately not folded. Tokenization split on whitespace plus an explicit separator list including the forward slash, without which "attention-deficit/hyperactivity" is one token and every word-level metric reads it as a total miss.

Eight measures were then applied. Exact Match tests character identity of the normalized strings; Normalized Levenshtein Similarity is 1 minus edit distance over the longer string length; chrF2 is sacreBLEU CHRF (char order 6, word order 0, beta 2, whitespace excluded) on untokenized normalized text; BLEU is sacreBLEU sentence BLEU-4 with exponential smoothing and effective order on Sindhi word tokens, smoothing being necessary because segments run from two to eleven tokens; Translation Edit Rate (TER) is uncapped sacreBLEU TER on Sindhi word tokens, the one measure where lower is better; Jaccard Similarity is intersection over union of unique word types; ROUGE-L is the F1 of the longest common subsequence over Sindhi word tokens; and Cosine Similarity is character n-gram TF-IDF ($n = 2-4$, sublinear term frequency, L2-normalized) fitted on the 90-segment study corpus; it is a lexical-semantic proxy rather than a neural sentence embedding, because no widely used pretrained encoder lists Sindhi with published evaluation and an unvalidated one would yield irreproducible numbers.

3.7 Statistical analysis

Descriptive statistics, per-metric ranks and bias-corrected percentile bootstrap confidence

intervals (10,000 resamples, seed 20260830) were calculated for each model. Because the same eighteen terms were translated by all four systems and the data did not meet parametric assumptions, Friedman's test was run separately for each metric at $\alpha = .05$ with $n = 18$ blocks and $k = 4$ systems; effect sizes are reported as Kendall's W , calculated as χ^2 divided by $n(k - 1)$. Holm-Bonferroni correction was applied across the seven testable metrics. Exact Match is excluded: it returned 0.00 for every system on every term, so it has no rank variance, Friedman is undefined on it, and including it would inflate the family size and weaken every genuine test. Pairwise Wilcoxon signed-rank tests with Holm correction within each metric followed any significant Friedman result. No omnibus test was run on the averaged ranks, because the metrics are strongly intercorrelated and are not independent blocks. Analyses used Python 3.13.4 with scipy 1.18.1, sacreBLEU 2.5.1 and scikit-learn 1.9.0.

3.8 Ethical considerations

No human participants, patient records or identifiable personal data were involved. The source terminology is publicly available diagnostic classification, and the three language experts contributed voluntarily and consented to research use of their translations, so the study carried minimal ethical risk and required no formal ethics board review.

4. Results

4.1 Descriptive performance across metrics

Table 1 reports mean scores across the eighteen terms. Copilot led on chrF2 (23.24), Jaccard Similarity (0.095) and Cosine Similarity (0.271); ChatGPT on BLEU (6.90), Translation Edit Rate (88.48) and ROUGE-L F1 (0.155); Gemini on Normalized Levenshtein Similarity (0.445); DeepSeek led on none. Exact Match was 0.00 for all four models on all eighteen terms, a floor effect rather than a scoring artifact: no system reproduced the expert consensus character-for-character even after normalization.

Table 1. Mean translation performance of the four large language models across eighteen terms

Metric	DeepSeek	Gemini	ChatGPT	Copilot
Exact Match (%)	0.00	0.00	0.00	0.00
Normalized Levenshtein Similarity	0.402	0.445	0.429	0.431
chrF2	17.56	20.72	22.09	23.24
BLEU	4.95	4.89	6.90	6.29
TER	91.50	93.57	88.48	88.99
Jaccard Similarity	0.069	0.079	0.094	0.095
Cosine Similarity	0.209	0.254	0.250	0.271
ROUGE-L F1	0.110	0.121	0.155	0.146

Note: values are means across the eighteen terms. Higher is better on every metric except TER, where lower is better. chrF2, BLEU and TER are reported on a 0–100 scale; the remaining metrics on 0–1. Exact Match is 0.00 for every system on every term – a floor effect, not a scoring artifact.

4.2 Composite ranking by mean rank

Averaging each system's Friedman mean rank across the seven discriminating metrics gives a composite summary, lower being better: ChatGPT 2.34, Copilot 2.37, Gemini 2.60, DeepSeek 2.69 (Table 2). The spread is 0.35 of a rank across four systems, so the composite separates them very weakly and is not an overall verdict.

Table 2. Composite standing by mean rank across the seven discriminating metrics

Model	Mean rank	Overall rank
ChatGPT	2.34	1
Copilot	2.37	2
Gemini	2.60	3
DeepSeek	2.69	4

Note: values are Friedman mean ranks (1 = best on that metric) averaged across the seven metrics that discriminate between systems. TER is inverted before ranking so that a low rank always means better translation. Exact Match is excluded because all four systems tie at 2.50 on it by construction.

4.3 Inferential tests and effect sizes

Only chrF2 separated the four systems, $\chi^2(3) = 14.33$, $p = .002$, Holm-adjusted $p = .017$, Kendall's $W = 0.265$ – a small effect; the other six testable metrics returned W between 0.022 and 0.059, all negligible (Table 3). Pairwise Wilcoxon post-hoc tests on chrF2 found no significant pair after Holm correction within the metric, the closest being ChatGPT against DeepSeek ($p = .072$, $r = 0.59$): the systems differ as a group on character-

level overlap without any specific pair being resolvable at this sample size.

Table 3. Friedman test results by metric, with Holm-Bonferroni correction

Metric	χ^2	df	p	p (Holm)	W
Exact Match	0.00	3	1.000	excluded	0.000
Normalized Levenshtein Similarity	1.43	3	.698	1.000	0.027
chrF2	14.33	3	.002	.017	0.265
BLEU	1.94	3	.586	1.000	0.036
TER	1.58	3	.665	1.000	0.029
Jaccard Similarity	2.44	3	.486	1.000	0.045
Cosine Similarity	1.18	3	.759	1.000	0.022
ROUGE-L F1	3.16	3	.367	1.000	0.059

Note: $\alpha = .05$. Holm-Bonferroni correction is applied across the seven testable metrics; Exact Match is excluded because zero rank variance leaves Friedman undefined, and including it would inflate the family size and weaken every other test. W is Kendall's W, calculated as the metric's χ^2 divided by $n(k - 1)$; bands are $< .10$ negligible, $.10-.29$ small, $.30-.49$ moderate, $\geq .50$ large. Where a metric is not significant after correction, the reading is that eighteen terms cannot resolve a difference on it, not that the systems are equivalent.

4.4 Summary of findings

All four systems produced recognizable Sindhi renderings, but the descriptive leads are split across systems and only chrF2 separated them significantly, with no resolvable pairwise difference. No model matched the expert consensus exactly on any of the eighteen terms, which tempers any claim that the problem is solved.

5. Discussion

All four systems produced measurable lexical and character-level overlap with the expert benchmark, notable given how little dedicated Sindhi data they

are likely to have seen in training: general multilingual capability does transfer to languages it was not built for. But the leads are split rather than concentrated, only chrF2 separated the systems significantly, and no pairwise comparison survived correction. For a healthcare technology manager the honest reading is not that any system is best but that the four are too close to separate on this benchmark, character-level overlap excepted – and there Copilot scores highest without a margin eighteen terms can attribute to any specific rival.

The zero Exact Match rate is not straightforwardly a failure, since clinical terminology tolerates more than one acceptable rendering and verbatim agreement with a single consensus is a strict test. It does show that no system defaults to the terminology experts judged most appropriate, and for constructs tied to assessment criteria and service eligibility, a plausible Sindhi phrase is not interchangeable with the one clinicians use. That chrF2 alone separates the systems fits its sensitivity to partial structural overlap, which word-based metrics miss and which matters most in morphologically complex languages.

The most useful lesson is methodological: the eight metrics rank the same four systems differently, and none leads on more than three. A study reporting only BLEU would have named ChatGPT the winner; only chrF2, Copilot; only Levenshtein similarity, Gemini – precisely the risk of mistaking a narrow strength for overall superiority that low-resource translation research warns about (Han et al., 2024; Tafa et al., 2025). The likeliest explanation for the overall performance level is training-data scarcity: Sindhi clinical terminology is poorly represented in these models' training corpora and no standardized Sindhi medical glossary exists to have been absorbed indirectly, so a model works from general multilingual association rather than clinical knowledge – a pattern documented repeatedly across other low-resource pairs (Haddow et al., 2022; Ranathunga et al., 2023). Terminology accuracy and professional reliability accordingly remain unresolved even as fluency improves (Kasperè et al., 2023; Mehandru et al., 2022), and if isolated terminology cannot be translated reliably, full clinical narratives are unlikely to be easier.

The technology-management implication is that these tools are not ready as a standalone translation step for Sindhi neurodevelopmental terminology but are useful as first-draft generators, giving a bilingual reviewer a starting point rather than a blank page. The appropriate adoption model is human-in-the-loop – machine output corrected by a qualified Sindhi-speaking professional before it reaches a patient, student or family – consistent with recommendations for high-stakes settings generally (Khoong and Rodriguez, 2022; Mehandru et al., 2022). Future work could test terminology-aware prompting, retrieval-augmented approaches drawing on validated medical ontologies, or fine-tuning on available Sindhi medical text; the gaps found here may reflect insufficient domain adaptation rather than a hard architectural ceiling.

6. Conclusion

This study tested ChatGPT, Gemini, Copilot and DeepSeek on eighteen neurodevelopmental disorder terms translated from English into Sindhi, scored against an expert-consensus benchmark on eight metrics. All four systems demonstrated real but limited capability. The descriptive leads split across three of the four systems and only chrF2 separated them significantly; no pairwise comparison survived correction, and the composite ranking spans just 0.35 of a rank from first to last. Exact Match was zero throughout – a floor effect, not a measurement failure – and that result captures the state of the technology better than any single score: useful, but not yet reliable enough to remove human review.

For organizations weighing large language models for clinical or educational translation into under-resourced languages like Sindhi, the evidence supports draft generation under human review rather than autonomous deployment. That calculus may shift as systems evolve, but investment in validated terminology standards and Sindhi-specific evaluation infrastructure will matter more than waiting for general model improvements to solve the problem.

7. Strengths and limitations of the study

The study addresses a genuine gap: Sindhi clinical terminology translation has had essentially no systematic evaluation, and no prior study has compared multiple current large language models on this task. The benchmark rests on expert consensus rather than automated alignment, the eight-metric design avoids a one-sided read, and the neurodevelopmental focus ties the findings to a domain with real consequences for diagnosis, education planning and family communication.

Eighteen terms with four systems is a small design: at this sample size Friedman detects only large, consistent rank separations, which is exactly the pattern observed – one metric significant, six not, no pairwise comparison resolvable. The study therefore establishes the benchmark and method and estimates effect sizes for a larger run rather than settling which system performs best. The

three-expert reference is a consensus rendering rather than the correct one: every metric measures distance from that consensus, not from correctness. The independent pre-consensus translations were retained, so inter-rater agreement can be computed, but it is not reported here. Testing isolated terms also removes the context a model would use to disambiguate, so these scores are a lower bound on in-context performance. The evaluation used automated metrics without usability ratings from practicing clinicians, educators or caregivers; model versions move quickly, so the results are a snapshot; and the scope was one language pair and one clinical domain, so generalization should be cautious.

Declarations

Funding: No external funding was received for this study.

Conflict of interest: The authors declare no conflicts of interest.

Ethical approval: This study involved no human participants, patient records, biological materials or identifiable personal information, and therefore did not require formal ethical approval.

Informed consent: Not applicable in respect of patients, as the study involved no human participants, patient records or identifiable personal data. The three Sindhi-language experts who produced the reference translations participated voluntarily, consented to the research use of their translations, and were exposed to no risk of harm.

Data and materials availability:

The dataset, the eighteen source terms, each expert's independent pre-consensus translations and the consensus reference set, the verification record from the reconciliation stage, the raw model outputs, and the normalization and analysis scripts are available from the corresponding author on reasonable request.

Use of AI-assisted technologies: Generative AI tools were used to assist with manuscript language and clarity and to help write and check the normalization and analysis scripts. They were not

used for study design, data collection, interpretation of results or the drawing of conclusions. Specifically, portions of the manuscript were paraphrased and language-edited using generative AI tools to improve readability and consistency of expression. All intellectual content is the authors' own: the research question, study design, the eighteen source terms, the expert translations and consensus reference set, the model outputs, the statistical analysis, and the interpretation and conclusions were produced entirely by the authors. All citations and references were identified, retrieved and verified by the authors from the primary sources; no reference, quotation, statistic or data point was generated, suggested or supplied by an AI system. All AI-assisted output was reviewed and verified by the authors, who accept full responsibility for the content and integrity of the manuscript. Note that the four systems evaluated in this study are the object of the research and are described in Section 3.4.

Author contributions:

Aoun Ali conceived the study, designed the methodology, curated the data, conducted the formal analysis, and drafted the manuscript. Barkat Ali Dahri contributed to the study design and the interpretation of the findings, and reviewed and edited the manuscript. Syed Hussain Ali Shah Jillani contributed to the clinical framing of the terminology and the interpretation of the findings, and reviewed and edited the manuscript. All authors read and approved the final version.

REFERENCES

- American Psychiatric Association (2022) *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition, text revision, Washington, DC, American Psychiatric Association Publishing.
<https://doi.org/10.1176/appi.books.9780890425787>

- Das, S.B., Biradar, A., Mishra, T.K. and Patra, B.K. (2023) 'Improving multilingual neural machine translation system for Indic languages', *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 22, No. 6, Article 169. <https://doi.org/10.1145/3587932>
- Dew, K.N., Turner, A.M., Choi, Y.K., Bosold, A. and Kirchhoff, K. (2018) 'Development of machine translation technology for assisting health communication: a systematic review', *Journal of Biomedical Informatics*, Vol. 85, pp.56-67. <https://doi.org/10.1016/j.jbi.2018.07.018>
- Haddow, B., Bawden, R., Miceli Barone, A.V., Helcl, J. and Birch, A. (2022) 'Survey of low-resource machine translation', *Computational Linguistics*, Vol. 48, No. 3, pp.673-732. https://doi.org/10.1162/coli_a_00446
- Han, L., Gladkoff, S., Erofeev, G., Sorokina, I., Galiano, B. and Nenadic, G. (2024) 'Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning', *Frontiers in Digital Health*, Vol. 6, Article 1211564. <https://doi.org/10.3389/fdgth.2024.1211564>
- Kasperé, R., Mikelionienė, J. and Venckienė, D. (2023) 'Medical terminology issues: a feasibility study of machine translation in a low-resource language', *SKASE Journal of Translation and Interpretation*, Vol. 16, No. 2, pp.5-22. <https://doi.org/10.33542/JTI2023-2-2>
- Khoong, E.C. and Rodriguez, J.A. (2022) 'A research agenda for using machine translation in clinical medicine', *Journal of General Internal Medicine*, Vol. 37, No. 5, pp.1275-1277. <https://doi.org/10.1007/s11606-021-07164-y>
- Mehandru, N., Robertson, S. and Salehi, N. (2022) 'Reliable and safe use of machine translation in medical settings', *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, pp.2016-2025. <https://doi.org/10.1145/3531146.3533244>
- Merx, R., Phillips, C. and Suominen, H. (2024) 'Machine translation technology in health: a scoping review', *Studies in Health Technology and Informatics*, Vol. 318, pp.78-83. <https://doi.org/10.3233/SHTI240895>
- Mujadia, V. and Sharma, D.M. (2023) 'Towards speech to speech machine translation focusing on Indian languages', *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Dubrovnik, pp.161-168. <https://doi.org/10.18653/v1/2023.eacl-demo.19>
- Mujadia, V. and Sharma, D.M. (2024) 'BhashaVerse: translation ecosystem for Indian subcontinent languages', *arXiv:2412.04351*. <https://doi.org/10.48550/arXiv.2412.04351>
- Noll, R., Frischen, L.S., Boeker, M., Storf, H. and Schaaf, J. (2023) 'Machine translation of standardized medical terminology using natural language processing: a scoping review', *New Biotechnology*, Vol. 77, pp.120-129. <https://doi.org/10.1016/j.nbt.2023.08.004>

- Noll, R., Berger, A., Kieu, D., Mueller, T., Bohmann, F.O., Müller, A., Holtz, S., Stoffers, P., Hoehl, S., Guengoeze, O., Eckardt, J.-N., Storf, H. and Schaaf, J. (2025) 'Assessing GPT and DeepL for terminology translation in the medical domain: a comparative study on the Human Phenotype Ontology', *BMC Medical Informatics and Decision Making*, Vol. 25, Article 237. <https://doi.org/10.1186/s12911-025-03075-8>
- Online Sindhi Dictionaries (2025) 'Autism spectrum disorders', *Online Sindhi Dictionaries [online]*, Sindhi Language Authority, Hyderabad, <https://dic.sindhila.edu.pk/index.php?txtsrch=Autism%20spectrum%20disorders> (accessed 3 September 2026).
- Ranathunga, S., Lee, E.S.A., Skenduli, M.P., Shekhar, R., Alam, M. and Kaur, R. (2023) 'Neural machine translation for low-resource languages: a survey', *ACM Computing Surveys*, Vol. 55, No. 11, Article 229. <https://doi.org/10.1145/3567592>
- Tafa, T.O., Hashim, S.Z.M., Othman, M.S., Alhussian, H., Nasser, M., Abdulkadir, S.J., Huspi, S.H., Adeyemo, S.O. and Bena, Y.A. (2025) 'Machine translation performance for low-resource languages: a systematic literature review', *IEEE Access*, Vol. 13, pp.72486-72505. <https://doi.org/10.1109/ACCESS.2025.3562918>
- Wei, B., Zhen, J., Li, Z., Wu, Z., Wei, D., Guo, J., Rao, Z., Li, S., Luo, Y., Shang, H., Yang, J., Xie, Y. and Yang, H. (2024) 'Machine translation advancements of low-resource Indian languages by transfer learning', *arXiv:2409.15879*. <https://doi.org/10.48550/arXiv.2409.15879>
- Zappatore, M. and Ruggieri, G. (2024) 'Adopting machine translation in the healthcare sector: a methodological multi-criteria review', *Computer Speech and Language*, Vol. 84, Article 101582. <https://doi.org/10.1016/j.csl.2023.101582>