

A REPRODUCIBLE FRAMEWORK FOR EVALUATING ROMANIZED SINDHI OUTPUT CONSISTENCY IN CONVERSATIONAL LARGE LANGUAGE MODELS: PILOT EVIDENCE FROM REPEATED PROMPTING

Irum Naz Sodhar¹, Abdul Hafeez Buller^{*2}, Asadullah Laghari³, Umair Ramzan Sheikh⁴

^{1,4}Department of Information Technology, Faculty of Computing, Shaheed Benazir Bhutto University, Shaheed Benazirabad, Sindh, Pakistan.

^{*2}Department of Building and Architectural Engineering, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Sindh, Pakistan.

³Postgraduate Scholar, Chengdu University of Technology, Chengdu-China.

¹irumnaz@sbbusba.edu.pk, ^{*2}engrabdulhafeezbuller@gmail.com / ah.buller@quest.edu.pk, ³asadullahlaghari@gmail.com, ⁴umairsheikh@sbbusba.edu.pk

DOI: <https://doi.org/10.5281/zenodo.23387252>

Keywords

Romanized Sindhi, large language model evaluation, low-resource NLP, output consistency, reproducibility, transliteration

Article History

Received: 25 June 2026

Accepted: 16 September 2026

Published: 15 October 2026

Copyright @Author

Corresponding Author: *

Abdul Hafeez Buller

Abstract

The informal digital communication is in the form of Romanized Sindhi which is not well represented in the benchmarks of natural language processing. This study suggests an evaluation method which is clear and concise to check whether a conversational large language model retains Romanized Sindhi in repeated prompting or not and tests it on the only source record recovered, one instruction-free sentence, repeated 10 times. Coding was done with the use of non-exclusive indicators for Sindhi script, Romanized Sindhi, Urdu script, Romanized Urdu and English. The main outcome measure was retention of Romanized-Sindhi, defined as having any Romanized-Sindhi span in a response. Romanized Sindhi was used in 6 out of 10 responses (60%; Wilson 95% confidence interval 31.3%-83.2%) while Sindhi script was used in 7 responses. Urdu, Urdu Romanized and English were used once each. There were 5 different language/script combinations observed. The results indicate that there is variability in output format for the recorded interface condition, but not for semantic accuracy, learning or overall model performance. Limited interpretation due to single prompt, no model and decoding metadata, uncertain session independence and duplicated screenshot assets and no independent native-speaker annotation. The main difference is the inclusion of a reproducible evaluation specification that differentiates between script retention and semantic adequacy and retains the original of the evidence in an attached annexure. A multi-sentence benchmark, controlled model snapshots, machine-readable outputs and blinded annotation by qualified Sindhi speakers are required for a full-scale confirmatory study.

INTRODUCTION

Sindhi is an Indo-Aryan language that is mainly spoken in India and Pakistan. Despite the progress in corpora, part-of-speech annotation, embeddings and dependency parsing [1]-[3] the language remains under-resourced from a computational point of view. This trend has further been continued by recent studies that have proposed data augmentation for Sindhi text classification [15] and culturally sensitive instruction tuning for structured reasoning [16]. The Perso-Arabic script, an extension of the Arabic script, is used to write the standard Sindhi language, and Latin characters are used in informal digital communication in Pakistan. So, Romanized Sindhi is Sindhi linguistic content in Latin script and not a mixture of English and Sindhi.

Informal Romanization is very diverse. Spellings are often phonetic, idiosyncratic, platform dependent and the same lexeme can have several forms in Latin script. In the research of Romanized Sindhi, problems of transliteration, spacing, and punctuation and character-mapping have been reported [4, 5]. Naturalness, pronunciation transparency, conventionality, invertibility and input convenience are competing design objectives, as is observed in the work across the languages of south Asia [6].

Romanization aware datasets and modelling approaches provide a more general methodological base. Dakshina data set contains parallel data in native script for 12 languages of the South Asian region [7] and Aksharantar contains large-scale Indic transliteration data and multilingual transliteration models [17]. Later studies involved Romanized language identification [8], multilingual adaptation through Romanization [9] and context-aware transliteration [10] and identification under informal spellings [11]. These research studies have shown that script choice, language identity, transliteration quality and semantic adequacy are interrelated but are different targets of evaluation. The second source of complexity is from conversational large language models. If the desired task is not mentioned, the same surface input might elicit translation, transliteration,

explanation, continuation or paraphrase. Prompt templates can significantly alter the results of evaluations [12], synonymous or typographical perturbations can negatively affect the stability of the responses [24] and hosted systems can vary between repeated evaluations even if the settings are supposed to be deterministic [13]. A repeated-prompt experiment thus demands precise model metadata, controlled sessions, a predetermined task and a standardized coding protocol.

The procedure was referred to in the source manuscript as casual or casual learning, and the binary label resemblance was considered as the primary result. There's no justification for either label. No model training, fine tuning, intervention or cause-effect estimation was done. It was a process of multiple inferences through a conversational interface. Thus, repeated prompting, response variability, language/script presence and Romanized-Sindhi retention are used in the present manuscript.

1.1 Research gap, questions, and contribution

Existing Romanized Sindhi studies do not have a controlled trial level framework for auditing the retention of script in conversational model responses. The first research question is to know how many times the recorded interface kept the Romanized Sindhi in 10 submissions of the same sentence. The second research question is: What language/script categories emerged in these responses? The third research question is: What methodological controls are needed to make the pilot a repeatable comparative benchmark?

The contribution is in three parts. The second is that the study uses an undefined resemblance score instead of observable, auditable endpoints. Second, all the trial matrix, uncertainty interval, category combinations, and limits of evidence are provided, rather than just one headline percentage. Third, the study offers a confirmatory protocol and an extra provenance archive of all seven different response screenshots utilized in the source manuscript, as well as their trial mappings.

2. Related Work

2.1 Sindhi language resources

The resources of Sindhi NLP have grown slowly. Motlani [1] presented descriptions of early language-technology tools for Sindhi. Later, a large manually annotated part-of-speech corpus was released in the SiPOS benchmark [2] and dependency annotations and language-specific resources have been made available in the Universal Dependencies initiative [3]. These are useful to the standard-script processing, but do not remove the problems of informal Latin-script communication.

The growth and the scarcity of Sindhi resources have been confirmed by recent studies. DAUGSindhi evaluated augmentation techniques for Sindhi news classification [15] and Sindhi-Reasoning-Instruct suggested culturally-based instruction data for deductive, inductive and causal reasoning [16]. Task-specific data and explicit evaluation criteria and native-language quality assessment are relevant in both studies but neither can be used as a benchmark for repeated Romanized-Sindhi prompting.

2.2 Romanized Sindhi and South Asian Romanization

Communication rules, transliteration problems and downstream analysis have been studied in Romanized Sindhi. Character mapping inconsistencies, spelling variations, token boundaries and punctuation [4] [5] are known issues. In a later sentiment analysis research, a hybrid method was used for Romanized Sindhi text [14]. The advantages of parallel resources, representative test sets and sentence level context are clearly seen at the regional level in the works of Dakshina [7] and Aksharantar [17] and later the context-aware transliteration works [10]. Language-identification studies reveal that romanization takes away strong cues in script and causes confusion between related languages [8] [11].

2.3 Repeated prompting and evaluation reproducibility

The wording of prompts has been shown to be an important factor for the capability of the model that can be reported [12]. Moreover, the results of hosted-system experiments indicate that nominally equivalent runs may produce different results [13]. When task wording, model snapshot, system instructions, decoding controls, session state and annotation rules are documented, repeated prompting is therefore informative. All such controls cannot be recovered in the present pilot and analysis is limited to evidence provenance and bounded interpretation.

2.4 Evaluation taxonomy for Romanized-language responses

If more than one construct is labeled and evaluated under the same label in the Romanized language, the evaluation is often affected. Script Identification: Questions if a span is in Latin, Perso-Arabic, Devanagari or another script. Language identification is the question asking for the language of the span. There is a distinction between transliteration (an evaluation of the linguistic material translated between scripts) and translation (an evaluation of the meaning translated between languages). Semantic similarity, grammaticality, naturalness and task compliance are other constructs. Script choice alone cannot be used to reliably infer any of these dimensions.

The current structure is thus based on a hierarchical endpoint structure. The most apparent end point is Romanized-Sindhi retention which can be found from the given matrix. The five language/script indicators are secondary descriptive endpoints. The outcomes of semantic adequacy and grammatical quality are not measured, as the source record does not have machine-readable transcripts or independent native-speaker ratings. This hierarchy assists in preventing pushing the measured endpoint out to unsupported claims.

The reliability and reproducibility is another coat. Reliability is whether repeated runs, annotators or prompt formulations result in consistent measurements. Reproducibility is the capacity of

another research group to repeat the procedure with the prompt, model snapshot, parameters, and coding manual provided in the documentation. The original record does not contain sufficient information to allow for rerunning the same computational condition.

2.5 Recent multilingual and Indic LLM benchmarks

However, in recent benchmark design, there is a trend towards more comprehensive language coverage and explicit multilingual generation tasks. IndicGenBench is an evaluation platform for user-facing generation in 29 Indic languages and various scripts [18]. Belebele offers parallel reading comprehension tasks for 122 language variations [19] and MEGEVERSE tests a large number of models and tasks for 83 languages [20]. The benchmarks offer useful comparative frameworks, but do not offer a solution to the script variation and informal romanisation or task ambiguity for Romanised Sindhi.

Language fidelity is now a standalone evaluation goal as well. MultiQ determines whether or not the model answers for open-ended questions remain in the prompt language in a typologically diverse language set [21]. Further care must be taken with multilingual model-based evaluation as scoring bias and lower reliability have been observed when compared with human judgments in low-resource and non-Latin-script languages [22]. Therefore, native-speaker annotation plays an important role in the semantic adequacy and naturalness in a confirmatory study of Sindhi.

Holistic evaluation frameworks suggest the use of standardized scenarios, multiple metrics, transparent prompts, and public outputs, instead

of a single aggregate score [23]. E-Bench is also a good example of the need to test synonymous, colloquial, simplified and typographically perturbed prompts [24]. Multilingual evaluation research also distinguishes between the use of English as an interface language and understanding of the target language directly [25]. These studies together provide for separate task conditions, controlled prompt families, language-fidelity measures and complete run documentation.

3. Materials and Methods

3.1 Study design and source record

An exploratory repeated-prompt case-study design was employed in the investigation. One Romanized Sindhi sentence was input into a conversational large language model (LLM) through the web. The source document had a trial-coding table, 10 figure captions for the responses, and 7 various response-image assets that were utilized in several responses. The interface is similar to ChatGPT, but the model name and model version, access date, account settings, system message and decoding parameters are not preserved. The unit of analysis was one recorded response.

There are 8 tokens, 38 non-space characters, and 45 characters (with spaces) in the input. There is no occurrence of a numeral or punctuation mark. No explicit instructions for translation, transliteration, explanation, grammatical analysis or for an answer in the same script. The lack of specification of the task is thus a prominent design feature and a possible explanation of format variation.

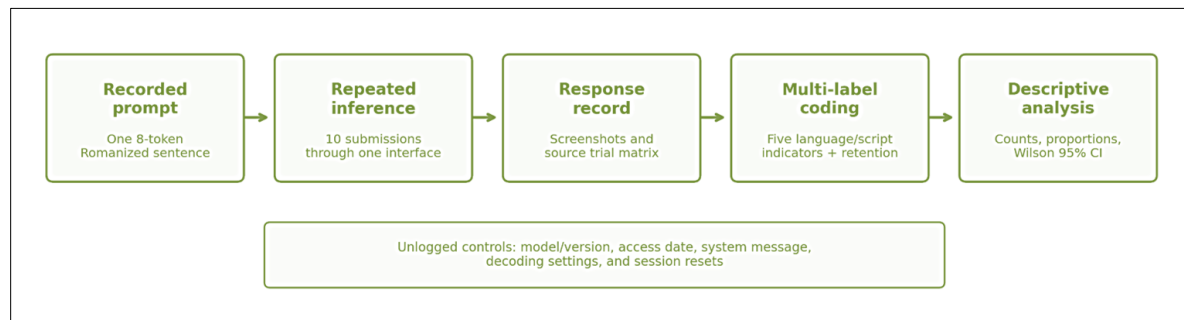


Figure 1. Auditable evaluation workflow reconstructed from the source record. The lower band identifies controls that were not logged.

Table 1. Reconstructed protocol and operational definitions

Field	Recorded or reconstructed specification
Design	Exploratory repeated-prompt case study
Unit of analysis	One generated response to the same Romanized Sindhi input
Trials	Ten recorded submissions in the order reported by the source manuscript
Primary endpoint	Romanized-Sindhi retention: Present when any Romanized Sindhi span was coded in the response; Absent otherwise
Secondary endpoints	Non-exclusive presence of Sindhi script, Romanized Sindhi, Urdu script, Romanized Urdu, and English
Quantitative source	The trial matrix reported in the supplied manuscript
Visual source	Seven unique response screenshots mapped to ten trial captions; duplicated assets are identified in Supplementary File S1
Unreported controls	Model/version, access date, system message, temperature, top-p, seed, output limit, session reset, and independent annotation

3.2 Coding framework

The source label resemblance was operationalized as Romanized-Sindhi retention as the source coding was based on the presence or absence of Romanized-Sindhi. The new endpoint is not semantically related. A semantically incorrect response could be a same-script response, and a native-script response could be a transliteration or translation. So, the notion of script retention and the semantic adequacy were conceptually separated.

Five language/script indicators were coded as binary (not mutually exclusive) variables. Responses may have more than one positive label. Sindhi script, Urdu script, Romanized Sindhi, Romanized Urdu and English were the script-language categories assigned in the source matrix.

No formal decision rules reported for mixed tokens, named entities or disputed spans in source document.

3.3 Statistical analysis

The count, percentage, combination of categories and ordered binary record is provided for the ten trials. The first ratio was $x/n = 6/10$. The two-sided Wilson 95% confidence interval was constructed, using $z = 1.96$ since the Wald interval is not good for small binary samples. The resulting interval was 31.3% - 83.2%. The record does not have a pre-specified comparator, one prompt, or one uncertain session independence; therefore, a hypothesis test or regression model or trend test was not used.

The order of the trials was not altered for transparency. The sequence of retained outputs from Trials 2 to 7 was called an observed sequence instead of a learning phase. Record can't identify stochastic decoding, prompt ambiguity, conversational context, server side behavior, or model updates as causal mechanisms.

3.4 Evidence-integrity audit

The quantitative reconstruction has been checked with the tables and the embedded relationships of images in the Word file provided. A total of seven different response-image resources were shared in 10 response captions identified in the audit. Trials 1, 8 and 9 were performed with the same asset and

Trials 2 and 3 were performed with a different asset. The Trial 9 caption also referred to Romanized Sindhi and the quantitative table indicated that Romanized Sindhi was not present. The trial matrix was retained as a quantitative source as it contained the entire 10 row coding record. All original mapping is kept on the supplementary annexure and discrepancies noted. There was no claim for re-annotation by an independent native-speaker. No machine-readable transcriptions were available so that no reliable re-analysis of the semantics or grammar was possible. Thus the present results should be regarded as a clear re-analysis of the record given, and not as a new model experiment.

(a) Sindhi-script response

Aseen dae rhya ahyoon udhara paisa tawhan khe

لڳي ٿو توهان چئي رهيا آهيو:
"اسين ڏئي رهيا آهيون اوڏر پنسا توهان کي" (يعني "اسان توهان کي قرض/اوڏر پنسا ڏئي رهيا آهيون").
ڇا توهان چاهيو ٿا ته مان ان جملو کي سنڌي ۾ صحيح لکڻ، ترجمو ڪرڻ يا ڪنهن خاص انداز ۾ ترتيب ڏيان؟

(b) Romanized-Sindhi response

Aseen dae rhya ahyoon udhara paisa tawhan khe

Lagaye tho ahyo ta tawhan udhar paisa wapas kanda chahyo ahyo 😊 — par ma samajhi na sagyo ta tawhan jo matlab chha? Tawhan chahyo tho ma likhan mein madad karyan (jaise message ya letter) je lae?

(c) Mixed language/script response

Aseen dae rhya ahyoon udhara paisa tawhan khe

Laghe tho ta tawhan Sindhi mein likhyo aa:
"Aseen dae rhya ahyoon udhara paisa tawhan khe"

Is jo matlab Urdu mein thindo:
| "ہم آپ کو ادھار پیسے دے رہے ہیں"
انگریزی ۾:
| "We are giving you money on credit."
چاهيو ته مان هي جملي کي ٿورو وڌيڪ رسمي يا ڪاروباري انداز ۾ به لکي ڏيان؟

Figure 2. Sample images of representative source responses: (a) Sindhi script response (b) Romanized-Sindhi response and (c) mixed response with Sindhi script, Romanized Sindhi, Urdu script, and English. The panels are about the variability of output formats, NOT about semantic correctness. The seven unique response assets and the complete mapping from trials to assets are provided in Supplementary File S1.

3.5 Reproducibility reporting framework

The source record was subjected to a reproducibility framework prior to interpretation. The framework includes four domains: input specification, system specification, execution protocol, and outcome measurement. Input specification is the particular instruction and task wanted. System specification consists of provider, model identifier, snapshot or access date and system instructions. The execution protocol is the decoding parameters, output limits, session boundaries, and trial order. Outcome measurement includes the raw responses, rules for annotation, annotator qualifications, adjudication and analysis code.

Each domain was rated as available, partially available or unavailable. Available item = can be recovered and used directly. Partially available item was not fully or distinctly presented, e.g., response screenshots. The item was not saved in the document provided, it is not available. The classification is stated in the Results section as it directly limits the claims that can be made based on the pilot data.

4. Results

4.1 Primary outcome: Romanized-Sindhi retention

Six respondents referred to Romanized Sindhi while four did not. The observed retention proportion was 60% (Wilson 95% CI 31.3%, 83.2%). The large gap is because of the small number of trials and doesn't represent performance at the population level for Romanized Sindhi prompts.

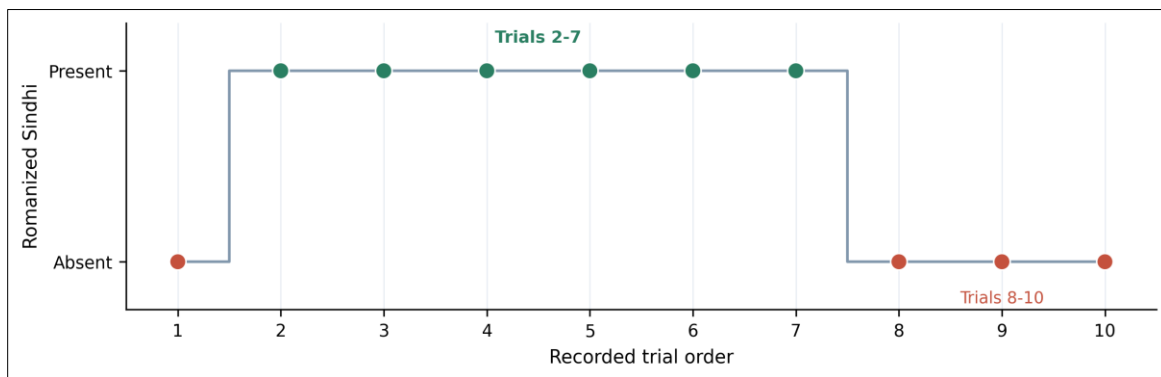


Figure 3. Romanized-Sindhi retention across the ten recorded trials.

Table 2. Language/script presence and Romanized-Sindhi retention by trial

Trial	Sindhi script	Romanized Sindhi	Urdu script	Romanized Urdu	English	RS retention
1	1	0	0	0	0	No
2	0	1	0	0	0	Yes
3	0	1	0	1	0	Yes
4	0	1	0	0	0	Yes
5	1	1	0	0	0	Yes

Trial	Sindhi script	Romanized Sindhi	Urdu script	Romanized Urdu	English	RS retention
6	1	1	1	0	1	Yes
7	1	1	0	0	0	Yes
8	1	0	0	0	0	No
9	1	0	0	0	0	No
10	1	0	0	0	0	No
Count	7	6	1	1	1	6 Yes

Note. A value of 1 indicates category presence. Categories are non-exclusive. RS = Romanized Sindhi.

4.2 Secondary outcomes: category presence and combinations

The largest number of responses was in Sindhi script, seven. Six of the responses were written in Romanized Sindhi. Urdu script, Romanized Urdu and English were used one time each. Mixed responses were classified as positive in more than

one category. Five combinations of the categories were mentioned: Sindhi script alone (4 trials), Romanized Sindhi alone (2 trials), Sindhi script and Romanized Sindhi (2 trials), Romanized Sindhi and Romanized Urdu (1 trial) and Sindhi script and Romanized Sindhi, Urdu script and English (1 trial).

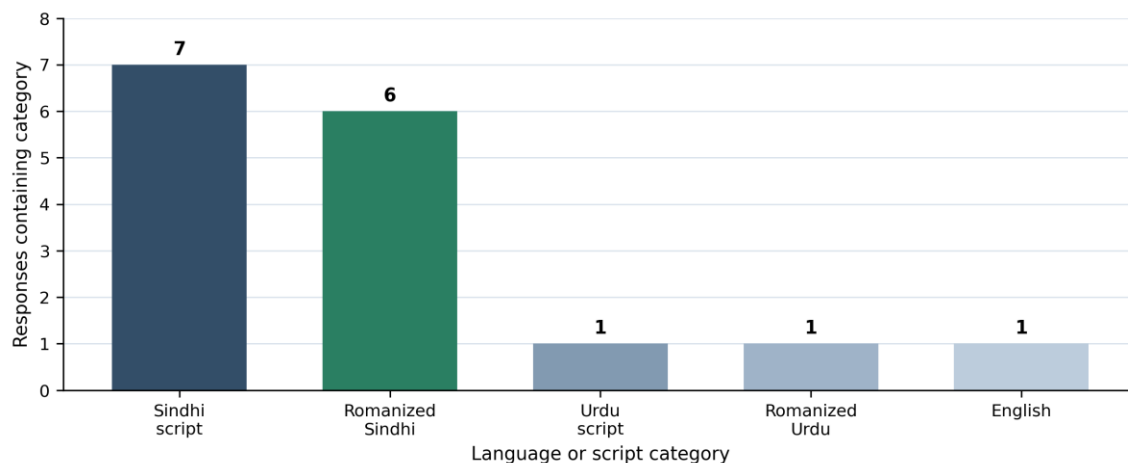


Figure 4. Frequency of language/script categories across ten responses.

Table 3. Aggregate descriptive outcomes

Outcome	Estimate	Interpretive boundary
Romanized-Sindhi retention	6/10 (60%)	Script-presence endpoint; not semantic accuracy
Wilson 95% confidence interval	31.3%-83.2%	Sampling interval only; independence remains uncertain
Sindhi-script presence	7/10 (70%)	Non-exclusive with other categories
Distinct category combinations	5	Describes observed heterogeneity in one prompt
Longest retention run	6 trials	Trials 2-7; no learning or phase claim

4.3 Reproducibility-completeness results

The audit on the reproducibility showed that the exact surface prompt, the number of trials, the order of trials, the source coding matrix and embedded visually assets were all accessible. Task, model-id, model-snapshot, access-date, system-message, decoding-controls, session-reset-

procedure, raw-text-outputs, annotation-manual, annotator-identities, and agreement-statistic were not included. The evidence from the screenshots was rated as partial because 10 captions were used for 7 different assets, one caption was repeated in the quantitative matrix.

Table 4. Reproducibility-completeness audit of the supplied record

Reporting item	Status	Consequence for interpretation
Exact surface prompt and trial count	Available	Supports reconstruction of the reported input condition
Trial matrix and category labels	Available	Supports descriptive re-analysis of the supplied coding
Response screenshots	Partial	Duplicated assets and one caption-table conflict limit verification
Requested task or output instruction	Unavailable	Task compliance cannot be defined
Model identifier, snapshot, and access date	Unavailable	The computational system cannot be reproduced
Decoding settings and session resets	Unavailable	Run independence and stochastic controls are unknown
Machine-readable output transcripts	Unavailable	Automated linguistic and semantic re-analysis is blocked
Annotation manual and independent raters	Unavailable	Measurement reliability cannot be estimated

Discussion**5.1 Interpretation of the observed variability**

The key finding is that the interface scored was not always Romanized Sindhi, which can be replicated in the record provided. Six responses were in Romanized Sindhi and four responses had completely shifted from the input Romanized form. The 60% is a script-retention rate, not an accuracy score. The choice of the script cannot determine the understanding, the quality of translation, the grammaticality of the translation and the semantic faithfulness.

The first and most apparent is the prompt ambiguity. Only a sentence was provided as input, no task directive. In a conversational system,

transliteration to Sindhi script, translation, explanation, paraphrase and grammatical analysis or continuation are all reasonable inferences. The five combinations of categories observed thus are not a pure linguistic capability but a combination of the under specification of the task and the behaviour of the system.

There may also be variability in the system level. Prompt-template changes [12] and hosted-system variation [13] can alter the results of the evaluation. The present evidence does not differentiate between those mechanisms, because metadata of the models, the decoding settings, session boundaries and timestamps were not

captured. The series of successes from Trials 2 to 7 are therefore a descriptive string.

5.2 Implications for Romanized Sindhi evaluation

Four dimensions should be distinguished in the Romanized Sindhi script: task compliance, output script, semantic adequacy and linguistic quality. If a validated reference is available, then character error rate / word error rate can be used for transliteration tasks. Semantic equivalence and naturalness require qualified human judgment, language identification is only achieved using explicit label rules and agreement rules. Research on Romanization of South Asian languages has revealed that the spelling variation is not a hindrance, but an inherent part of the task which needs to be assessed contextually [6, 10, 17]. Multilingual benchmark research also indicates that language fidelity and task level reporting should be made explicit, rather than being a single undifferentiated score [18]-[21].

The high frequency of Sindhi script may be a trend towards normalizing Latin-script input in native script, but this is based on one prompt only and cannot be a general trend. No information on spelling variants, morphology, code switching, named entities, negation, politeness, tense, question forms or sentence length in the source sentence. There is a need for a broader benchmark before model level claims can be supported.

5.3 Confirmatory protocol for full-scale evidence

A confirmatory study needs to be done using a set of several hundred Romanized Sindhi sentences, validated by native speakers, and the sample size worked out on the basis of precision. Some common domains of everyday life, named entities, negation, tense, declarative and interrogative sentences, code switching, short and long utterances, and some frequent spelling variations should be included in the set. Different task conditions should be asked for same script responses, transliteration to Sindhi script, translation and linguistic analysis.

Families of models and individual snapshots of models at a specific date should be considered for

each condition. The provider, the model ID, the access date, the system message, the full prompt, the temperature, top-p (if available), seed (if available), max output length, and the procedure for session-reset should be included in each run. Fresh sessions should be used for independent trials. Raw prompts and outputs should be represented as fixed identifier UTF-8 text.

Language, script, compliance of the task, semantic adequacy and grammaticality should be independently assessed by at least two qualified native Sindhi annotators. The labels should include an annotation manual, pilot calibration and adjudication protocol, and agreement statistic. Model-based evaluators should only be used following calibration in the language by human evaluators [22]. Pre-specified primary and secondary outcomes should be defined and clustering by sentence and by model should be accounted for, with the reporting of confidence intervals for each model-level estimate. The provision of prompts, outputs, annotations and analysis code would boost the reproducibility in a large way as per the transparent multi-metric evaluation practice [23].

5.4 Relation to prior research

These category variations are similar to the problem that Romanized-language studies have found: Latin-script input diminishes script-based language cues and allows for multiple plausible spellings [8], [11]. Studying context-aware transliteration also shows that sentence context plays an important role in determining the potential transliterations [10]. The methodological concern is also transferred to the assessment of conversational models in the current pilot and it is shown that the underspecified input can also vary in terms of requested task and output script.

The results are also in line with the multi-prompt and hosted-system evaluation research. Mizrahi et al. [12] demonstrated that the promptness of the form can impact the reported performance, Zhang et al. [24] demonstrated that stability can be measured under natural prompt perturbations, and Atil et al. [13] demonstrated that output variation can be measured in hosted systems. Another design consideration is the difference

between understanding in the target language and interfaces in English [25]. This control doesn't exist in the current record, but it's an example of why control is important for a low-resource language. If a task directive is not supplied and the model record is not fixed, output diversity cannot be clearly attributed to linguistic weakness or stochastic generation.

The most important methodological development is the distinction between observation and inference. The Trial Matrix is used to make statements about the presence of recorded categories. The matrix is not appropriate for a causal explanation, for an estimate of accuracy, or for a comparison of model families. This is particularly important in the context of low-resource-language studies, where small data sets and limited annotation capacity can result in over-interpretation of descriptive results.

5.5 Scope of contribution

The present article is considered to be a methodological pilot, brief report or research note. The novelty is not due to a big benchmark result, but to the evaluation framework, evidence audit, and low-resource-language focus. In a normal, full-length comparative article, the confirmatory experiment above would be required. The lack of observations, model metadata or independent annotation cannot be made up for by the editorial improvement.

6. Threats to Validity and Reproducibility

The Romanized-Sindhi retention is not a test of semantic correctness, which limits the construct validity. There are no controlled prompts, decoding settings, session resets or times. Small number of observations (10) and independence is doubtful. One sentence, one undocumented interface condition and no model comparison limit external validity. The factors that restrict the reliability of measurement are: source supplied labels, duplicated screen shot assets, discrepancy in caption table, and lack of inter annotator agreement.

The Wilson interval is only an approximation to the binomial sampling uncertainty. The interval

does not account for design bias or lack of independence. No causal mechanism, accuracy claim, model ranking or population level generalization is supported. The labels provided are currently not validated linguistic annotations, but will need a full raw transcript and native speaker audio check.

An extra threat is evidence provenance. The screenshot resources are reused and therefore a one-to-one visual verification of the ten coded trials is not possible and the caption for Trial 9 is in conflict with the reported matrix. The present study helps to reduce that threat by identifying the quantitative source, retaining all original mapping in Supplementary File S1, and documenting the discrepancy instead of taking a screen shot-based interpretation for granted. The mitigation improves transparency but not the uncertainty.

7. Conclusion

Ten contributions were made, of one Romanized Sindhi sentence, using different combinations of language and script. Six responses were in Romanized Sindhi and seven responses were in Sindhi script. So, the evaluated interface could not always keep the script entered under the recorded condition. The evidence does not evaluate the semantic correctness, nor can it be used to draw conclusions about learning or the general capabilities of large language models.

The updated framework enhances scientific defensibility with observable endpoints, uncertainty reporting, evidence provenance auditing, and a separation of pilot findings from confirmatory requirements. Next step in research is a multi-sentence, multi-model, independently annotated benchmark with controlled inference settings and open data.

Declarations

Acknowledgment. Acknowledgments: This research project titled Romanization of Sindhi text (SRSP 2024-2025) has been funded by the Sindh Higher Education Commission (SHEC), Ka-ra-chi under Sindh Research Support Program (SRSP) with project number 482. The authors are thankful to SHEC for financial support and to

Shaheed Benazir Bhutto University, Shaheed Benazirabad (SBBUSBA), Sindh, Pakistan for institutional support.

Ethics. The source study reports no human-participant, personal-data, or animal component. Data availability. Supplementary File S1 contains the seven unique response screenshots recovered from the supplied source manuscript, together with the complete trial-to-asset provenance map and discrepancy notes. Machine-readable response transcripts and complete model metadata were not available.

Conflict of interest. A completed declaration from every listed author is required before submission. Author contributions. Contributor roles and corresponding-author details must be completed by the author group before submission.

Supplementary material. Supplementary File S1 preserves each unique response screenshot once, identifies every source trial mapping, and records the caption-table discrepancies. Duplicate displays, superseded methodological graphics, and unsupported legacy dashboards are excluded from the curated file.

REFERENCES

- [1] R. Motlani, "Developing language technology tools and resources for a resource-poor language: Sindhi," in Proc. NAACL Student Research Workshop, 2016, pp. 51-58. doi: 10.18653/v1/N16-2008
- [2] W. Ali, Z. Xu, and J. Kumar, "SiPOS: A benchmark dataset for Sindhi part-of-speech tagging," in Proc. Student Research Workshop Associated with RANLP 2021, 2021, pp. 22-30. Available: ACL Anthology record
- [3] J. Bauer et al., "Universal Dependencies for Sindhi," in Proc. Eighth Workshop on Universal Dependencies, 2025, pp. 105-118. Available: ACL Anthology record
- [4] I. N. Sodhar, A. H. Jalbani, M. I. Channa, and D. N. Hakro, "Identification of issues and challenges in Romanized Sindhi text," International Journal of Advanced Computer Science and Applications, vol. 10, no. 9, pp. 229-233, 2019. doi: 10.14569/IJACSA.2019.0100929
- [5] I. N. Sodhar, A. H. Jalbani, M. I. Channa, and D. N. Hakro, "Romanized Sindhi rules for text communication," Mehran University Research Journal of Engineering and Technology, vol. 40, no. 2, pp. 298-304, 2021. doi: 10.22581/muet1982.2102.04
- [6] I. Demirşahin, C. Johny, A. Gutkin, and B. Roark, "Criteria for useful automatic romanization in South Asian languages," in Proc. 13th Conference on Language Resources and Evaluation, 2022, pp. 6662-6673. Available: ACL Anthology record
- [7] B. Roark et al., "Processing South Asian languages written in the Latin script: The Dakshina dataset," in Proc. 12th Language Resources and Evaluation Conference, 2020, pp. 2413-2423. Available: ACL Anthology record
- [8] Y. Madhani, M. M. Khapra, and A. Kunchukuttan, "Bhasa-Abhijnaanam: Native-script and romanized language identification for 22 Indic languages," in Proc. 61st Annual Meeting of the Association for Computational Linguistics, vol. 2, 2023, pp. 816-826. doi: 10.18653/v1/2023.acl-short.71
- [9] S. Purkayastha, S. Ruder, J. Pfeiffer, I. Gurevych, and I. Vulić, "Romanization-based large-scale adaptation of multilingual language models," in Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 7996-8005. doi: 10.18653/v1/2023.findings-emnlp.538
- [10] C. Kirov, C. Johny, A. Katanova, A. Gutkin, and B. Roark, "Context-aware transliteration of Romanized South Asian languages," Computational Linguistics, vol. 50, no. 2, pp. 475-534, 2024. doi: 10.1162/coli_a_00510

- [11] A. Benton, A. Gutkin, C. Kirov, and B. Roark, "Improving informally Romanized language identification," in Proc. 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 2318-2336. doi: 10.18653/v1/2025.emnlp-main.117
- [12] M. Mizrahi, G. Kaplan, D. Malkin, R. Dror, D. Shahaf, and G. Stanovsky, "State of what art? A call for multi-prompt LLM evaluation," Transactions of the Association for Computational Linguistics, vol. 12, pp. 933-949, 2024. doi: 10.1162/tacl_a_00681
- [13] B. Atıl et al., "Non-determinism of deterministic LLM system settings in hosted environments," in Proc. 5th Workshop on Evaluation and Comparison of NLP Systems, 2025, pp. 135-148. doi: 10.18653/v1/2025.eval4nlp-1.12
- [14] I. N. Sodhar, S. Sulaiman, A. H. Buller, and A. N. Sodhar, "Hybrid approach used to analyze the sentiments of Romanized text (Sindhi)," International Journal of Advanced Computer Science and Applications, vol. 14, no. 3, pp. 538-544, 2023. doi: 10.14569/IJACSA.2023.0140362
- [15] R. Vavekanand, B. Das, and T. Kumar, "DAugSindhi: A data augmentation approach for enhancing Sindhi language text classification," Discover Data, vol. 3, Art. no. 22, 2025. doi: 10.1007/s44248-025-00040-8
- [16] M. Mehak, K. Zeinalipour, P. Soomro, C. Chesi, M. Gori, and M. Maggini, "Enabling structured reasoning in Sindhi with culturally grounded instruction tuning," in Proc. 2nd Workshop on Language Models for Low-Resource Languages, 2026, pp. 239-258. doi: 10.18653/v1/2026.loreslm-1.22
- [17] Y. Madhani et al., "Aksharantar: Open Indic-language transliteration datasets and models for the next billion users," in Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 40-57. doi: 10.18653/v1/2023.findings-emnlp.4
- [18] H. Singh, N. Gupta, S. Bharadwaj, D. Tewari, and P. Talukdar, "IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages," in Proc. 62nd Annual Meeting of the Association for Computational Linguistics, vol. 1, 2024, pp. 11047-11073. doi: 10.18653/v1/2024.acl-long.595
- [19] L. Bandarkar et al., "The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants," in Proc. 62nd Annual Meeting of the Association for Computational Linguistics, vol. 1, 2024, pp. 749-775. doi: 10.18653/v1/2024.acl-long.44
- [20] S. Ahuja et al., "MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks," in Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2024, pp. 2598-2637. doi: 10.18653/v1/2024.naacl-long.143
- [21] C. Holtermann, P. Rottger, T. Dill, and A. Lauscher, "Evaluating the elementary multilingual capabilities of large language models with MultiQ," in Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 4476-4494. doi: 10.18653/v1/2024.findings-acl.265
- [22] R. Hada et al., "Are large language model-based evaluators the solution to scaling up multilingual evaluation?" in Findings of the Association for Computational Linguistics: EACL 2024, 2024, pp. 1051-1070. doi: 10.18653/v1/2024.findings-eacl.71

- [23] P. Liang et al., "Holistic evaluation of language models," Transactions on Machine Learning Research, 2023. Available: OpenReview record
- [24] Z. Zhang, B. Hao, J. Li, Z. Zhang, and D. Zhao, "E-Bench: Towards evaluating the ease-of-use of large language models," in Proc. 31st International Conference on Computational Linguistics, 2025, pp. 2329-2339. Available: ACL Anthology record
- [25] W. Poelman and M. de Lhoneux, "The roles of English in evaluating multilingual language models," in Proc. Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies, 2025, pp. 492-498. Available: ACL Anthology record



International Journal of
Advanced Research