



CORRECTING ROMAN URDU-ENGLISH ONLINE: A REFLEXIVE CASE STUDY OF ONE AI-ASSISTED SESSION

Hafiz Muhammad Sikandar

College of International Studies, Southwest University, Chongqing, China
<https://orcid.org/0009-0009-8305-7873>

Abbas Jihad

College of International Studies, Southwest University, Chongqing, China
<https://orcid.org/0009-0006-7348-8862>

Sohaib Ullah

Center for Studies of Education and Psychology of Ethnic Minorities in Southwest China,
Southwest University, Chongqing, China
<https://orcid.org/0009-0004-6892-1858>

ARTICLE INFO



Open access

JEL Category:
O33, Z13

Keywords:

Roman Urdu
translanguaging
digital discourse
generative AI
language ideology
Pakistan

ABSTRACT

AI writing assistants explain multilingual digital writing, but explanation and correction are distinct operations, a distinction underexamined in evaluations of generative AI and vernacular online language. This study examines whether one AI-assisted session treats semantic explanation and interactional preservation as separate tasks in Roman Urdu-English material. The data comprise sixteen source-verified textual units from six public Reddit threads, analyzed as a reflexive qualitative case study of one recorded ChatGPT Work session. Six correction operations, from orthographic normalization to monolingualization, are compared with observable interactional features such as address terms, quoted voice, idiom, and humor cues. Two authors coded every unit under deliberately contrasting criteria, and their disagreements are reported instead of being collapsed into a reliability score. The session generally recovered propositional content but treated interactional form unevenly: some English renderings retained jokes, quotations, or advice sequences, while others removed idiom, cadence, or locally meaningful address, and mixed-language corrections sometimes altered social positioning. These findings describe one conversation and establish neither a stable corrective default nor a causal effect of correction. Multilingual evaluation of AI writing tools should therefore distinguish semantic explanation, corrective operation, and feature-level transformation.

Address of the corresponding author:

Sohaib Ullah

sohaibullah09@gmail.com

Received: 01.10.2026

Revised: 12.10.2026

Accepted: 17.10.2026

Available online: 18.10.2026



1 INTRODUCTION

A writing assistant received the line “system down hai, baad mai ayen,” explained it as a request to return later, and revised it to “System down hai, baad mein aayen.” In the same recorded session, it rendered a longer workplace reprimand entirely in English, replacing the Urdu idiom *kursi garam karna* with “sitting here doing nothing.” Both responses were intelligible. They differed in what they treated as editable and in how much of the source interaction remained available in the revised text.

This difference cannot be reduced to grammatical accuracy. Roman Urdu is not a standardized orthography, and English material in Pakistani online discourse may operate as a borrowing, technical term, quotation, stance marker, clause, or longer passage. A request to “correct” such writing leaves several objectives unresolved. The user may want spelling normalization within Roman Urdu, translation into English, adaptation for a new audience, grammatical editing of English segments, or minimal intervention that preserves an online voice. A system that chooses among these objectives does not necessarily reveal a stable internal ideology. It produces one response under a particular prompt, interface, conversation history, and system configuration.

The initial version of this study described 24 units as if they formed a small audit of a model’s generic corrective policy. That description was not defensible. Three purported source texts had been materially rewritten, the model responses were generated sequentially inside the same research conversation, the units were clustered within threads, and the coding made language choice too closely determine pragmatic retention. These defects require a change of research object. The present article does not estimate model behavior. It examines a documented conversational trajectory and methodological problems that become visible when semantic explanation is mistaken for preservation of interactional form.

Sixteen source-verified units from six public Reddit threads remain in the revised corpus. They include posts, comments, replies, quoted speech, recipes, advice, and narrative codas. The analysis

treats each thread as a case and each textual unit as an embedded segment. It asks two questions:

RQ1: How do the verified Roman Urdu-English units organize stance, address, quotation, humor, expertise, and institutional voice within their thread positions?

RQ2: How did the recorded AI session explain and reformulate those units, and which observable interactional features were retained, altered, omitted, or left uncertain?

The claim is deliberately limited. In this archive, propositional explanation and corrective transformation were separable. An English rendering could retain a joke or advice sequence, while a mixed-language revision could still alter gender marking, interpersonal force, or the distribution of voices. That finding is methodological rather than causal. It supports feature-level analysis and controlled replication, not a general conclusion about ChatGPT, generative AI, or what students learn from automated feedback. Throughout, the article keeps apart three kinds of statements: transformations that can be observed by comparing a source string with its recorded correction; the authors’ interpretations of what those transformations do to the interaction; and hypotheses that only a controlled design could test (Table 4).

2 LITERATURE REVIEW

2.1 Roman Urdu as a digitally mediated practice

Roman Urdu is Urdu represented in Latin script, but that description does not identify a single orthographic system. Spellings vary with pronunciation, region, schooling, keyboard practice, platform convention, and individual history. Bilal et al. (2019), working with text messages from Pakistani university students, documented adaptation to available mobile technologies and a preference for Roman Urdu in everyday textual interaction. Computational studies likewise treat spelling variation as a central property of Roman Urdu data rather than as incidental noise. Ameer et al. (2022), for example, built an English Roman Urdu code-mixed emotion corpus from SMS messages, while Saeed et al. (2021) developed a large Roman Urdu toxic-

comment resource. These projects establish the scale and technical difficulty of Roman Urdu processing. Their task designs, however, usually require stable labels for language identity, sentiment, emotion, or toxicity. Such labels do not by themselves explain what a particular alternation accomplishes in an exchange.

Pakistan's broader language order matters because resources do not circulate with equal institutional value. Rahman (1996) traces how Urdu and English became connected to state formation, education, administration, and social differentiation. Shamim (2017) shows that English is repeatedly positioned as a language of development and mobility even as access to it remains unequal. These histories do not license a direct inference from any English word to class privilege. They do show why English, Urdu, and regional languages cannot be treated as neutral alternatives. A word such as system, board, teacher, or payment may be an ordinary element of Pakistani Urdu usage, a domain-linked borrowing, or a resource that helps stage institutional authority. Its force must be reconstructed from sequence, co-text, speaker relation, and uptake.

The Roman script adds another layer. It may widen access for users accustomed to Latin keyboards while excluding readers who rely on Perso-Arabic Urdu. It can index informality, technological convenience, youth practice, or diasporic participation, but none of those associations is automatic. A mixed Reddit comment is therefore better approached as a situated digital literacy event than as a defective version of a named language. The multiliteracies framework locates meaning in the design of available semiotic resources across media and communities (New London Group, 1996). On Reddit, those resources include reply position, capitalization, emoji, quotation, punctuation, English platform idioms, Romanized Urdu, and the visible organization of a thread.

2.2 Alternation, borrowing, translanguaging, and indexicality

Code-switching research has long shown that alternation can contextualize footing, quotation, alignment, topic, and participant relation (Auer, 1998; Gumperz, 1982). Yet "switch" can imply that analysts already know where one bounded code

ends and another begins. Translanguaging theory contests that assumption by distinguishing a speaker's repertoire from named-language boundaries produced through political and institutional histories (Otheguy et al., 2015). Li Wei (2018) treats translanguaging as a practical theory concerned with how multilingual speakers assemble resources to make meaning and social space. The two traditions need not be collapsed. Sequential code-switching analysis is useful when participants orient to a contrast; translanguaging is useful when rigid language partitioning obscures repertoire-based practice.

This article therefore avoids coding every English-looking token as a switch. It uses alternation descriptively and records the form involved: an address term, discourse marker, technical borrowing, clause, quoted voice, platform expression, or English-dominant passage containing Roman Urdu. The distinction is empirical. System down in a service formula differs from the English reassurance "I have huge hopes for you" following a mock Urdu threat. The former helps voice an institutional explanation. The latter changes the affective trajectory of a teasing exchange.

Indexicality clarifies why lexical retention is not identical to pragmatic retention. Linguistic forms acquire social meanings through repeated associations with speakers, activities, and stances; those meanings remain variable and context-dependent (Eckert, 2008). The term burger, for instance, can participate in locally classed and language-related positioning, but an analyst cannot assign one fixed definition without examining its use. Audience design adds a related complication. Speakers adapt style to actual and imagined recipients (Bell, 1984), and digital platforms complicate that process because multiple audiences can converge around the same post (Androutsopoulos, 2014). Translation or register adaptation may therefore be an intentional audience-design choice. A preservation-oriented analysis must not presume that every user wants the source form to remain unchanged.

2.3 Digital discourse and the problem of context

Computer-mediated discourse analysis treats online language as interaction shaped by

technological and social conditions rather than as a collection of isolated sentences (Herring, 2004). A reply can depend on a preceding question, a quoted phrase, the identity continuity of a pseudonymous speaker, or a narrative established several turns earlier. Removing usernames may reduce direct exposure, but removing reply structure can also erase evidence. This tension is particularly sharp in public-platform research because verbatim text and direct URLs remain searchable.

Ethical guidance for internet research rejects the idea that public availability alone settles questions of consent, vulnerability, redistribution, or contextual harm (Franzke et al., 2020; Markham & Buchanan, 2012). The present corpus includes accounts of schooling, workplace conflict, and medical labor. Sending those texts to an external AI service constituted additional processing beyond passive observation. A defensible archive must therefore state what was transmitted, whether the text remains searchable, what links are redistributed, and what institutional ethics determination was obtained. Where such documentation is absent, a limitation statement does not retroactively create approval.

2.4 AI correction as task design

Multilingual model evaluation often measures performance on predetermined tasks. Ahuja et al. (2023) demonstrate substantial variation across languages and prompting strategies in a benchmark covering many multilingual datasets. Such work is essential for comparative performance claims, but benchmark accuracy does not settle whether an edited utterance remains socially fitted to its original interaction. Human evaluation research similarly warns that constructs, instructions, annotator selection, and reporting practices determine what a score can support (van der Lee et al., 2021).

Research on AI and multilingual writing has begun to name the tension directly. Kuteeva and Andersson (2024) examine diversity and standards in AI-mediated writing for publication. Jeon et al. (2025) frame generative AI through translanguaging and identify dilemmas produced when flexible multilingual practices meet systems organized around standardized output. De Roock (2024) links generative AI, writing, and linguistic

justice, while Smith et al. (2026) formulate standard language ideology as a problem in AI-generated language. These arguments establish reasons to inspect correction as a normative task. They do not justify inferring a stable model policy from one opaque conversational session.

Prompt wording and interaction history matter. Single-prompt evaluations can be brittle across semantically similar instructions, models, and tasks (Mizrahi et al., 2024). A conversation also creates a trajectory: later responses may draw on earlier turns and on system instructions unavailable to the researcher. Consequently, the expression “generic prompt” is insufficient when the surrounding context is not generic. The proper unit of analysis may be the whole session rather than each output treated as an independent trial.

Automated correction also contains distinct operations. Translation changes language distribution. Orthographic normalization regularizes spellings. Grammatical editing changes morphosyntax. Register adaptation changes relations of formality or audience. Compression removes material. Monolingualization eliminates alternation, whether or not the resulting English corresponds to any particular codified standard. Grouping all these operations under “standardization” obscures what changed and what evidence would be needed to explain why.

The distinction also prevents a false opposition between preservation and accessibility. Canagarajah (2011) uses code-meshing to examine how multilingual resources enter writing conventionally governed by monolingual expectations. That work treats rhetorical choice and negotiation as central, not the unchanged reproduction of every form. In an editing task, a writer might reasonably prefer an English version for one audience and a minimally normalized Roman Urdu-English version for another. Evaluation must therefore attach preservation to a specified communicative purpose. A correction is not pragmatically adequate merely because mixed tokens remain, and a translation is not pragmatically inadequate merely because they do not.

This requirement changes the role of the evaluator. Rather than deciding whether an output

is globally faithful, the evaluator identifies which relations and formal cues are necessary for the stated purpose, then documents transformations and uncertainty. When no purpose is stated, competing evaluations should remain visible. One reader may prioritize propositional accessibility; another may prioritize quoted voice or peer intimacy. A controlled study can compare those judgments. A single analyst cannot collapse them into a precise percentage without making the construct appear more settled than it is.

2.5 Correction as a Metapragmatic Act

The feature families introduced above describe what changed in each correction. They do not explain why a request to correct a Roman Urdu-English sentence should touch address terms, idiom, and quoted voice at all, rather than staying confined to spelling and syntax. Silverstein (1979) supplies the missing vocabulary. He describes a metapragmatic act as an utterance, or here a machine-generated response, that regulates language use while appearing only to describe it. When the assistant glosses “system down hai, baad mai ayen” as a request to return later and then respells *mai* to *mein*, it reports what the sentence means and simultaneously ratifies a norm about which spelling counts as correct. An evaluation that checks only whether meaning survived will miss the second operation. In A07 that operation is slight; in A01, where the respelling is accompanied by the masculine *bolta hoon*, it changes how the writer is positioned. The concept is used here to describe what a response does to a text, not to attribute an intention to the system that produced it.

Of the semiotic processes that Irvine and Gal (2000) identify in the ideological making of linguistic boundaries, only erasure is used in this article, because it is the one for which the corpus offers textual evidence; iconization and fractal recursivity would require evidence about how readers attribute character to speakers, which the study did not collect. Erasure, in their account, is the process by which an ideology renders invisible the persons or practices that do not fit its picture of a bounded code. The corpus contains plausible candidates. *System* and *payment* sit comfortably inside Urdu syntax in A08 and A09, and the corrections render the surrounding clauses in English. Whether this is erasure in the ideological

sense, or translation of a kind the prompt permitted, cannot be settled from the outputs alone. A09 and A18 also show that a joke or a concession can survive the same operation largely intact, so the concept names a possible cost of translation, not an outcome observed in every English rendering.

Erasure matters most where the source stages more than one voice, and Agha’s (2005) account of footing and enregisterment, the concept most directly tied to the coded evidence, bridges form and social relation in a way the feature families otherwise leave implicit. A register exists because hearers and speakers repeatedly align talk with a particular figure and adjust their footing accordingly. The bank thread stages three figures (an employee’s formula, a manager’s reprimand, a narrator’s ironic aside), and each correction handles their footing differently. Replacing *kursi garam karna* with “sitting here doing nothing” keeps the manager’s authority but, on our reading, narrows the locally recognizable managerial register through which that authority was voiced. That second judgment is interpretive. It rests on the coders’ familiarity with how the idiom circulates in Pakistani workplace talk, and the two coders weighed it differently (see Results). Feature-level coding, in other words, already operationalizes a theory of footing; naming the theory clarifies why translation can leave one figure’s voice intact while narrowing another’s.

The word *correct* in the original prompt is therefore not a transparent request for accuracy. Flores and Rosa (2015) argue that appropriateness-based models of language relocate the labor of legitimation onto minoritized speakers while leaving the standard itself unexamined. Their argument gives a reason to treat an unqualified request to correct Roman Urdu-English material as normatively loaded, since such a request leaves unstated the standard against which error is measured. It does not supply evidence about this session. The corrections were uneven; A02, A12, and A15 kept mixed material largely intact, and the outputs alone cannot show whether the English renderings elsewhere followed from the wording of the prompt, from the conversational history, or from system instructions the authors could not see. The raciolinguistic critique is accordingly used here to frame a question for

controlled research, and not as an explanation of the outputs analyzed below.

The concepts therefore carry unequal evidentiary weight in what follows. Footing and enregisterment, together with Silverstein's account of metapragmatic regulation, are applied directly to coded features and can be checked against the quoted material. Erasure is used more cautiously, as a name for one possible cost of translation. The critique of appropriateness remains a framing for future work. None of them competes with the feature-level method described below; they explain why that method needs seven separable families rather than one retained-or-lost judgment. The same institutional vocabulary (*system*, *payment*) can voice bureaucratic threat in one line and employee explanation in the next, and no lexical inventory settles which reading holds without attention to footing and to the erasure or retention of the alternation that staged it.

2.6 Analytic Framework

The analysis separates three stages. First, semantic explanation concerns whether the response reconstructed the event, participant relations, and relevant stance. Second, corrective operation records what the response did to the text without attributing intention to the model. Third, feature-level transformation compares observable elements across source and correction.

Seven feature families organize the comparison: address and affiliation; quoted voice; idiom or locally indexical wording; humor or irony cues; sequential contrast between discourse segments; emoji, emoticon, capitalization, and punctuation; and technical or institutional vocabulary. Each salient feature is marked as retained, altered, omitted, or uncertain. An overall case memo is then written from the feature pattern. It is not derived from whether the correction remains multilingual. An English translation can retain quotation, humor, and sequencing. A mixed correction can alter a pronoun, gender marking, footing, or degree of intimacy.

The categories in Table 1 are intentionally separable. "Translation to English" records an operation. "Quoted voice altered" records a feature-level comparison. Neither predetermines the other.

Table 1. Separable Analytic Layers and Decision Rules

Analytic layer	Values	Decision rule
Explanation	Adequate, partial, uncertain	Does the gloss recover event, participants, and relevant stance?
Corrective operation	Orthographic normalization, grammatical editing, translation, register adaptation, compression, monolingualization	What observable change does the correction perform? Multiple values are allowed.
Feature comparison	Retained, altered, omitted, uncertain	Compare each salient feature in source and correction without using language choice as a proxy.
Synthesis memo	Largely preserved, mixed, substantially transformed, uncertain	Summarize the feature pattern and record a competing reading where warranted.

Note: Categories are descriptive and do not imply independent observations or population estimates.

This separation follows the metapragmatic and enregisterment theory developed above: an operation such as translation changes the code in which a figure is voiced, while a feature-level judgment such as altered footing changes what that voicing accomplishes, and the two require independent evidence.

This framework treats pragmatic interpretation as contestable. The source writer was not interviewed, and no coder can recover intention as ground truth. The analysis instead makes a warranted claim about how textual form supports interactional reading. Alternative readings are recorded where evidence permits more than one conclusion. Where a reading depends on a culturally specific expression, such as *kursi garam karna* in A08 or *chordo* in A22, the analysis states the reading the coders adopted, the textual cue on which it rests, and the competing reading where one is available. The coders' competence as Pakistani users of Urdu and English makes these readings possible, but it is a resource for interpretation and not a substitute for the writers' own accounts. The aim is auditability: another

analyst should be able to identify the feature at issue and disagree with the interpretation without first accepting a global label such as “flattening.”

3 METHODOLOGY

3.1 Design and unit of analysis

The study is a retrospective, reflexive qualitative case analysis of one recorded AI-assisted research session. It is not a controlled model audit, experiment, or estimate of default behavior. The six Reddit threads are the case units. Sixteen utterances or passages are embedded segments used for close comparison. Segments within a thread are dependent because they share topic, participants, adjacency, and sometimes authorship. Model outputs are also dependent because they were generated sequentially in one conversation.

This redesign answers a narrower question than the original study. It documents how one underspecified correction request was resolved in a particular interactional archive. It does not identify the causal effect of the word *correct*, compare prompts, or generalize to other sessions, checkpoints, products, users, or languages.

3.2 Corpus provenance and exclusions

The original workbook listed 24 units from eight r/pakistan threads, with three units per thread. The threads were identified by browsing r/pakistan and by keyword searching within it for publicly accessible threads on everyday topics (Table 2) in which posts or comments visibly combined Roman Urdu and English. Within each thread, three units were chosen in which alternation appeared to do interactional work, for instance in quotation, joking, advice, or reprimand. This was purposive selection for pragmatic interest; it was not a sampling procedure. The initial collection was recorded between 3 June and 7 September 2026. The revision audit revisited the linked pages and compared visible source text character by character where the page exposed it. A unit was retained only when its exact text was visible on the live thread and matched the workbook string. This produced four provenance states: verified against the visible live thread; not visible or not reverified;

materially inconsistent with the visible source; and excluded.

Sixteen units were verified and retained: A01–A03, A07–A13, A15, A17–A18, and A22–A24. Five units were excluded because the exact passage could not be reverified on the accessible live page: A04–A06, A14, and A16. Three were excluded for material alteration. A19 and A20 had converted predominantly English source passages into Roman Urdu, and A21 replaced “Na yaar” with “Dude.” Because those changes concern the linguistic form under analysis, their earlier model outputs and codes are not used.

The audit cannot reconstruct the complete sampling frame. The search terms, the dates of individual searches, the number of threads and comments screened, and the reasons for passing over candidate material were not logged before the original spreadsheet was created; they are reported here as unavailable and are not estimated. The retained material is therefore an availability-based analytic corpus, not a corpus selected independently of pragmatic salience. This limitation bears on interpretation in three ways. The corpus probably over-represents units in which alternation is conspicuous, humorous, or institutionally marked, because those were the features that drew the authors’ attention, and corrections of more routine mixed text may behave differently. The exclusions also fall unevenly across threads (Table 2), so that two topics, English-language attitudes and board examinations, contribute nothing to the analysis. Finally, nothing in the corpus licenses a statement about how often a given correction operation occurs, either on Reddit or in AI output. RQ1 is correspondingly descriptive: it asks what the retained cases do, not how frequently Roman Urdu-English alternation performs those functions on Reddit.

The exact source strings, page links, verification date, thread position, and exclusion reason appear in the restricted audit matrix described under Researcher position and ethical decisions. The workbook does not call the data de-identified. Usernames are omitted, but verbatim searchable text and thread links make re-identification possible, and access to the matrix is limited for that reason.

Table 2 summarizes the revised case structure. Counts describe embedded segments, not independent observations.

Table 2. Revised Thread-Level Case Structure

Thread case	Retained segments	Interactional material	Audit disposition
Language repertoires	3	Post, comment, reply	Live text verified
Banking and customer service	3	Two quotations and narrative coda	Live text verified
Food and cooking	3	Instruction, reply, recipe	Live text verified
Exam preparation	2	Advice replies	One additional unit not reverified
Education and language policy	2	Comment and reply	One additional unit not reverified
Medical work	3	Reply, satire, advice	Live text verified
English-language attitudes	0	Three listed comments	Excluded because exact text was not reverified
Board examinations	0	Post passages and reply	Excluded because listed text materially differed from source

3.3 Thread context and segment boundaries

Segment boundaries follow interactional action rather than a uniform word count. Short quoted formulas, such as A07, are retained because the surrounding comment explicitly presents them as reported speech. Longer recipe and advice passages are retained when their internal sequence produces the action being analyzed. No segment is treated as statistically independent.

An anonymized map assigns thread-specific speaker labels such as T1-S1 and T1-S2. It records whether a segment is a post, comment, reply, quotation, or narrative coda and identifies the relevant preceding action. This retains speaker continuity without publishing account names. For A07–A09, for example, one

commenter narrates a bank encounter: an employee's formula is followed by a manager's reprimand and then by the narrator's ironic coda. Analyzing these as three independent statements would erase the escalation that gives each segment its force.

3.4 Recorded model condition

Every retained segment was submitted sequentially in the same assisted research session with the user-level prompt: "Please explain what this text means, and correct any errors you find." The interface identified the product as OpenAI Codex in ChatGPT Work and described it as based on GPT-5. The exact checkpoint, sampling parameters, hidden system prompt, and full prior conversational context were not available. Responses were recorded between 3 June and 7 September 2026.

These constraints are part of the object. The phrase "recorded session" is used throughout. No response is described as a stable model output, and no claim is made about an unconditioned corrective policy. A controlled replication would require isolated sessions or an API, immutable identifiers, randomized order, multiple runs, complete configuration records, and comparison across prompt conditions.

3.5 Analytic procedure

The original human reading and model response were separated before revision. The recoding then proceeded in four steps. First, each retained thread was read as an interactional sequence. Second, the source segment was annotated for the seven observable feature families. Third, the explanation and correction portions of the recorded response were separated. Fourth, each feature was compared across source and correction using retained, altered, omitted, or uncertain. The second and fourth steps were carried out separately by two coders, as described below.

Max and Alex are pseudonymous labels for the two coders, the first and second authors, each of whom coded all sixteen units in a separate column of the revision workbook without seeing the other's codes until both passes were complete. The passes were deliberately asymmetrical. Max applied a conservative semantic criterion,

counting a function as retained when the correction kept its interactional sequence even if language distribution changed. Alex applied a form-sensitive criterion, treating changes to address, idiom, rhythm, and voicing as more consequential. Because the coders worked from different criteria, some of their disagreements are a designed product of the procedure, and they are reported as evidence of where the construct of preservation is contestable, not as coding error. For the same reason, no agreement coefficient was computed: a coefficient presupposes a shared rule, and sixteen dependent segments could not yield a stable estimate in any case.

Disagreements were documented feature by feature, with the textual cue on which each coder relied. The coders then discussed each disputed unit. Where a shared reading emerged, the synthesis memo records it with the reasoning; where it did not, the memo keeps both readings and assigns the label mixed or uncertainty instead of adopting one coder's view. A08 and A23, discussed in the Results, show a disagreement carried into a mixed synthesis in this way. The independence of the two passes was limited in other respects. Both coders are authors, both knew the research questions, and both had seen the codes assigned to the original 24 units. The procedure makes competing readings visible and auditable; it does not provide the independent verification that blind coding by annotators outside the research team would offer, and the article does not claim it.

3.6 Researcher position and ethical decisions

The first and second authors, who carried out the coding, are Pakistani users of Urdu and English who read and write Roman Urdu in everyday digital communication and take part in online spaces of the kind the corpus draws on. The third author did not code the Urdu material; his contribution concerned the analytic framework and the scrutiny of claims against the evidence. Insider competence makes available readings of idioms such as *kursi garam karna* and of address terms such as *bhai* and *yaar* that an outside analyst would have to reconstruct. It also brings its own regional, class, and generational habits of reading, and it cannot stand in for what the writers themselves meant. The original 24-unit workbook

did not record these details; they are stated here for the first time.

No institutional ethics review was sought, and no determination was obtained from the authors' institution or any other body. At the time of collection, the authors judged that publicly posted text, analyzed without contact with its writers, fell outside the scope of human-participant review. That judgment was not checked against a written institutional policy, and it is reported here as the authors' decision, not as an exemption. Guidance for internet research holds that public availability does not settle questions of consent, contextual harm, or redistribution (Franzke et al., 2020; Markham & Buchanan, 2012), and the absence of review is accordingly treated as a limitation of the study.

The revision considered three risks, and the measures available were limited to what could still be done after collection. The first is searchable quotation. The object of analysis is what a correction did to the exact form of a string, so paraphrase would remove the evidence. The article therefore quotes only the shortest string needed for each point, omits usernames, thread titles, and URLs, and refers to speakers by thread-specific labels. On the authors' reading, the retained units concern language preference, a banking anecdote, cooking, examination advice, schooling, and the working lives of doctors, and none discloses a health condition, a legal matter, or a personal detail that would identify the writer beyond the text itself. The schooling and medical-work material is nonetheless quoted no more extensively than the argument requires. The second risk is the processing of third-party text through an external AI service. Only the unit strings and the prompt were submitted, without account names. The data-use settings of the ChatGPT Work account at the time were not recorded, so the authors cannot state whether the submitted text was excluded from model training; this, too, is reported as a limitation.

The third risk, redistribution, governs the form of supplementary material. The public supplement contains the coding scheme, the decision rules in Table 1, the anonymized speaker map, the feature-level codes for each unit, and a description of each unit that does not reproduce its wording; it contains no URLs. The audit matrix, which holds

the verbatim strings, page links, and verification records, is available to the editor and reviewers on request and, after publication, to researchers who state a research purpose and agree not to republish the strings or links. A unit is withdrawn from the matrix if its source post is deleted. An unrestricted deposit would require formal institutional determination and a review of the platform's terms for redistributing user content, neither of which has been completed.

4 RESULTS

4.1 The thread as the unit of interpretation

The bank-service case demonstrates why the thread is the unit. A07 voices an employee through "system down hai, baad mai ayen." The English technical phrase *system down* appears inside Urdu syntax and, on the coders' reading, shifts the immediate explanation from an employee's choice to an unavailable infrastructure. The recorded correction regularized *mai* to *mein* and *ayen* to *aayen* while leaving the formula intact. At the segment level, this is minimal orthographic normalization.

A08 is the manager's reply within the same story. The source demands proof, uses *kursi garam karna* to accuse the employee of occupying a seat without working, and threatens dismissal. The English correction retains the accusation, demand, and threat, but substitutes "sitting here doing nothing" for the idiom. Max therefore read the action sequence as retained with an altered idiom; Alex treated the idiom and Urdu reprimand as central to the manager's footing. The synthesis records mixed preservation, not total loss.

A09 supplies the narrator's coda: "His face was red with anger but kya karta payment clear krni pari." The correction, "His face was red with anger, but what could he do? He had to clear the payment," removes alternation yet retains the concession and ironic inevitability. This is a disconfirming case for any rule that equates English translation with pragmatic loss. The language distribution changes, but the narrative turn survives.

The medical-work thread creates a similar dependence. A22, "chordo bhai 🤔 (me too),"

replies to a prospective doctor who describes studying intensely. The phrase can express affiliative resignation, a warning, or comic self-recognition. The model glossed it as advice to "leave it" and preserved "brother," the crying emoji, and "me too" in its correction. Familiarity with the phrase narrows the possible readings but does not choose among them; only the writer, or uptake later in the thread, could do that. Because *chordo* remains ambiguous without participant validation, the revised code is uncertain rather than a forced register-misjudgment label.

4.2 English renderings varied in feature retention

The English corrections did not form a homogeneous category. A03 moves from "Urdu only" to *tooti phooti* and back to "No language flex for me here." Its correction retains self-deprecation and the platform term *flex*, but compresses the English-Urdu-English cadence into one English sentence. The relevant transformation is rhythmic and sequential, not simply monolingualization.

A10 also complicates the earlier binary. The source develops a mock theory of Pakistani cooking by rapidly changing water, tomatoes, and spices to produce *karahi*, *salan*, *qorma*, and *white handi*. The explanation explicitly identifies humor, and the English correction preserves the dish taxonomy and procedural parallelism. It reduces repetition and removes the Roman Urdu imperative rhythm. The joke is not wholly erased, but its performance is less compressed and less locally voiced. Both audit passes therefore reject the old "lost" judgment and record mixed preservation.

A11 extends that cooking sequence. The speaker begins with *haan na*, moves through *pulao*, *khichdi*, and *biryani*, and marks a limit to expertise by placing pressure-cooker use outside "my scope." The English version preserves the recipes, the self-limitation, and the explanation of tenderizing. It removes the alignment marker and the recurring "game" metaphor. The correction retains epistemic structure while altering affiliation and style.

A13 begins with the hybrid evaluation "Ye pretty delusional" and follows it with encouragement.

The English correction preserves the blunt assessment and the repair, “Do not give up, though,” while changing the conversational opening. A17 similarly retains an experiential counterclaim about adapting from Urdu-medium schooling to English books and translates *chill out* as “relax.” Neither case supports a claim of complete pragmatic deletion. They show that stance can remain while local form and degree of informality shift.

A18 asks why English speakers seem impressive and answers, “Kyun k un ki samajh nahin ati!” The model preserves the two-voice question-and-punchline sequence in English. The bilingual contrast disappears, but the joke remains legible. The revised analysis therefore separates removal of multilingual form from loss of humor. They can coincide, but this case shows they need not.

4.3 Minimal and mixed-language corrections were not neutral copies

Corrections that retained both Roman Urdu and English still involved choices. In A01, the response preserves *yaar* and the language list but changes spellings, agreement, and the final self-description. It also selects masculine *bolta hoon*, although the source does not establish the writer’s gender. The earlier “retained” label concealed this added gender marking. Feature-level coding records affiliation as retained, and gender positioning as altered.

A02 preserves the movement from a mock threat to the English reassurance “I have huge hopes for you :P.” It regularizes the Urdu and retains the emoticon. Here the correction keeps the affective sequence with relatively little transformation. A12 likewise maintains a mixed recipe, domestic evaluation, and *bhai*, even while replacing some ingredients with English equivalents and regularizing Roman Urdu. These are the clearest cases in which the session’s corrections remained close to the source action.

A15 retains *bhai*, numerical targets, the contrast between arts and sciences, and an offer of private help. The correction changes *hit me in dms* to “message me,” reducing a platform-specific idiom while keeping the mentoring sequence. This is audience adaptation at a small scale. It cannot be

described as preservation without qualification, but it also does not amount to major pragmatic loss.

These cases matter theoretically because multilingual surface form is not a sufficient preservation metric. A correction may keep tokens from both named languages yet introduce gender, change platform register, or redistribute authority. Conversely, a monolingual English correction may preserve quotation, sequencing, and humor. The relevant question is which features the communicative objective requires.

4.4 Quotation and institutional vocabulary sometimes survived translation

A23 predicts that increasing a doctor’s fee will convert a *complaint* into a *compliment*, then quotes an imagined patient praising the doctor’s attention. The English correction retains the complaint-to-compliment transformation, the quoted patient voice, and the relation among price, reputation, and listening. It removes Roman Urdu and smooths the satirical voice. The original “lost” code was therefore too strong. Max read the central satire as retained; Alex treated the local quoted voice as materially altered. The synthesis records mixed preservation.

A24 moves from English encouragement into Roman Urdu explanation of how experience becomes reputation. The correction preserves the advice, causal account, and the address term “Bro,” but makes the passage continuously English. The result is stylistically smoother and less visibly negotiated across repertoire. What changes is the distribution of coaching voice and locally framed explanation, not the proposition that recognition produces trust.

The workplace cases also show that institutional vocabulary can survive. *System* and *payment* remain in A08 and A09, while the idiomatic frame around them changes. It would therefore be misleading to say that translation necessarily removes institutional critique. In this archive, it sometimes recasts critique as explicit report. Whether that recasting is helpful depends on the intended audience and purpose, which the prompt did not specify.

Table 3. Evidentiary Logic for Selected Disagreements

ID	Source feature	Correction evidence	Revised synthesis
A01	<i>yaar</i> and modest repertoire listing	<i>Yaar</i> remains, but masculine <i>bolta hoon</i> is introduced	Affiliation retained; gender positioning altered
A03	English-Urdu-English comic cadence	Self-deprecation and <i>flex</i> remain in continuous English	Meaning retained; rhythmic alternation altered
A08	<i>kursi garam karna</i> within a managerial threat	Accusation and threat remain; idiom becomes “doing nothing”	Action sequence retained; idiomatic force altered
A09	Alternation marks an ironic narrative coda	“What could he do?” preserves concession in English	Multilingual form removed; irony largely retained
A10	Repetitive imperatives build a cooking joke	Dish taxonomy and parallelism remain; cadence is compressed	Humor partly retained; performance altered
A18	Question and reply produce a punchline	Two-voice joke remains in English	Humor retained; bilingual contrast omitted
A22	<i>chordo bhai</i> plus emoji and “me too”	Literal “leave it” reading remains contestable	Interpretation uncertain
A23	Complaint-to-compliment wordplay and quoted patient voice	Both are retained in smoother English	Satirical structure retained; local voicing altered

Note: Categories are descriptive and do not imply independent observations or population estimates.

Table 3 displays the evidentiary logic for selected disagreements. It does not rank outputs or convert the memos into a scale.

5 DISCUSSION

The recorded session usually produced explanations that were plausible at the level of event and stance, but its corrections redistributed interactional features in different ways. Language choice predicted neither total preservation nor total loss. The strongest evidence concerns specific transformations: an idiom paraphrased, a

cadence compressed, an address term retained, a gendered form introduced, a quotation preserved, or a joke carried across translation. Each of these is an observation about one output. Table 4 separates such observations from the interpretations and hypotheses built on them.

Table 4. Levels of Claim in the Analysis

Level of claim and basis	Examples in this study	What it does not establish
Observed transformation. Direct comparison of the source string with the recorded correction	<i>mai</i> respelled <i>mein</i> (A07); <i>kursi garam karna</i> replaced by “sitting here doing nothing” (A08); masculine <i>bolta hoon</i> introduced (A01); “hit me in dms” rendered “message me” (A15); alternation removed (A09, A18)	Why the change occurred, or whether it would recur in another session
Interpretation. The coders’ reading of the interactional effect, using thread context and insider competence, with disagreements recorded	Managerial footing narrowed (A08); irony retained (A09); humor retained (A18); comic performance altered (A10); sense of <i>chordo</i> left uncertain (A22)	The writer’s intention, or how other community members would read the revision
Hypothesis. Inference across units, framed by theory and untested	An unqualified request to “correct” leaves the target variety unstated and so permits English resolution; a clarification question or labeled alternatives would change which features are kept; outcomes would differ across isolated sessions	A stable model behavior or a causal effect of the prompt; testing requires the controlled design outlined below

Note: Examples refer to the recorded session only. Hypotheses are listed for testing and are not findings of this study.

This finding changes how correction should be conceptualized. “Standardization” is too broad because the archive contains orthographic normalization, grammatical editing, English translation, compression, and register adaptation. “Monolingualization” accurately describes cases where alternation disappears, but it does not identify a particular English standard or demonstrate a causal ideology. The session sometimes retained colloquial expressions such

as *bro* and *flex*. The most precise description is that the session resolved an underspecified editing request through several corrective operations, some of which reduced multilingual form.

The strongest counterargument is practical. Many users who ask for correction may want standard-English rendering, especially when writing for a wider audience. Translation can increase accessibility, clarify institutional narratives, and preserve interactional action, as A09 and A18 show. Audience design is not betrayal of the source. It is one legitimate objective among others. The problem in this session is therefore not that English translation occurred. It is that the prompt did not establish whether translation, normalization, or audience adaptation was wanted, and the response rarely labeled the choice as one among several possible revisions.

Translanguaging theory remains useful once it is not made to predetermine the result. It directs attention to the repertoire through which speakers organize relations, while sequential analysis identifies where contrasts, quotations, and reply positions matter. The framework also retains distinctions that the earlier manuscript blurred. Technical borrowings do not automatically constitute switches. Code-meshing in writing, clause-level alternation, quoted speech, platform slang, and English-dominant passages with Urdu segments require separate descriptions before they can be compared.

Read against these findings, the concepts introduced earlier do limited and specific work. Erasure, in Irvine and Gal's (2000) sense, names what may be lost when a correction renders a mixed clause in English, as in A03, A08, and A24, although A09 and A18 show that such rendering need not discard the interactional action. The corpus cannot show that the session tends toward English resolution in general; it shows that English resolution occurred in a number of units under an unspecified request, alongside corrections that kept mixed material. Enregisterment (Agha, 2005) offers a reason, which this study proposes but cannot test, why some resolutions cost more than others: a figure voiced through a locally recognizable register loses more in paraphrase than one whose institutional vocabulary already travels comfortably in English.

The Pakistan-specific literature prevents another shortcut. English carries historical connections to administration, education, and mobility (Rahman, 1996; Shamim, 2017), but those associations cannot be assigned mechanically to every English token. In the corpus, English institutional vocabulary can voice bureaucracy, but English also performs reassurance, recipe scaffolding, humor, and peer advice. Roman Urdu can intensify intimacy or complaint, yet it can also carry institutional reported speech. The relation between resource and function is locally organized.

For multilingual AI evaluation, the practical consequence is a design requirement for future research rather than a pedagogical conclusion. A controlled study should compare at least explanation only, unspecified correction, repertoire-preserving correction, standard-English translation, audience-specific editing, and correction following a clarification question. Each item should appear in an isolated session, order should be randomized, and multiple runs should estimate response variability. Model identity, date, parameters, system configuration, and full prompts should be archived. Texts should be stratified by genre, length, English proportion, quoted speech, humor, address terms, and institutional vocabulary.

Qualified Pakistani Urdu-English coders from outside the research team, working without sight of one another's codes, should annotate observable features before judging broader interactional fit. Pre-consensus disagreements should remain available, and community participants should evaluate whether revisions suit specified purposes. An agreement coefficient may be informative for tightly defined categories, but it cannot substitute for documenting why a term such as *burger*, *bhai*, or *chordo* is interpreted in a particular way. Original writers' judgments would be especially valuable where ethical and practical conditions permit contact. Those are testable hypotheses. Classroom research could compare user choices when systems provide a single English correction, two labeled alternatives, or a clarification question. The present data establishes only the need to distinguish those conditions.

6 CONCLUSION

The bank anecdote that opened this article did not yield one uniform pattern. The session preserved “system down hai” with minor spelling changes, translated the manager’s reprimand while paraphrasing an idiom, and translated the narrator’s coda while retaining its ironic concession. That sequence is more informative than a percentage of outputs labeled retained or lost.

In this recorded conversation, semantic explanation, corrective operation, and interactional preservation were distinct analytic objects. English translation sometimes removed cadence or local voicing and sometimes carried quotation, humor, or advice across languages. Mixed-language correction sometimes preserved affiliation and sometimes introduced new positioning. The claims about footing, irony, and register rest on two authors’ readings, recorded together with their disagreements, and they are warranted interpretations of the texts, not verified facts about what the writers intended. A defensible evaluation must therefore specify the user’s objective, analyze observable features, retain thread context, and report uncertainty. Whether these patterns recur is unresolved. Answering that question requires a new controlled corpus and independent community evaluation, not stronger claims from the present session.

STATEMENTS

Conflict of interest. The authors declare no conflict of interest.

Funding. This research received no external funding.

Author contributions. H.M.S. and S.U. carried out the coding as the two coders (Max and Alex). A.J. contributed to the analytic framework and to the review of claims against the evidence. All authors contributed to designing the study and writing the manuscript, and have read and approved the final version for submission.

Ethics statement. The study analyzes publicly available Reddit posts and comments together with one recorded ChatGPT Work session; it does not involve interaction with human participants. No institutional ethics review was sought, and no determination was obtained. The authors judged at the time of collection that public, non-interactive text fell outside human-participant review, but did not verify that judgment against a written institutional policy, and they report the absence of review as a limitation, not as an exemption. Usernames and account-identifying details were omitted from the analysis, from the text submitted to the AI service, and from the public supplement; verbatim strings and links are held in an access-controlled audit matrix whose conditions of access are stated in the Methodology. The authors recognize, following Franzke et al. (2020), that public availability alone does not eliminate the risks created by searchable quotation, vulnerable contexts, third-party processing, or redistribution.

REFERENCES

- Agha, A. (2005). Voice, footing, enregisterment. *Journal of Linguistic Anthropology*, 15(1), 38–59. <https://doi.org/10.1525/jlin.2005.15.1.38>
- Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., Bali, K., & Sitaram, S. (2023). MEGA: Multilingual evaluation of generative AI. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 4232–4267). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.258>
- Ameer, I., Sidorov, G., Gomez-Adorno, H., & Nawab, R. M. A. (2022). Multi-label emotion classification on code-mixed text: Data and methods. *IEEE Access*, 10, 8779–8789. <https://doi.org/10.1109/ACCESS.2022.3143819>
- Androutsopoulos, J. (2014). Linguaging when contexts collapse: Audience design in social networking. *Discourse, Context & Media*, 4–5, 62–73. <https://doi.org/10.1016/j.dcm.2014.08.006>
- Auer, P. (Ed.). (1998). *Code-Switching in Conversation: Language, Interaction and Identity*. Routledge.

- Bell, A. (1984). Language style as audience design. *Language in Society*, 13(2), 145–204. <https://doi.org/10.1017/S004740450001037X>
- Bilal, A., Rextin, A., Kakakhail, A., & Nasim, M. (2019). Analysing emergent users' text messages data and exploring its benefits. *IEEE Access*, 7, 2870–2879. <https://doi.org/10.1109/ACCESS.2018.2885332>
- Canagarajah, A. S. (2011). Codemeshing in academic writing: Identifying teachable strategies of translanguaging. *The Modern Language Journal*, 95(3), 401–417. <https://doi.org/10.1111/j.1540-4781.2011.01207.x>
- de Roock, R. S. (2024). To become an object among objects: Generative artificial “intelligence,” writing, and linguistic justice. *Reading Research Quarterly*, 59(4), 590–608. <https://doi.org/10.1002/rrq.569>
- Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, 12(4), 453–476. <https://doi.org/10.1111/j.1467-9841.2008.00374.x>
- Flores, N., & Rosa, J. (2015). Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review*, 85(2), 149–171. <https://doi.org/10.17763/0017-8055.85.2.149>
- Franzke, A. S., Bechmann, A., Zimmer, M., Ess, C., & Association of Internet Researchers. (2020). *Internet Research: Ethical Guidelines 3.0*. <https://aoir.org/reports/ethics3.pdf>
- Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge University Press.
- Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S. A. Barab, R. Kling, & J. H. Gray (Eds.), *Designing for Virtual Communities in the Service of Learning* (pp. 338–376). Cambridge University Press. <https://doi.org/10.1017/CBO9780511805080.016>
- Irvine, J. T., & Gal, S. (2000). Language ideology and linguistic differentiation. In P. V. Kroskrity (Ed.), *Regimes of Language: Ideologies, Politics, and Identities* (pp. 35–84). School of American Research Press. <https://web.stanford.edu/~eckert/PDF/IrvineGal2000.pdf>
- Jeon, J., Li, W., Tai, K. W. H., & Lee, S. (2025). Generative AI and its dilemmas: Exploring AI from a translanguaging perspective. *Applied Linguistics*, 46(4), 709–717. <https://doi.org/10.1093/applin/amaf049>
- Kuteeva, M., & Andersson, M. (2024). Diversity and standards in writing for publication in the age of AI: Between a rock and a hard place. *Applied Linguistics*, 45(3), 561–567. <https://doi.org/10.1093/applin/amae025>
- Li, W. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, 39(1), 9–30. <https://doi.org/10.1093/applin/amx039>
- Markham, A., & Buchanan, E. (2012). *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee* (Version 2.0). Association of Internet Researchers. <https://aoir.org/reports/ethics2.pdf>
- Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12, 933–949. https://doi.org/10.1162/tacl_a_00681
- New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review*, 66(1), 60–92. <https://doi.org/10.17763/haer.66.1.17370n67v22j160u>

- Otheguy, R., García, O., & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review*, 6(3), 281–307. <https://doi.org/10.1515/applirev-2015-0014>
- Rahman, T. (1996). *Language and Politics in Pakistan*. Oxford University Press.
- Saeed, H. H., Ashraf, M. H., Kamiran, F., Karim, A., & Calders, T. (2021). Roman Urdu toxic comment classification. *Language Resources and Evaluation*, 55(4), 971–996. <https://doi.org/10.1007/s10579-021-09530-y>
- Shamim, F. (2017). English as the language for development in Pakistan: Issues, challenges and possible solutions. In H. Coleman (Ed.), *Dreams and Realities: Developing Countries and the English Language* (pp. 291–310). British Council. https://www.researchgate.net/publication/313722489_English_as_the_language_of_development_in_Pakistan_Issues_challenges_and_possible_solutions
- Silverstein, M. (1979). Language structure and linguistic ideology. In P. R. Clyne, W. F. Hanks, & C. L. Hofbauer (Eds.), *The Elements: A Parasession on Linguistic Units and Levels* (pp. 193–247). Chicago Linguistic Society.
- Smith, G., Fleisig, E., Bossi, M., Rustagi, I., & Yin, X. (2026). Standard language ideology in AI-generated language. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. <https://doi.org/10.1145/3805689.3806504>
- van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67, 101151. <https://doi.org/10.1016/j.csl.2020.101151>