Code-Switching Is Not Noise: Evaluating LLMs on the Language People Actually Speak

Anonymous ACL submission

Abstract

For many multilingual users, code-switching is not degraded language but the ordinary communicative register of everyday interaction. Yet multilingual LLM evaluation often treats languages as separable monolingual conditions, while prior code-switching benchmarks primarily evaluate classification or comprehension tasks. We argue that code-switched assistant interaction is a first-class human-centered NLP evaluation setting. We introduce a diagnostic probe of 100 Hinglish and Spanglish prompt groups, each paired with English and local-language controls, and evaluate GPT-4o, Claude, Qwen2.5-72B, and Llama-3.3-70B on generative assistant responses. Responses are scored for task success, register preservation, pragmatic intent preservation, non-translation compliance, and naturalness, with binary failure labels for silent monolingualisation, register collapse, translation-over-assistance, pragmatic cue loss, and over-formalisation. Results show that code-switched prompts do not primarily break task completion: mean task success remains high (1.77/2), close to English controls (1.86/2). Instead, they break interactional fit: register preservation drops to 0.88/2 and naturalness to 0.94/2. Even GPT-4o and Claude silently monolingualise 38% and 34% of code-switched prompts. We argue that multilingual assistant evaluation must measure not only whether models understand users, but whether they respect how users actually speak.

1 Introduction

For many bilingual and multilingual speakers, everyday communication does not occur in neatly separated monolingual blocks. A user may ask an assistant to "write this politely but *zyada corporate mat banana*," or request a reply that sounds "*tranqui pero respetuoso*." Such prompts are not corrupted English, incomplete Hindi, or informal Spanish. They are ordinary bilingual registers through which speakers encode tone, social distance, cultural context, and pragmatic constraints (Gumperz, 1982; Auer, 1999).

Despite this, language-model evaluation remains shaped by a largely monolingual paradigm. Multilingual benchmarks such as XTREME and XGLUE evaluate models across many languages and tasks, but usually treat languages as separable evaluation conditions rather than mixed interactional registers (Hu et al., 2020; Liang et al., 2020). Code-switching benchmarks such as LinCE and GLUECoS make important progress, but their primary tasks are language identification, POS tagging, NER, sentiment analysis, QA, NLI, or related static-text evaluations (Aguilar et al., 2020; Khanuja et al., 2020). These tasks do not directly answer a human-centered question: when users naturally code-switch while asking an assistant to help them, does the model preserve the user's register, tone, and pragmatic intent?

We argue that code-switched assistant interaction exposes a gap between *task completion* and *interactional fit*. A model may perform the literal task while still failing the user interactionally: it may silently normalise the response into polished English, over-formalise a casual bilingual request, miss a pragmatic cue expressed in the non-English portion of the prompt, or treat the prompt as material to translate rather than as a request to act on. These are not merely multilingual comprehension failures. They are sociolinguistic and interactional failures.

We introduce a paired diagnostic probe of 100 prompt groups across Hinglish and Spanglish. Each group contains a code-switched prompt, an English control, and a local-language control. This design lets us compare responses to the same underlying task across different communicative registers. We evaluate four LLM assistants: GPT-4o, Claude, Qwen2.5-72B-Instruct, and Llama-3.3-70B-Instruct.

Our contributions are: (1) positioning code-switched assistant interaction as a first-class human-centered NLP evaluation setting; (2) a paired diagnostic probe for evaluating Hinglish and Spanglish prompts against monolingual controls; and (3) an annotation taxonomy that separates semantic task success from interactional fit.

2 Related Work

Multilingual evaluation. Multilingual benchmarks have expanded NLP evaluation beyond English, but often evaluate languages as distinct conditions. XTREME evaluates cross-lingual generalisation across 40 languages and nine tasks (Hu et al., 2020); XGLUE evaluates cross-lingual understanding and generation (Liang et al., 2020). These resources are essential, but they differ from everyday bilingual interaction, where users may mix languages within the same turn.

Code-switching in NLP. Code-switching has a long history in sociolinguistics and NLP. LinCE centralises evaluation across Spanish-English, Nepali-English, Hindi-English, and Arabic varieties for language identification, NER, POS tagging, and sentiment analysis (Aguilar et al., 2020). GLUECoS evaluates English-Hindi and English-Spanish code-switched NLP across language identification, POS tagging, NER, sentiment, QA, and NLI (Khanuja et al., 2020). More recent work studies code-mixed language understanding with multilingual encoders (Winata et al., 2021). Our focus is different: generative assistant interaction, where the model must preserve task intent, language choice, register, and pragmatic meaning.

3 Diagnostic Probe

3.1 Prompt design

We constructed 100 prompt groups covering everyday assistant tasks where tone and pragmatic nuance matter. The probe includes 50 Hinglish groups and 50 Spanglish groups. Each group contains three semantically matched variants: (i) a natural code-switched request, (ii) an English control, and (iii) a local-language control, using romanised Hindi for Hinglish groups and Spanish for Spanglish groups. This yields 300 prompt variants.

Prompts span five task types: writing and rewriting, advice and planning, emotional or social support, workplace communication, and locally situated pragmatic meaning. The probe is diagnostic rather than exhaustive: it is designed to test whether common LLM assistants preserve interactional fit under naturalistic mixed-language input, not to represent all code-switching practices.

3.2 Models and prompting

We evaluate GPT-4o, Claude, Qwen2.5-72B-Instruct (Yang et al., 2024), and Llama-3.3-70B-Instruct (Grattafiori et al., 2024).[1] All models receive the same system instruction:

> You are a helpful assistant. Respond naturally to the user's request. Preserve the user's intended language choice and tone where appropriate. Do not translate unless the user asks for translation.

This is a conservative prompt: it explicitly instructs the model to preserve language choice and tone. Thus, failures to preserve register or unnecessary translation are not simply due to missing instructions. All generations use temperature 0, and each model answers all 300 prompt variants, yielding 1,200 generations in total.

3.3 Annotation rubric

Each response is scored on five 0-2 dimensions: **task success**, **register preservation**, **pragmatic intent preservation**, **non-translation compliance**, and **naturalness**. We also annotate seven binary failure modes: **silent monolingualisation**, **register collapse**, **translation-over-assistance**, **pragmatic cue loss**, **local context loss**, **over-formalisation**, and **wrong language choice**; Table 3 reports five of these, and the full set is recorded in the released annotations. The key distinction is that a response can score well on task success while failing register preservation or naturalness. Responses were scored by a single annotator in this diagnostic release; we treat this as a limitation and discuss it in the Limitations section. Full dimension anchors and failure-mode definitions appear in Appendix A.

4 Results

4.1 Task success remains high, but interactional fit collapses

Table 1 and Figure 1 show mean scores by prompt variant. Code-switched prompts do not primarily degrade task completion: task success remains high at 1.77/2, close to English controls at 1.86/2 and

[1] GPT-4o and Claude were queried through the OpenAI and Anthropic commercial APIs; the open-weight models were run from their public instruction-tuned checkpoints. Model access and decoding details are summarised in Appendix C.

Variant	Task	Reg.	Prag.	Non-trans.	Nat.
English	1.86	1.82	1.78	1.95	1.84
Local-language	1.72	1.61	1.64	1.88	1.58
Code-switched	1.77	0.88	1.05	1.52	0.94

Table 1: Mean response scores by prompt variant on a 0-2 scale. Task = task success; Reg. = register preservation; Prag. = pragmatic intent preservation; Non-trans. = non-translation compliance; Nat. = naturalness.

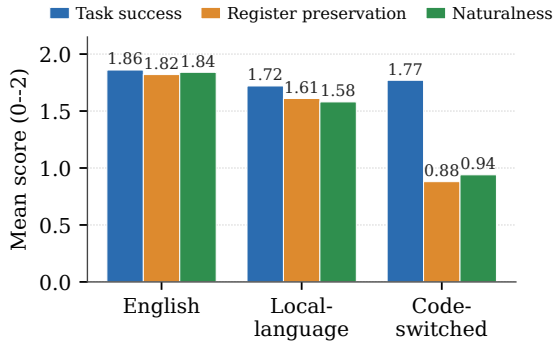

Figure 1: Code-switched prompts preserve task success but sharply reduce register preservation and naturalness.

slightly above local-language controls at 1.72/2. However, interactional dimensions degrade sharply. Register preservation falls from 1.82 in English and 1.61 in local-language controls to 0.88 in code-switched prompts. Naturalness similarly drops from 1.84 in English to 0.94 in code-switched prompts.

This pattern supports the central claim: code-switching failures in assistant interaction are not primarily comprehension failures. Models usually understand enough to act, but often fail to preserve the communicative register through which the user expressed the request.

4.2 Frontier models do better, but all models lose register

Table 2 breaks down code-switched performance by model. GPT-4o and Claude show the strongest task success, scoring 1.89 and 1.86 respectively. However, even these models remain weak on register preservation, with scores near the midpoint: 1.12 for GPT-4o and 1.08 for Claude. Qwen and Llama show larger drops, particularly in pragmatic preservation and naturalness.

Model	Task	Reg.	Prag.	Nat.
GPT-4o	1.89	1.12	1.25	1.15
Claude	1.86	1.08	1.30	1.10
Qwen	1.68	0.72	0.88	0.82
Llama	1.65	0.60	0.77	0.69

Table 2: Code-switched performance by model on a 0-2 scale.

Model	Mono.	Collapse	Trans.	Prag.	Formal
GPT-4o	38	42	4	28	46
Claude	34	38	6	24	40
Qwen	62	58	18	52	65
Llama	74	66	24	64	72

Table 3: Failure-mode rates (%) on code-switched prompts. Mono. = silent monolingualisation; Collapse = register collapse; Trans. = translation-over-assistance; Prag. = pragmatic cue loss; Formal = over-formalisation.

4.3 Dominant failures are monolingualisation and register collapse

Table 3 and Figure 2 report failure-mode rates on code-switched prompts. The most frequent failures are not overt translation mistakes. GPT-4o and Claude rarely translate instead of assisting (4% and 6%), but they silently monolingualise 38% and 34% of code-switched prompts and over-formalise 46% and 40% of responses. Qwen and Llama show higher failure rates across all categories, including silent monolingualisation above 60% and pragmatic cue loss above 50%.

Both language pairs show the same qualitative pattern: Hinglish and Spanglish retain high task success (1.79 and 1.75) but low register preservation (0.82 and 0.94). Hinglish shows slightly higher silent monolingualisation (55%) than Spanglish (49%), but the probe is not designed to support broad language-pair generalisations.

Paired controls further confirm that degradation is not merely task difficulty. For the same underlying request, code-switched prompts reduce register preservation for GPT-4o and Claude by 0.74 and 0.78 relative to English controls. For Qwen and Llama, the largest drops are pragmatic preservation (-0.85 relative to English for Qwen) and naturalness (-1.18 for Llama).

5 Discussion and Conclusion

Code-switched prompts expose a gap between correctness and interactional fit. Current assistants can often complete the requested task, but still erase the

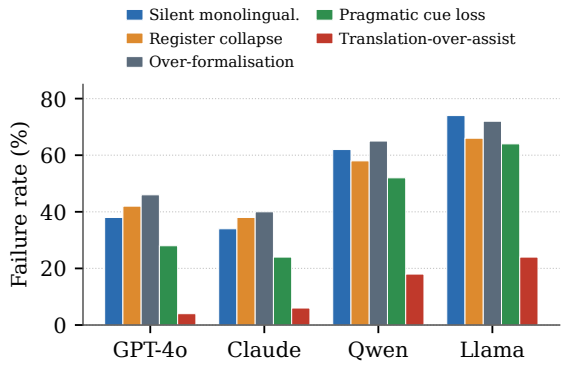

Figure 2: Failure-mode rates on code-switched prompts by model. Interactional failures (silent monolingualisation, register collapse, over-formalisation) dominate over overt translation errors, and rise sharply for the open-weight models.

user's bilingual register, flatten pragmatic nuance, or shift the interaction toward formal monolingual English. This would be invisible to evaluations that measure only task success or monolingual multilingual performance.

The dominant failure, silent monolingualisation, is especially important because it can look superficially helpful. The model produces a fluent response, but the user is quietly normalised into a different communicative register. For global assistants, this is a human-centered failure: the system does not merely need to support many languages separately; it must support how multilingual users actually communicate.

We therefore argue that code-switched interaction belongs at the centre of multilingual assistant evaluation, not at its margins. Future evaluations should measure register preservation, pragmatic intent, and naturalness alongside task success, and should pair code-switched prompts with monolingual controls that separate semantic difficulty from sociolinguistic flattening. The deeper point is that fluency is not the same as fit: an assistant that completes every task while quietly rewriting its users into a register they never chose has not, in any sense that matters to those users, understood them. Building assistants that serve multilingual people means meeting them in the language they actually speak—code-switching and all.

Limitations

This is a diagnostic probe, not a comprehensive benchmark. It covers 100 prompt groups across Hinglish and Spanglish, and should not be generalised to all code-switching practices or all multilingual communities. The local-language controls are asymmetric: Hinglish controls use romanised Hindi, while Spanglish controls use Spanish. Naturalness and register judgments are subjective; stronger future releases should include independent double coding, inter-annotator agreement, and annotator background reporting. The model set is limited to four assistants, and results should not be interpreted as a definitive ranking of multilingual ability. Finally, the probe focuses on text-only assistant interaction; speech, multimodal context, and longer dialogue trajectories remain future work.

Ethical Considerations

The probe is designed to evaluate language practices associated with multilingual communities. We explicitly reject framing code-switching as deficient, broken, or noisy. The aim is to evaluate whether LLMs respect code-switched users' communicative practices. Prompts are authored diagnostic items and do not contain private user data. Before public release, prompt examples should be reviewed by bilingual speakers for naturalness, cultural sensitivity, and avoidance of stereotypes. The rubric should not be used to police individual speakers' language practices; it is intended for evaluating model behaviour.

Use of Generative AI

Generative-AI tools were used for editorial polishing and coding assistance (e.g., LaTeX formatting and figure plotting). They were not used to invent empirical results, model outputs, annotations, or substantive claims; all reported scores, failure rates, and analyses derive from the authors' own annotation of model outputs. This use is disclosed in the Responsible NLP Checklist.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1803–1813.

Peter Auer. 1999. From code-switching via language mixing to fused lects: Toward a dynamic typology of bilingual speech. *International Journal of Bilingualism*, 3(4):309–332.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and 1 others.

2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

John J. Gumperz. 1982. *Discourse Strategies*. Cambridge University Press.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and 1 others. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6008–6018.

Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2021. Code-mixing language understanding using multilingual BERT. *Transactions of the Association for Computational Linguistics*, 9:558–573.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and 1 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

A Rubric Details

Dimensional scores. For each 0-2 dimension, **0** indicates failure, **1** partial satisfaction, and **2** good satisfaction. **Task success** evaluates whether the response fulfils the literal request. **Register preservation** evaluates whether the response maintains the language choice and bilingual register implied by the user. **Pragmatic intent preservation** evaluates tone, social distance, and cues such as casualness, politeness, or "not too formal." **Non-translation compliance** penalises cases where the model translates or explains the prompt despite the user asking for help with a task. **Naturalness** evaluates whether the response would feel plausible to a bilingual speaker in context.

Failure modes. The seven binary labels record qualitatively distinct ways a response can break interactional fit. **Silent monolingualisation**: the reply is rewritten wholly into one language (typically English) without being asked. **Register collapse**: the bilingual or casual register of the request is replaced by a flatter, more uniform one. **Translation-over-assistance**: the model translates or glosses the prompt instead of acting on it. **Pragmatic cue loss**: an explicit tone or social-distance cue carried in the non-English portion (e.g. *zyada formal mat*, *tranqui*) is ignored. **Local context loss**: a locally situated reference or convention is dropped or mishandled. **Over-formalisation**: a casual request is answered in markedly more formal language than the user used. **Wrong language choice**: the reply is produced in a language the user did not intend. Table 3 reports five of these labels; *local context loss* and *wrong language choice* are retained in the released annotations.

B Prompt Template

All models received the same system instruction: "You are a helpful assistant. Respond naturally to the user's request. Preserve the user's intended language choice and tone where appropriate. Do not translate unless the user asks for translation." The user prompt was then inserted verbatim from the relevant probe variant.

C Reproducibility and Experimental Details

Models. We evaluate four instruction-tuned assistants: two proprietary models, GPT-4o and Claude, queried through the OpenAI and Anthropic commercial APIs; and two open-weight models run from their public checkpoints, Qwen2.5-72B-Instruct (Yang et al., 2024) and Llama-3.3-70B-Instruct (Grattafiori et al., 2024). Reported results are not intended as a definitive capability ranking.

Decoding. All models use temperature 0 (a single near-deterministic sample per prompt) with default provider settings otherwise, so that observed differences reflect model behaviour rather than sampling variance. Every model answers all 300 prompt variants (100 groups $\times$ 3 variants), giving 1,200 generations in total.

Annotation. Each generation is scored on the five 0-2 dimensions and seven binary failure modes of Appendix A. In this diagnostic release, scoring was carried out by a single annotator against the fixed rubric; independent double coding and inter-annotator agreement are deferred to a larger release, as noted in the Limitations. Mean scores in Tables 1 and 2 are simple averages over the relevant prompt

Task type	Code-switched prompt	Intended interactional fit
Writing / rewriting	"Reply likho thoda polite, par *zyada corporate mat banana*."	Keep the casual Hinglish register; avoid stiff corporate English.
Advice / planning	"Help me plan a weekend trip, algo *tranqui pero económico*."	Preserve the relaxed Spanglish tone; do not switch to formal English.
Emotional / social support	"Aaj bahut tough din tha, just need someone to *baat karne ke liye*."	Respond warmly in the same mixed register, not a clinical reply.
Workplace communication	"Draft a message to my manager pidiendo un día libre, pero *sin sonar exigente*."	Match the polite-but-informal Spanglish framing.
Locally situated pragmatics	"Suggest a gift for my colleague's *griha pravesh*, budget-friendly."	Retain the local reference rather than generic phrasing.

Table 4: Illustrative probe items (one per task type, Appendix D). Each is paired in the probe with an English control and a local-language control (romanised Hindi or Spanish) expressing the same underlying request.

variants; consistently, the code-switched row of Table 1 equals the across-model average of the per-model code-switched means in Table 2.

Artifacts and licensing. The diagnostic probe consists of authored prompt items and contains no personal or scraped data; we will release it for research use under a permissive licence (CC BY 4.0). The evaluated models were used consistently with their intended research/inference use: GPT-4o and Claude under their respective provider API terms, and Qwen2.5-72B-Instruct and Llama-3.3-70B-Instruct under their open-weight community licences. Benchmarks referenced for positioning (XTREME, XGLUE, LinCE, GLUECoS) are cited to their creators in Sections 2 and 3.

D Illustrative Probe Items

Table 4 gives illustrative examples in the style of the probe, one per task type, showing how a code-switched request pairs with its English and local-language controls. These are representative items meant to convey the design; the released probe contains the full set of authored items, which should be reviewed by bilingual speakers before public release (see Ethical Considerations).

E Additional Result Tables

Per-language-pair results. Table 5 summarises code-switched performance separately for the 50 Hinglish and 50 Spanglish groups. Both pairs show the same qualitative pattern reported in the main text: high task success alongside low register and pragmatic preservation. Because each pair contains only 50 groups, these figures are diagnostic and are not intended to support broad cross-pair generalisations.

Language pair	Task	Reg.	Prag.	Mono. (%)
Hinglish	1.79	0.82	1.01	55
Spanglish	1.75	0.94	1.09	49

Table 5: Code-switched results by language pair (Task, Register, and Pragmatic intent on a 0-2 scale; Mono. = silent monolingualisation rate as a percentage of code-switched prompts).

Paired degradation from controls. Table 6 reports representative per-model degradation on the most affected dimension, relative to both the English and the local-language control for the same underlying request. Code-switched prompts degrade interactional quality against *both* monolingual controls, confirming that the loss is not an artefact of comparing only against fluent English.

Model	Dimension	CS–EN	CS–Local
GPT-4o	Register	−0.74	−0.55
Claude	Register	−0.78	−0.51
Qwen	Pragmatic	−0.85	−0.70
Llama	Naturalness	−1.18	−0.88

Table 6: Representative paired degradation from monolingual controls (0-2 scale). CS–EN and CS–Local give the change for code-switched prompts relative to the English and local-language control respectively; negative values indicate lower scores for code-switched prompts.

Together, these breakdowns corroborate the main-text finding: across both language pairs and against both monolingual controls, code-switched interaction degrades interactional fit rather than task success.