Enabling Structured Reasoning in Sindhi with Culturally Grounded Instruction Tuning

Mehak Mehak[1], Kamyar Zeinalipour[1]*, Pireh Soomro[2], Christiano Chesi[1], Marco Gori[1], Marco Maggini[1]
[1]University of Siena
[2]Mehran University of Engineering & Technology
kamyar.zeinalipour@unisi.it

Abstract

While Large Language Models (LLMs) excel in high-resource contexts, reasoning capabilities in low-resource languages (LRLs) like Sindhi remain limited. To bridge this gap, we introduce *Sindhi-Reasoning-Instruct*, the first culturally grounded Sindhi instruction corpus. We fine-tuned six LLaMA and Mistral models (1B–24B) to evaluate if parameter-efficient tuning enables deductive, inductive, and causal reasoning. Results demonstrate that linguistically authentic data is the decisive factor. Fine-tuning effectively restored Sindhi's Perso-Arabic orthography and SOV structure, with the Mistral-Small-24B model achieving a massive 141% relative improvement in human quality ratings over its base version. Furthermore, structured reasoning capabilities were found to scale with model size; while smaller models achieved high fluency, Mistral-Small-24B achieved top performance across logical categories, reaching 83% on inductive reasoning tasks. This study provides empirical evidence that expert-curated, native instruction data allows LRL models to move beyond simple translation toward robust, structured reasoning. The dataset and models are publicly available at https://github.com/KamyarZeinalipour/Sindhi-Reasoning.

1 Introduction

The advent of Large Language Models (LLMs) has revolutionized Natural Language Processing (NLP), yet a significant "digital divide" persists. Progress remains concentrated in high-resource languages, while Low-Resource Languages (LRLs) face severe data scarcity (Zhang et al., 2025). Crucially, while instruction tuning has improved basic LRL tasks, it has largely neglected complex reasoning capabilities.

This gap is acute in Sindhi, an Indo-Aryan language spoken by over 40 million people. Sindhi presents unique computational challenges, including a complex Perso-Arabic script and strict Subject-Object-Verb (SOV) morphology (Rahman Arain and Bhatti, 2010). Existing resources are limited to lower-level processing like word embeddings (Ali et al., 2024), lacking frameworks for structured deductive or causal inference. Consequently, current multilingual models fail to generalize to Sindhi's specific logical structures (Ponti et al., 2020), creating a "resource-reasoning gap."

To address this, this study aims to answer the following main research question: *How can a holistic framework, integrating linguistically grounded resource creation and parameter-efficient model adaptation, bridge the gap to enable structured reasoning in Sindhi?* This overarching inquiry is

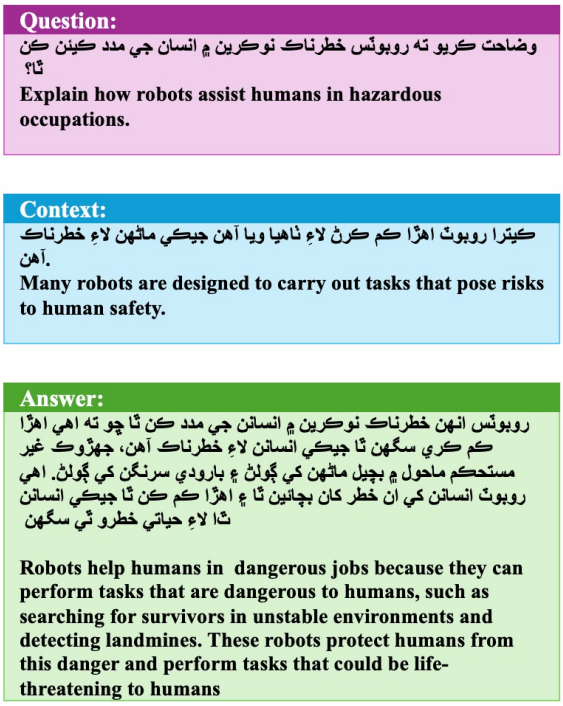

Figure 1: A sample from the Sindhi-Reasoning-Instruct dataset illustrating a commonsense reasoning task in Sindhi.

*Corresponding author

investigated through four specific Research Questions (RQs):

- **RQ1 (Data Efficacy):** What is the role of culturally grounded, native-instruction data in mitigating the "resource-reasoning gap" for morphologically complex languages like Sindhi?
- **RQ2 (Adaptation Capabilities):** To what extent can parameter-efficient instruction tuning (LoRA) enable English-centric architectures (LLaMA/Mistral) to perform structured deductive and causal reasoning in a non-Latin script?
- **RQ3 (Linguistic Alignment):** How does instruction tuning impact the model's ability to restore authentic syntactic properties, specifically Sindhi's SOV word order and Perso-Arabic orthography?
- **RQ4 (Scaling Laws):** How do structured reasoning capabilities scale with model parameter size in a low-resource setting, and does a critical threshold exist for complex inference?

Our approach centers on *Sindhi-Reasoning-Instruct*, the first fully annotated Sindhi reasoning corpus (2,190 pairs), derived from encyclopedic and native news sources. As illustrated in Figure 1, this dataset transforms factual contexts into structured logic tasks. We fine-tuned six LLaMA and Mistral models (1B–24B) using Low-Rank Adaptation (LoRA).

Results demonstrate that linguistically authentic data is decisive. Fine-tuning effectively restored Sindhi's orthography and SOV structure, with the Mistral-Small-24B model achieving a 141% increase in human quality ratings. Furthermore, reasoning capabilities scaled with model size; Mistral-Small-24B attained 83% accuracy on inductive reasoning tasks. This study provides empirical evidence that expert-curated instruction data allows LRL models to move beyond simple translation toward robust, structured reasoning.

2 Related Work

This work intersects three areas: instruction tuning for low-resource languages, cross-lingual reasoning, and Sindhi NLP.

Instruction Tuning for Low-Resource Languages Instruction tuning aligns LLMs with human intent (Vaillancourt and Thompson, 2024), yet data scarcity hinders its application to low-resource languages (LRLs). Kohli et al. (2023) fine-tuned LLaMA for Bengali (BengaliGPT) but faced grammatical and reasoning limitations. Similarly, MURI (Köksal et al., 2024) generated datasets via reverse instructions, though often prioritizing generic tasks over logical inference. While frameworks like LinguaLIFT (Zhang et al., 2024) propose two-stage tuning to enhance reasoning, such methods remain untested on languages with complex morphosyntax like Sindhi.

Cross-Lingual and Multilingual Reasoning Transferring reasoning capabilities to LRLs is a major challenge. Sheng et al. (2025) demonstrated that models struggle to generalize abductive, deductive, and inductive rules to unseen languages. Ponti et al. (2020) confirmed this via XCOPA, showing poor zero-shot transfer for causal reasoning. Furthermore, deductive reasoning remains imperfect for languages with ambiguous structures (Ramji and Ramji, 2025).

Sindhi Natural Language Processing Sindhi NLP is currently limited to foundational resources. Early work focused on rule-based morphology (Rahman Arain and Bhatti, 2010), while recent studies developed word embeddings (Ali et al., 2021, 2024) and surveyed available tools (Sodhar et al., 2023). Crucially, no prior work addresses higher-order reasoning for Sindhi; existing resources are confined to basic tasks like POS tagging and sentiment analysis.

Gap Analysis and Contribution We address these gaps by: (1) targeting structured reasoning (deductive, inductive, causal, commonsense) rather than generic tuning; (2) generating native, culturally grounded data to mitigate zero-shot transfer failures (Ponti et al., 2020); and (3) introducing the first Sindhi reasoning dataset, demonstrating that parameter-efficient fine-tuning successfully adapts LLMs to Sindhi's unique orthography and morphosyntax (Sodhar et al., 2023).

3 Dataset Creation

To address the scarcity of reasoning-focused resources for Sindhi, we developed the *Sindhi-Reasoning-Instruct* corpus. Figure 2 presents an overview of the full pipeline, from data sourcing

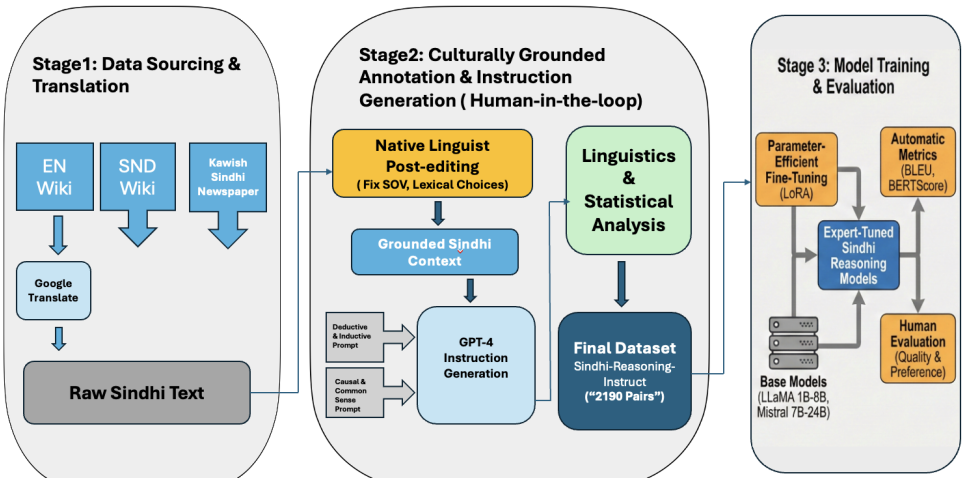

Figure 2: The proposed methodological pipeline for Sindhi-Reasoning-Instruct. The workflow consists of three stages: (1) Data Sourcing Translation, (2) Human-in-the-loop Annotation Instruction Generation, and (3) Parameter-Efficient Fine-Tuning (LoRA) and Evaluation.

to model evaluation. Unlike prior datasets based on direct translation of English benchmarks, our approach emphasizes cultural grounding and linguistic authenticity through a human-in-the-loop pipeline.

3.1 Data Sourcing and Composition

The dataset is constructed from 1,200 high-quality text passages sourced from two distinct domains to ensure diverse reasoning contexts:

1. **Encyclopedic Domain (80%):** We selected articles from Sindhi Wikipedia; however, due to its limited size, we additionally sourced content from English Wikipedia covering Science, History, and Sociology. These texts provide the factual density required for *deductive* and *inductive* reasoning.

2. **News Domain (20%):** We curated articles from *Daily Kawish*, a prominent Sindhi newspaper. These texts capture contemporary discourse and local events, essential for *causal* and *commonsense* reasoning in a native context.

3.2 The Translation and Adaptation Pipeline

We adopted a "Translate-then-Edit" approach for the English encyclopedic portion.

1. **Initial Translation:** English source texts were translated into Sindhi using **Google Translate** to establish a baseline draft.

2. **Human Post-Editing:** Native Sindhi linguists manually reviewed the machine-translated outputs. This stage was critical to correct:

   - **Lexical Errors:** Replacing loanwords with authentic Sindhi terminology (e.g., correcting "school" to *"madrasa"* depending on context).
   - **Grammatical Alignment:** Enforcing the correct Subject-Object-Verb (SOV) structure, which machine translation models often drift from when translating from SVO languages like English.

The necessity of our human-in-the-loop pipeline is evident when analyzing the raw machine translation errors, which are detailed alongside our correction protocols in Appendix C.

3.3 Instruction Generation Framework

To transform these raw texts into reasoning tasks, we employed a semi-automated generation method using GPT-4, guided by specific prompts for four reasoning schemas:

- **Deductive Reasoning**: Applying a general rule to a specific case.
- **Inductive Reasoning**: Deriving a general rule from specific cases.
- **Causal Reasoning**: Identifying a cause and its effect.
- **Commonsense Reasoning**: Using everyday knowledge to infer something.

To ensure the generation of distinct logical structures, we employed a strict prompting strategy using two distinct templates: one tailored for **Induc-**

tive and Deductive reasoning, and a separate template for **Causal and Commonsense** tasks. The exact system prompts used for generation are provided in Appendix A. This process yielded a dataset of 2,190 instruction-context-output triplets, all of which were used for evaluation.

3.4 Human Annotation and Quality Assurance

To ensure the dataset represents authentic Sindhi rather than synthetic translations and preserve Sindhi reasoning structure, we implemented a rigorous annotation pipeline involving two native Sindhi speakers, both holding university degrees in Linguistics. This academic background ensured a deep theoretical understanding of Sindhi morphosyntax and phonology beyond native intuition. The process followed a strict cross-verification protocol: one linguist performed the initial corrections, the second verified the output, and any disagreements were resolved through collaborative discussion to reach consensus. The annotation proceeded in two distinct stages:

Stage 1: Logical Verification Annotators assessed the logical validity of the generated pairs. They removed "hallucinated" constraints where the model invented facts not present in the context. Approximately 12% of the generated samples were discarded or rewritten at this stage due to logical inconsistencies.

Stage 2: Linguistic Refinement The accepted pairs underwent a final linguistic pass. Annotators focused on *orthographic normalization*. Sindhi is written in a Perso-Arabic script with 52 characters, including unique implosives (e.g., */bb/, /dd/, /gg/*) and aspirates (e.g., */bh/, /dh/*). We identified specific phonological markers that base models frequently mishandle; a complete inventory of these special characters is listed in Appendix B. Automated models often substitute these with their Urdu or Arabic equivalents. Annotators manually corrected these character substitutions to ensure the dataset serves as a gold standard for Sindhi orthography.

3.5 Cultural Alignment

To ensure the dataset reflects the Sindhi sociolinguistic reality rather than Western biases, we employed a "Localization-over-Translation" strategy during annotation:

- **Entity Adaptation:** Terms were contextually mapped to local equivalents (e.g., generic "school" to *"madrasa"* where appropriate).
- **Geographic Grounding:** Physical and temporal constraints (e.g., seasons, agricultural cycles) were adjusted to match the geography of the Sindh region (see Appendix C).
- **Normative Reasoning:** By integrating local news (*Daily Kawish*), causal reasoning tasks implicitly encode native social norms rather than defaulting to the individualistic perspectives common in English training data.

4 Linguistic, Semantic, and Statistical Analysis

To ensure the *Sindhi-Reasoning-Instruct* corpus provides a robust foundation for reasoning tasks, we conducted a comprehensive quantitative analysis. This examination investigates both the global linguistic properties of the dataset and the specific structural variations across different reasoning components.

4.1 Automated Analysis Framework

We employed **Stanza v1.5** (Qi et al., 2020) (model `sd`) to process the corpus. Structural complexity was quantified using **Syntactic Tree Depth** (max dependency path). This uniform pipeline generated all token counts and Type-Token Ratios (TTR) presented in Table 1.

4.2 Corpus Statistics

The dataset comprises three components: Context (*Text*), Instruction (*Question*), and Response (*Final Answer*). As shown in Table 1, each component has a distinct statistical profile aligned with its role. **Contexts** are information-dense (mean 232.4 tokens), **Questions** are concise (mean 29.8 tokens) and serve as focused reasoning prompts, and **Responses** are substantive (mean 169.1 tokens), reflecting multi-step reasoning rather than extractive labels. Figure 4 further confirms this separation, showing clear distributional differences between concise instructions and verbose, high-context reasoning components.

4.3 Task-Specific Analysis

We stratified the analysis by reasoning type to understand the structural demands of each task, as

Metric	Context	Instruction	Response
Total Tokens	508,785	65,373	370,318
Mean Tokens	232.43	29.86	169.17
Median Tokens	223.00	28.00	170.00
Std. Deviation	98.70	17.31	98.66
Unique Lemmas	14,497	2,641	12,236
MATTR (w=50)	0.798	0.763	0.785

Table 1: Global corpus statistics. The high token count in responses reflects the generative nature of the reasoning tasks.

detailed in Table 2. The dataset maintains a relatively balanced distribution, with Deductive and Inductive reasoning comprising the largest segments (N=597 each), followed by Causal and Commonsense tasks (N=498 each). Regarding input complexity, Deductive and Inductive tasks utilize significantly longer contexts (≈242 tokens) compared to Causal or Commonsense tasks. Deductive tasks also require the most verbose instructions (15.2 tokens), reflecting the need for explicit rule specification. Syntactically, Inductive prompts exhibit the highest complexity (Tree Depth: 5.55), a linguistic necessity for embedding multiple specific observations before soliciting a general hypothesis. Conversely, Causal reasoning demands the most complex outputs; its answers exhibit the highest tree depth (3.80) and substantial length (62.2 tokens), indicating that articulating cause-and-effect relationships requires more complex sentence structures (e.g., subordination) than the constrained outputs of logical entailment.

Reasoning Type	Count	Mean Tokens			Tree Depth		
		Instr	Ctx	Ans	Instr	Ctx	Ans
Deductive	597	15.2	242.0	53.1	5.41	4.55	3.10
Inductive	597	12.8	242.0	58.4	5.55	4.55	3.25
Causal	498	14.5	221.2	62.2	5.09	4.77	3.80
Commonsense	498	11.3	220.6	64.4	5.28	4.77	3.15

Table 2: Comprehensive dataset statistics ($N = 2,190$). We report the sample size (**Count**), average length (**Mean Tokens**), and linguistic complexity (**Tree Depth**) for Instructions (*Instr*), Contexts (*Ctx*), and Answers (*Ans*).

4.4 Lexical Richness and Diversity

To evaluate vocabulary diversity independent of text length, we employed the Moving-Average Type-Token Ratio (MATTR) with a window of 50 tokens. Global analysis reveals high richness across components: **Context** scores highest (0.798) due to dense informational sources, followed closely by **Response** (0.785), reflecting descriptive reasoning rather than repetition. **Instruction** diversity is lowest (0.763), as reasoning prompts rely on a closed set of directive verbs (e.g., *"budhayo"* - explain, *"tijziyo karyo"* - analyze) to unambiguously trigger tasks.

Table 3 details task-wise variations. Wikipedia-based tasks (*Deductive*, *Inductive*) share high Context richness (0.809), surpassing News-based tasks (≈0.783) which utilize accessible journalistic language. regarding instructions, *Deductive Reasoning* shows the highest diversity (0.761) by incorporating specific entities and constraints. Conversely, *Commonsense* prompts score lowest (0.726) due to highly standardized phrasing (e.g., *"Inhen amal jo samaaji natijo cha thindo?"* - What will be the social result of this action?).

Reasoning Type	Context	Instruction	Response
Deductive	0.809	0.761	0.796
Inductive	0.809	0.740	0.801
Causal	0.783	0.748	0.790
Commonsense	0.783	0.726	0.794

Table 3: Lexical Richness (MATTR, w=50) by Component and Task.

4.5 Morpho-Syntactic Profile

We analyzed Part-of-Speech (POS) distribution to verify linguistic authenticity. Across the corpus, the distribution reflects Sindhi's typological signature: **Nouns** consistently constitute ≈30% of tokens, confirming the dataset is entity-centric and informationally dense. **Adpositions** (ADP) account for ≈17.3% of Contexts, validating that our pipeline preserved Sindhi's postpositional nature (e.g., *khān, wat, te*) rather than English prepositions. Finally, the combined presence of **Verbs** (8.8%) and Auxiliaries (5.0%) supports the prevalence of compound verb structures (e.g., *karyo wye* - was done) essential for tense and aspect.

Component-specific analysis (Figure 3) reveals functional shifts. **Instructions** exhibit a marked spike in Verbal density (12.1%) compared to Contexts (8.8%), aligning with the use of imperative verbs (e.g., *"budhayo"* - explain, *"pish-goi karyo"* - predict) to trigger reasoning. Instructions also show the highest **Adposition** frequency (20.2%), highlighting the relational nature of reasoning where postpositional markers are required to define complex logical dependencies (e.g., "From X, deduce Y").

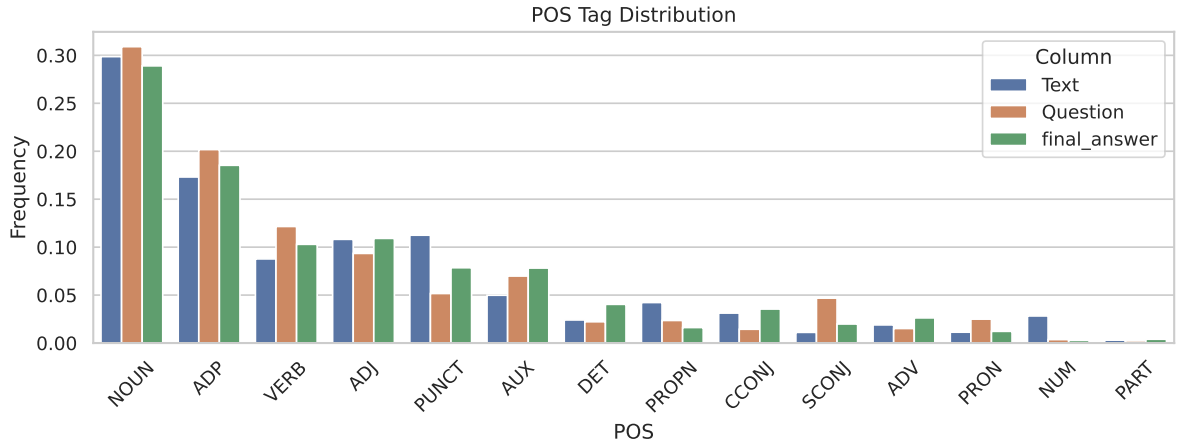

Figure 3: Comparative Part-of-Speech (POS) distribution showing the functional distinction between Entity-heavy Contexts (Noun) and Action-oriented Instructions (Verb/ADP).

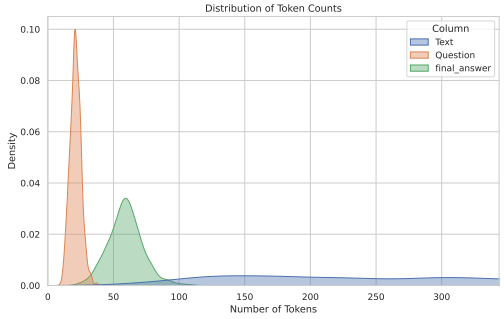

Figure 4: Token length distributions showing the separation between concise instructions and verbose reasoning contexts.

4.6 Qualitative Analysis

To ensure the dataset serves as a gold standard for reasoning, we conducted a manual linguistic verification. **Morphologically**, the corpus preserves Sindhi's rich inflectional system, maintaining consistent verbal agreement and auxiliary constructions (e.g., aspectual markers) essential for descriptive reasoning. **Syntactically**, the data strictly adheres to the canonical Subject-Object-Verb (SOV) order, preserving verb-final structures even within complex, multi-clause explanations. **Semantically**, logical relations are explicitly anchored using distinct discourse markers specific to deductive, inductive, and causal inference. This ensures that the model learns reasoning through explicit linguistic cues rather than implicit context. A comprehensive linguistic breakdown, including suffix frequency analysis and discourse marker distributions, is provided in Appendix F.

5 Experiments and Results

To validate our hypotheses on Sindhi reasoning, we conducted a comprehensive experimental study comparing base (off-the-shelf) large language models with their instruction-tuned counterparts, evaluating improvements in linguistic fidelity (morphology and syntax), reasoning capabilities (deductive, inductive, causal, and commonsense), and semantic coherence.

5.1 Experimental Setup

Model Selection To address RQ4 regarding scaling laws, we selected six open-weight models spanning 1B to 24B parameters. The lineup includes the **LLaMA** family (3.2-1B, 3.2-3B, 3.1-8B) to test efficiency in constrained settings, and the **Mistral** family (7B v0.3, Nemo-12B, Small-24B) to evaluate the emergence of deeper reasoning capabilities at higher capacities.

Fine-Tuning Methodology We fine-tuned all models on the *Sindhi-Reasoning-Instruct* corpus using Low-Rank Adaptation (LoRA) to ensure parameter efficiency. Experiments were conducted on three NVIDIA A6000 (48GB) GPUs using the AdamW optimizer with a cosine learning rate schedule ($1e^{-4}$). To ensure robust adaptation to the Sindhi script and logic, we applied LoRA adapters not only to attention layers but to all linear modules, including embeddings and the language model head. Training was performed for three epochs with BF16 precision and Flash Attention 2 to optimize throughput. For reproducibility, the complete configuration of hyperparameters, LoRA settings,

and optimization details are provided in Appendix E.

5.2 Evaluation Methodology

We adopted a dual evaluation strategy combining automatic metrics with rigorous human assessment.

Automatic and Human Protocols For automatic evaluation, we employed standard Natural Language Generation (NLG) metrics to quantify improvements, utilizing ROUGE-L, BLEU, and TF-IDF for word overlap, alongside E5 Similarity and BERTScore to capture embedding-based semantic alignment. To further assess reasoning nuances and cultural appropriateness, we conducted a blind human evaluation on 600 randomly sampled outputs from the dataset with two native Sindhi speakers. [1] This protocol comprised two distinct tasks: an absolute **Quality Rating** on a 5-point scale (A to E) based on linguistic accuracy and logic, and a **Preference Selection** where annotators blindly compared the Base and Expert-Trained models. For a detailed breakdown of the annotation guidelines and the specific rating rubric used, please refer to Appendix G.

5.3 Results

Quantitative Analysis (Automatic Metrics) Automatic evaluation revealed substantial gains across all model sizes. As detailed in Table 4, word overlap metrics saw the most dramatic increases. For instance, the BLEU score for the smallest model (LLaMA-1B) increased by over 136%, while the largest model (Mistral-24B) achieved the highest absolute scores across all metrics. Semantic metrics like BERTScore also showed consistent improvements, confirming that fine-tuned models generated responses semantically closer to the reference gold standard.

Qualitative Analysis (Human Evaluation) Human annotators observed base models either failed to generate any response or produced numerous errors, such as incorrect gender agreement and verb–argument agreement, whereas the fine-tuned model's outputs preserved meaning, logical consistency, and contextual relevance. Annotators worked collaboratively and cross-checked all evaluations to ensure accuracy, consistently preferring the fine-tuned models. As shown in Figure 5, expert-trained models achieved a significantly higher proportion of "A" (Excellent) and "B" (Good) ratings, while the base models frequently fell into the "D" (Poor) or "E" (Unacceptable) categories.

The preference evaluation further corroborated this trend. Figure 6 illustrates that annotators favored the expert-trained model's response in the vast majority of cases. Notably, for the Mistral-Small-24B model, the expert-tuned version was preferred in nearly 100% of instances, reflecting its superior ability to generate coherent and logically sound Sindhi text.

To quantify these improvements, we mapped the categorical human ratings to a numerical scale, assigning a value of 5 to 'A' (Excellent) down to 1 for 'E' (Unacceptable). Table 5 presents the relative improvement in mean rating scores derived from this mapping. The Mistral-Small-24B model achieved the highest relative improvement of 140.94%, moving from a base score of 1.49 to 3.59.

Performance by Reasoning Type Evaluating specific reasoning capabilities revealed that model size played a critical role. As shown in Table 6, Mistral-Small-24B emerged as the top performer across all categories, validating the capability of larger models to handle complex inference even in low-resource languages.

6 Conclusion

This study introduces *Sindhi-Reasoning-Instruct*, the first expert-curated dataset designed to bridge the reasoning gap in Sindhi NLP. Addressing our main research question, we conclude that parameter-efficient instruction tuning successfully enables large language models to perform structured reasoning in Sindhi, provided the training data is linguistically grounded.

Regarding linguistic alignment (**RQ3**), results confirm that fine-tuning is essential for linguistic restoration; the elimination of script-mixing and the achievement of near-native morphological accuracy demonstrate that parameter-efficient tuning prevents the garbled multilingual output observed in base models. Furthermore, addressing the "resource-reasoning gap" (**RQ1 & RQ2**), the sharp decline in hallucinations provides empirical evidence that culturally grounded, expert-curated data is the decisive factor in enabling models to move beyond probabilistic guessing toward grounded logical inference. Finally, regarding scaling laws

[1] We observed substantial agreement between annotators, with a Cohen's Kappa score of $\kappa = 0.78$ for quality ratings and $\kappa = 0.82$ for pairwise preferences.

Model	ROUGE-L		BLEU		TF-IDF		E5 Sim		BERTScore	
	Base	Tuned	Base	Tuned	Base	Tuned	Base	Tuned	Base	Tuned
Mistral-Small-24B	0.288	**0.441**	0.123	**0.257**	0.170	**0.437**	0.892	**0.948**	0.864	**0.922**
Mistral-Nemo	0.307	0.419	0.133	0.234	0.225	0.395	0.893	0.941	0.871	0.915
Mistral-7B v0.3	0.317	0.426	0.126	0.247	0.266	0.421	0.886	0.940	0.877	0.916
LLaMA-3.1-8B	0.363	0.431	0.170	0.247	0.304	0.418	0.914	0.943	0.895	0.918
LLaMA-3.2-3B	0.296	0.364	0.119	0.182	0.198	0.322	0.892	0.918	0.867	0.901
LLaMA-3.2-1B	0.225	0.331	0.067	0.159	0.141	0.264	0.863	0.894	0.850	0.890

Table 4: Comparative Automatic Evaluation results showing Base vs. Tuned performance across six models. Best performing scores are highlighted in bold.

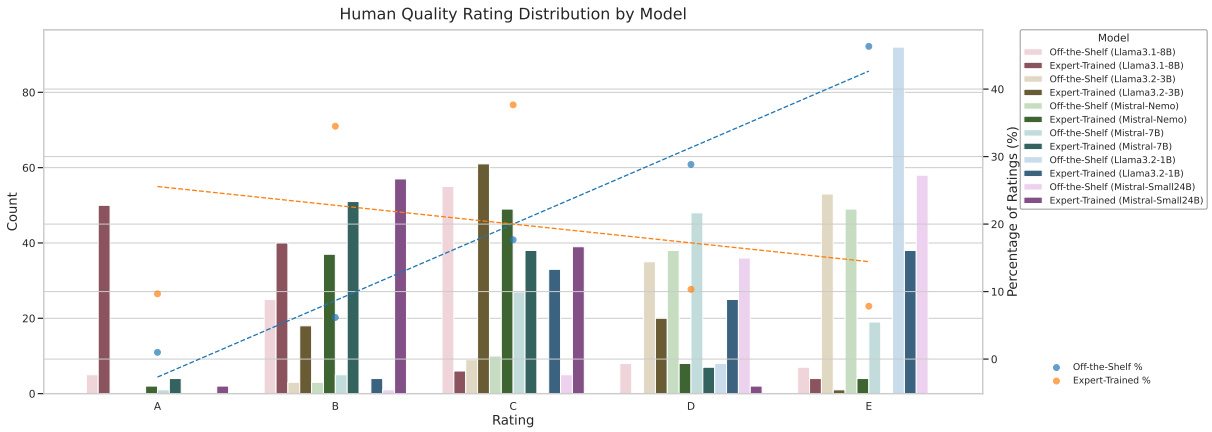

Figure 5: Rating Distribution by Model. Expert-Trained models (solid bars) consistently achieve higher quality ratings (A/B) compared to Off-the-Shelf baselines (faded bars).

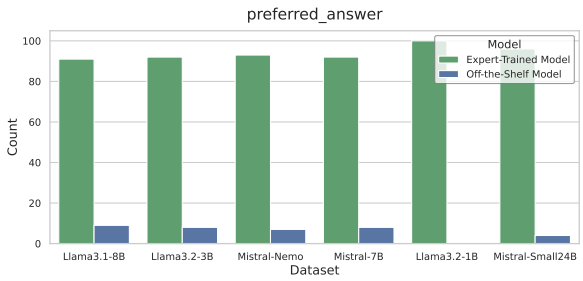

Figure 6: Annotator Preference Counts. The green bars indicate the overwhelming preference for Expert-Trained models over Off-the-Shelf models (blue bars) across all architectures.

Model	Tuned Score	Base Score	Δ	% Improv.
LLaMA 3.1 8B	4.32	3.13	1.19	38.02%
LLaMA 3.2 1B	2.03	1.08	0.95	87.96%
LLaMA 3.2 3B	2.96	1.62	1.34	82.72%
Mistral 7B	3.52	2.21	1.31	59.28%
Mistral Nemo	3.25	1.67	1.58	94.61%
Mistral 24B	3.59	1.49	2.10	140.94%

Table 5: Percentage of improvement in human ratings after fine-tuning.

Reasoning Type	Best Model (Mistral 24B)	Runner-up
Inductive	83%	Mistral 7B (77%)
Deductive	80%	LLaMA 8B (70%)
Commonsense	82%	Mistral 7B (67%)
Causal	77%	Mistral Nemo (63%)

Table 6: Performance of models across different reasoning types.

(**RQ4**), our findings clarify that while surface linguistic features can be mastered by smaller models, deep reasoning capabilities scale linearly with model size, with Mistral-Small-24B demonstrating superior performance.

Future work will focus on expanding this corpus and developing a standardized static benchmark to facilitate reproducible research in low-resource language reasoning.

7 Limitations

While this study establishes a foundational framework for reasoning in Sindhi, several limitations must be acknowledged.

Dataset Scale and Domain Diversity Although *Sindhi-Reasoning-Instruct* is the first of its kind, its size ($N = 2,190$) is relatively modest compared to high-resource instruction sets. Consequently,

the models may struggle with long-tail entities or highly specialized domains (e.g., medical or legal reasoning) not covered by our Wikipedia and news sources. Furthermore, while we categorized reasoning into four types, the dataset does not yet cover symbolic logic, mathematics, or code generation.

Translation Artifacts Despite our rigorous "Human-in-the-loop" and "Localization-over-Translation" protocols, a significant portion of the dataset (80%) originates from English encyclopedic texts. While we successfully adapted the linguistic surface form and cultural references, the underlying logical flows may still reflect Western rhetorical structures rather than indigenous modes of argumentation. True cultural grounding requires a shift toward data generated *ab initio* by native speakers, rather than translated content.

Tokenizer Inefficiency We utilized English-centric base models (LLaMA and Mistral). These architectures utilize tokenizers optimized for Latin scripts. Consequently, the tokenization of the Sindhi Perso-Arabic script is highly inefficient, often fragmenting single words into multiple sub-tokens. This increases the computational cost of inference and reduces the effective context window, potentially limiting the model's ability to handle extremely long reasoning chains compared to English inputs.

Evaluation Metrics Finally, while we employed both automated and human evaluation, standard NLG metrics (BLEU, ROUGE) are known to be poor proxies for logical validity. Our human evaluation, though rigorous, relied on a small pool of expert linguists. Future work requires larger-scale crowdsourced evaluation to test generalization across diverse Sindhi dialects.

References

Wazir Ali, Jay Kumar, Junyu Lu, and Zenglin Xu. 2021. Word embedding-based new corpus for low-resource language: Sindhi. *arXiv preprint arXiv:1911.12579*.

Wazir Ali, Saifullah Tumrani, Jay Kumar, and Tariq Rahim Soomro. 2024. An evaluation of sindhi word embedding in semantic analogies and downstream tasks. *arXiv preprint arXiv:2408.15720*.

Guneet Singh Kohli, Arghyadeep Sen, Sambit Sekhar, Shashikanta Sahoo, Shantipriya Parida, Satya Ranjan Dash, and Ondřej Bojar. 2023. Bengaligpt: An instruction following llama model for bengali. *arXiv preprint arXiv:2312.12879*.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2024. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *arXiv preprint arXiv:2409.12958*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Mutee U. Rahman Arain and Mohammad Iqbal Bhatti. 2010. Finite state morphology and sindhi noun inflections. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24)*, pages 669–676.

Raghav Ramji and Keshav Ramji. 2025. Inductive linguistic reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22783–22810.

Yu Sheng, Wanting Wen, Linjing Li, and Daniel Zeng. 2025. Evaluating generalization capability of language models across abductive, deductive and inductive logical reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 4945–4957.

Irum Naz Sodhar, Suriani Sulaiman, and Abdul Hafeez Buller. 2023. Morphology-assisted sindhi text analysis for natural language processing applications. *Indian Journal of Science and Technology*, 16(35):2894–2901.

Emily Vaillancourt and Christopher Thompson. 2024. Instruction tuning on large language models to improve reasoning performance. *TechRxiv*.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024. Lingualift: An effective two-stage instruction tuning framework for low-resource language reasoning. *arXiv preprint arXiv:2412.12499*.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2025. Lingualift: An effective two-stage instruction tuning framework for low-resource language reasoning. *arXiv preprint arXiv:2412.12499*.

A Methodology: Instruction Generation

To generate diverse reasoning tasks, we employed a rigid prompting strategy that forced the model to distinguish between Inductive and Deductive logic. Figure 7 shows the exact system prompt used in our generation pipeline.

B Linguistic Resources: Special Characters

A major challenge in Sindhi NLP is the correct representation of its 52-character alphabet, which includes unique implosive sounds. Figure 8 lists the specific characters that were frequently hallucinated or replaced by Urdu equivalents in base models, necessitating the creation of our gold-standard dataset.

C Data Annotation and Quality Control Examples

Our "Human-in-the-Loop" pipeline was essential for correcting machine translation errors.

Figure 9 highlights common issues with direct Google Translation, such as incorrect font rendering and morphological errors. Figures 10 and 11 demonstrate how human annotators corrected these issues to create the final *Sindhi-Reasoning-Instruct* dataset.

D Qualitative Model Outputs Examples

We provide a direct qualitative comparison between Base models and Fine-Tuned models to demonstrate the impact of instruction tuning.

D.1 Causal Reasoning

Figure 12 demonstrates the model's ability to explain the causes of Smog. The Base models (LLaMA/Mistral) failed to generate Sindhi text entirely or hallucinated, whereas the Fine-Tuned Mistral-Small-24B generated a coherent, factually accurate explanation.

D.2 Inductive Reasoning

Figure 13 tests the model's ability to predict consequences based on historical data (War). The Fine-Tuned model successfully inferred complex socio-economic consequences, demonstrating abstract reasoning capabilities.

The fine-tuning process adapted multilingual pretrained models to Sindhi reasoning tasks using parameter-efficient techniques.

E Training Methodology and Implementation Details

The fine-tuning process adapted multilingual pre-trained models to Sindhi reasoning tasks using parameter-efficient techniques. This section details the theoretical objective function and the specific implementation configurations.

E.1 Training Objective

Given an instruction–answer pair in Sindhi, tokenized as a sequence $x_{1:T}$, the model predicts the probability of each token conditioned on the instruction prompt c. The training objective is the cross-entropy loss:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\frac{1}{Z} \sum_{(c,x_{1:T})} \sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}, c) \tag{1}$$

Perplexity (PPL) was used as the intrinsic evaluation metric: $\text{PPL} = \exp(\mathcal{L}_{\text{NLL}})$.

To ensure the model learns to reason rather than just model the prompt distribution, we employed a masked loss where only the Answer tokens are optimized:

$$\mathcal{L}_{\text{masked}} = -\frac{1}{\sum_t w_t} \sum_{t=1}^{T} w_t \log p_\theta(x_t \mid x_{<t}, c) \tag{2}$$

Where $w_t \in \{0, 1\}$ indicates whether token t belongs to the answer.

E.2 Implementation and Hyperparameters

Training was implemented using the `accelerate` library with DeepSpeed integration to optimize throughput on limited hardware. We utilized a standard causal language modeling objective with a maximum sequence length of 2048 tokens.

To ensure robust adaptation to the Sindhi script and logic, we applied LoRA adapters not only to attention layers but to all linear modules, including embeddings and the language model head. Parameter updates were performed using the AdamW optimizer. The learning rate followed a schedule with a linear warm-up phase followed by cosine decay. Regularization included dropout (0.1) and weight decay to prevent overfitting on the limited dataset.

Table 7 details the comprehensive configuration used across all model architectures.

Hyperparameter	Value
LoRA Configuration	
Rank (r)	64
Alpha (α)	128
Dropout	0.1
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, embed_tokens, lm_head
Optimization	
Optimizer	AdamW
Learning Rate	1×10^{-4}
Scheduler	Cosine
Weight Decay	1×10^{-4}
Warmup Ratio	0.0
Batch Size (per device)	4
Gradient Accumulation	2 steps
Precision	BF16
Sequence Length	2048

Table 7: Hyperparameters and configuration used for fine-tuning via LoRA.

F Linguistic Analysis Details

F.1 Morphological Analysis

The Sindhi reasoning dataset reflects the language's rich and productive morphological system, particularly in verbal inflection and auxiliary constructions. Verbs in the dataset encode tense, aspect, gender, and number primarily through suffixation and compound verb structures, with frequent auxiliaries such as (*"tho," "the," "payo," "paye,"* and *"tha"*).

Analysis of suffix frequency (see Figure 14) shows a strong dominance of aspectual markers, indicating that the dataset contains a high proportion of explanatory and descriptive sentences, which are essential for reasoning tasks. Nominal morphology also exhibits consistent marking of plurality and gender through suffixes such as *{ -un, -a, in/een }*, often followed by postpositions encoding grammatical relations. The regular recurrence of these inflectional patterns across instructions, contexts, and answers suggests that the dataset maintains morphosyntactic consistency and provides sufficient morphological coverage for learning agreement patterns inherent to Sindhi.

F.2 Syntactic Analysis

From a syntactic perspective, the dataset strongly adheres to Sindhi's canonical Subject–Object–Verb (SOV) word order. Manual inspection of sentence-final constructions confirms that the majority of clauses terminate in finite verb forms or auxiliary complexes, reflecting verb-final structure. This is particularly evident in multi-clause reasoning answers, where subordinate and coordinate clauses are linked using discourse markers such as (*cho ta, tehen kre, inkare*) while preserving verb-final order in each clause. The dataset also consistently employs postpositions rather than prepositions, aligning with Sindhi's head-final typology. The presence of complex sentence structures, including conditional, causal, and explanatory clauses, demonstrates that the dataset captures syntactic patterns required for logical reasoning, rather than relying on simplified sentence forms.

F.3 Semantic Analysis

Semantically, the dataset is characterized by explicit encoding of logical relations through discourse markers and reasoning-oriented constructions (see Figure 15).

- **Deductive reasoning** samples frequently employ justificatory markers such as (*cho ta* - because) and (*zaruri ahe ta* - it is necessary that).
- **Inductive reasoning** relies on inferential expressions like (*andazo lagai sighje tho* - It can be guessed) and (*Ina ma zahir the tho* - this shows that).
- **Causal reasoning** is signaled through result-oriented markers such as (*ina je nateeje men* - as a result) and (*sabab ihyo ahe ta* - the reason is that).
- **Commonsense reasoning** systematically invokes shared knowledge using phrases like (*aam fehm ji bunyad ty* - according to commonsense).

The consistent and category-specific use of these markers indicates that the dataset encodes semantic relations in a linguistically explicit manner. This makes reasoning structures transparent at the surface level, ensuring that semantic inference is grounded in language rather than implicit cues.

G Human Evaluation Rubric

The human evaluation utilized a strict grading rubric to ensure consistency across the 600 sampled outputs. The specific criteria for the 5-point Quality Rating scale were defined as follows:
Rating A (Excellent): Comprehensive understanding; precise, cohesive, and contextually relevant

with no grammatical errors.
Rating B (Good): Accurate and comprehensive; contains minor errors in flow or clarity that do not impact the main idea.
Rating C (Average): Partially accurate or slightly incomplete; may contain grammatical errors or ambiguous phrasing.
Rating D (Poor): Mostly incorrect or difficult to understand; exhibits significant logic or grammar issues.
Rating E (Unacceptable): Illogical, factually hallucinated, or completely unrelated content.

Fixed prompt text (instruction) as provided:

fixed_prompt = '''
You are a helpful assistant that generates reasoning-based instructions in Sindhi. Your task is to generate tasks involving *Inductive* and *Deductive* reasoning based on user-provided factual text.

Reasoning Types:
1. *Inductive Reasoning*: Derive general conclusions from specific facts.
2. *Deductive Reasoning*: Draw specific conclusions from general facts.

For each reasoning type, generate:
- *Instruction* (in Sindhi): A reasoning-based question or task, not just factual recall.
- *Input* (in Sindhi): A relevant scenario or fact.
- *Output* (in Sindhi): A well-reasoned response (4–5 lines), showing logical inference.

Example 1 — Inductive Reasoning
Fact: "سنڌو نديءَ جي ڪنارن تي هزارين سال کان ماڻهو آباد آهن."

Generated Instruction:

· Instruction: ڏنل معلومات جي بنياد تي اهو ٻڌايو ته دريائن جي ويجهو رهڻ ماڻهن لاءِ ڪيئن فائديمند ٿي سگهي ٿو؟
· Input: "سنڌو نديءَ جي ڪنارن تي هزارين سال کان ماڻهو آباد آهن."
· Output: "ڇو ته دريائن جي ويجهو پاڻي، زراعت، ۽ آمد و رفت لاءِ بهتر موقعا ملن ٿا."

Example 2 — Deductive Reasoning
Fact: "شاهه عبداللطيف ڀٽائي 18هين صديءَ جو شاعر هو."

Generated Instruction:

· Instruction: ڏنل معلومات جي بنياد تي فيصلو ڪريو ته شاهه عبداللطيف ڀٽائي 17هين صديءَ ۾ موجود هو يا نه؟
· Input: "شاهه عبداللطيف ڀٽائي 18هين صديءَ جو شاعر هو."
· Output: "نه، ڇو ته هو 18هين صديءَ ۾ پيدا ٿيو، 17هين صديءَ ۾ نه."

Output Format (in JSON):
Use the following fact to generate both Inductive and Deductive reasoning tasks:

Figure 7: The system prompt used for generating the instruction dataset. It explicitly defines the requirements for Inductive vs. Deductive reasoning tasks in Sindhi.

Character	Sound (IPA Approx.)	Description
ڈھ	/dʱ/	Aspirated voiced dental stop
ڈ	/ɖ/	Retroflex voiced stop
ڇ	/tʃʰ/	Aspirated "ch"
ڳ	/ɠ/	Implosive or retroflex voiced stop
ٿ	/tʰ/	Aspirated voiceless dental stop
ڍھ	/ɖʱ/	Aspirated retroflex voiced stop
ڱ	/ŋ/	Velar nasal
ک	/kh/	Voiceless velar stop, Used slightly differently than Urdu
گ	/k/	Voiceless velar stop
ڀ	/bʰ/	Aspirated voiced bilabial stop
ڦ	/pʰ/	Aspirated voiceless bilabial stop
ٹ	/ʈ/	Retroflex voiceless stop
ٹھ	/ʈʰ/	Aspirated retroflex voiceless stop
ٻ	/b/	Implosive or special voiced bilabial
ج	/ʤ/	Voiced retroflex affricate
ڃ	/ʒ/ or /ɳ/	palatal nasal
ڻ	/ɳ/	Voiced Retroflex nasal
ڄ	/ʤʰ/	Aspirated "j" sound (as in “judge”)
ڑ	/ɽ/	Retroflex flap

List of Sindhi Special Characters

Figure 8: List of Sindhi special characters (Implosives and Nasals) with IPA approximations. These characters were the primary focus of the "Orthographic Normalization" phase.

1. **Examples from Google translation and after human annotation**

English Text	Translated by Google	Improvement	Comment
Acid rain is any type of rain that is rapidly moving towards the surface of the earth, which is affected by the action of the sun (low pH). While water in the air is also included in the inactive pH, the pH level of acid rain is lower than that at a pH between 6.5 and 8.5 and its highest concentration is between 4 and 5. Acid rain is rapidly moving towards the surface, which is affected by the action of the sun. Acid rain has a harmful effect on the atmosphere, water vapor and air. Acid rain can quickly release sulfur dioxide and nitrogen oxides, which are produced by the reaction of water molecules in the atmosphere to accelerate the process of oxidation.	تيزابي مينهن ڪنهن به قسم جو مينهن آهي جيڪو تيزي سان زمين جي متاڻري ڏانهن وڌي رهيو آهي، جيڪو سج جي عمل (گهٽ پي ايڇ) کان متاثر ٿئي ٿو. جڏهن پي هوا ۾ پاڻي پي غير فعال پي ايڇ ۾ شامل آهي، تيزابي مينهن جي پي ايڇ جي سطح 6.5 ۽ 8.5 جي وچ ۾ پي ايڇ کان گهٽ آهي ۽ ان جي سپ کان وڌيڪ ڪنسنٽريشن 4 ۽ 5 جي وچ ۾ آهي. تيزابي مينهن تيزي سان سطح ڏانهن وڌي رهيو آهي، جيڪو سج جي عمل کان متاثر ٿئي ٿو. تيزابي مينهن جو ماحول، پاڻي جي بخارات ۽ هوا تي نقصانڪار اثر پوي ٿو. تيزابي مينهن جلدي سلفر ڊاءِ آڪسائيڊ ۽ نائٽروجن آڪسائيڊ خارج ڪري سگهي ٿو، جيڪي آڪسائيڊيشن جي عمل کي تيز ڪرڻ لاءِ فضا ۾ پاڻي جي مالڪيولن جي رد عمل سان پيدا ٿين ٿا.	تيزابي مينهن ڪنهن به قسم جو مينهن آهي جيڪو تيزي سان زمين جي متاڻري ڏانهن وڌي رهيو آهي، جيڪو سج جي عمل (گهٽ پي ايڇ) کان متاثر ٿئي ٿو. جڏهن ته هوا ۾ پاڻي پڻ غير فعال پي ايڇ ۾ شامل آهي، تيزابي مينهن جي پي ايڇ جي سطح 6.5 ۽ 8.5 جي وچ ۾ پي ايڇ کان گهٽ آهي ۽ ان جي سپ کان وڌيڪ ڪنسنٽريشن 4 ۽ 5 جي وچ ۾ آهي. تيزابي مينهن تيزي سان سطح ڏانهن وڌي رهيو آهي، جيڪو سج جي عمل کان متاثر ٿئي ٿو. تيزابي مينهن جو ماحول، پاڻي جي بخارات ۽ هوا تي نقصانڪار اثر پوي ٿو. تيزابي مينهن جلدي سلفر ڊاءِ آڪسائيڊ ۽ نائٽروجن آڪسائيڊ خارج ڪري سگهي ٿو، جيڪي آڪسائيڊيشن جي عمل کي تيز ڪرڻ لاءِ فضا ۾ پاڻي جي مالڪيولن جي رد عمل سان پيدا ٿين ٿا.	Some fonts were incorrect, such as (۽ئي instead of ئتي).

Figure 9: Error Analysis: Comparison between Google Translation and Human Annotation. Note the correction of "Active pH" and specific scientific terminology.

Examples generated by GPT and annotated by a human

Deductive	Sindhi	English Translation	After Human Annotation	Comment
Instruction	ڏنل حقيقتن جي بنياد تي **کيا** وضاحت ڪريو ته ايندڙ عام چونڊن ۾ ووٽرن جي شرڪت گهٽ ٿيڻ کا امڪان آهي؟	Based on the given facts, explain whether voter turnout is likely to decrease in the upcoming general elections.	ڏنل حقيقتن جي بنياد تي وضاحت ڪريو ته **چا** ايندڙ عام چونڊن ۾ ووٽرن جي شرڪت گهٽ ٿيڻ جو امڪان آهي؟	Usage of urdu word in کیا instead of چا
Context	**اگر** ڪنهن **علاقے** ۾ امن امان جون حالات خراب ٿين، ته **ووٹرن** جي شرڪت گهٽ ٿيندي آهي۔ هن وقت سنڌ جي ڪيترن ئي ضلعن ۾ دهشتگردی ۽ بدامنی وڌي وئي آهي۔ پوليس ۽ **سيڪيورٹی** ادارا انتخابات لاءِ مڪمل انتظام ڪرڻ ۾ **مشڪلات** کي منهن ڏئي رهيا آهن۔	If law and order deteriorates in a region, voter turnout decreases. Currently, terrorism and insecurity have increased in several districts of Sindh. Police and security institutions are facing difficulties in making complete arrangements for elections.	**جيڪڏهن** ڪنهن **علاقي** ۾ امن امان جون **حالتون** خراب ٿين، **ته ووٹرن** جي شرڪت گهٽ ٿيندي آهي. هن وقت سنڌ جي ڪيترن ئي ضلعن ۾ دهشتگردي ۽ بدامني وڌي وئي آهي. پوليس ۽ سيڪيورٽي ادارا چونڊن لاءِ مڪمل انتظام ڪرڻ ۾ مشڪلاتن کي منهن ڏئي رهيا آهن	Mentioned bold letters are in urdu
Output	ڏنل حقيقتن مان هڪ واضح منطقي نتيجو نڪري ٿو. پهرين ڳالهه اها آهي ته جتي امن امان خراب ٿئي، اتي عوام خوف سبب گهٽ ووٽ ڏيڻ لاءِ باهر نڪرندو آهي. ٻي ڳالهه اها آهي ته سنڌ جي ڪيترن ئي ضلعن ۾ واقعي بدامني وڌي وئي آهي، جنهن ڪري ماڻهن جو اعتماد اڳ ئي گهٽجي رهيو آهي. ٽين ڳالهه اها آهي ته سيڪيورٽي ادارا مڪمل انتظام ڪرڻ ۾ ناڪام ڏسجي رهيا آهن، جنهن سان **پولنگ اسٽيشنن** تي سلامتي جا خدشا وڌيڪ وڌندا. انهن ٽن حقيقتن کي گڏجي ڏسڻ سان اهو لازمي نتيجو نڪري ٿو ته ايندڙ عام چونڊن ۾ ووٽرن جي شرڪت گهٽ ٿيندي، جو ته بدامني، دهشتگردي ۽ سيڪيورٽي جي ڪوٽ گڏجي ووٹرن کي رڪاوٽ بڻائيندي.	From the given facts, a clear logical conclusion emerges. The first point is that wherever law and order deteriorates, people come out less to vote due to fear. The second point is that in many districts of Sindh, unrest has indeed increased, which has already reduced public confidence. The third point is that the security agencies appear to have failed to make complete arrangements, which will further heighten concerns about safety at polling stations. Looking at these three facts together, it necessarily follows that voter turnout in the upcoming general elections will decline, since unrest, terrorism, and lack of security collectively act as obstacles for voters.	ڏنل حقيقتن مان هڪ واضح منطقي نتيجو نڪري ٿو. پهرين ڳالهه اها آهي ته جتي امن امان خراب ٿئي، اتي عوام خوف سبب گهٽ ووٽ ڏيڻ لاءِ باهر نڪرندو آهي. ٻي ڳالهه اها آهي ته سنڌ جي ڪيترن ئي ضلعن ۾ واقعي بدامني وڌي وئي آهي، جنهن ڪري ماڻهن جو اعتماد اڳ ئي گهٽجي رهيو آهي. ٽين ڳالهه اها آهي ته سيڪيورٽي ادارا مڪمل انتظام ڪرڻ ۾ ناڪام ڏسجي رهيا آهن، جنهن **سان پولنگ اسٽيشنن** تي سلامتي جا خدشا وڌيڪ وڌندا. انهن ٽن حقيقتن کي گڏجي ڏسڻ سان اهو لازمي نتيجو نڪري ٿو ته ايندڙ عام چونڊن ۾ ووٽرن جي شرڪت گهٽ ٿيندي، جو ته بدامني، دهشتگردي ۽ سيڪيورٽي جي ڪوٽ گڏجي ووٽرن کي رڪاوٽ بڻائيندي	

Figure 10: Dataset Example 1 (Deductive Reasoning): The annotator corrected the Urdu word *"kiya"* to the Sindhi *"cha"*, ensuring dialectal purity.

Causal	Sindhi	English Translation	Human Annotation	Comment
Instruction	وضاحت ڪريو ته جو **درخت** ڇو ڏينهن عام طور تي بهار ۾ ملهايو ويندو آهي؟	Explain why Arbor Day is usually celebrated in the spring.	وضاحت ڪريو ته ڇو وڻن جو ڏينهن عام طور تي بهار ۾ ملهايو ويندو آهي؟	Changed bold letter word from Urdu into Sindhi
Context	وڻن جو ڏينهن عام طور تي بهار ۾ ملهايو ويندو آهي.	Arbor Day is usually celebrated in the spring.		
Output	بهار جي موسم ۾ مٽي ۾ نمي هوندي آهي، ۽ پوکيل وڻن کي وڌڻ لاءِ **موزون** ماحول ملي ٿو. پڻ، هن موسم ۾ ڏينهن ڊگها ٿيندا آهن ۽ گرمي پد هوندو آهي، معتدل جيڪو وڻ پوکڻ لاءِ بهترين وقت آهي.	In the spring season, the soil contains moisture, providing a suitable environment for planted trees to grow. Additionally, the days are longer and the temperatures are moderate during this season, making it the best time for tree planting.	بهار جي موسم ۾ مٽي ۾ نمي هوندي آهي، ۽ پوکيل وڻن **بهتر** ماحول کي وڌڻ لاءِ ملي ٿو. هن موسم ۾ ڏينهن ڊگها ٿيندا آهن ۽ گرمي پد هوندو آهي، جيڪو معتدل وڻ پوکڻ لاءِ بهترين وقت آهي.	**موزون** into **بهتر**

Figure 11: Dataset Example 2 (Causal Reasoning): Annotators refined the instruction to explain "Arbor Day", correcting the context to fit the spring season description.

Example 1 : Causal Reasoning

Instruction : وضاحت ڪريو ته ڌوڙ جي ٺهڻ جا اهم سبب ڇا آهن ۽ انهن جي ڪري ڪهڙا اثر ٿي سگهن ٿا؟

Explain what the main causes of smog formation are and what effects they can cause?

Models	Base Model's Output	Finetuned Model's Output	Comment
Meta_Llama-3.2-1B	**No any output,**	ڌوڙ جا سبب بٽجندڙ ذرات جي درجه بندي سان ڌوڙ جي صورتحال تبديل ٿي سگهي ٿي. جيڪڏهن دنيا کان بهتر هوا واري هلندڙ آسمان يا ٺاهن ٿيون ته اهي دنيا ۾ دنيا کان بهتر هوا واري آسمان جي پيداوار سان واڌارو آڻي سگهون ٿيون. The classification of the particles that cause smog can change the smog situation. If better-ventilated skies are created, they can increase the production of better-ventilated skies in the world.	Base Model **Rating E** Fine tuned **Rating D**
Meta_Llama-3.2-3B	ڌوڙ جي ٺهڻ جا اهم سبب جوتهان هوندڙ ذرڙا ۽ ري، ۽ آتش فشاني سرگرميindustrialزيڪ، آهن. انهن ڇڏڻ ڪري ڌوڙ جي ٺاهي سگهي سگهي، ڌوڙ جو ٺاهي وار ڏي سگهي سگهي. The main causes of smog formation are traffic, industry, and volcanic activity, which release particulate matter into the atmosphere. These compounds can cause smog to form, which can **cause hair loss.**	ڌوڙ جي ٺهڻ جو سبب زراعت ۾ خشڪ موسم سبب ٿي سگهي ٿو جنهن سبب مائي ۽ دونهيون هوا ۾ لٽڪيل نظر رکن ٿيون. اهي سبب ٿيندڙ فضائي تراشن ہين زميني عنصرن مان محروم هوندا آهن. The formation of haze can be caused by dry weather in agriculture, which causes dust and fumes to remain suspended in the air. The resulting atmospheric particles are deprived of other terrestrial elements.	Base model : Rating D, some hallucination. Finetuned model : Rating D
Meta- Llama- 3.1-8B	آتش فشاني سرگرمي، جهنگلي باهه، زرعي ڪاروبار (لڙ سازن ۽ جوتن)، ڦرنفڪ، صنعتي سرگرمين ۽ هوائون وارو موسم ڌوڙ جو سبب بٽجندڙ ذرڙن جا ذريعا آهن. هي ذرڙا بخار جو نيوڪليئي طور ڪم ڪندي ڌوڙ ٺاهي چڪرن ٿا، جو آخر ۾ بخار ۽ برف ٻيخود ٿي وڃي ٿا. Volcanic activity, wildfires, agricultural activities (burning and plowing), traffic, industrial activities, and windy weather are all sources of smog-causing particles. These particles act as nuclei for vapor, forming smoke, which eventually condenses into vapor and ice. (**R**	ڌوڙ جي ٺهڻ جا اهم سبب ٿي سگهن ٿا جو ته اهو مٽي، دونهون، ۽ ٻيا خشڪ ذرات هوا ۾ لٽڪي ٿو. اهي ذرات هوائي جهازن کي بدران نيرو پيرو نظر اچڻ جو سبب بڻجن ٿا ۽ انهن جي ڊرڻ سان ڦرنفڪ ۽ جهنگلي باهه جي رهنمائي ۾ حڪومتشون ٿي سگهن ٿيون. Smog can be a major cause of haze because it causes dust, smoke, and other dry particles to hang in the air. These particles can make airplanes appear blue-gray, and their dispersion can hinder traffic and wildfire control efforts.	Base Model : Rating A Finetuned : Rating A
Mistral-7B-Instruct-v0.3	ڌوڙ جي ٺهڻ جا اهم سبب انهن جي ڪري ڪهڙا اثر ٿي سگهن ٿا سبب جنهن ته زراعت اور صنعت پري کان ڏنو وبو اور وبو جي حوالي سان ڏسڻ جي هدايت ٿي ملحصر ڪري بولدو آهي. The main causes of smog formation and their effects depend on the direction of view, which is seen from a distance and from agriculture and industry.	ڌوڙ جي ٺهڻ جا اهم سبب ٿي سگهن ٿا چو ته اهو مٽي، دونهون، ۽ ٻيا خشڪ ذرات هوا ۾ لٽڪي ٿو. اهي ذرات هوائي جهازن کي بدران نيرو پيرو نظر اچڻ جو سبب بڻجن ٿا ۽ انهن جي ڊرڻ سان ڦرنفڪ ۽ جهنگلي باهه جي رهنمائي ۾ حڪومتشون ٿي سگهن ٿيون. Smog can be a major cause of haze because it causes dust, smoke, and other dry particles to hang in the air. These particles can make airplanes appear blue-gray, and their dispersion can hinder traffic and wildfire control efforts.	Base Model: Rating D Finetuned : Rating C
Mistral-Nemo-Instruct-2407	главнымsoot سبب ٿا جا ضرورجي يڪاري رنجش (تهرمان) ۽ بارومٽري ڳولار حوالي ڳلي ڌوڙ ۾ سجو وجو، باف .ondiڏسڻ ۽ وارو ڌوڙ ۽ برفاني ڌوڙ جي پير رنجشي ۽ بارومٽري ڳولار ۾ اضافي تغير سي رنجش ۽ بارومٽري ڳولار کي وجو ٿي ضرورتا هي. Response is not in Sindhi	ڌوڙ جي ٺهڻ جا اهم سبب زراعت، ڦرنفڪ، صنعتي آهن، جيڪي مٽي، ۽ گيسن ۾ حردار ادا ڪن ٿيون. اهي سبب گاديء ۾ ڌوڙ جي ٺهڻ کي وڌائين ٿا، جيڪي بصري وضاحت کي بي اثر بڻائي سگهن ٿا ۽ ڪيترين کي متاثر ڪري سگهن ٿا. The main causes of smog are agriculture, traffic, and industry, which contribute to dust and gases. These causes increase the formation of smog in the capital, which can impair visual clarity and affect the quality of life.	Base Model: Rating E Finetuned : Rating B
Mistral-Small-24B	ڊاسيورن او چڪيدو ۾ ٿو ٿي، خلون ۽ ڌوڙ ۾ تيارو ۽ وڌهو آي نڙ ڌوڙ جي ٺهڻ، جو عوام ٿي سسٽيم. Response is not in Sindhi ,	ڌوڙ جي ٺهڻ جا اهم سبب زراعت، ڦرنفڪ، صنعتي عملون، ۽ هوائون واري موسم آهن. انهن سبب ڪري هوا ۾ مختلف قسم جا ذرڙا جڙي وڃن ٿا، جيڪي انسان جي صحت تي منفي اثر وجهي سگهن ٿا، جنهن ته سانس جي مسئلن ۽ بصارت جي مسئلن جو سبب بڻجي سگهن ٿا. The main causes of smog are agriculture, traffic, industrial processes, and windy weather. These cause various types of particles to be suspended in the air, which can have negative effects on human health, such as causing respiratory problems and vision problems.	

Figure 12: Qualitative Comparison on Causal Reasoning (Topic: Smog Formation). The Fine-Tuned model (Right Column) successfully generates valid Sindhi script, while base models (Left Column) fail.

Example 2: Inductive Reasoning

Instruction

ڏنل معلومات جي بنياد تي اهو اندازو لڳايو ته جنگ دوران شهري آباديءَ تي حملن جا ڪهڙا نتيجا نڪري سگهن ٿا؟

Based on the information given, predict what consequences attacks on civilians during war might have?

Models	Base Model's Output	Finetuned Model's Output
Meta_Llama-3.2-1B	No any Output	جنگ دوران شهري آباديءَ تي حملن سان ڀگهي وقت جي زوال جي امداد ۽ ان مان نشانو اخراج جو ساموندي آسمان سان به نقل ۽حمل کي بچاءَ جو سبب بڻجي سگهي ٿو. During war, the long-term fallout from attacks on civilian populations and the resulting targeted emissions can also be a factor in protecting transportation over the seas.
Meta_Llama-3.2-3B	انقلابي فوجو جو ڏنل معلومات جي بنياد تي Shahi Japani Fauji ke pakki taqat ۾ ادبي فوجي ڪرامات نانڪنگ پليچر رڇي هتي بي اڳوڻي سڃاتي ويندو هيڪون، وري تي شهري آباديءَ تي حملون جو نتيجه لڳي سگهي. Based on the information provided by the revolutionary forces, the Shahi Japanese Fauji ke Pakki taqat in Nanking, the literary military miracles, were created here, and the second previous one will be known, and the result may be attacks on the civilian population.	شهري آباديءَ تي حملن جا نتيجا انهي پويان اچي سگهن ٿا ته اهي شهر ۾ آباد علائقائي معاشي ۽ سماجي نظم متاثر ڪندا. اهي حملا ماڻهن کي تباهه ڪرڻ کان علاوه امن ۽ استحڪام کي برقرار رکڻ ۾ مددگار ثابت ٿي سگهجي ٿو. The consequences of attacks on urban populations can come from the fact that they affect the regional economic and social order of the city. These attacks can be helpful in maintaining peace and stability, in addition to destroying people.
Meta-Llama-3.1-8B	ڳهڙيل ڏکڻ ۽ هيٺيان نتيجا جنم ورتا هوسن ته شهري آبادي کي بهجي، نانجنڪ جنگ دوران هي ٽيا هٽا. The south and the following consequences would have been born if the civilian population had been evacuated, as happened during the Nanking War.	جنگ دوران شهري آباديءَ تي حملن سان ماڻهن کي وڏو اخلاقي ۽ سماجي نقصان ٿئي ٿو، ان سان گڏوگڏ مقامي معاشرتي نظام ۽ اقتصادي نظام به تباهه ٿي سگهن ٿا. Attacks on civilian populations during war cause great moral and social harm to people, and can also destroy local social and economic systems.
Mistral-7B-Instruct-v0.3	نانجنگ جي جنگ دوران شهري آباديءَ تي حملن جا ڪهڙا نتيجا سگهن نڪي نمي شي، جيسڙ نانجنگ جي جنگ دوران قومي انقلابي فوج جي پوئتي هٽڻ کان فوري پوءِ جمهوريه چين جي گاديءَ واري هنڌ نانجنگ جي جنگ جو خاصيتي جوڇي نا نڪند. The consequences of attacks on civilians during the Battle of Nanjing cannot be predicted, as the specific circumstances of the Battle of Nanjing, the capital of the Republic of China, were not immediately clear after the retreat of the National Revolutionary Army during the Battle of Nanjing.	جنگ دوران شهري آباديءَ تي حملن سان عام طور تي خوف ۽ عدم استحڪام پيدا ٿيندو. اهڙي حالتن ۾ ماڻهو پنهنجي زندگين کي محفوظ ڪرڻ لاءِ پريشان ٿي سگهن ٿا ۽ گهرن کي خالي ڪري ڇڏي سگهن ٿا. Attacks on civilian populations during war will usually create fear and instability. In such situations, people may become anxious to save their lives and evacuate their homes.
Mistral-Nemo-Instruct-2407	इसें اندازو لا نانجنگ جو خونيانا ڪار جوئي تي شهري عوام ڪي حمله ڪا اٿي نظر اَنهن نو جنگ دوران شهری آبادی کے ليے حمله ڪرنے کا خطرناک نتائج ہو سکيں. حمله ڪرنے والو ته جنگ دوران شهری آبادی پر حمله ڪرنا غلط Pascual. Output is in urdu language	جنگ دوران شهري آباديءَ تي حملن سان گورننس جي نقصان، انساني نقصان ۽ شهري زندگيءَ جي بحران شامل ٿي سگهي ٿو. شهرين کي بي گهر ڪرڻ ۽ انهن جي روزگار واري وسيلن تي قبضو ڪرڻ سان معاشي نظام ۾ به درد بڻجي سگهي ٿو. Attacks on civilian populations during war can lead to loss of governance, human losses, and a crisis in civilian life. Displacing civilians and seizing their livelihoods can also cause economic pain.
Mistral-Small-24B	ہنل اطلاعات جو نيالوندرو منڌ ۾ ايک تر تيسره جهار اچو ہنڌ اُهن شهرين جا آبادي جو قتل عام ہويا اچي. جودا جنگي بلاڪات ۾ دنيا بهر جا قوم، جنهن بڏائيون ۾ جنهڪو آني، آنجهي، امير غليبي قتل عام هوا. Output in Mutli languages	جنهن طرح نانجنگ قتل عام مان ظاهر ٿئي ٿو ته جنگ دوران شهري آبادي کي سنگين نقصان پهچائي سگهي ٿو، جن ۾ جانون گم ٿين، گهرن جي تباهيءَ ۽ عوامي خوف شامل آهن. اهڙيون حالتون انساني حقن جي بدترين صورتحال پيدا ڪن ٿيون ۽ معاشرتي توازن کي متاثر ڪن ٿيون. As the Nanjing Massacre demonstrates, war can inflict serious harm on the civilian population, including loss of life, destruction of homes, and public fear. Such situations create a worsening human rights situation and disrupt social balance.

Figure 13: Qualitative Comparison on Inductive Reasoning (Topic: Consequences of War). The expert model moves beyond translation to provide structured logical inferences.

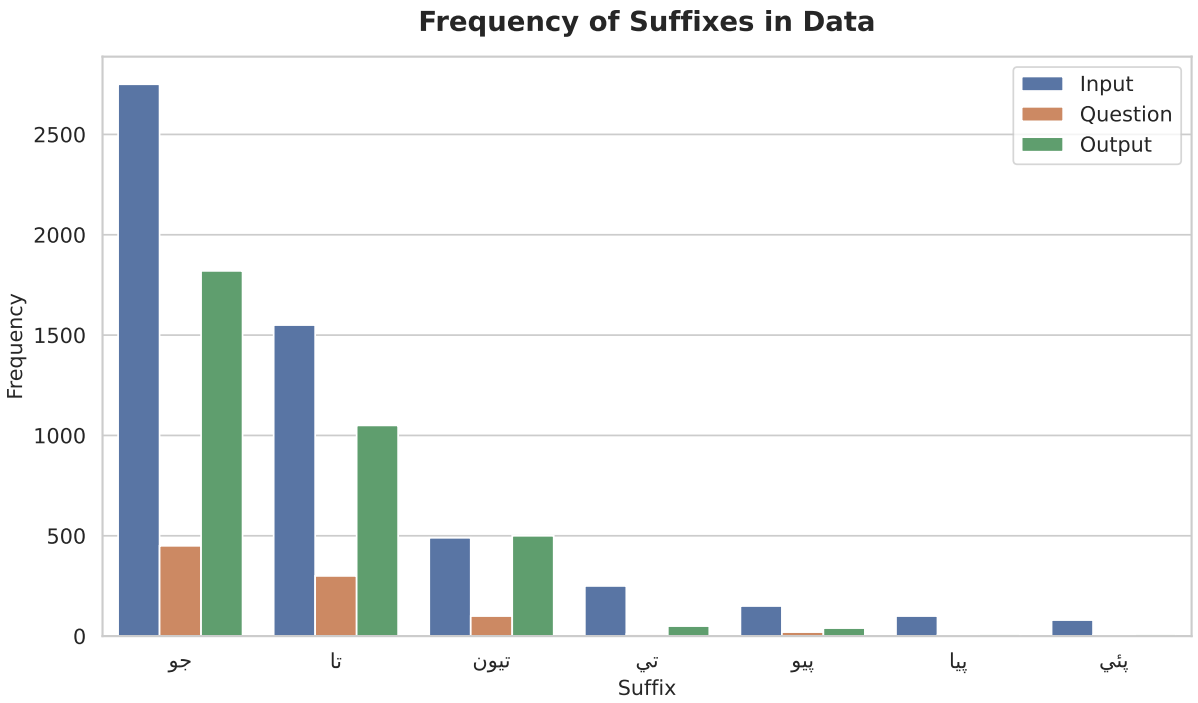

Figure 14: Frequency of Suffixes in Data. This chart illustrates the distribution of input, question, and output tokens.

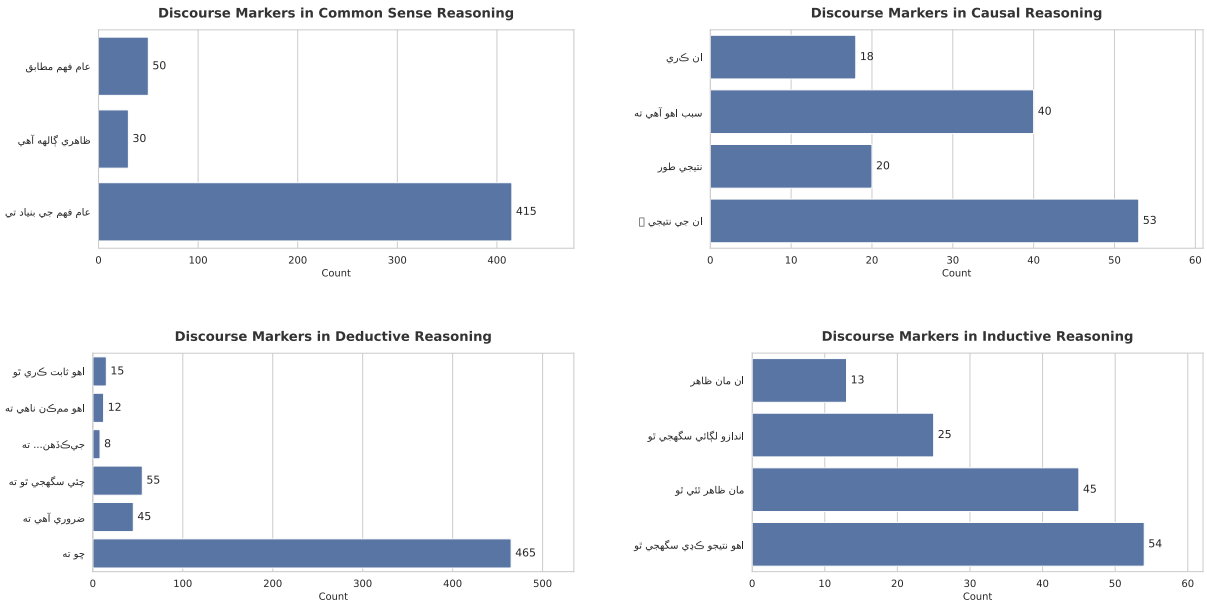

Figure 15: Analysis of Discourse Markers across four reasoning types: Common Sense, Causal, Deductive, and Inductive. The horizontal bars indicate the frequency of specific Sindhi markers.