IndicSentEval: How Effectively do Multilingual Transformer Models encode Linguistic Properties for Indic Languages?

Akhilesh Aravapalli[1], **Mounika Marreddy**[2],
Radhika Mamidi[1], **Manish Gupta**[1,3], **Subba Reddy Oota**[4]
[1]IIIT Hyderabad, India, [2]University of Bonn, Germany, [3]Microsoft, India, [4]Inria, France
aforakhilesh@gmail.com, mmarredd@uni-bonn.de, radhika.mamidi@iiit.ac.in,
gmanish@microsoft.com, subbareddyoota@gmail.com

Abstract

Transformer-based models have revolutionized the field of natural language processing. To understand why they perform so well and to assess their reliability, several studies have focused on questions such as: Which linguistic properties are encoded by these models, and to what extent? How robust are these models in encoding linguistic properties when faced with perturbations in the input text? However, these studies have mainly focused on BERT and the English language. In this paper, we investigate similar questions regarding encoding capability and robustness for 8 linguistic properties across 13 different perturbations in 6 Indic languages, using 9 multilingual Transformer models (7 universal and 2 Indic-specific). To conduct this study, we introduce a novel multilingual benchmark dataset, IndicSentEval, containing approximately ∼47K sentences. Our probing analysis of surface, syntactic, and semantic properties reveals that, while almost all multilingual models demonstrate consistent encoding performance for English, surprisingly, they show mixed results for Indic languages. As expected, Indic-specific multilingual models capture linguistic properties in Indic languages better than universal models. Intriguingly, universal models broadly exhibit better robustness compared to Indic-specific models, particularly under perturbations such as dropping both nouns and verbs, dropping only verbs, or keeping only nouns. Overall, this study provides valuable insights into probing and perturbation-specific strengths and weaknesses of popular multilingual Transformer-based models for different Indic languages. We make our code and dataset publicly available[1].

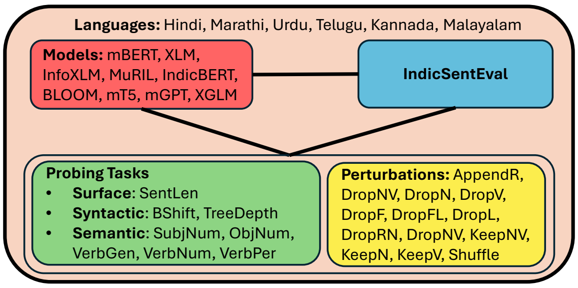

Figure 1: We evaluate 9 multilingual Transformer models on 8 probing tasks in 6 Indic languages using our IndicSentEval dataset. We analyze the effects of 13 perturbations on the performance of these models.

1 Introduction

Transformer-based language models (Vaswani et al., 2017), pretrained for both mono-lingual and multilingual contexts using millions of text documents, have demonstrated substantial enhancements in the performance of various natural language processing (NLP) tasks (Kenton and Toutanova, 2019; Pires et al., 2019; Conneau et al., 2020; Radford et al., 2019; Brown et al., 2020; Wang et al., 2019, 2018). To understand what types of linguistic properties (surface, syntactic and semantic) are encoded across layers of Transformer-based models, initial studies have investigated the layer-wise representations via a range of probing tasks (Adi et al., 2017; Hupkes et al., 2018; Conneau et al., 2018; Rogers et al., 2020; Jawahar et al., 2019; Mohebbi et al., 2021). However, these studies focus solely on English. Although, several studies have examined the presence of shared representation aspects across widely spoken languages in multilingual models (Chi et al., 2020; Acs et al., 2024), unfortunately, there is no work that investigates the extent to which multilingual Transformer-based models encode linguistic properties for different Indic languages. Further, recent neuro-AI studies have revealed that the brain uses different parsing strategies for different linguistic properties, which further differ across languages (Zhang et al., 2022; Oota et al., 2023). This inspired us to study the nuances of how multilingual Transformer-based

[1]https://github.com/aforakhilesh/IndicBertology

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1881–1905
December 20-24, 2025 ©2025 Association for Computational Linguistics

models capture linguistic properties across layers and languages.

Indic languages offer a rich tapestry of linguistic features that contribute to the global linguistic diversity. Hence, in this paper, we focus on Indic languages. We study the degree to which linguistic properties of Indic languages are captured by two kinds of multilingual models: universal and Indic-specific models. Universal multilingual models have been pretrained using a variety of pretraining objectives and using data which contains a small and varying fraction of Indic languages across languages and models (Tables 19, 20, and 21 in Appendix F). Indic-specific models (Kakwani et al., 2020; Dabre et al., 2021; Kumar et al., 2022) have been specifically trained on Indic language data only. Specifically, we focus on 6 Indic languages: three Indo-European languages (Hindi, Marathi, Urdu) and three Dravidian languages (Telugu, Kannada, Malayalam).

Even if multilingual models encode linguistic properties of Indic languages effectively, such encodings may not be robust to input text perturbations for particular models. Lack of such robustness may make some models less reliable than others for real-world applications. Although there exist many such studies (Wang et al., 2021; Jin et al., 2020; Li et al., 2020; Garg and Ramakrishnan, 2020; Sanyal et al., 2022; Neerudu et al., 2023) on robustness analysis of Transformer-based models, they focus on English, and on downstream tasks. Robustness analysis for prediction of linguistic properties in a multilingual setting for Indic languages is under explored. Hence, we systematically study how various multilingual Transformer-based models may differ in their robustness for different linguistic properties with respect to different kinds of input text perturbations across languages. We provide detailed related work in Appendix A, focusing on three aspects: (i) probing in non-English and multilingual contexts, (ii) differences from English-centric findings, and (iii) critiques of probing methodology and recent advances.

We analyze 7 universal multilingual language models, each pretrained on data spanning $\sim$100+ languages, with only a small amount of Indic language data. These include mBERT (Pires et al., 2019), XLM-R (Conneau et al., 2020), InfoXLM (Chi et al., 2021), mGPT (Shliazhko et al., 2024), XGLM (Lin et al., 2022), BLOOM (Scao et al., 2022), and mT5 (Xue et al., 2021). Additionally, we examine 2 Indic-specific models, IndicBERT (Kakwani et al., 2020) and MuRIL (Khanuja et al., 2021), which are trained using corpora for Indic languages along with English. While it is expected that these Indic-specific models would be better at encoding and robustness for Indic language input, are there some linguistic properties which are better encoded by universal models? Are the universal models more robust to particular perturbation types? How effectively and robustly are English language properties encoded by these universal and Indic-specific models?

To perform such detailed analyses, we curate a novel multilingual dataset, IndicSentEval, from resources generated by the "Indian Language Machine Translation" (ILMT) initiative. This dataset contains information about three types of linguistic properties per Indic language: surface, syntactic and semantic. To investigate encoding and robustness ability of multilingual models, we design probing tasks for prediction of each property. The surface task probes whether the model learns a representation which is predictive of sentence length (SentLen). Syntactic tasks test for sensitivity to word order, i.e., bigram shift (BShift) and the depth of the syntactic tree (TreeDepth). Semantic tasks check for the subject and the direct object number in the main clause (SubjNum and ObjNum, respectively). Tasks mentioned so far were discussed in English focused studies too. However, morphology for Indic languages is significantly different from English, primarily in regard to the main verb used in a sentence. This prompts us to expand our analysis to encompass three additional semantic probing tasks related to the main verb within the sentence: verb gender (VerbGen), verb number (VerbNum), and verb person (VerbPer). Fig. 1 shows the conceptual diagram of our study.

Overall, the main contributions of this paper are as follows. (1) We perform an extensive study of the degree to which 9 multilingual Transformer-based models capture 8 linguistic properties across 6 Indic languages. (2) We contribute a novel dataset, IndicSentEval, with $\sim$47K sentences across the 6 languages. (3) We find that Indic-specific models like MuRIL and IndicBERT best capture linguistic properties for Indic languages, while universal models like mBERT, InfoXLM, mGPT and BLOOM show mixed results across properties. (4) Surprisingly, our robustness analysis with respect to 13 text perturbations shows that universal multilingual models (InfoXLM, BLOOM, mGPT, XGLM and mT5) demonstrate higher resilience to perturbations compared to BERT-like models (mBERT, IndicBERT and MuRIL).

2 IndicSentEval Dataset

We curate the IndicSentEval dataset from resources generated by the ILMT initiative, which serves as an Indic language counterpart to SentEval (Conneau et al., 2018) and offers labeled data for the eight probing tasks. We utilize the morph and chunk level Indic languages data (Tandon and Sharma, 2017; Bhatt et al., 2009; Xia et al., 2008) available in Shakti Standard Format (SSF) (Bharati et al., 2007, 1995). SSF is a highly readable representation for storing Indic language data with linguistic annotations. Fig. 4 in Appendix B shows an example of a Hindi sentence in SSF format.

Probing Tasks. Probing tasks (Adi et al., 2017; Hupkes et al., 2018; Jawahar et al., 2019; Mohebbi et al., 2021) help unpack the linguistic features possibly encoded in neural language models. These probing tasks are formulated as prediction tasks and focus on several aspects of sentence structure. We experiment with eight probing tasks to evaluate how effectively multilingual models encode linguistic properties across six Indic languages: Hindi (`hi`), Telugu (`te`), Marathi (`mr`), Kannada (`kn`), Urdu (`ur`), and Malayalam (`ml`). These eight probing tasks are grouped into three categories: surface (SentLen), syntactic (BShift and TreeDepth), and semantic (SubjNum, ObjNum, VerbGen, VerbPerson and VerbNumber). We selected these tasks because they cover different aspects of language and require different levels of abstraction and generalization. These tasks involve 3 binary and 5 multi-class classification problems. The specifics of the initial five probing tasks are thoroughly outlined in (Conneau et al., 2018) as well. We also provide brief descriptions for each probing task in Appendix D, with a summary of class labels for each task in Table 1.

IndicSentEval curation details. For each property, we gather data per language as follows. (1) *SentLen:* We iterate through all the nodes in the SSF format representation of the sentence and count number of words in each chunk. (2) *TreeDepth:* We utilize the data from the dependency tree to perform a traversal, specifically employing breadth first search. This traversal enables us to calculate the tree depth of the sentence. (3) *BShift:* For this task, we randomly (probability=0.2) select the sentences from the dataset, and then a randomly selected bigram (equal probability for all bigrams) is inverted. Sentences with inverted bigrams are marked as 1, and the rest as 0. (4) *SubjNum/ObjNum:* We identify the subject/object (a noun that can be singular or plural) of a sentence using the assigned semantic roles in the SSF format, and use the NN/NNS annotations. (5) *Verb Gender/Person/Number:* We first locate the chunk containing the main verb using the annotated chunk label from the SSF format. Then, we extract the gender/person/number information from the annotated morph output. Detailed statistics of number of samples across 6 Indic languages for 8 probing tasks are provided in Tables 10 and 11 in Appendix C. Further, Table 12 in Appendix D shows examples for each probing task per language.

Task	|C|	Labels
SentLen	8	(0-5),(6-8),(9-12),(13-16),(17-20),(21-25),(26-28),(29-32)
TreeDepth	5	(0-2),(3-5),(6-8),(9-11),(12-20)
BShift	2	0,1
SubjNum	2	singular, plural
ObjNum	2	singular, plural
VerbGen	4	masculine, feminine, neutral, any
VerbNum	3	singular, plural, any
VerbPer	7	1^{st} person, 2^{nd} person, 3^{rd} person, 1^{st} person honorific, 2^{nd} person honorific, 3^{rd} person honorific, any

Table 1: Probing task details: number of classes ($|C|$) and class labels.

3 Text Perturbation Analysis

While probing reveals what linguistic features are present in representations, perturbation-based robustness tests address a different question: "how robust is a model to various types of noise and whether it can still understand and process the core meaning of the text despite the introduced variation?" Since our current IndicSentEval dataset lacks noise examples, hence, in this study, we conducted our perturbation analysis on the input dataset to evaluate its robustness. To answer this question, we experiment with three different categories of perturbations: AppendR, DropText and Positional. We chose these perturbations because they simulate types of noise found in real datasets by introducing different degrees of noise and variation in the input text. Particularly, we experiment with the following text perturbations.

AppendR. We append a random (R) phrase to original sentence. This mimics real scenarios where additional, irrelevant data is included in text input.

DropText. DropText perturbations reflect situations where critical information is missing or only certain types of words are retained, which is common in incomplete or corrupted datasets. This includes *DropNV* (dropping words based on their part-of-speech tag, specifically both nouns (N) and verbs (V)), *DropN* (dropping all nouns), *DropV* (dropping all verbs), *DropRN* (dropping one random noun), *DropRV* (dropping one random verb), *KeepNV* (dropping all words except nouns and verbs), *KeepN* (dropping all words except nouns), and *KeepV* (dropping all words except verbs). DropText perturbations are designed to provide deeper insights into word-level attention mechanisms within models, specifically aiming to determine whether models focus more on objects (nouns), actions (verbs), or the contextual elements surrounding these key parts-of-speech.

Positional. This includes *DropF/DropL/DropFL* (replacing first/last/both words by "[UNK]" to maintain the original phrase length) and *Shuffle* (randomly shuffling the words in a sentence). These position-based text perturbations help us understand the extent to which words at specific positions (first/last) or relative positions impact the language structure encoding capabilities of various models.

Overall, these text perturbations help in understanding the contribution of specific word types and sentence structures to the encoding capabilities of multilingual models. Tables 13-18 in Appendix E display examples of perturbations for each language.

4 Methodology

Multilingual language models. We experiment with nine multilingual Transformer-based models (listed in Table 19 in Appendix F). First seven have been trained across 100+ languages; IndicBERT and MuRIL support 12 and 17 Indic languages, respectively. Representations are extracted from the encoder layers of mBERT-base, IndicBERT-base, mT5-base, XLM-R, InfoXLM and MuRIL; and from the decoder layers of BLOOM, mGPT and XGLM. We use pretrained model checkpoints from Hugging Face (Wolf et al., 2020).

Probing tasks classifier. To evaluate each probing task using a multilingual model representation, we use logistic regression (Wright, 1995) classifier with sentence representations as input and the probing task label as target. The base model is frozen. We use mean pooling across tokens to get the sentence representation. Details of the hyperparameters are reported in Appendix F.

Dataset splits. We use a stratified five-fold cross-validation approach which involves splitting the dataset into five equal parts, where four parts are used for training and the remaining part is used for testing. This process is repeated five times, with each part used for testing once. To report our results, we calculate the average performance of the model across all five folds.

Evaluation metrics. Similar to earlier studies (Conneau et al., 2018; Jawahar et al., 2019; Mohebbi et al., 2021), for all the probing tasks, we use *classification accuracy* as the evaluation metric. Let A_c and A_p be accuracy of a model on the clean and perturbed test sets, respectively. To evaluate the perturbation results for probing tasks, we use *robustness score (RS)* defined as $RS = 1 - \frac{A_c - A_p}{A_c}$. Typically, *RS* of a model ranges between 0 and 1 where 0 indicates that the model is not robust, and 1 indicates that the model is completely robust. Score > 1 suggests that the model's performance improves when the perturbation is applied.

5 Experimental Results

We measure probing accuracy independently for each multilingual model, within each layer separately. Along with six Indic languages, we measure the probing accuracy for English language across all multilingual models and compare the findings for Indic languages against English. Further, we investigate the robustness of these multilingual models by perturbing the input sentences.

5.1 Probing Results

How effectively do multilingual models encode hierarchy of linguistic structure for Indic languages? We assess the linguistic structure by contrasting universal models trained on 100+ languages with those specifically trained on Indic languages only. Unless otherwise specified, the results presented in the main paper reflect an average accuracy across encoder-based universal models, decoder-based universal models and Indic-specific models.

Surface-level tasks. We show the accuracy scores obtained for surface level task (i.e. SentLen) in the first row in Fig. 2. Analyzing the performance across languages, we observe the following patterns: (i) For encoder-based universal models as well as for Indic models, there is a trend of higher

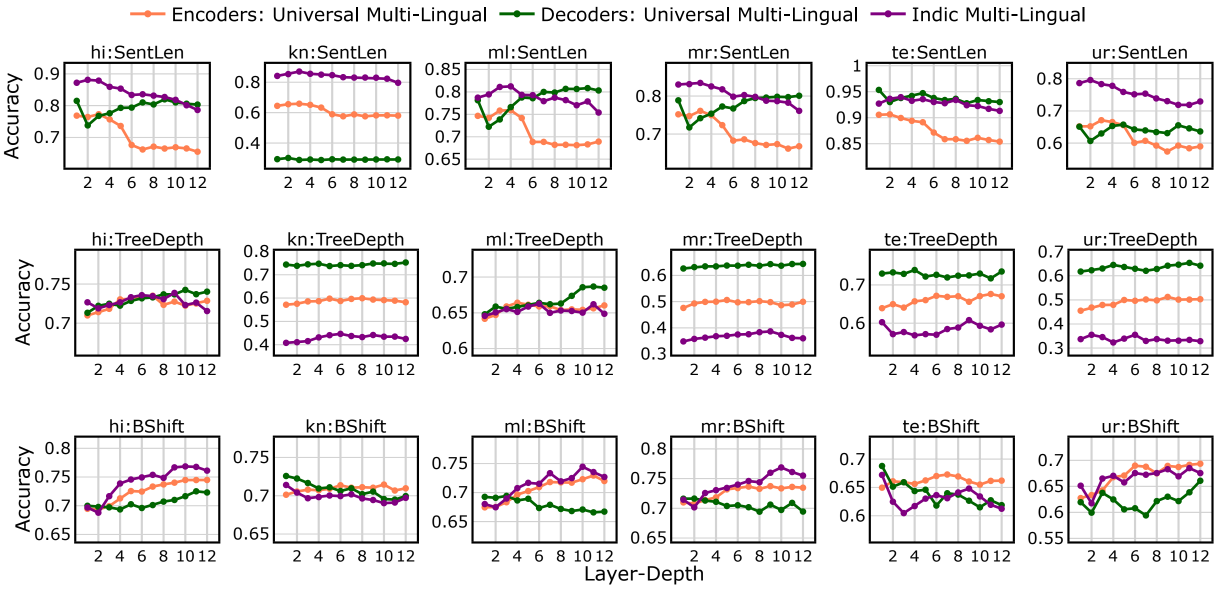

Figure 2: Probing task results: Layerwise accuracy comparisons between various multilingual representations on surface (top row) and syntactic (bottom two rows) probing tasks. We report the layerwise probing accuracies for individual multilingual models in Figs. 5 to 10 in Appendix G.

accuracy in the early (or lower) layers, which decreases in the later (or higher) layers. This pattern is expected, as surface-level tasks generally require minimal processing. (ii) Notably, the decoder-based universal multilingual models deviates from this trend, showing lower accuracy in the early layers and higher accuracy in the later layers. This seemingly unusual pattern in decoder-based universal models is actually intuitive because masked self attention in their autoregressive architectural design implies that only deeper layers in such models can effectively grasp the input length. (iii) Overall, among all the models, Indic models report best accuracy, while encoder-based universal models display poorer performance. This is likely because Indic models are specifically trained on Indic languages, making them more attuned to the nuances and idiosyncrasies of these languages. On the other hand, universal models, which are designed for universal applicability, might struggle with specific linguistic features unique to Indic languages. These features include script differences, morphological complexity, and visual factors such as orthography and word length. Tokenizers of universal models tokenize Indic language inputs to many tokens with little correlation with actual input length in words.

We present the individual model-specific results across languages in Figs. 5 to 10 in Appendix G. We observe that among all the models, IndicBERT reports the best accuracy, while universal multilingual model XLM-R displays poor performance. More detailed analysis is reported in Appendix G.

Syntactic tasks. In the bottom two rows of Fig. 2, we display the accuracy scores for syntactic tasks. For TreeDepth, we observe that probing accuracy tends to be higher in the middle layers for various (model, language) combinations. This trend is particularly notable in both encoder-based universal models and Indic models. Moreover, this pattern is consistent for three languages: `hi`, `kn` and `ml`. However, decoder-based multilingual models do not exhibit any clear layer-wise trend, and the same applies to the other three Indic languages. Encoders, with their bidirectional attention, might be inherently better at capturing hierarchical structures, while decoders, with their unidirectional attention, might excel in tasks requiring sequential processing. The lack of a clear trend in decoder-based multilingual models might indicate that these models distribute syntactic processing more evenly across layers. The varying performance across languages underscores the importance of considering linguistic diversity in model training.

In contrast to TreeDepth, for BShift task, we generally observe higher probing accuracy in the later layers for both encoder-based universal and Indic models across various languages. Notably, decoder-based universal models exhibit a decreasing trend in accuracy for `kn`, `ml` and `te`, while showing an increasing trend for `hi` and `ur`. This suggests that

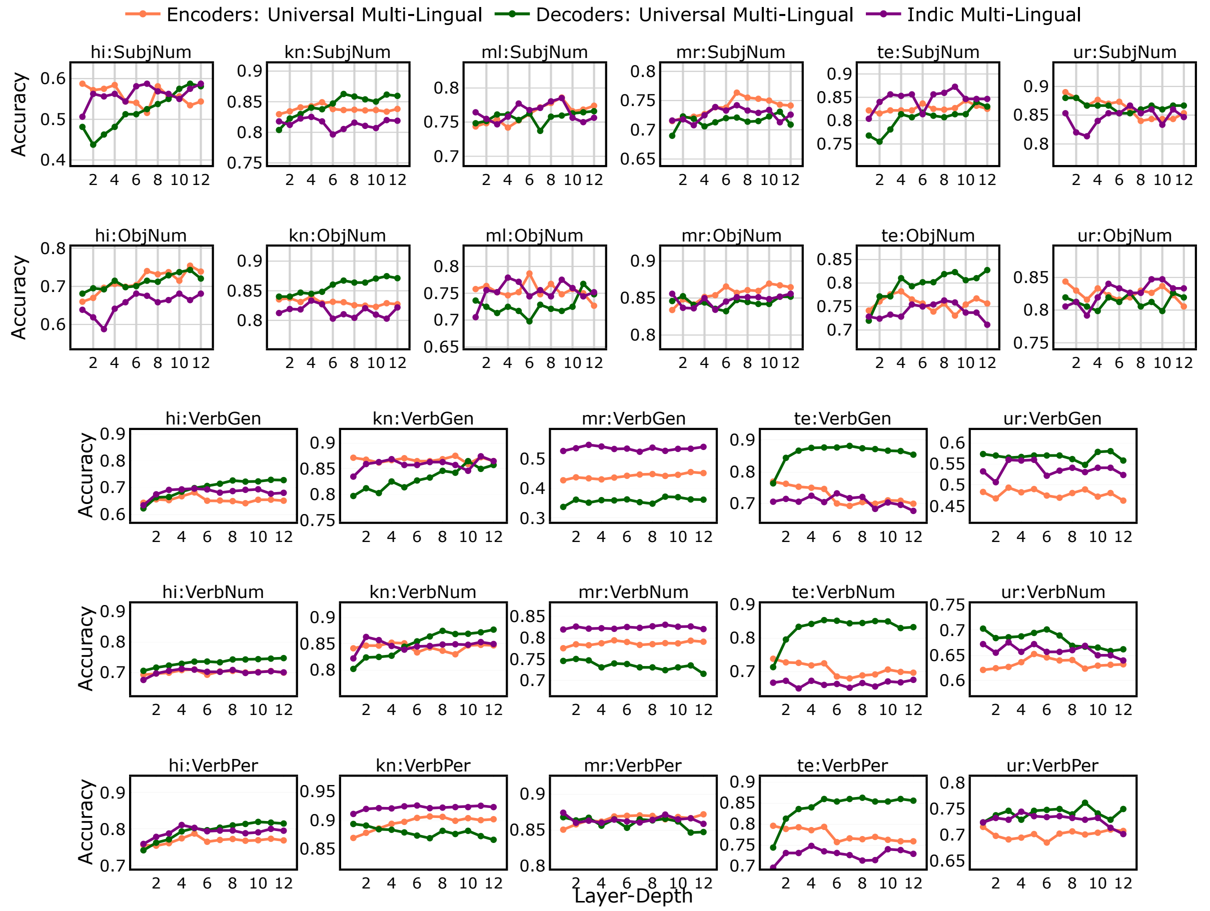

Figure 3: Probing task results: Layerwise accuracy comparisons between various multilingual representations on semantic probing tasks. For Malayalam, there is an absence of SSF data for the VerbGen, VerbPer, and VerbNum tasks. We report the layerwise probing accuracies for individual multilingual models in Figs. 5 to 10 in Appendix G.

languages with different syntactic structures may require different layers to process syntactic tasks effectively.

Overall, when comparing performance across models and tasks, decoder-based universal models stand out for their superior accuracy in capturing tree depth information across different languages. Conversely, Indic models show notable proficiency in BShift. These findings highlight the need for comprehensive evaluation across multiple tasks and languages to understand model capabilities fully. Evaluating models on diverse syntactic tasks can reveal strengths and weaknesses that might not be apparent from a single task. These observations contribute to a deeper understanding of how various multilingual models process syntactic information, demonstrating both model-specific and language-specific trends and capabilities in linguistic tasks.

From the individual model specific results in Figs. 5 to 10 in Appendix G, we observe that among all the models, InfoXLM and mT5 stand out for their superior accuracy in capturing tree depth information across different languages. Conversely, Indic model MuRIL shows notable proficiency in BShift. More detailed analysis is reported in Appendix G.

Semantic tasks. We plot the accuracy scores obtained for semantic tasks in Fig. 3: SubjNum and ObjNum in the first 2 rows, and VerbGen, VerbNum and VerbPer in the last 3 rows. The last three rows do not have results for Malayalam (`ml`) since we do not have labeled data for `ml` for those tasks.

From SubjNum and ObjNum results, we make the following observations. For decoder-based universal models, we observe an increasing trend from lower to higher layers for `hi`, `kn` and `te`. Dravidian languages often have more complex morphological systems for marking plurality, with a variety of suffixes and sometimes even changes in the noun stem itself. Hence, this increasing trend makes sense for `kn` and `te`. When considering other languages and models, we note that both encoder-based and Indic models exhibit an increasing trend for the ObjNum task, specifically for `hi` and `mr`. This highlights the variability in model performance based on the language. Interestingly, the middle layers of the encoder-based universal models and Indic models show higher probing accuracy for languages such as `hi`, `ml`, `mr` and `te` in the SubjNum task.

For the verb-related tasks, decoder-based universal models perform the best for most languages and tasks. Also, in most cases, the last layer is the most predictive, except in `mr` and `ur` for gender, number and person detection, where initial layers provide better results. This indicates that gender, number and person detection in `mr` and `ur` is straightforward and does not need deep processing. We observe a decreasing trend for encoder-based universal models for language `te`. For the Indic models, we observe an increasing trend for two languages `hi` and `kn` across tasks. Overall, both encoder-based universal and Indic models show mixed trend in performance across layers.

From the individual model specific results in Figs. 5 to 10 in Appendix G, we observe that XLM-R exhibits lower accuracy and lacks a discernible trend when it comes to capturing semantics. MuRIL has the highest probing accuracy although it performs the worst for `kn` ObjNum task. More detailed analysis is reported in Appendix G.

Comparison of encoding performance of multilingual models for linguistic properties for English vs Indic languages. We conduct probing for English across nine multilingual models for five tasks: SentLen, TreeDepth, BShift, SubjNum, and ObjNum using SentEval dataset (Conneau et al., 2018). Fig. 11 in Appendix G reports the probing accuracy. Surprisingly, across all multilingual models, we observe that surface features show a decreasing trend from lower to higher layers. For syntactic TreeDepth feature probing accuracy is higher in the middle layers, while BShift has an increasing trend from lower to higher layers. Finally, the semantic tasks SujbNum and ObjNum are best encoded in the later layers. This implies that English is encoded in the same manner in both universal and Indic multilingual models, whereas Indic languages show mixed results across models.

Overall insights from probing experiments. While Indic-specific models like MuRIL and IndicBERT are likely the best at capturing language properties within the realm of Indic languages due to their targeted training, both encoder and decoder-based universal models like mBERT, InfoXLM, BLOOM and mGPT show mixed results. Their broader training might enable them to capture more general properties across many languages, but they may lack a deep understanding specific to each language, particularly those less represented in their training corpus. The effectiveness of these models thus depends on specific linguistic features and tasks, as well as the range of languages being considered. mBERT and MuRIL capture linguistic features for `hi` very well. For `mr` and `te`, mT5 and MuRIL show better accuracy for surface, syntactic and semantic tasks. This is in line with the fact that `hi`, `mr`, and `te` are better represented in pretraining datasets for these models.

	hi	kn	ml	mr	te	ur
mBERT	0.794	0.736	0.583	0.626	**0.980**	0.715
IndicBERT	0.807	0.777	0.576	0.662	0.921	**0.781**
XLM-R	0.883	0.883	0.675	0.692	**0.981**	**0.865**
InfoXLM	**0.921**	**1.093**	0.698	0.702	0.900	0.662
MuRIL	0.778	0.728	0.575	0.604	**0.990**	0.704
BLOOM	**0.957**	**0.903**	**0.731**	**0.742**	0.966	0.707
mT5	**0.961**	0.790	**0.773**	**0.767**	0.952	**0.757**
mGPT	0.950	**1.040**	**0.728**	**0.736**	0.946	0.723
XGLM	0.954	0.972	0.724	0.730	0.956	0.715

Table 2: Comparison of robustness scores on probing tasks: multilingual models across languages, averaged across layers, highlighting top-3 scores.

	Sent Len	Tree Depth	Subj Num	Obj Num	Verb Gen	Verb Num	Verb Per
mBERT	0.398	0.487	0.916	0.924	0.839	0.903	**0.939**
IndicBERT	0.364	0.504	0.928	0.931	0.874	**0.947**	**0.969**
XLM-R	0.492	0.564	**1.030**	**0.996**	**0.968**	**1.001**	**1.006**
InfoXLM	**0.855**	0.836	0.895	0.916	0.738	0.808	0.926
MuRIL	0.385	0.471	0.900	**0.935**	0.833	**0.906**	0.918
BLOOM	0.604	**0.870**	**0.933**	0.930	**0.892**	0.849	0.899
mT5	0.582	**0.905**	0.915	0.931	**0.898**	0.853	0.914
mGPT	**0.918**	**0.872**	**0.932**	**0.936**	0.820	0.879	0.931
XGLM	**0.761**	0.865	0.930	0.925	0.834	0.864	0.904

Table 3: Comparison of robustness scores on probing tasks: multilingual models vs. probing tasks, averaged across layers, highlighting top-3 scores.

5.2 Perturbation Results

We perform 13 different text perturbations to understand the contribution of specific word types and sentence structures to the encoding capabilities of multilingual language models[2]. We analyze the impact of such text perturbations for every pair of (model, language), (model, probing task), (perturbation, model) and (language, probing task) in Tables 2, 3, 4 and 5 respectively.

We report results for the layerwise perturbation analysis and other pairs like (perturbation, language) and (perturbation, probing task) in Tables 6, 7 and 8. All these tables show weighted

[2]Perturbations do not make sense for BShift probing task.

	mBERT	Indic BERT	XLM-R	Info XLM	MuRIL	BLOOM	mT5	mGPT	XGLM
AppendR	0.810	**0.764**	0.886	0.877	0.790	0.839	0.837	0.938	0.888
DropNV	**0.735**	**0.780**	**0.870**	**0.793**	**0.735**	**0.810**	**0.802**	**0.865**	**0.838**
DropN	0.819	0.847	**0.876**	0.865	0.808	0.826	0.825	0.886	0.856
DropV	**0.747**	**0.779**	0.882	**0.822**	**0.743**	0.831	0.838	0.914	0.873
DropF	0.828	0.847	0.884	0.892	0.814	0.847	0.849	0.950	0.898
DropFL	0.760	0.794	0.886	0.844	0.754	0.849	0.859	0.952	0.901
DropL	0.768	0.796	0.887	0.848	0.759	0.849	0.852	0.953	0.901
DropRN	0.820	0.845	0.883	0.890	0.811	0.845	0.837	0.938	0.891
DropRV	0.768	0.793	0.886	0.849	0.760	0.847	0.851	0.949	0.898
KeepNV	0.812	0.842	0.890	0.853	0.803	**0.817**	**0.813**	0.881	0.849
KeepN	**0.734**	**0.779**	0.889	**0.780**	**0.731**	0.823	**0.817**	**0.872**	**0.847**
KeepV	0.824	0.866	**0.849**	0.851	0.825	**0.820**	**0.809**	**0.858**	**0.839**
Shuffle	0.812	**0.746**	0.889	0.884	0.798	0.845	0.844	0.935	0.890

Table 4: Comparison of robustness scores on probing tasks: multilingual models vs. perturbation types, averaged across layers, highlighting top-3 scores.

averages across marginalized dimensions.

Which multilingual models are more robust to perturbations in Indic languages? Table 2 shows that universal models like InfoXLM, BLOOM, mT5 and mGPT show greater resilience to perturbations in at least four languages. In contrast, the universal model (mBERT) and the Indic-specific models (IndicBERT and MuRIL) display a more significant accuracy drop across all the Indic languages. Accuracy drop for BERT-specific models is perhaps because cross-lingual transfer might be less effective, resulting in decreased accuracy compared to other multilingual models.

Which multilingual models are more robust across probing tasks? Table 3 shows that universal models have greater robustness compared to Indic models and mBERT. Additionally, surface and syntactic probing tasks are significantly impacted by perturbations compared to semantic properties.

Which text perturbations have the greatest impact on multilingual models? From Table 4, we observe: (i) Dropping both nouns and verbs has an adverse effect on all models. (ii) Similarly, dropping only verbs or retaining only nouns affects the performance drop across models. Thus, eliminating nouns and verbs can lead to losing vital information necessary for accurate predictions. Discarding nouns or verbs can disrupt the syntactic coherence of text, making it more challenging for the model to comprehend and process.

Which Indic languages are more robust across probing tasks? Table 5 shows that models exhibit greater robustness across probing tasks for `hi` and `te`, especially for semantic properties. Conversely, surface and syntactic properties are more affected by perturbations, except for similar language structure observed for `te` and `kn`. Additionally, languages such as `ur` and `mr` are more impacted by perturbations due to relatively lower training token counts compared to other languages.

	Sent Len	Tree Depth	Subj Num	Obj Num	Verb Gen	Verb Num	Verb Per
`hi`	0.436	0.338	**1.104**	**1.009**	**0.931**	**0.951**	**0.972**
`kn`	**0.793**	**0.632**	0.902	0.896	0.818	0.860	0.967
`ml`	0.362	0.367	0.859	0.897	**0.968**	**1.001**	**1.006**
`mr`	0.277	0.578	0.836	**0.938**	0.738	0.808	0.926
`te`	**0.895**	**0.939**	**0.952**	**0.968**	**0.915**	**1.003**	**0.994**
`ur`	0.232	0.580	**0.949**	0.935	0.738	0.837	0.875

Table 5: Comparison of robustness scores on probing tasks: Indic languages vs. probing tasks, averaged across layers, highlighting top-3 scores.

	`hi`	`kn`	`ml`	`mr`	`te`	`ur`
SentLen	1, 2, 3	4, 3, 2	3, 4, 1	1, 3, 2	1, 2	3, 1, 2
TreeDepth	7, 5, 6	5, 4, 9	4, 5, 3	4, 8, 9	10, 11, 12	3, 2, 5
VerbGen	10, 8, 5	11, 4, 12	-	11, 10, 12	6, 8, 9	3, 8, 9
VerbNum	Equal	3, 4, 5	-	2, 4, 5	10, 12	5, 4, 3
VerbPer	Equal	Equal	-	3, 4, 5	5, 11, 12	4, 5
SubjNum	N/A	Equal	9, 8, 7	7, 10, 9	10, 11	1, 2
ObjNum	11, 12	4, 3, 9	Equal	11, 10, 12	11, 12	10, 9, 11

Table 6: Summary of the layerwise robustness analysis averaged across multilingual models, considering 13 perturbations. Each cell reports the most affected layers after text perturbations for each probing task and language. "Equal" denotes that all layers are affected, while '-' indicates the absence of a probing dataset for that particular language.

Which layers are more affected due to text perturbations for Indic languages? Table 6, we observe: (i) across all Indic languages, early layers are impacted for surface properties. (ii) for TreeDepth syntactic property, middle layers are affected for all languages except `te`. (iii) Surprisingly, later layers are impacted more than early and middle layers. Regarding semantic properties, for `ur`, early to middle layers are impacted more than later layers, except for ObjNum. Similarly, for `kn`, SubjNum and ObjNum are impacted on middle layers; for `mr`, VerbNum and VerbPer are affected.

Which text perturbations have the greatest impact on probing tasks across six Indic languages? Tables 7 and 8 show that dropping nouns and verbs significantly affects accuracy across all six languages. Specifically, dropping verbs impacts accuracy of three verb-based semantic tasks. Similarly, position perturbations, such as keeping nouns in specific positions, affect verb tasks, while keeping verbs or nouns affects surface and syntactic tasks.

Overall insights from perturbation experiments. Text perturbation analysis reveals that universal models such as InfoXLM, BLOOM, mGPT,

	hi	kn	ml	mr	te	ur
AppendR	**0.788**	0.854	0.615	0.666	1.000	0.743
DropNV	**0.812**	**0.826**	**0.605**	**0.664**	**0.833**	**0.740**
DropN	0.866	**0.828**	**0.609**	0.666	1.023	0.743
DropV	**0.813**	0.847	0.619	0.666	**0.878**	**0.742**
DropF	0.865	0.853	0.618	0.666	1.067	**0.742**
DropFL	0.847	0.851	0.618	0.666	0.906	0.746
DropL	0.848	0.855	0.616	**0.665**	0.908	0.751
DropRN	0.862	0.855	**0.612**	0.667	1.045	0.745
DropRV	0.842	0.856	0.618	0.666	0.913	0.748
KeepNV	0.848	0.833	0.619	0.673	1.022	**0.741**
KeepN	**0.771**	**0.827**	0.634	0.672	**0.857**	0.755
KeepV	0.870	**0.828**	0.678	**0.540**	1.021	0.746
Shuffle	0.845	0.854	0.615	0.666	0.935	0.747

Table 7: Comparison of robustness scores on probing tasks: Indic languages vs. perturbation types, averaged across layers, Lowest robustness values per column are highlighted in bold.

	Sent Len	Tree Depth	Subj Num	Obj Num	Verb Gen	Verb Num	Verb Per
AppendR	0.525	0.573	**0.899**	**0.898**	0.885	0.932	0.964
DropNV	**0.446**	0.564	**0.922**	0.931	**0.771**	**0.861**	**0.916**
DropN	0.472	0.573	0.934	0.924	0.921	0.967	0.993
DropV	0.508	0.572	0.941	0.955	**0.775**	**0.861**	**0.901**
DropF	0.544	0.585	0.940	0.960	0.908	0.955	0.984
DropFL	0.546	0.587	0.946	0.964	0.784	0.863	**0.916**
DropL	0.541	0.589	0.945	0.952	0.805	0.874	0.920
DropRN	0.537	0.586	0.935	0.933	0.907	0.958	0.990
DropRV	0.538	0.588	0.942	0.954	0.806	0.876	0.919
KeepNV	0.464	0.562	0.944	0.952	0.909	0.953	0.987
KeepN	**0.443**	**0.559**	0.963	0.972	**0.746**	**0.854**	**0.893**
KeepV	**0.417**	**0.554**	0.924	**0.915**	0.949	0.978	1.023
Shuffle	0.506	**0.550**	**0.905**	**0.917**	0.891	0.939	0.967

Table 8: Comparison of robustness scores on probing tasks: perturbation types vs. probing tasks, averaged across layers. Lowest robustness values per column are highlighted in bold.

XGLM and mT5 demonstrate higher resilience to perturbations compared to BERT-like models mBERT, IndicBERT, and MuRIL. Perhaps, the larger multilingual models are more robust as they rely less on language specific word order compared to Indic models (Dufter and Schütze, 2020; Liang et al., 2023). Specifically, dropping both nouns and verbs proves to be particularly sensitive across all languages, impacting semantic and syntactic properties significantly. Perturbations, such as position alterations, also affect model performance, emphasizing the importance of considering linguistic nuances in robustness testing.

5.3 Correlation Analysis of Probing with Downstream Tasks

We conducted a correlation analysis between our IndicSentEval probing results and downstream task performance from the IndicGLUE benchmark. Specifically, we analyze how the performance of the IndicSentEval probing tasks correlates with the results of the IndicGLUE downstream tasks and other benchmarks. This includes models such as IndicBERT, MuRIL, mBERT, and XLM-R across syntactic and semantic tasks like NER, POS tagging, sentiment analysis, and natural language inference. The Table 9 shows the aggregated scores across multiple Indic languages. We make the following observations from Table 9: (i) Models like MuRIL and IndicBERT, which scored highest on syntactic and semantic probes, also performed best on corresponding tasks like POS tagging, NER (syntactic), and sentiment classification, NLI (semantic). (ii) Universal models like mBERT showed weaker probe performance and correspondingly lower scores on most IndicGLUE tasks, especially for morphologically rich languages. (iii) Tasks requiring deeper understanding (e.g. NLI) correlated more strongly with semantic probes, while syntactic probes aligned better with tagging tasks. This implies alignment between what probes measure and what different task categories demand. These findings suggest that the linguistic properties measured by probing are indeed predictive of real-world task performance, particularly in morphologically rich Indic languages.

Model	Syntactic Probes	Semantic Probes	POS Tagging	NER	Sentiment Analysis	NLI
IndicBERT	0.82	0.80	0.88	0.84	0.81	0.79
MuRIL	0.85	0.86	0.89	0.83	0.85	0.87
mBERT	0.72	0.70	0.76	0.88	0.70	0.69
XLM-R	0.80	0.78	0.85	0.85	0.83	0.82

Table 9: Performance comparison of multilingual models across various probing and downstream tasks.

6 Discussion and Conclusion

We evaluated 9 multilingual Transformer-based models on 8 probing tasks to understand linguistic structures in 6 Indic languages, using our contributed INDICSENTEVAL dataset. Indic-specific models like MuRIL and IndicBERT excel due to targeted training, while universal models like mBERT, InfoXLM, BLOOM, mGPT, and XGLM show mixed results. Perturbation analysis reveals decoder-based models are the most robust. Overall, our findings highlight the importance of language model interpretability. Language proficiency is seen in models with comprehensive training datasets. Encoding ability and perturbation impact vary across languages and models, underscoring the need for robust training strategies and tailored architectures to handle linguistic variations and perturbations effectively.

7 Limitations

The current work focused on only 6 Indic languages. It would be interesting to expand this to more Indic languages.

We performed experiments with base versions of various models. It would be interesting to see if larger variants perform better at these tasks and if the trends differ.

We experimented with a basic set of linguistic properties. It would be nice to explore more complex linguistic properties like morphological tagging, syntactic parsing, and semantic similarity.

8 Ethics Statement

All the models used in this work are publicly available on Hugging Face and free for research.

We utilized publicly accessible resources in SSF format from https://ltrc.iiit.ac.in/showfile.php?filename=downloads/kolhi/ and https://ltrc.iiit.ac.in/showfile.php?filename=downloads/lingResources/newreleases.html. The datasets are licensed under Creative Commons by Non-Commercial 4.0 (CC BY-NC 4.0). No anticipated risks are associated with using the data from these provided links. We adapted the accessible resources to generate diverse probing and perturbation datasets as required.

References

Ahmed Abdelali, Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. 2022. Post-hoc analysis of arabic transformer models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 91–103.

Judit Acs, Endre Hamerlik, Roy Schwartz, Noah A Smith, and Andras Kornai. 2024. Morphosyntactic probing of multilingual bert models. *Natural Language Engineering*, 30(4):753–792.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Akshar Bharati, Vineet Chaitanya, Rajeev Sangal, and KV Ramakrishnamacharyulu. 1995. *Natural language processing: a Paninian perspective*. Prentice Hall India Learning Private Limited.

Akshar Bharati, Rajeev Sangal, and Dipti M Sharma. 2007. Ssf: Shakti standard format guide. *Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India*, pages 1–25.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 186–189.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $ &!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for natural language generation of indic languages. *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863.

Anh Dang, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In *13th edition of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2024)*, pages 177–188. Association for Computational Linguistics (ACL).

Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, et al. 2023. Indicmt eval: A dataset to meta-evaluate machine translation metrics for indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for bert's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437.

Daniel Edmiston. 2020. A systematic analysis of morphological content in bert models for multiple languages. *arXiv preprint arXiv:2004.03032*.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.

Mary Godunova and Ekaterina Voloshina. 2024. Probing of pretrained multilingual models on the knowledge of discourse. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 78–90.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8018–8025.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, page 2.

Mohammed Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, Mitesh M Khapra, et al. 2024. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M Khapra, and Pratyush Kumar. 2022. Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for indic languages. In *Annual Meeting of the Association for Computational Linguistics*.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. 2021. Exploring the role of bert token representations to explain sentence probing results. In *The 2021 Conference on Empirical Methods in Natural Language Processing*, pages 792–806.

Pavan Kalyan Reddy Neerudu, Subba Oota, Mounika Marreddy, Venkateswara Kagita, and Manish Gupta. 2023. On robustness of finetuned transformer-based nlp models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7180–7195.

Subba Reddy Oota, Mounika Marreddy, Manish Gupta, and Raju Bapi. 2023. How does the brain process syntactic structure while listening? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6624–6647.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Soumya Sanyal, Zeyi Liao, and Xiang Ren. 2022. RobustLR: A diagnostic benchmark for evaluating logical robustness of deductive reasoners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9614–9631.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, BigScience Workshop, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Juhi Tandon and Dipti Misra Sharma. 2017. Unity in diversity: A unified parsing strategy for major indian languages. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 255–265.

Maria Tikhonova, Vladislav Mikhailov, Dina Pisarevskaya, Valentin Malykh, and Tatiana Shavrina. 2023. Ad astra or astray: Exploring linguistic knowledge of multilingual bert through nli task. *Natural Language Engineering*, 29(3):554–583.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Raymond E Wright. 1995. Logistic regression. *Reading and Understanding Multivariate Statistics*, pages 217–244.

Fei Xia, Owen Rambow, Rajesh Bhatt, Martha Palmer, and Dipti Misra Sharma. 2008. Towards a multi-representational treebank. *LOT Occasional Series*, 12:159–170.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Xiaohan Zhang, Shaonan Wang, Nan Lin, and Chengqing Zong. 2022. Is the brain mechanism for hierarchical structure building universal across languages? an fMRI study of Chinese and English. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7852–7861.

Jianyu Zheng and Ying Liu. 2022. Probing language identity encoded in pre-trained multilingual models: a typological view. *PeerJ Computer Science*, 8:e899.

Overview of Appendix Sections

A Related Work

Our work is most closely related to that of Jawahar et al. (2019) and Mohebbi et al. (2021), who focus on understanding the interpretability of language models via probing and observing linguistic structures in English. Our study on multilingual language models for six Indic languages complements their English-focused studies.

Our work also relates to the growing literature that creates resources for low-resource Indic languages. Several recent studies have developed resources including, Indic language models (Khan et al., 2024), Indic NLP suite (Kakwani et al., 2020), Indic BART (Dabre et al., 2021), Indic NLG (Kumar et al., 2022), IndicMT Eval (Dixit et al., 2023), and Naamapadam (Mhaske et al., 2023). Our research is the first to interpret both universal and Indic-specific models for Indic languages through probing representations and assessment of robustness, while also providing implications for future research directions. Overall, we complement these works by studying the linguistic structure of a wide range of low-resource Indic languages on multilingual models.

Non-English and Multilingual Probing Studies: Research on probing language models has increasingly moved beyond English to multilingual and non-English contexts, particularly focusing on morphologically rich and low-resource languages (Pires et al., 2019; Edmiston, 2020; Abdelali et al., 2022; Acs et al., 2024). For instance, Pires et al. (2019) is one of the earliest probing studies on mBERT, showing that mBERT enables zero-shot cross-lingual transfer using shared representations, but performs best on typologically similar languages, struggling with distant or low-resource ones. Extending from previous work, Edmiston (2020) probe BERT's hidden layers for discrete morphological features in five languages (e.g. gender, tense, case) and observe that BERT encodes discrete morphological features across multiple languages, but lacked explicit analysis of model robustness across language families. More recently, Abdelali et al. (2022) probed several Arabic BERT-based models (and mBERT) and observed that Arabic BERT models capture morphology in lower layers and syntax in higher layers, but had limited focus on dialectal variation and cross-model comparison. Very recently, Acs et al. (2024) introduced a large multilingual probing dataset across 42 languages and found strong morphological encoding in mBERT/XLM-R, but did not analyze model robustness or compare to language-specific models. While these earlier studies primarily focused on morphological encoding in universal multilingual models (like mBERT and XLM-R), IndicSentEval uniquely evaluates both universal multilingual models and Indic-specific models, explicitly comparing their robustness and representational capabilities across six morphologically rich and script-diverse Indic languages.

Critiques of Probing Methodology and Recent Advances Probing methods have faced significant criticism in recent years, primarily for overinterpreting what models "know" from linear classifier performance (Hewitt and Liang, 2019; Voita and Titov, 2020; Ravichander et al., 2021). For instance, Hewitt and Liang (2019) showed that complex probes may "memorize" tasks, not reveal what's encoded in the representations, ensuring probes extract meaningful, not spurious, information. Extended to previous work, Voita and Titov (2020) proposed using Minimum Description Length (MDL) to evaluate how efficiently linguistic features can be extracted. They show that MDL probes reward simpler models and fewer training samples, improving interpretability and robustness. Further, Ravichander et al. (2021) demonstrate that models can encode linguistic properties even if these properties are not necessary for the task the model was trained on. Specifically, they highlight the importance of careful controls when designing probing experiments, as high probing accuracy does not necessarily indicate that the probed information is utilized by the model during its primary task. In our IndicSentEval work, we address these critiques by: (i) using lightweight probes with standardized architecture across all models and languages; (ii) focusing on relative comparisons (across layers, languages, and perturbations) rather than absolute performance; (iii) including input-level perturbations to test mechanistic interpretability and model robustness; and (iv) complementing standard probing with morpho-semantic tasks tailored to Indic grammar, allowing deeper insight into model behavior for typologically diverse languages.

Differences from English-Centric Findings: Probing studies in non-English and multilingual contexts have revealed notable differences from English-centric findings, due to linguistic diversity and the multilingual training regime (Zheng and Liu, 2022; Tikhonova et al., 2023; Godunova and Voloshina, 2024; Dang et al., 2024). For instance, Zheng and Liu (2022) showed that multilingual models encode universal syntactic features well but struggle with fine-grained morphology (e.g., verb agreement), unlike their performance on English tasks. Similarly, Tikhonova et al. (2023) reported that mBERT lacks the clear layer-wise linguistic hierarchy found in English BERT, suggesting architectural behavior does not generalize across languages. More recently, studies including Godunova and Voloshina (2024) found that discourse-level understanding remains uniformly weak across high- and low-resource languages, contrasting with better performance on lower-level tasks in English. Also, Dang et al. (2024) found that GPT models generalize well to inflect unseen words in morphologically simple languages, but performance declines sharply as morphological complexity increases. In contrast to prior studies that either focused primarily on English or reported high-level trends across many languages, IndicSentEval reveals that while multilingual models exhibit consistent representational hierarchies for English, they show mixed and often unstable behavior across Indic languages, especially in encoding syntactic and morphological features-highlighting language-specific gaps that broader multilingual benchmarks often obscure.

B SSF format

We curate the INDICSENTEVAL dataset from resources generated by the ILMT initiative, which serves as an Indic language counterpart to the SentEval dataset and offers labeled data for the eight

```
<Sentence id='4'>
1     ((   NP  <fs name='NP' drel='k1:VGF'>
1.1 डेंगू   N_NNP    <fs af='डेंगू,n,m,sg,3,d,0,0' name='डेंगू' posn='10'>
      ))
2     ((   NP  <fs name='NP2' drel='k7t:VGF'>
2.1 मलेरिया N_NNP    <fs af='मलेरिया,n,m,sg,3,o,0,0' name='मलेरिया' posn='20'>
2.2 के     PSP <fs af='के,psp,,,,,,' name='के' posn='30'>
2.3 बाद    N_NST    <fs af='बाद,nst,m,sg,3,d,,' name='बाद' posn='40'>
      ))
3     ((   NP  <fs name='NP3' drel='r6:NP4'>
3.1 दूसरे   QT_QTO   <fs af='दूसरा,num,m,sg,,o,,' name='दूसरे' posn='50'>
3.2 नंबर   N_NN     <fs af='नंबर,n,m,sg,3,o,0,0' name='नंबर' posn='60'>
3.3 की     PSP <fs af='का,psp,f,sg,,d,,' name='की' posn='70'>
      ))
4     ((   NP  <fs name='NP4' drel='k1s:VGF'>
4.1 सबसे   RP_INTF <fs af='सबसे,avy,,,,,,' name='सबसे' posn='80'>
4.2 महत्वपूर्ण     JJ   <fs af='महत्वपूर्ण,adj,any,any,,d,,' name='महत्वपूर्ण' posn='90'>
4.3 उष्णकटिबंधीय  JJ   <fs af='उष्णकटिबंधीय,adj,any,any,,d,,' name='उष्णकटिबंधीय' posn='100'>
4.4 बीमारी N_NN     <fs af='बीमारी,n,f,sg,3,d,0,0' name='बीमारी' posn='110'>
      ))
5     ((   VGF <fs name='VGF' stype='declarative' voicetype='active'>
5.1 है     V_VM     <fs af='है,v,any,sg,3,,है,hE' name='है' posn='120'>
      ))
6     ((   BLK <fs name='BLK' drel='rsym:VGF'>
6.1 ।     RD_PUNC <fs af='।,punc,,,,,,' name='।' posn='130'>
      ))
</Sentence>
```

Figure 4: A sample of an SSF formatted sentence in Hindi language.

probing tasks. We utilize the morph and chunk level Indic languages data (Tandon and Sharma, 2017; Bhatt et al., 2009; Xia et al., 2008), which is available in Shakti Standard Format (SSF) (Bharati et al., 2007, 1995). SSF is a highly readable representation for storing Indic language data with linguistic annotations. We refer the reader to read (Bharati et al., 2007) for more details about the SSF format.

An example of a Hindi sentence in the SSF format is illustrated in Fig. 4. Each line in Fig. 4 delineates a word within a sentence. Every line of a sentence representation in SSF comprises four components: Address, token, Category, and Attribute-value pairs. The Address encompasses two numbers, denoting the chunk number the word resides in and its relative position within the chunk. Token signifies the POS tag corresponding to the word. The feature list, articulated as "<fs af = root, category, gender, number, person, case, tense, aspect>", encapsulates linguistic feature information of a word. The attribute fields are fixedly positioned and separated by commas. Attributes remain blank if a property is inapplicable to the word. Properties like root, category, gender, number, person, and case are delineated within the feature list.

C INDICSENTEVAL dataset statistics

Lang	Sent Len	BShift	Tree Depth	Subj Num	Obj Num	Verb Gen	Verb Num	Verb Per
hi	12202	12202	8911	398	812	8911	8897	8897
te	3192	3192	3192	763	578	2243	2635	2501
ur	2363	2363	2363	374	356	1463	1465	1446
ml	7667	7667	7667	1556	642	-	-	-
kn	9806	9806	9806	4790	3690	1329	5371	5448
mr	12029	12029	12029	2488	3815	5934	7637	6791

Table 10: Number of samples for different probing tasks per language in INDICSENTEVAL. The symbol '-' signifies the absence of SSF data for the Malayalam language for the VerbGen, VerbPer, and VerbNum tasks.

Language	Vocabulary Size
Hindi	19589
Kannada	25310
Malayalam	28723
Marathi	35607
Telugu	5834
Urdu	5593

Table 11: Vocabulary size of each language from INDICSENTEVAL.

Language	Sentence	Probing Task labels
Marathi	शिंपल्याच्या शेतीने जगाचे लक्ष इकडे आकर्षित केले आहे . Mussel farming has attracted the attention of the world here.	SentLen → 8; TreeDepth → 3; SubjNum →singular(sg); ObjNum → singular(sg); BShift → शिंपल्याच्या शेतीने जगाचे लक्ष इकडे ***केले आकर्षित*** आहे (Mussel farming **attracted has** the attention of the world here); VerbGen → male(m); VerbNum → singular(sg); VerbPer → 1st person (1)
Kannada	ಅಳಿಲು ಭಕ್ತಿ ಮಳಲು ಸೇವೆ ' ಎಂಬಂತೆ ಶಾಲಾ ಬಾಲಕರು ಸಹ ಯಥಾಶಕ್ತಿ ಓದುಬರೆಹ ಕಲಿಸಿದರು . School boys were also taught reading and writing to the best of their ability as 'Alilu Bhakti Malalu Seva'.	SentLen → 11; TreeDepth → 3; SubjNum →singular(sg); ObjNum → singular(sg); BShift → ಅಳಿಲು ಭಕ್ತಿ ಮಳಲು ಸೇವೆ ' ಎಂಬಂತೆ ***ಬಾಲಕರು ಶಾಲಾ*** ಸಹ ಯಥಾಶಕ್ತಿ ಓದುಬರೆಹ ಕಲಿಸಿದರು (**boys school** were also taught reading and writing to the best of their ability as 'Alilu Bhakti Malalu Seva'); VerbGen → any(any); VerbNum → plural(pl); VerbPer → 3rd person (3)
Malayalam	കാർഷിക പദ്ധതികളുടെ അടങ്കൽ 600കോടി രൂപ ആയി പുതിയ സർക്കാർ ഉയർത്തി . The new government has increased the amount of agricultural projects to 600 crore rupees.	SentLen → 10; TreeDepth → 3; SubjNum →singular(sg); ObjNum → singular(sg); BShift → കാർഷിക പദ്ധതികളുടെ അടങ്കൽ 600കോടി രൂപ ***പുതിയ ആയി*** സർക്കാർ ഉയർത്തി. (The **government new** has increased the amount of agricultural projects to 600 crore rupees.) ;
Urdu	شروع شروع میں ریڈیو پروگرام ریڈیوں سیٹوں تک محدود تھے ۔ Initially, radio programs were limited to radio sets.	SentLen → 11; TreeDepth → 2; SubjNum →singular(sg); ObjNum → plural(pl); BShift → شروع شروع میں ریڈیو پروگرام سیٹوں ریڈیوں تک محدود تھے ۔ (Initially, radio programs were limited to **sets radio**) ; VerbGen → male (m); VerbNum → singular (sg); VerbPer → any (any)
Telugu	నీరు పడటంతో ఏర్పడే పాలనురుగు మరియు నీటి బిందువుల స్వాభావిక ఫౌంటెన్లు పర్యాటకులను ముద్దముద్దగా తడిపేస్తాయి . The inherent fountains of milky foam and water droplets created by falling water drenched tourists.	SentLen → 13; TreeDepth → 7; SubjNum → plural(pl); ObjNum → plural(pl); BShift → నీరు పడటంతో ఏర్పడే పాలనురుగు మరియు నీటి బిందువుల ***ఫౌంటెన్లు స్వాభావిక*** పర్యాటకులను ముద్దముద్దగా తడిపేస్తాయి . (The **fountains inherent** of milky foam and water droplets created by falling water drenched tourists) ; VerbGen → neuter (n); VerbNum → plural(pl); VerbPer → 3rd person (3)
Hindi	जोर से आवाज लगाने पर भी यह नजारा देखा जा सकता है । This scene can be seen even after making loud noise.	SentLen → 13; TreeDepth → 3; SubjNum → plural (pl); ObjNum → singular (sg); BShift → जोर से आवाज लगाने पर भी यह नजारा ***जा देखा*** सकता है । (This scene can **seen be** even after making loud noise) ; VerbGen → male (m); VerbNum → singular (sg); VerbPer → any (any)

Table 12: Examples for probing tasks w.r.t each language.

D Probing Tasks

Surface level tasks (1) *Sentence Length (SentLen):* Here, the objective is to predict the number of words in sentences, which has been grouped into 8 categories as shown in Table 1 (see main paper).

Syntactic tasks (2) *Tree Depth (TreeDepth):* The goal of this task is to predict the maximum depth of the sentence's syntactic tree, which is categorized into five options based on depth intervals as shown in Table 1. As constituency data in Indic languages is unavailable, we utilize dependency tree data to determine the tree depth. This task provides valuable insights into the structural complexity and organization of sentences. (3) *Bigram Shift (BShift):* This task involves binary classification aimed at predicting whether two consecutive tokens within a sentence are inverted or not.

Semantic tasks (4) *Subject Number (SubjNum):* This task evaluates sentences to determine the number of the subject in the main clause. It categorizes the subjects as NN (singular) or NNS (plural or mass, such as "colors," "waves," etc.). (5) *Object Number (ObjNum):* This task involves identifying whether the object of the main clause in a sentence is singular or plural/mass. It uses the label NN for singular objects and NNS for plural or mass objects. (6) *Verb Gender (VerbGen):* This task involves categorizing the grammatical gender of the main verb in a sentence as masculine, feminine, neutral, or any. (7) *Verb Number (VerbNum):* This task assesses the number of the main verb in a sentence, determining whether it is singular, plural, or any. (8) *Verb Person (VerbPer):* This task involves categorizing the grammatical person of the sentence's main verb into one of 7 classes as shown in Table 1. Honorifics, especially common in Indic languages, are forms of address that convey respect and politeness based on social status, age, or relationship. A typical example from Hindi is the use of "ji" to respectfully address elders, as seen in "dada-ji".

E Text Perturbation Examples

Tables 13-18 display an illustrative example of text perturbations for each perturbation type per language. For each language, the tables present the original sentence followed by a perturbed sentence resulting from each of the thirteen perturbations. Additionally, English translations are provided for enhanced comprehension.

F Details of Multilingual Models

F.1 Training data proportion of the tested languages

Table 21 shows the proportion of training data for each Indic language within each model's total Indic language tokens. From Table 21, we make the following observations:

Balanced vs. Imbalanced Distribution:

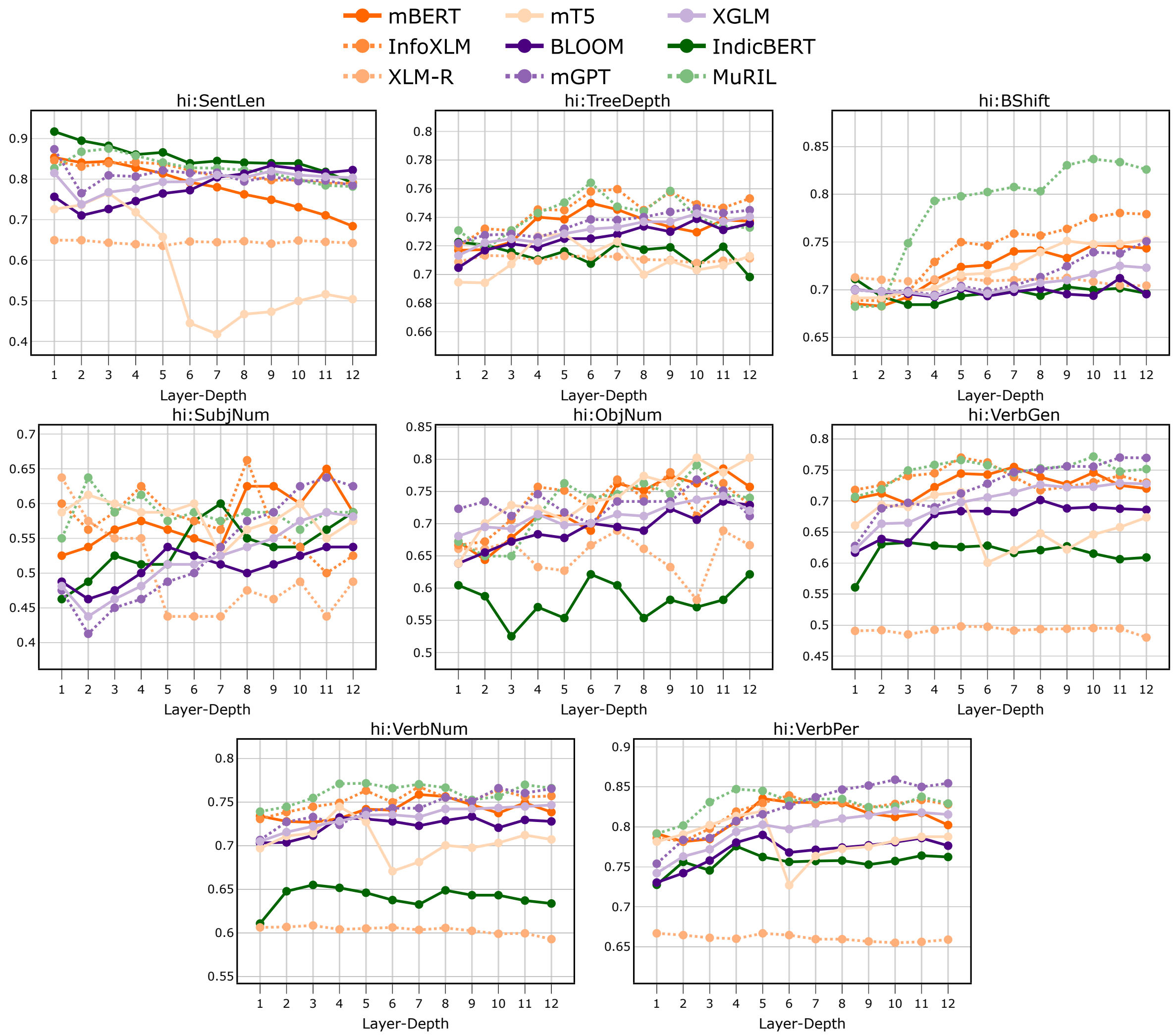

Figure 5: Hindi language probing task results: Layerwise accuracy comparisons between various multilingual representations on 8 probing tasks.

Model	Pretraining objectives	MSL
mBERT-base (Kenton and Toutanova, 2019)	MLM, NSP	512
XLM-R-base (Conneau et al., 2020)	MLM, CLM, TLM	512
InfoXLM-base (Chi et al., 2021)	MLM, TLM, CLM, XLCO	512
BLOOM-base (Scao et al., 2022)	CLM	512
mT5-base (Xue et al., 2021)	MLM, CLM	512
mGPT (Shliazhko et al., 2024)	CLM	512
XGLM (Lin et al., 2022)	CLM	512
IndicBERT-base (Kakwani et al., 2020)	MLM	128
MuRIL-base (Khanuja et al., 2021)	MLM, TLM	512

Table 19: Details of multilingual Transformer-based models used in this study: MLM (masked language modeling), CLM (causal LM), TLM (translation LM), XLCO (cross-lingual contrastive learning), MSL (maximum sequence length).

- MuRIL model stands out with the most balanced distribution among the six languages shown (12.7% for Hindi, with other languages between 6-9%).
- In contrast, mGPT shows high Hindi focus (50.5%) with much smaller proportions for other languages.

	hi	kn	ml	mr	te	ur	Total Tokens (for all Indic languages)
mBERT (11)	Detailed information not known publicly						184M
IndicBERT (23)	1.84B	712M	767M	560M	671M	-	7.59B
XLM-R (15)	1.71B	169M	313M	175M	249M	730M	3.99B
InfoXLM (13)	1.17B	95.6M	327M	130M	225M	289M	3.467B
MuRIL(16)	4.8B	2.4B	2.7B	2.6B	2.66B	3.3B	37.76B
BLOOM (13)	Detailed information not known publicly						2.7B
mT5 (11)	24B	1.1B	1.8B	14B	1.3B	2.4B	58.3B
mGPT	1.1B	0.11B	0.11B	0.11B	0.11B	0.2B	2.18B
XGLM	3.45B	446M	458M	935M	689M	1.35B	11.0B

Table 20: Tokens for pretraining multilingual models.

Hindi Dominance:

- Most models allocate the largest share of their Indic language tokens to Hindi.
- This varies dramatically from 12.7% in MuRIL to 50.5% in mGPT.

Coverage Beyond Major Languages:

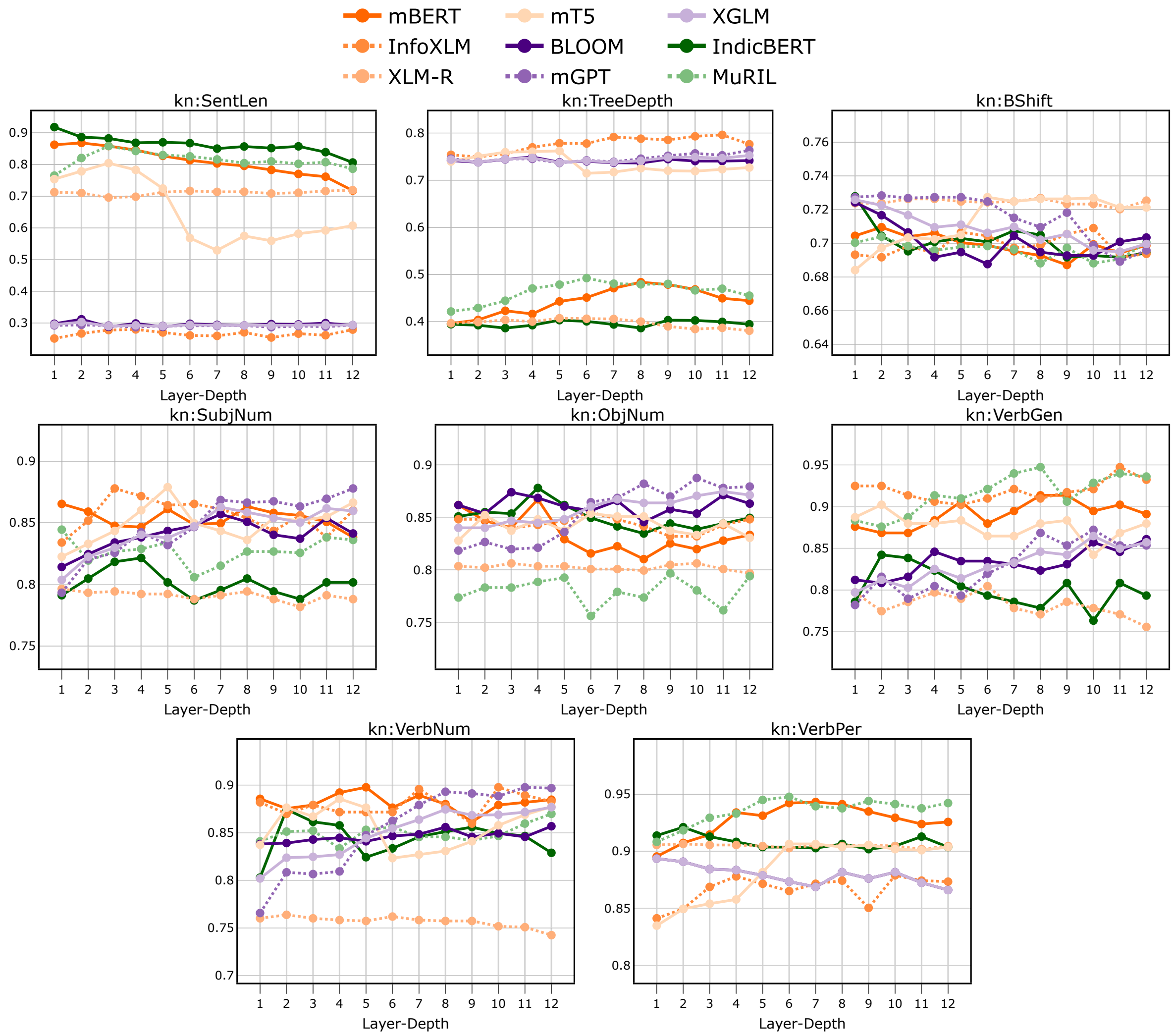

Figure 6: Kannada language probing task results: Layerwise accuracy comparisons between various multilingual representations on 8 probing tasks.

- MuRIL allocates 51.1% of its tokens to "Other Indic" languages beyond the six shown, suggesting much broader coverage.
- This aligns with research showing MuRIL supports 17 Indian languages in total.

Model	hi (Hindi)	kn (Kannada)	ml (Malayalam)	mr (Marathi)	te (Telugu)	ur (Urdu)	Other Indic
IndicBERT	24.2%	9.4%	10.1%	7.4%	8.8%	–	40.1%*
XLM-R	42.9%	4.2%	7.8%	4.4%	6.2%	18.3%	16.2%*
InfoXLM	33.7%	2.8%	9.4%	3.7%	6.5%	8.3%	35.6%*
MuRIL	12.7%	6.4%	7.2%	6.9%	7.0%	8.7%	51.1%*
mT5	41.2%	1.9%	3.1%	24.0%	2.2%	4.1%	23.5%*
mGPT	50.5%	5.0%	5.0%	5.0%	5.0%	9.2%	20.3%*
XGLM	31.4%	4.1%	4.2%	8.5%	6.3%	12.3%	33.2%*

Table 21: Training data proportion of the six Indic languages for both universal and Indic multilingual language models.

Hyper-parameters. We train logistic regression with a regularization parameter C=20 and use a L2 penalty term. For multi-class tasks, we use the "multinomial" setting while training. All experiments were done on a machine with a T4 GPU.

G Probing Results

We report the layerwise probing accuracies for individual multilingual models in Figs. 5 to 10.

Surface-level tasks Analyzing the performance across languages, we observe the following patterns: (i) For universal multilingual models (mBERT, mT5 and InfoXLM) as well as for Indic models (IndicBERT and MuRIL), there is a trend of higher accuracy in the early (or lower) layers, which decreases in the later (or higher) layers. This pattern is expected, as surface-level tasks generally require minimal processing. (ii) Notably, the universal model BLOOM deviates from this trend, showing lower accuracy in the early layers and higher accuracy in the later layers. This unusual pattern in BLOOM could be attributed to its unique autoregressive architectural design, and its approach to representing languages, especially

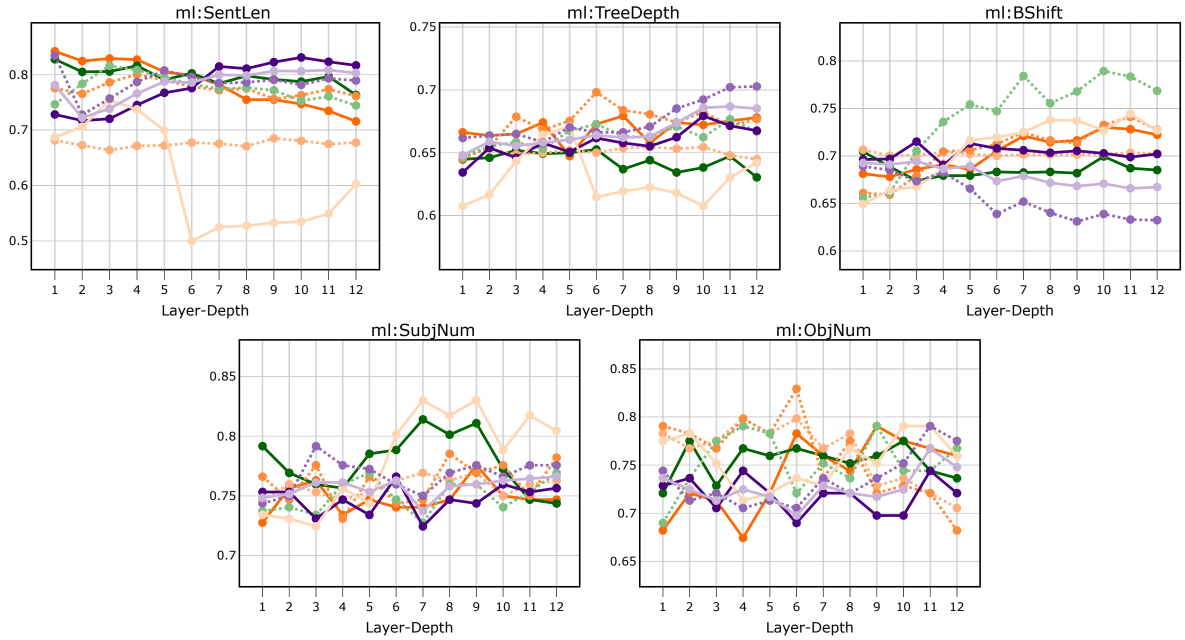

Figure 7: Malayalam language probing task results: Layerwise accuracy comparisons between various multilingual representations on 5 probing tasks.

Text-Perturbation	Perturbed Sentence	Perturbed Sentence (Translation in English)
Original	ప్రపంచంలో ఎక్కడైనా ఇంతే అనుకోనేరు.	You can't think like this anywhere in the world.
AppendR	చెత్త ప్రపంచంలో ఎక్కడైనా ఇంతే అనుకోనేరు.	Garbage You can't think like this anywhere in the world.
DropNV	ఎక్కడైనా ఇంతే.	You can’t like this anywhere in the.
DropN	ఎక్కడైనా ఇంతే అనుకోనేరు.	You can't think like this anywhere in the.
DropV	ప్రపంచంలో ఎక్కడైనా ఇంతే.	You can't like this anywhere in the world.
DropF	ఎక్కడైనా ఇంతే అనుకోనేరు.	can't think like this anywhere in the world.
DropFL	ఎక్కడైనా ఇంతే.	can't think like this anywhere in the.
DropL	ప్రపంచంలో ఎక్కడైనా ఇంతే.	You can't think like this anywhere in the.
DropRN	ఎక్కడైనా ఇంతే అనుకోనేరు.	You can't think like this anywhere in the.
DropRV	ప్రపంచంలో ఎక్కడైనా ఇంతే.	You can't like this anywhere in the world.
KeepNV	ప్రపంచంలో అనుకోనేరు	think world
KeepN	ప్రపంచంలో	world
KeepV	అనుకోనేరు	think
Shuffle	ప్రపంచంలో ఇంతే అనుకోనేరు ఎక్కడైనా.	think can’t anywhere in world the you like this.

Table 13: Text perturbation examples for Telugu language.

those less common in the dataset. This suggests that BLOOM may require its deeper layers to effectively grasp the nuances of these languages, even for tasks at the surface level. (iii) Overall, among all the models, IndicBERT reports best accuracy, while XLM-R displays poorer performance. This is likely because IndicBERT is specifically trained on Indic languages, making it more attuned to the nuances and idiosyncrasies of these languages. On the other hand, models like XLM-R, which are designed for universal applicability, might struggle with specific linguistic features unique to Indic languages. These features include script differences, morphological complexity, and visual factors such as orthography and word length.

Syntactic Tasks For TreeDepth, we observe that probing accuracy tends to be higher in the middle layers for various (model, language) combinations. This trend is particularly notable in three universal models, mBERT, mT5 and InfoXLM, and an Indic model, MuRIL. Moreover, this pattern is consistent for three languages: `hi`, `kn` and `ml`. However, other multilingual models do not exhibit any clear layer-wise trend, and the same applies to the other three Indic languages. This suggests a model-specific and language-specific affinity in handling syntactic complexity, where certain models are more adept at processing syntactic information in specific layers, and this proficiency varies across different languages. In contrast to TreeDepth, for BShift task, we generally observe higher probing accuracy in the later layers for four of the models across various

Text-Perturbation	Perturbed Sentence	Perturbed Sentence (Translation in English)
Original	200 ശിൽപികൾ രണ്ടു് വർഷം കൊണ്ട് ആണ് പ്രതിമയുടെ നിർമാണം പൂർത്തിയാക്കിയത്.	200 sculptors completed the construction of the statue in two years.
AppendR	200 ശിൽപികൾ രണ്ടു് വർഷം കൊണ്ട് ആണ് പ്രതിമയുടെ കൂത്തമ്പലങ്ങളിൽ നിർമാണം പൂർത്തിയാക്കിയത്.	200 sculptors completed the construction of the statue in the hills in two years.
DropNV	200 രണ്ടു് വർഷം കൊണ്ട് ആണ് നിർമാണം.	200 the of the in two.
DropN	200 രണ്ടു് വർഷം കൊണ്ട് ആണ് നിർമാണം പൂർത്തിയാക്കിയത്.	200 completed the of the in two.
DropV	200 ശിൽപികൾ രണ്ടു് വർഷം കൊണ്ട് ആണ് പ്രതിമയുടെ നിർമാണം.	200 sculptors the construction of the statue in two years.
DropF	ശിൽപികൾ രണ്ടു് വർഷം കൊണ്ട് ആണ് പ്രതിമയുടെ നിർമാണം പൂർത്തിയാക്കിയത്.	sculptors completed the construction of the statue in two years.
DropFL	ശിൽപികൾ രണ്ടു് വർഷം കൊണ്ട് ആണ് പ്രതിമയുടെ നിർമാണം.	sculptors completed the construction of the statue in two.
DropL	200 ശിൽപികൾ രണ്ടു് വർഷം കൊണ്ട് ആണ് പ്രതിമയുടെ നിർമാണം.	200 sculptors completed the construction of the statue in two.
DropRN	200 ശിൽപികൾ രണ്ടു് വർഷം കൊണ്ട് ആണ് നിർമാണം പൂർത്തിയാക്കിയത്.	200 sculptors completed the construction of the in two years.
DropRV	200 ശിൽപികൾ രണ്ടു് വർഷം കൊണ്ട് ആണ് പ്രതിമയുടെ നിർമാണം.	200 sculptors the construction of the statue in two years.
KeepNV	ശിൽപികൾ പ്രതിമയുടെ പൂർത്തിയാക്കിയത്	sculptors completed construction statue years
KeepN	ശിൽപികൾ പ്രതിമയുടെ	sculptors construction statue years
KeepV	പൂർത്തിയാക്കിയത്	completed
Shuffle	രണ്ടു് ആണ് 200 കൊണ്ട് ശിൽപികൾ പൂർത്തിയാക്കിയത് പ്രതിമയുടെ നിർമാണം വർഷം.	Two the 200 in sculptors completed of the statue construction years.

Table 14: Text perturbation examples for Malayalam language.

Text-Perturbation	Perturbed Sentence	Perturbed Sentence (Translation in English)
Original	डलहौजी उस धौलाधार पर्वतमाला के सामने पड़ता है जो साल भर बर्फ़ की नई नई परतें ओढ़ती है ।	Dalhousie lies in front of the Dhauladhar range which is covered with new layers of snow throughout the year.
AppendR	डलहौजी उस धौलाधार पर्वतमाला के सामने पड़ता है जो लपकने साल भर बर्फ़ की नई नई परतें ओढ़ती है ।	Dalhousie lies in front of the Dhauladhar range which is covered with new layers of snow throughout the year catch.
DropNV	उस के जो भर की नई नई ।	in of the which with new of throughout the.
DropN	उस के पड़ता है जो भर की नई नई ओढ़ती है ।	lies in of the which is covered with new of throughout the.
DropV	डलहौजी उस धौलाधार पर्वतमाला के सामने जो साल भर बर्फ़ की नई नई परतें ।	Dalhousie in front of the Dhauladhar range which with new layers of snow throughout the year.
DropF	उस धौलाधार पर्वतमाला के सामने पड़ता है जो साल भर बर्फ़ की नई नई परतें ओढ़ती है ।	lies in front of the Dhauladhar range which is covered with new layers of snow throughout the year.
DropFL	उस धौलाधार पर्वतमाला के सामने पड़ता है जो साल भर बर्फ़ की नई नई परतें ओढ़ती ।	lies in front of the Dhauladhar range which is covered with new layers of snow throughout the.
DropL	डलहौजी उस धौलाधार पर्वतमाला के सामने पड़ता है जो साल भर बर्फ़ की नई नई परतें ओढ़ती ।	Dalhousie lies in front of the Dhauladhar range which is covered with new layers of snow throughout the.
DropRN	डलहौजी उस धौलाधार पर्वतमाला के पड़ता है जो साल भर बर्फ़ की नई नई परतें ओढ़ती है ।	Dalhousie lies in front of the range which is covered with new layers of snow throughout the year.
DropRV	डलहौजी उस धौलाधार पर्वतमाला के सामने पड़ता जो साल भर बर्फ़ की नई नई परतें ओढ़ती है ।	Dalhousie lies in front of the Dhauladhar range which is with new layers of snow throughout the year.
KeepNV	डलहौजी धौलाधार पर्वतमाला सामने पड़ता है साल बर्फ़ परतें ओढ़ती है	Dalhousie lies front Dhauladhar range is covered layers snow year.
KeepN	डलहौजी धौलाधार पर्वतमाला सामने साल बर्फ़ परतें	Dalhousie front Dhauladhar range layers snow year.
KeepV	पड़ता है ओढ़ती है	lies is covered.
Shuffle	। के उस की है नई पड़ता जो भर डलहौजी नई बर्फ़ साल ओढ़ती परतें सामने पर्वतमाला धौलाधार है	.with which new lies Dalhousie snow year covered layers in front of the range Dhauladhar is.

Table 15: Text perturbation examples for Hindi language.

languages. Notably, BLOOM exhibits a decreasing trend in accuracy for kn and te, while showing an increasing trend for ur. This suggests different models' layers may specialize in different types of syntactic processing, with some models better handling tasks like BShift in their later layers. Overall, when comparing performance across models and tasks, InfoXLM and mT5 stand out for their superior accuracy in capturing tree depth information across different languages. Conversely, MuRIL shows notable proficiency in BShift. These distinctions highlight how different models may be better suited for different types of syntactic analyses. These observations contribute to a deeper understanding of how various multilingual models process syntactic information, demonstrating both model-specific and language-specific trends and capabilities in linguistic tasks.

Semantic Tasks These include SubjNum and ObjNum, and VerbGen, VerbNum and VerbPer. We do not have results for ml for VerbGen, VerbNum and VerbPer, since we do not have labeled data for ml for those tasks.

From SubjNum and ObjNum results, we make

Text-Perturbation	Perturbed Sentence	Perturbed Sentence (Translation in English)
Original	त्यामुळे कर्मचारी युनियनने न्यायालयात धाव घेतली होती.	Therefore, the employee union had approached the court.
AppendR	सत्कारास त्यामुळे कर्मचारी युनियनने न्यायालयात धाव घेतली होती.	Congratulations Therefore, the employee union had approached the court.
DropNV	त्यामुळे.	Therefore, the the.
DropN	त्यामुळे घेतली होती.	Therefore, the had approached the.
DropV	त्यामुळे कर्मचारी युनियनने न्यायालयात धाव.	had approached.
DropF	कर्मचारी युनियनने न्यायालयात धाव घेतली होती.	, the employee union had approached the court.
DropFL	कर्मचारी युनियनने न्यायालयात धाव घेतली.	, the employee union had approached the.
DropL	त्यामुळे कर्मचारी युनियनने न्यायालयात धाव घेतली.	Therefore, the employee union had approached the.
DropRN	त्यामुळे कर्मचारी युनियनने धाव घेतली होती.	Therefore, the employee union had approached the.
DropRV	त्यामुळे कर्मचारी युनियनने न्यायालयात धाव घेतली.	Therefore, the employee union approached the court.
KeepNV	कर्मचारी युनियनने न्यायालयात धाव घेतली होती	, employee union had approached court.
KeepN	कर्मचारी युनियनने न्यायालयात धाव	, employee union court.
KeepV	घेतली होती	had approached.
Shuffle	कर्मचारी धाव त्यामुळे न्यायालयात. होती युनियनने घेतली	the employee approached therefore the court. Had union

Table 16: Text perturbation examples for Marathi language.

Text-Perturbation	Perturbed Sentence	Perturbed Sentence (Translation in English)
Original	ಶೈಕ್ಷಣಿಕ ಪ್ರವಾಸ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಅರಿಯಲು ಸುಲಭ ಸಾಧನ.	Educational travel is an easy tool for students to understand the importance of travel right from the start.
AppendR	ಶೈಕ್ಷಣಿಕ ಪ್ರವಾಸ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಅರಮನೆಯಲ್ಲೂ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಅರಿಯಲು ಸುಲಭ ಸಾಧನ.	Educational travel is an easy palace tool for students to understand the importance of travel right from the start.
DropNV	ಶೈಕ್ಷಣಿಕ ಸುಲಭ.	is an easy for to the of right from the.
DropN	ಶೈಕ್ಷಣಿಕ ಅರಿಯಲು ಸುಲಭ.	is an easy for to understand the of right from the.
DropV	ಶೈಕ್ಷಣಿಕ ಪ್ರವಾಸ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಸುಲಭ ಸಾಧನ.	Educational travel is an easy tool for students to the importance of travel right from the start.
DropF	ಪ್ರವಾಸ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಅರಿಯಲು ಸುಲಭ ಸಾಧನ.	travel is an easy tool for students to understand the importance of travel right from the start.
DropFL	ಪ್ರವಾಸ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಅರಿಯಲು ಸುಲಭ ಸಾಧನ.	travel is an easy tool for students to understand the importance of travel right from the.
DropL	ಶೈಕ್ಷಣಿಕ ಪ್ರವಾಸ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಅರಿಯಲು ಸುಲಭ.	Educational travel is an easy tool for students to understand the importance of travel right from the.
DropRN	ಶೈಕ್ಷಣಿಕ ಪ್ರವಾಸ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಅರಿಯಲು ಸುಲಭ ಸಾಧನ.	Educational travel is an easy tool for students to understand the of travel right from the start.
DropRV	ಶೈಕ್ಷಣಿಕ ಪ್ರವಾಸ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಸುಲಭ ಸಾಧನ.	Educational travel is an easy tool for students to the importance of travel right from the start.
KeepNV	ಪ್ರವಾಸ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಅರಿಯಲು ಸಾಧನ	Educational travel tool students understand importance travel start.
KeepN	ಪ್ರವಾಸ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಆರಂಭದಿಂದಲೇ ಪ್ರವಾಸದ ಮಹತ್ವವನ್ನು ಸಾಧನ	Educational travel tool students importance travel start.
KeepV	ಅರಿಯಲು	understand.
Shuffle	ಸಾಧನ ಅರಿಯಲು ಆರಂಭದಿಂದಲೇ. ಶೈಕ್ಷಣಿಕ ಪ್ರವಾಸದ ಪ್ರವಾಸ ಸುಲಭ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಮಹತ್ವವನ್ನು	tool to understand right from the start educational of travel travel easy for students the importance is an.

Table 17: Text perturbation examples for Kannada language.

Text-Perturbation	Perturbed Sentence	Perturbed Sentence (Translation in English)
Original	ابتدا میں رسائل کی زبان سست تھی اور اس میں مقامیت بھی رہتی تھی -	In the beginning, the language of the letters was slow and there was a lack of space in it.
AppendR	ابتدا میں رسائل کی زبان سست تھی اور اس انٹرنیٹ میں مقامیت بھی رہتی تھی -	In the beginning, the language of the letters was slow and there was a lack of space in it internet.
DropNV	میں کی سست اور اس میں بھی -	In the, the of the slow and there a of in it.
DropN	میں کی سست تھی اور اس میں بھی رہتی تھی -	In the, the of the was slow and there was a of in it.
DropV	ابتدا میں رسائل کی زبان سست اور اس میں مقامیت بھی -	In the beginning, the language of the letters slow and there a lack of space in it.
DropF	میں رسائل کی زبان سست تھی اور اس میں مقامیت بھی رہتی تھی -	the beginning, the language of the letters was slow and there was a lack of space in it.
DropFL	میں رسائل کی زبان سست تھی اور اس میں مقامیت بھی رہتی -	the beginning, the language of the letters was slow and there was a lack of space in.
DropL	ابتدا میں رسائل کی زبان سست تھی اور اس میں مقامیت بھی رہتی -	In the beginning, the language of the letters was slow and there was a lack of space in.
DropRN	ابتدا میں کی زبان سست تھی اور اس میں مقامیت بھی رہتی تھی -	In the beginning, the language of the was slow and there was a lack of space in it.
DropRV	ابتدا میں رسائل کی زبان سست تھی اور اس میں مقامیت بھی تھی -	In the beginning, the language of the letters was slow and there a lack of space in it.
KeepNV	ابتدا رسائل زبان تھی مقامیت رہتی تھی	beginning, language letters was lack space.
KeepN	ابتدا رسائل زبان مقامیت	beginning, language letters lack space.
KeepV	تھی رہتی تھی	was was.
Shuffle	سست تھی میں رسائل بھی اس زبان رہتی میں تھی اور ابتدا کی مقامیت -	space in the beginning, and was the language of the letters was slow a lack of in it.

Table 18: Text perturbation examples for Urdu language.

the following observations. For mBERT, InfoXLM, mT5, BLOOM as well as MuRIL, we observe an increasing trend from lower to higher layers for hi, mr and te. This suggests that for these languages, the models become more proficient in handling semantic tasks related to SubjNum and

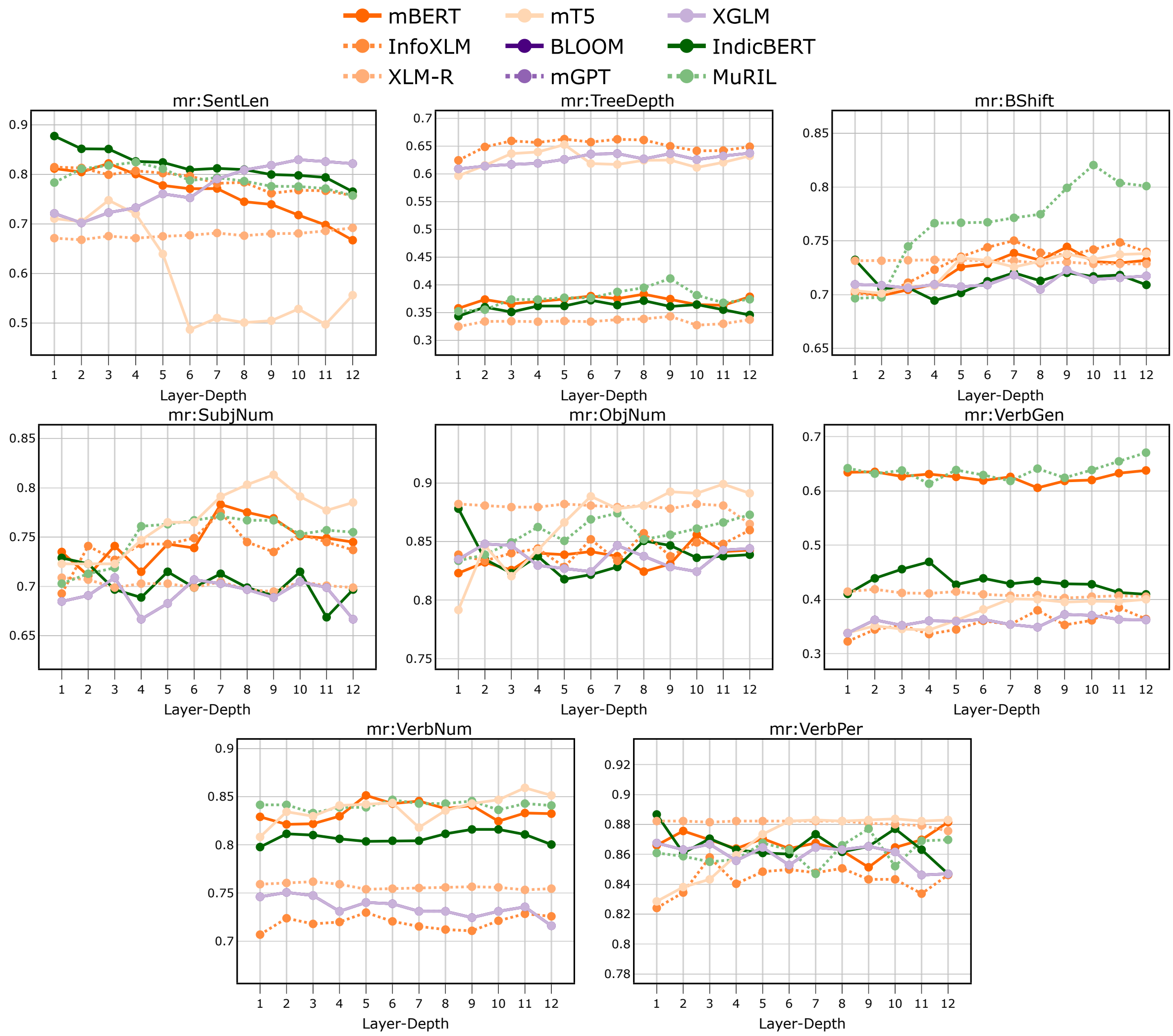

Figure 8: Marathi language probing task results: Layerwise accuracy comparisons between various multilingual representations on 8 probing tasks.

ObjNum as we move to the higher layers. When considering other languages and models, we note that IndicBERT exhibits an increasing trend for the ObjNum task, specifically for `ur`. This highlights the variability in model performance based on the language. Interestingly, the middle layers of the models show higher probing accuracy for languages such as `hi` and `ml` in the SubjNum task. Similar to its performance in surface and syntactic tasks, XLM-R exhibits lower accuracy and lacks a discernible trend when it comes to capturing semantics. MuRIL has the highest probing accuracy although it performs the worst for `kn` ObjNum task.

For the verb-related tasks, Indic model MuRIL, performs the best for most languages and tasks. Also, in most cases, the last layer is the most predictive, except in `te` for gender and person detection, where initial layers provide better results. This indicates that gender and person detection in `te` is straightforward and does not need deep processing. For the universal models, mBERT and BLOOM report an increasing trend across languages and tasks. On the other hand, mT5 showcases an increasing trend for `kn` and `mr` languages and decreasing trend for `hi`, `te` and `ur` languages. Models, including, IndicBERT, XLM-R and InfoXLM do not show any trend and have constant performance across layers.

H Extended Discussion

We evaluated 9 multilingual Transformer-based models on 8 probing tasks to understand linguistic structures in 6 Indic languages, using our contributed INDICSENTEVAL dataset. Indic-specific models like MuRIL and IndicBERT excel due to targeted training, while universal models like

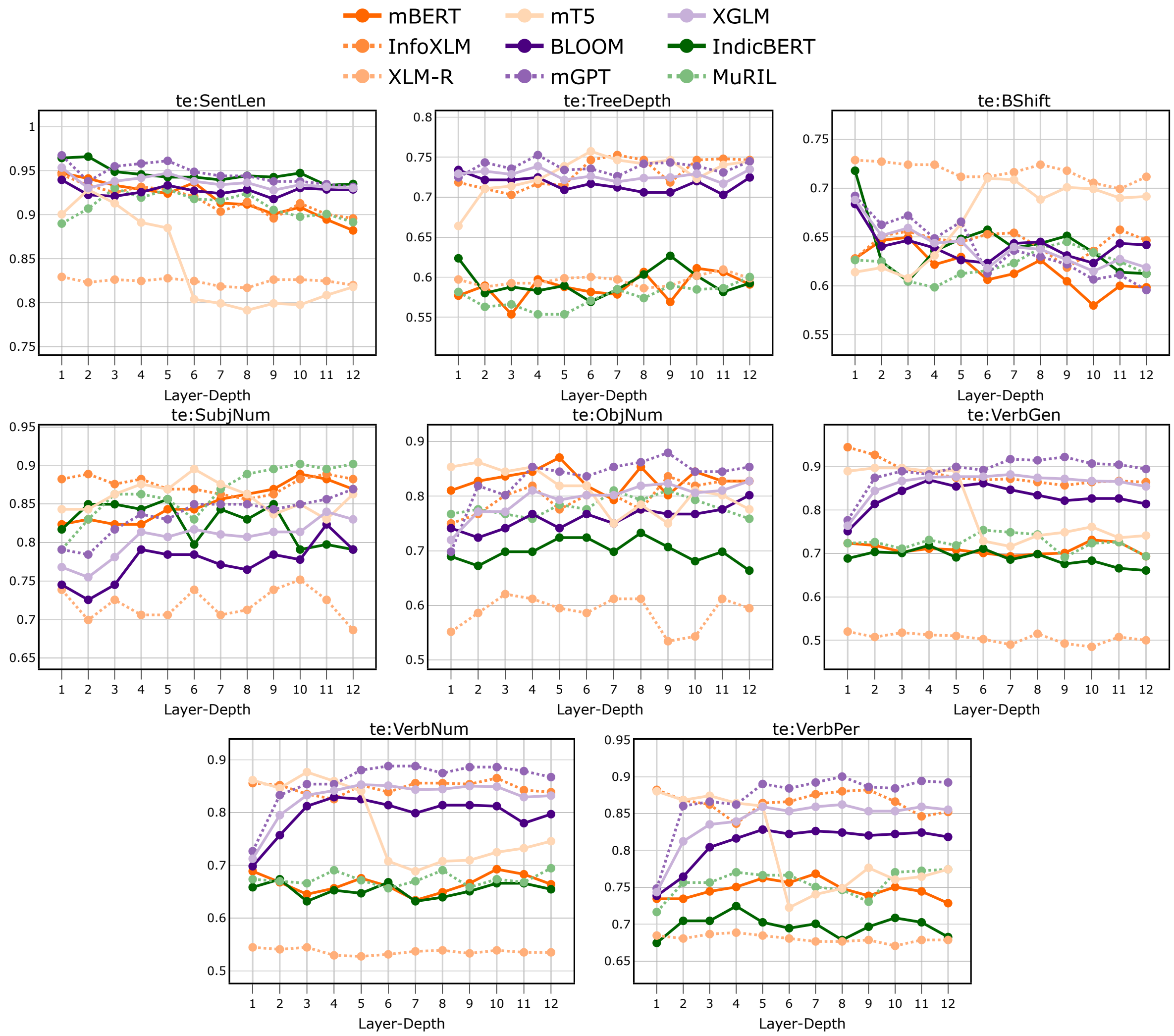

Figure 9: Telugu language probing task results: Layerwise accuracy comparisons between various multilingual representations on 8 probing tasks.

mBERT, InfoXLM, BLOOM, mGPT, and XGLM show mixed results. Perturbation analysis reveals decoder-based models are the most robust, as also noted by Neerudu et al. (2023). Verbs and word order are key signals for encoding linguistic structures. TreeDepth is the most sensitive to perturbations, while SubjNum and ObjNum are the most resilient.

Overall, our study represents the first analysis of the interpretability of both Universal and Indic multilingual language models across six Indic languages where several languages have large training corpora while some have less. Our scientific findings from this model interpretability analysis via both probing and perturbations shed light on how language models capture language hierarchy and how training data influences the language understanding across layers in these models. Surprisingly, universal models show greater resilience to perturbations in at least four Indic languages. In contrast, the universal model (mBERT) and the Indic-specific models (IndicBERT and MuRIL) display a more significant accuracy drop across all the Indic languages. This suggests the necessity for multilingual models that are robust to perturbations and can capture language hierarchy regardless of their training data. Overall, our findings demonstrate significant variability in the ability of current multilingual models to capture surface, syntactic, and semantic structures across different Indic languages. A hierarchy of language proficiency is discernible primarily for languages with more extensive training datasets. This highlights the necessity for innovative approaches in multilingual modeling that can accurately capture language structures, even in low-resource languages.

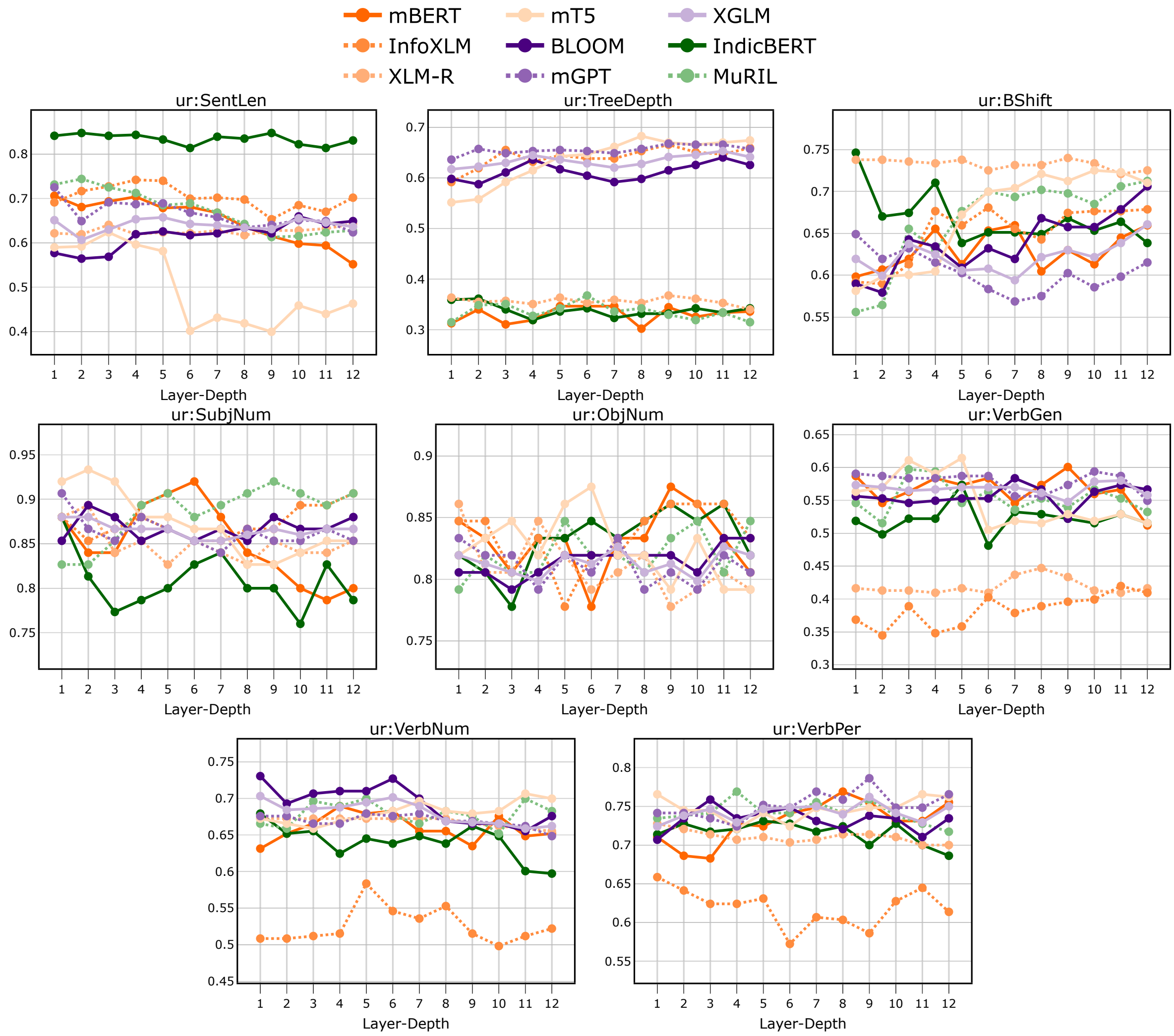

Figure 10: Urdu language probing task results: Layerwise accuracy comparisons between various multilingual representations on 8 probing tasks.

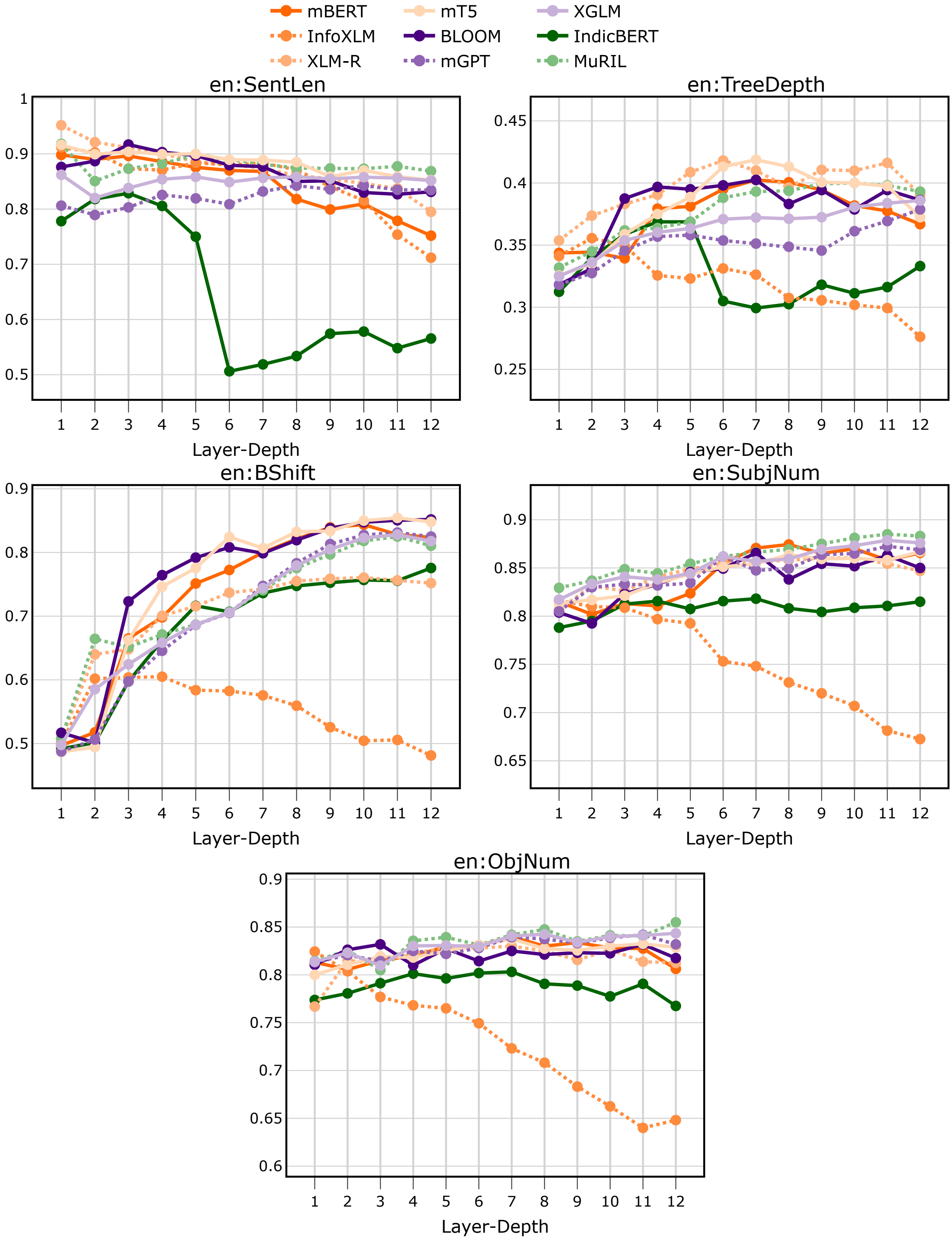

Figure 11: English language probing results: Layerwise accuracy was computed across Universal multilingual (mBERT, XLM-R, InfoXLM, BLOOM, mT5, mGPT and XGLM) and Indic multilingual (IndicBERT, MuRIL) representations on surface-level, syntactic probing and semantic probing tasks.