A Benchmark for Hindi Verb-Argument Structure Alternations

Kanishka Jain and Ashwini Vaidya
Indian Institute of Technology Delhi
{kanishka, avaidya} @hss.iitd.ac.in

Abstract

In this paper we introduce a Hindi verb alternations benchmark to investigate whether pretrained large language models (LLMs) can infer the frame-selectional properties of Hindi verbs. Our benchmark consists of minimal pairs such as *Tina cut the wood/∗Tina disappeared the wood*. We create four variants of these alternations for Hindi to test knowledge of verbal morphology and argument case-marking. Our results show that a masked monolingual model performs the best, while causal models fare poorly. We further test the quality of the predictions using a cloze-style sentence completion task. While the models appear to infer the right mapping between verbal morphology and valency in the acceptability task, they do not generate the right verbal morphology in the cloze task. The model completions also lack pragmatic and world knowledge, crucial for making generalizations about verbal alternations. Our work points towards the need for more cross-linguistic research of verbal alternations.

1 Introduction

A question that has been investigated repeatedly is whether large language models (LLMs) are able to learn the syntactic and semantic generalizations of a natural language given the diverse data they are trained on. A number of studies have created linguistic benchmarks consisting of syntactic phenomena (e.g. active-passives, syntactic agreement) using minimal pairs. LLMs are then tested on acceptability judgement tasks, comparing their performance with human judgements (Warstadt et al., 2020; Xiang et al., 2021; Someya and Oseki, 2023; Song et al., 2022).

Recent work evaluated transformer LLMs on Hindi syntactic agreement (Kryvosheieva and Levy, 2025). LLMs' performance was robust despite Hindi's complex split-ergative system. With respect to verb argument structure alternations, cross-linguistic results are mixed. For English as well as Chinese, experiments show that model performance is relatively poor for argument structure (Warstadt et al., 2020; Xiang et al., 2021). For Japanese on the other hand, models seem to match human accuracy (Someya et al., 2024). There is no previous work evaluating LLMs' knowledge of verb argument structure for Hindi.

The core meaning of an event is contributed by the verb in a sentence or context. It comes densely packed with information about the number of arguments (or participants), their role, and how they are related to each other. This information comprises syntactic knowledge: mapping the verbal morphology to the correct number of arguments in the sentence. It also contains semantic knowledge where the verb and its arguments contribute to the event meaning.

In this paper, we use both acceptability judgements and cloze-style sentence completions following Ettinger (2020). We evaluate both masked and causal models, and also compare multilingual and monolingual models (Martin et al., 2020; Song et al., 2022). Results from our acceptability task indicate knowledge of the mapping between verbs and syntactic frames. At the same time, the best performing models from this task are not able to predict the correct verb forms in a cloze-style sentence completion. We show that verb alternations require LLMs to make generalizations that are different from other syntactic phenomena.

2 Alternations in Hindi

Hindi verbs carry morphosyntactic information that signals the change in arguments. In the following examples, the base form of an intransitive verb /ubəl/ 'boil' changes to transitive in /ub**a**l/ and then to the indirect causative in /ubəl**va**/. While there is variation in the way each of these alternations are realized (e.g. some verbs have a null transitive alternation), there is a surface *form-function map-*

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17542–17549

ping unlike English. For example, *John broke the window* and *The window broke* are causative and intransitive, respectively but without any surface differences.

(1) pani **ubəl** rəha t^{h}a
water.M boil PROG.SG.M AUX.PST.SG.M
'The water was boiling.'

(2) ləɽka pani **ubal** rəha
boy.3.SG.M water.M boil.DCAUS PROG.SG.M
t^{h}a
AUX.PST.SG.M
'The boy was boiling the water.'

(3) ləɽka bəcce-se pani
boy.3.SG.M child.3.SG.M-AGT water.M
ubəl-va rəha t^{h}a
boil-ICAUS PROG.SG.M AUX.PST.SG.M
'The boy made/had the child boil the water.'

Begum et al. (2008) groups Hindi verbs together on the basis of this morphological relatedness. In this paper, we aim to investigate whether LLMs learn such a mapping between the morphological form and its corresponding argument frame.

One challenge in developing such an evaluation dataset for Hindi is that arguments are regularly dropped (elided), and case markers on the nouns exhibit *case syncretism*. For example in (5) the case /-se/ describes a source (Mira) and takes a transitive form. In example (4), the same case marker /-se/ is instrumental, occurring with a causative form of the verb /bədəl/ 'change'.

(4) amɪt-ne mira-se
amit.3.SG.M-ERG mira.3.SG.F-INST
g^{h}əDi **bədəl-va-i**
watch.3.SG.F change-ICAUS-PST.PERF.SG.F
'Amit made/had Mira change the watch.'

(5) amɪt-ne mira-se
amit.3.SG.M-ERG mira.3.SG.F-SOURCE
g^{h}əDi **bədəl-i**
watch.3.SG.F change-PST.PERF.SG.F
'Amit exchanged the watch from Mira.'

For our benchmark, we choose sentences where all argument and adjunct slots are filled. In our minimal pairs, the acceptable sentence has the /-va/ causative as in (3), with three arguments (causer, agent, and patient). An additional instrumental argument is also added to restrict the choice to causatives and avoid ambiguity. We then replace the grammatically correct verb with an incorrect form to test for awareness of the correct frame.

3 Benchmark construction

To examine the extent to which pretrained models effectively leverage syntactic and semantic information from the context, we introduce a benchmark of minimal pairs in Hindi. We construct minimal pairs such that both sentences have a common sentential prefix and a grammatical or ungrammatical verb (which occurs in SOV order in Hindi). The last word in each sentence is a past tense auxiliary (the verb occurs at second last position). All examples are shown in Table 1.

Our benchmark consists of 56 verbs that have been selected on the basis of different criteria. We first chose verbs on the basis of their frequency using the Shabd database corpus (Verma et al., 2022). We have selected verbs that are high on the Zipf scale to maximize the chance of their occurrence across model training corpora. This ensures that these verbs are well represented and we minimize out-of-vocabulary effects. We then categorized verbs according to their valency. Since the goal of this work is to study how well pretrained models understand the verb argument structure of Hindi verbs, the final verb list maps to all three syntactic frames – intransitive (1 argument), transitive (2 arguments), and ditransitive (3 arguments). We also consider finer classifications, e.g. intransitive verbs which are further categorized into unergative and unaccusative verbs. Transitive verbs contain a sub-category of ingesto-reflexives. The final set has 28 intransitive verbs (13 unergatives and 15 unaccusatives), 23 transitive verbs (with 13 ingesto-reflexives), and 5 ditransitive verbs.

For our evaluation, we generate four variants of our benchmark that are described below:

Different Verb: the two verbs are morphologically unrelated forms, with different valency.

Same Verb: the two verbs are morphologically related, but with a different valency.

No Case(E): the two verbs are morphologically related, but the verbal aspect is habitual, which results in the ergative marker on the subject being removed[1].

No Case(I): the two verbs are morphologically related, but we remove the additional adjunct argument from both sentences.

We can think of the 'Different Verb' and 'Same Verb' variants of the dataset as being maximally specified in terms of the arguments and adjuncts, al-

[1]Hindi has split ergativity where /-ne/ marker on agents appear only when the verb is in past perfective.

Task	Exp	Sentence Prefix	Verb	Acceptability
Accept-ability	DV	mã-ne arjun-se kulhaDi-se ləkDi mother-ERG arjun-AGT axe-INST wood	kət-vai t^hi cut-DCAUS.PST be.PST	✓
			jəli t^hi burn.PST be.PST	✗
	SV	mã-ne arjun-se kulhaDi-se ləkDi mother-ERG arjun-AGT axe-INST wood	kət-vai t^hi cut-DCAUS.PST be.PST	✓
			kəTi t^hi cutPST be.PST	✗
	No Case(E)	mã arjun-se kulhaDi-se ləkDi mother arjun-AGT axe-INST wood	kəT-va-ti t^hi cut-DCAUS-HAB be.PST	✓
			kət-ti t^hi cut-HAB be.PST	✗
	No Case(I)	mã-ne arjun-se (...) ləkDi mother arjun-AGT (...) wood	kəT-va-i t^hi cut-DCAUS be.PST	✓
			kət-i t^hi cutPST be.PST	✗
Cloze		mã-ne arjun-se kulhaDi-se ləkDi mother-ERG arjun-INST axe-INST wood	_____ t^hi	NA

Table 1: Minimal pairs from our Hindi verb alternation benchmark. The example sentence is translated as *Mother made Arjun cut the wood with an axe*. DV=Different Verb, SV=Same Verb, No Case(E)= no ergative case on subject, and No Case(I)= no instrument case marked adjunct. The cloze task shows the sentential prefix, missing verb and the auxiliary. Argument /arjun-**se**/ is glossed as AGT 'AGENT' to distinguish it from the Instrumental case for *kulhaDi* 'axe'.

lowing us to test whether the mapping between morphological encoding and valency is learned. The 'No Case' variants compares the morphologically related verbs but the case information is changed. This is done primarily to test whether the models are robust to subtle changes in the surface forms of the arguments. Table 1 shows example for each variant.

Each set has 56 pairs for the acceptability task. To collect acceptability judgements, we conducted a forced choice acceptability judgment experiment using PCIBEX (Zehr and Schwarz, 2023). Participants were asked to choose the most acceptable sentence (see Appendix B.1 for all details). We present annotator accuracy along with LLMs' in Table 2. For all the variants of our dataset, human accuracy is quite high. We use the sentential prefix as shown in Table 1 for the cloze task.

4 Models

We test our dataset using six models via the HuggingFace Transformers library (Wolf et al., 2020) – four BERT-based masked language models (XLM-RoBERTa, MuRIL, IndicBERT$_{v2}$ and HindBERT) and two causal language models (mGPT and BLOOM). All models, except for HindBERT are multilingual models and differ primarily in terms of their size and the language(s) they are trained on. (An overview of models is presented in Appendix A). mGPT has 1.3B and 3B variants and BLOOM has 560M, 1.1B, 1.7B, 3B, 7.1B, 13B, and 176B variants. We found that as the parameters increased beyond 1B for the these models, performance worsened. On the 'Different Verb' variant of our benchmark the performance of the 1.7 million and 1.1 billion variants of the BLOOM model was the same (75% accuracy). However, for BLOOM 3 billion, the performance dropped to 62.5%. These results are similar to Kryvosheieva and Levy (2025)'s results for Hindi where the performance dropped for BLOOM's 3 billion variant. Hence, in this study we present results only from mGPT$_{1.3b}$, BLOOM$_{560m}$ and BLOOM$_{1.1B}$.

We evaluate models' performance using sentence score. For causal models, the score of a sentence is computed as the sum of the log-probabilities of each token conditioned on the sequence of preceding tokens. Whereas for masked models, we employ the pseudo-log-likelihood (PLL) scoring method introduced by Kauf and Ivanova (2023). The original PLL scoring method estimates sentence probability by masking words iteratively in a sentence, calculate the probability of each mask, and then multiplying probabilities of each word (Wang and Cho, 2019; Salazar et al., 2020). However, this method does not mask within word tokens of a multi-token word and results in inflated scores (Kauf and Ivanova, 2023). There-

Type	Models	Accuracy			
		DV	SV	No Case(E)	No Case(I)
masked	XLM-R_{base}	67.9	55.4	35.7	58.9
	XLM-R_{large}	89.3	62.5	53.6	69.6
	MuRIL	85.7	76.8	50.0	67.9
	$IndicBERT_{v2}$	92.9	**91.1**	67.9	83.9
(monolingual)	HindBERT	**98.2**	83.9	**83.9**	**91.1**
causal	$mGPT_{1.3b}$	53.6	21.4	16.1	30.4
	$BLOOM_{560m}$	58.9	42.9	8.9	42.9
	$BLOOM_{1.1b}$	75.0	58.9	23.2	62.5
Humans		99.0	90.9	96.4	99.7

Table 2: Average percentage accuracy of the LLMs and human performance on each experiment (chance probability is 50%). Overall, LLMs performance is comparable to humans and the monolingual model (HindBERT) performs better than the multilingual ones.

fore, we calculate the PLL score for each word by masking within word tokens as well.

We calculate the PLL score for each sentence individually. The sentence with the greater PLL score is deemed to be more acceptable than the other. We then evaluate these probabilities against the gold data to calculate accuracy.

The Syntactic Log-Odds Ratio (SLOR) (Pauls and Klein, 2012; Lau et al., 2017; Lu et al., 2024) is also another method that is used to score sentences, while controlling for sentence length and lexical frequency. We did not calculate this score in our work as the training data for all the models that we tested was not publicly available. We also note that in our dataset all the example sentences were of similar length (between 9-11 words).

5 Results

Acceptability Task: Table 2 shows results for the acceptability task. For the 'Different Verb' variant, all masked models performed above chance with the monolingual model close to the human accuracy. However, all causal models lag far behind humans with only $BLOOM_{1.1b}$ achieving 75% accuracy. mGPT and BLOOM have shown good results in Kryvosheieva and Levy (2025)'s experiments on Hindi syntactic agreement but performed poorly for our task. Our results suggest that verbal alternations are more challenging than syntactic agreement for causal models. We additionally tested the Llama 3.2-1B and Llama 3.3-3B models for our acceptability task, but found their performance to be similar to mGPT and BLOOM.

For the 'Same Verb' task, there is a drop in performance, which is also reflected in the human accuracy. But the performance drop is more prominent in XLM-R-large and MuRIL. For the 'No Case(I)', both IndicBERT and HindBERT are less accurate. This shows that using an additional instrument argument, and maximally filling all argument and adjunct slots does help LLMs to discriminate, while it makes little difference to humans. The weak performance for 'No Case(E)' variant is surprising. All models are less accurate, showing that case information like the ergative marker **/-ne/** is an important cue for models. Ravfogel et al. (2019) also report that overt morphological case marking makes model prediction easier for syntactic agreement phenomena.

As discussed in Section 2 Hindi verbs can be classified into different categories according to their valency and type. In order to understand whether these distinctions impact model performance, we further analyze our results for each of the different categories. For intransitives and transitives, models' performance across each task was uniform, however we do see a decrease in performance for ditransitives in all variants except for the 'Different Verb' task (see Table 5 in Section C in the Appendix).

Sentence Completion Task: We also carried out a cloze-style sentence completion task. We took the best performing models– the multilingual $IndicBERT_{v2}$ and monolingual HindBERT and asked them to complete the sentence as shown in Table 1. Both models were shown 56 sentential prefixes with the missing verb followed by the auxiliary signaling the end of the sentence. All the gold examples contain the morphological /-va/ causative.

Models rarely generated verbs with the /-va/ causative. Rather, the completions are usually transitive or ditransitive verbs. Sometimes these completions may be grammatical due to the ambigu-

Sentential Prefix	Expected	Predicted
mohən-ne bəcci-se pənk^{h}e-se mombətti ____ t^{h}i 'Mohan made/had the girl ____ the candle with the fan.'	buj^{h}vai (made to extinguish)	1. k^{h}əridi (bought) 2. nikali (removed)

Table 3: Example of cloze predictions from (1) HindBERT and (2) IndicBERT$_{v2}$

ity in the case markers on the nouns (see Section 2). Our qualitative analysis suggests that in 28% of the sentences, LLMs produce completions are ungrammatical. The errors show lack of commmonsense or pragmatic knowledge, in particular semantic content of the nominal argument and the case marker. Table 3 shows such an example where the most appropriate verb would be *extinguish*, but the models predict *buy* or *remove*. This shows that the models learn about valency and morphological forms (as shown by the acceptability tasks) but not about event semantics.

We also collected human judgements to see whether they prefer the gold completions or models' predictions using a forced choice task. Annotators were shown pairs of completions and asked to select the most grammatical option. We then calculated the percentage of times annotators agreed with the gold completions, finding a mean agreement rate of 85.9%, which indicates strong preference for the gold completions over the models outputs (see Appendix B.2 for the experiment details).

6 Discussion

In this work, we have created a benchmark of minimal pairs with four variants to test the knowledge of Hindi verbal alternations. Our benchmark has been publicly released.[2] We show that masked models are the closest to human performance for the acceptability task, but when these models are used in a cloze-style completion, their completions lack integration of both syntactic and semantic knowledge. This indicates an incomplete understanding of verb frames.

Hindi morphologically encodes its verbal argument structure, and this information seems to give the models a boost in the 'Different Verb' variant (Mueller et al., 2020). At the same time, case syncretism is a disadvantage, which makes the argument and adjunct distinction more challenging for 'No Case'. Both IndicBERT$_{v2}$ and HindBERT are fairly large models, trained on 20 billion and 1.8 billion tokens respectively. It is unlikely that increasing the size of the models will help to improve their event semantics knowledge.

We see that that current models have close to human performance for acceptability judgements but they are far less robust in a generation task. The ungrammatical completions indicate that the models have a surface understanding of valency but are unable to integrate this knowledge with event meaning. Our research points towards the need to investigate syntactic and semantic integration in LLMs.

Limitations

Our study focuses on one syntactic phenomenon, that is knowledge of verb frames in Hindi, unlike benchmarks like BLiMP (Warstadt et al., 2020) that includes many syntactic phenomena. Future research work covering other syntactic phenomena for Hindi and other languages will give a generalized idea of models' linguistic competence. Further, we carried out the cloze task only with top performing models and not others. There is a possibility that causal models may have better performance and we plan to explore this in future work.

Ethical Consideration

We collected informed consent from all individuals who volunteered to participate in the data collection, adhering to all relevant norms and regulations of our institution. We also obtained required permissions from our institute's ethics committee. All the participants for all the studies were adequately compensated for their time.

Acknowledgments

We gratefully acknowledge the Google Research Scholar Award (2024) to the second author, which helped support this research. We are thankful to the reviewers for their comments and valuable feedback. We also thank the annotators for their participation.

References

Rafiya Begum, Samar Husain, Lakshmi Bai, and Dipti Misra Sharma. 2008. Developing verb frames for Hindi. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*

[2] https://github.com/kjain93/verb-knowledge-in-LLMs

(LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Raviraj Joshi. 2022. L3Cube-HindBERT and DevBERT: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., 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, Online. Association for Computational Linguistics.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Daria Kryvosheieva and Roger Levy. 2025. Controlled evaluation of syntactic knowledge in multilingual language models. *LoResLM 2025*, page 402.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Jiayi Lu, Jonathan Merchan, Lian Wang, and Judith Degen. 2024. Can syntactic log-odds ratio predict acceptability and satiation? In *Proceedings of the Society for Computation in Linguistics 2024*, pages 10–19, Irvine, CA. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3532–3542, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

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.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2024. JCoLA: Japanese corpus of linguistic acceptability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9477–9488, Torino, Italia. ELRA and ICCL.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ark Verma, Vivek Sikarwar, Himanshu Yadav, Ranjith Jaganathan, and Pawan Kumar. 2022. Shabd: A psycholinguistic database for hindi. *Behavior Research Methods*, 54(2):830–844.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 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, Online. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Jérémy Zehr and Florian Schwarz. 2023. PennController for internet based experiments (IBEX).

A Models Evaluated

A.1 XLM-R

XLM-R (Conneau et al., 2019) is a multilingual masked language model (MLM) developed by Facebook. It is pretrained on trained on 2.5TB of filtered CommonCrawl data in 100 languages including Hindi. In this work, we are evaluating the base and large version of this model. XLM-R_{base} has 12 layers, 768 hidden units, 12 attention heads, and 270M parameters where as XLM-R large has 24 layers, 1024 hidden units, 16 attention heads, and 550M parameters.

Type	Model	Tokens	Par
maked	XLM-R_{base}	2.5TB	270M
	XLM-R_{large}	2.5TB	550M
	MuRIL	21B	236M
	$IndicBert_{v2}$	20.9B	278M
(monolingual)	HindBert	1.8B	
causal	mGPT	46B & 442B	1.3B
	$Bloom_{560m}$	341B	560M
	$Bloom_{1.1b}$	341B	1.1B

Table 4: Models evaluated by training data size (in tokens) and number of parameters (Par). We couldn't find the exact number of parameters for HindBERT.

A.2 MuRIL

MuRIL (Multilingual Representations for Indian Languages) (Khanuja et al., 2021) is a multilingual transformer-based language model developed by Google, specifically for Indian languages. It is based on the BERT architecture, with 12 layers, 12 attention heads, and 236 million parameters. MuRIL is trained on significantly large amounts of Indian text corpora across 16 Indian languages and English. It significantly outperforms mBERT on all tasks in XTREME benchmark (Hu et al., 2020).

A.3 IndicBERT

IndicBERT (Kakwani et al., 2020) is a multilingual ALBERT-based language model developed by AI4Bharat, optimized for Indian languages. It has two versions and we are testing the version 2. IndicBERT v2 is trained on IndicCorp v2, an Indic monolingual corpus of 20.9 billion tokens, covering 24 Indian languages. The model has 12 encoder layers, 12 attention heads, and 278 million parameters.

A.4 HindBERT

HindBERT (Joshi, 2022) is a monolingual BERT-based transformer model trained exclusively on Hindi by L3Cube. It is trained on around 1.8 billion Hindi tokens. The model has 12 layers and 12 attention heads, and the vocabulary size of 197285.

A.5 mGPT

Multilingual GPT (mGPT) (Shliazhko et al., 2024) is a causal language model based on the GPT-3 architecture. It supports 61 languages, including several Indian languages, and the pretraining corpus size is 46B (Wikipedia), and 442B UTF characters (C4). There are two variants available for

Models	DV			SV			No Case(E)			No Case(I)		
	Intran	Tran	Ditran	Intran	Tran	Ditran	Intran	Tran	Ditran	Intran	Tran	Ditran
XLM-R_{base}	64.3	69.6	80	75	43.5	0	57.1	17.4	0	75.0	52.2	0
XLM-R_{large}	85.7	91.3	100	82.1	47.8	20.0	60.7	47.8	40.0	89.3	56.5	20.0
MuRIL	78.6	95.6	80	78.6	78.3	60.0	53.6	47.8	40.0	71.4	69.6	40.0
IndicBERT	92.9	91.3	100	96.4	86.9	80.0	75	56.52	80.0	92.9	78.3	60.0
HindBERT	96.4	100	100	92.9	82.6	40.0	89.3	86.9	40.0	100	91.3	40.0
$mGPT_{1.3b}$	42.9	65.2	60.0	53.6	8.7	0	21.4	13.0	0	53.6	8.7	0
$BLOOM_{560m}$	50	69.6	60.0	53.6	39.1	0	14.3	4.3	0	53.6	39.1	0
$BLOOM_{1.1b}$	71.4	78.3	80.0	75.0	60.9	0	28.6	21.7	0	75.0	60.9	0

Table 5: Average percentage accuracy of the LLMs on each experiment for different class of verbs

this model. In this work, we are evaluating only the small one with 1.3 billion parameters

A.6 BLOOM

BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) (Workshop et al., 2022) is a multilingual autoregressive transformer model developed by the BigScience project. It supports 46 natural languages, including many low-resource ones, and 13 programming languages. BLOOM is trained on the ROOTS corpus. The full model has 176 billion parameters but also has 5 small size variants. For our study, we test the 560 millions variant and the 1.1 billions variant.

B Experiments with Humans

B.1 Acceptability Task

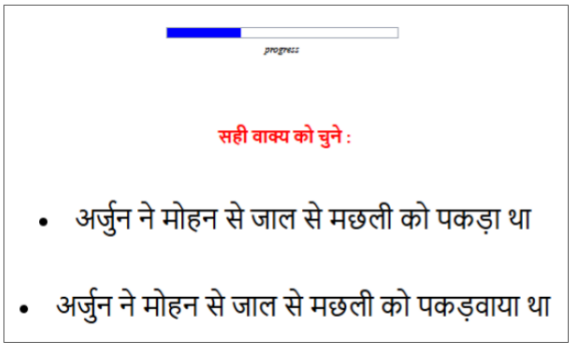

Figure 1: Example of a minimal. English translation: *Arjun made Mohan catch a fish with net.*

All the experiments for acceptability task were conducted using PCIBEX. Participants were given instruction about the task in both in Hindi and English. We explained that there are no risks involved in the task to each participant.

In each experiment they saw the minimal pair simultaneously as shown in Fig.1 and they were asked to choose the more grammatically acceptable sentence for each pair. We also included fillers and practice sets. The order of main sentences and fillers was shuffled.

Participants for first experiment, Different verb, were aged 18-40. We collected the data in person using anonymous id for each one of them. We have 15 judgements for each pair in this experiment. The participants were paid according to our institution policy. For the remaining variants we collected data on the crowdsourcing platform Prolific. For each of these experiments the dataset consisted of 28 randomly sampled sentences. We collected 20 judgements on each pair. All the participants were self reported native Hindi speakers and they were paid in accordance with Prolific's fair compensation policies.

B.2 Cloze Task

We collected human judgments on the completions produced by the two models. We presented each sentence prefix to 14 native speakers of Hindi on Prolific and provided them three options: the (gold) causative verb and the verbs predicted by IndicBERT and HindBERT. Participants were asked to choose the most appropriate completion for each sentence. The information sheet clearly mentioned that there are no risks involved in the study. All participants were self reported native speakers of Hindi and were paid in accordance with Prolific's fair compensation policies.

C Class wise analysis for Verbs

In Table 5, we present evaluation results of verbs categorized as intransitives (Intran), transitives (Tran) and ditransitives (Ditran) for all the models.