ID	Hindi AES Models	P1	P2	P3	P4	P5	P6	P7	P8	Average	Organic
1	SVR	0.799	0.612	0.605	0.657	0.797	0.630	0.400	0.380	0.610	0.579
2	Linear Regression	**0.800**	0.614	0.588	0.624	0.768	0.605	0.680	0.635	0.664	0.681
3	Random Forest	0.705	0.608	0.621	0.685	0.791	0.665	0.652	0.560	0.661	0.762
4	XGBoost	0.794	**0.667**	0.573	0.676	0.792	0.653	0.713	**0.641**	0.688	0.827
5	CNN	0.571	0.513	0.529	0.614	0.657	0.703	0.521	0.426	0.566	0.762
6	BiLSTM	0.631	0.517	0.612	0.703	0.643	0.713	0.607	0.443	0.608	0.842
7	CNN + LSTM + Attention	0.723	0.597	0.677	0.711	0.781	0.791	0.701	0.593	0.696	0.827
8	SKIPFLOW LSTM (Tensor)	0.742	0.621	0.695	0.731	0.804	0.777	0.717	0.619	0.713	0.812
9	mBERT	0.683	0.652	**0.711**	0.775	0.828	0.785	0.781	0.548	0.720	**0.852**
10	DistilmBERT	0.661	0.592	0.698	0.766	0.825	0.793	0.785	0.596	0.714	0.784
11	XLM-RoBERTa	0.758	0.585	0.692	**0.809**	**0.834**	**0.822**	**0.794**	0.639	0.741*	0.831
12	MuRIL	0.620	0.412	0.528	0.756	0.812	0.713	0.547	0.327	0.589	0.528
13	IndicBERT	0.651	0.489	0.659	0.751	0.799	0.784	0.708	0.412	0.656	0.796

Table 2: Experiment results of all models in terms of QWK on ASAP-Hindi corpus and the organic prompt. The bold number is the best performance for each prompt. The best average QWK is annotated with *.

ID	English AES Models	P1	P2	P3	P4	P5	P6	P7	P8	Average
1	EASE (SVR)	0.781	0.621	0.630	0.749	0.782	0.771	0.727	0.534	0.699
2	CNN + LSTM + Attention	0.822	0.682	0.672	0.814	0.803	0.811	0.801	0.705	0.764
3	SKIPFLOW LSTM(Tensor)	**0.832**	0.684	0.695	0.788	0.815	0.810	0.800	0.697	0.764
4	R^2BERT	0.817	**0.719**	**0.698**	**0.845**	**0.841**	**0.847**	**0.839**	**0.726**	**0.791**

Table 3: Published results on prominent models for AES in English. All results in terms of QWK score on the original ASAP Corpus.

from classical to advanced end-to-end methods. SKIPFLOW (Table 2, row-8) gave the highest average score amongst these, followed closely by the CNN + LSTM + Attention average. The SKIPFLOW average fell 0.051 points short off it's original implementation in English.

On average, the fine-tuned highly multilingual transformers (Table 2, rows 9-11) gave higher results across all prompts. The fine-tuned mBERT model gives the maximum score for prompt 3. The fine-tuned XLM-R model outperforms all other models on four of the eight ASAP-Hindi prompts (Prompts 4, 5, 6, 7), giving the maximum average QWK as well thereby, establishing a state-of-the-art for AES in Hindi. Compared to R^2BERT's average the fine-tuned XLM-R is 0.050 points short. Results on the mBERT model closely follow the results on the XLM-R model. The fine-tuned Indic transformers (Table 2, rows 12 and 13) did not perform comparably well. While IndicBERT did try to compete and learn, MuRIL's behavior during the training process was highly inconsistent, resulting in unpredictable results. This behavior could be attributed to a variety of factors which will be discussed in the next sub-section.

Results on the organic prompt were favourable (in comparison to results on the ASAP-Hindi set) with the fine-tuned mBERT model giving the highest QWK score. It is important to note that in contrast to the other prompts, the QWK scores obtained for organic prompt showed a slightly higher variance during training. It is likely that this is a result of the organic set's smaller magnitude compared to the ASAP-Hindi Dataset.

5.4 Analysis and Discussion

A general observation that is consistent with both previously established English AES results and our study, is that results on prompts 4,5, and 6 are higher than the other prompts for the ASAP dataset. Prompts requiring source-dependent responses perform better during training as compared to narrative, persuasive or expository prompts possibly due to a general consistency in syntax, coherence, and availability of source material. Such prompts are coupled with more balanced real-world rubrics allowing for consistency all throughout the writing and the scoring process, making such prompts ideal for generalization. In contrast generalization of ideas is slightly more difficult on prompts that allow for persuasive and expository discussions due to variance in human thought and cognition. Prompts with rubrics that are not consistent with real-world rubrics such as prompt-8, give results much worse than the aforementioned source-dependent prompts.

The more discrete nature of feature-based approaches allows for consistent performance, but their failure to under-