A Other tasks

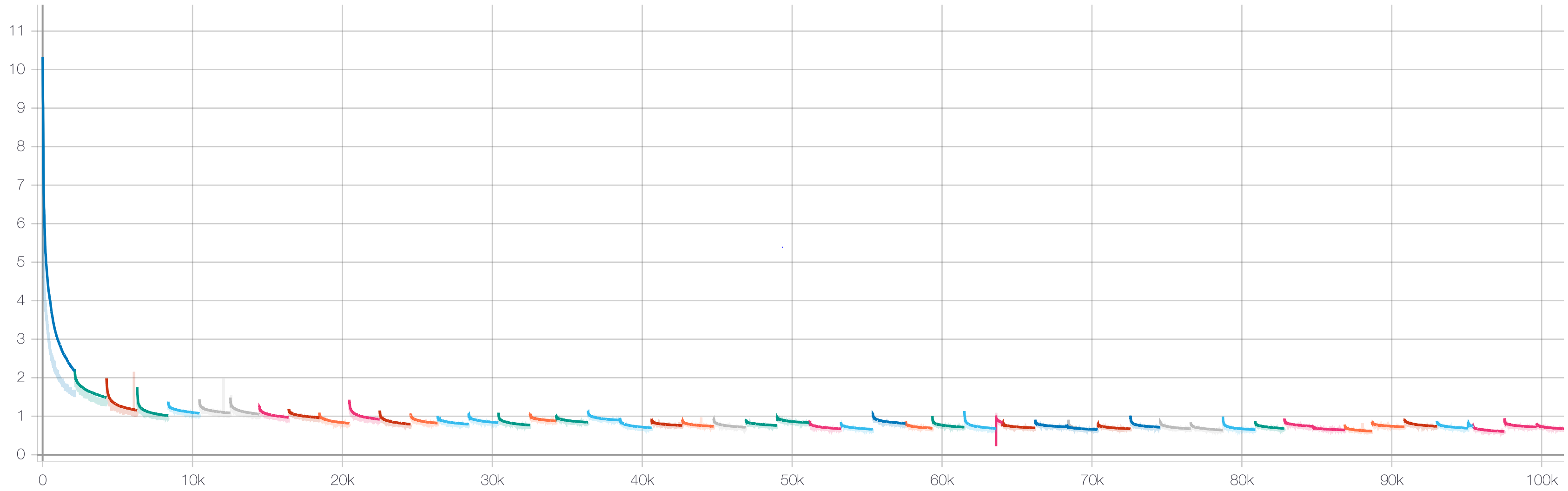
LG	IndicBert	mBERT	XLM-R	MuRIL	IA-TR	IA-O
Article-Genre Classification / bbc-articles (Accuracy)						
hi	0.7460	0.6055	**0.7552**	0.7367	0.7206	0.6963
Article-Genre Classification / inltk-headline (Accuracy)						
gu	**0.9291**	0.8916	0.8983	0.9226	0.8330	0.9044
mr	0.9430	0.8750	0.9248	**0.9545**	0.9355	0.9248
Article-Genre Classification / Soham-article (Accuracy)						
bn	0.7845	0.8023	0.8760	**0.9340**	0.5988	0.8029
Textual Entailment / copa-translated (Accuracy)						
hi	0.6250	**0.6590**	0.4318	0.5568	**0.6591**	0.5796
gu	0.5341	0.4318	0.4886	**0.5909**	**0.5909**	0.5113
mr	0.5909	0.5568	0.6136	0.5682	0.5909	**0.6590**
Textual Entailment / wnli-translated (F-Score)						
hi	0.3604	0.3604	0.3604	0.3604	0.3604	0.3616
gu	0.3604	0.3604	**0.5389**	0.3604	0.3107	0.3604
mr	0.3604	0.4167	**0.4496**	0.3914	0.4228	0.3604
Sentiment Analysis / iitp-movie-reviews (Accuracy)						
hi	0.5903	0.5677	0.6161	**0.7032**	0.6000	0.5967
Sentiment Analysis / iitp-product-reviews (Accuracy)						
hi	0.7132	0.7457	0.7897	**0.8183**	0.7705	0.7457
Discourse Mode Classification / midas-discourse (Accuracy)						
hi	0.7844	0.7120	0.7994	**0.8164**	0.7994	0.7943
Entity Classification / wikiann-ner (F-Score)						
hi	0.9031	0.8656	0.8962	**0.9237**	0.8720	0.8476
bn	0.9339	0.9181	0.9295	**0.9503**	0.9211	0.9486
gu	0.7021	0.6804	0.5532	0.8016	0.7306	**0.8318**
mr	0.8871	0.9127	0.8786	**0.9199**	0.8675	0.8482
or	0.3509	0.1905	0.2500	0.3882	0.3460	**0.5737**
pa	0.4444	0.5000	0.1786	0.8535	0.3491	**0.6313**
Title Prediction / wiki-section-title (Accuracy)						
hi	0.7780	0.8012	0.7692	**0.8528**	0.6779	0.6761
bn	0.8266	0.8253	0.8091	**0.8781**	0.6135	0.7062
gu	0.6879	0.7452	0.2739	**0.8465**	0.2614	0.4044
mr	0.7744	0.8049	0.7744	**0.8529**	0.5299	0.5031
or	0.6825	0.2222	0.6825	**0.8167**	0.2928	0.3147
pa	0.7754	0.7247	0.7029	**0.8240**	0.2817	0.6235
Part-of-Speech Tagging / ud-pos (F-Score)						
hi	0.9755	0.9693	0.9794	**0.9779**	0.9618	0.9562
mr	0.8024	**0.9024**	0.8249	0.5388	0.8114	0.7906
ur	0.9047	0.9102	**0.9280**	0.9168	0.9026	0.8915

Table 1: Comparison of Indo-Aryan LM with existing multilingual LMs.

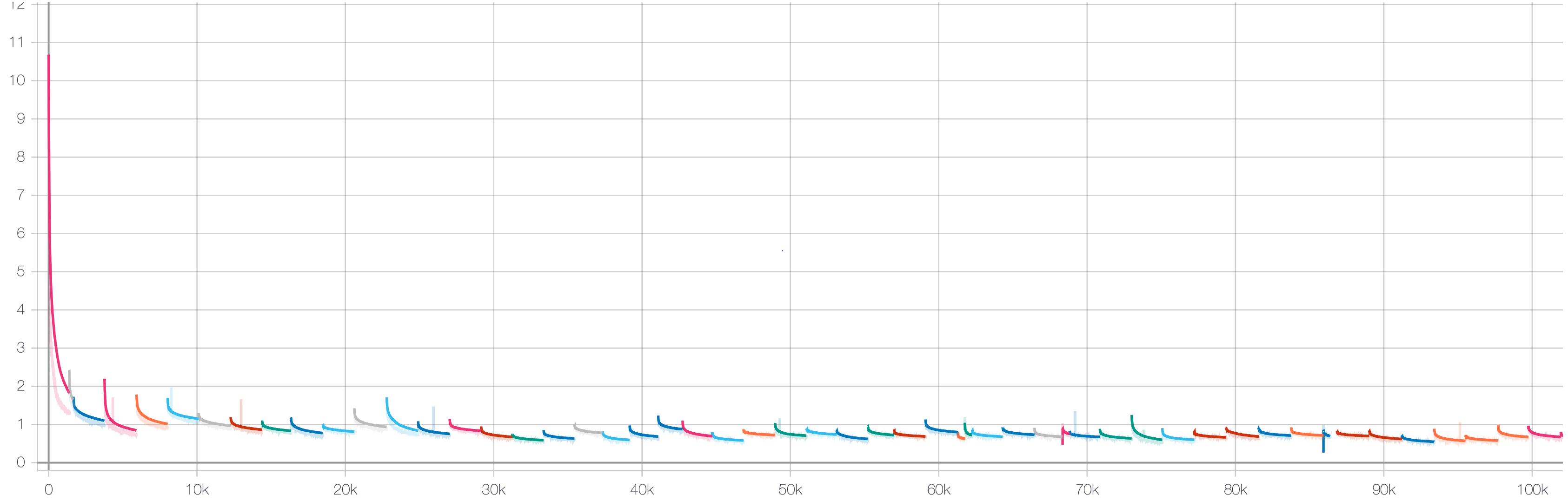
Here, the results are reported on eleven datasets spread across seven tasks. Notice that monolingual datasets namely bbc-articles, soham-articles, iitp-movie-reviews, iitp-product-reviews, and midas-discourse are added in Table 1 in addition to the four tasks reported in main manuscript. In total, 28 task-language combinations are considered. Although, inltk-headline dataset is multilingual, the prediction classes are different in both (Gujarati and Marathi) parts which is unsuitable for multilingual FT.

On 16 out of 28 experiments, MuRIL obtains the best results. Note that, pre-training of MuRIL enjoys an additional supervision using the parallel corpora of transliterated (romanized) and translated counterparts of original text. On 6 of the 28 experiments, either IA-TR or IA-O model improves over or equates to the state-of-the-art. This includes, 3-3 experiments pertaining to copa-translated and wikiann-ner each. Especially for the wikiann-ner, the improvement is significant for Gujarati (0.8318 vs 0.8016) and Oriya (0.5737 vs 0.3882).

Our fine-tuned models seems to be falling short on other tasks compared to the published results. Title Prediction task, in particular, seem to be the most difficult for the IA models.



(a) IA-Original



(b) IA-Transliterated

Figure 1: Loss graph pertaining to pre-training of language models. Steps and MLM loss are plotted on horizontal and vertical axes, respectively.