Languages	LLaMA-2-7B **3% (10K)**				LLaMA-2-13B **3% (10K)**			
	Base	Top	Bottom	Random	Base	Top	Bottom	Random
Arabic	6.771	115398.328	6.772	7.895	6.261	88678.016	6.315	7.112
Chinese	8.652	369027.531	8.564	9.837	7.838	70912.242	7.806	8.619
Czech	19.834	78480.219	19.837	24.005	17.744	53699.43	17.655	20.485
Danish	8.372	46503.742	8.373	9.929	7.402	39408.91	7.401	8.278
Dutch	16.959	55704.191	16.961	20.121	15.64	41159.938	15.572	18.295
English	7.653	15738.043	7.654	8.359	7.447	23678.322	7.234	7.821
Finnish	7.566	46616.094	7.568	8.934	6.887	47002.539	6.861	7.826
French	13.605	44385.668	13.609	15.003	12.765	38755.539	12.678	13.682
German	18.355	84497.234	18.361	21.404	17.29	43319.586	17.02	18.972
Greek	3.832	147740.5	3.833	4.527	3.599	70136.242	3.6	4.146
Hungarian	16.365	52652.363	16.367	20.039	14.756	48407.305	14.735	17.633
Indonesian	44.269	39055.945	44.267	48.175	37.909	36912.34	37.929	38.548
Italian	14.859	54297.523	14.865	17.341	13.694	42515.969	13.69	15.207
Japanese	10.888	358722.188	10.891	12.535	10.072	68055.984	10.118	11.661
Korean	4.965	102918.828	4.966	5.649	4.724	65209.328	4.736	5.241
Malay	66.581	23501.082	67.158	74.167	46.885	35517.879	47.191	58.947
Malayalam	5.133	314088.969	5.136	6.396	4.972	131629.438	4.971	5.654
Norwegian	14.425	38111.27	14.431	17.854	13.142	38500.664	13.138	15.041
Persian	6.509	78203.031	6.51	7.628	6.205	98292.281	6.22	7.009
Polish	12.629	81373.273	12.633	14.843	11.414	52403.461	11.393	12.987
Portuguese	15.318	47779.789	15.321	17.297	13.667	41184.457	13.86	15.376
Romanian	10.893	45836.578	10.897	13.061	9.652	51766.957	9.694	10.969
Russian	12.062	227916.828	12.061	13.728	11.048	103490.719	11.004	11.757
Spanish	17.079	57679.461	17.087	18.98	16.351	40338.426	16.265	17.292
Swahili	75.908	42977.977	75.93	89.38	70.519	40400.949	70.443	81.216
Swedish	14.714	55893.812	14.717	17.258	13.229	45396.66	13.301	14.933
Tamil	4.162	447989.969	4.162	5.04	4.028	141214.188	4.052	4.488
Turkish	11.214	57037.605	11.215	13.765	9.834	41566.105	9.791	11.374
Ukrainian	9.409	168085.672	9.408	10.875	8.295	94307.312	8.296	9.076
Vietnamese	5.824	36374.734	5.825	6.614	5.471	31730.328	5.467	5.995

Table 16: LLaMA-2 perplexity on 30 languages with 3% removal ratio. '10K' means that the region is selected from 10,000 samples. Here, we reduce training samples to 10, 000 during further pre-training across six languages.

Languages	LLaMA-2-7B **1% (100K)**				LLaMA-2-7B **5% (100K)**			
	Base	Top	Bottom	Random	Base	Top	Bottom	Random
Arabic	6.771	67579496	6.77	7.021	6.771	112504.609	6.774	10.823
Chinese	8.652	120887480	8.561	8.818	8.652	156026.938	8.565	12.775
Czech	19.834	24343856	19.835	21.176	19.834	96580.281	19.845	24.797
Danish	8.372	1631186.625	8.372	8.775	8.372	82876.266	8.375	13.565
Dutch	16.959	6845146	16.963	18.056	16.959	79497.211	16.961	27.01
English	7.653	512756	7.654	7.851	7.653	46197.477	7.656	9.289
Finnish	7.566	4727027.5	7.567	7.948	7.566	60183.328	7.56	13.005
French	13.605	4768049	13.608	14.198	13.605	87642.109	13.611	19.076
German	18.355	17940508	18.357	19.724	18.355	106160.992	18.364	28.772
Greek	3.832	14242545	3.833	3.972	3.832	141320.578	3.835	6.45
Hungarian	16.365	130584	16.366	17.35	16.365	77265.188	16.369	30.376
Indonesian	44.269	1654245	44.347	49.476	44.269	83353.344	44.298	64.743
Italian	14.859	5265871.5	14.863	15.607	14.859	83076.164	14.865	22.6
Japanese	10.888	28104000	10.88	11.196	10.888	124647.633	10.895	16.619
Korean	4.965	16449047	4.965	5.095	4.965	59954.559	4.967	7.831
Malay	66.581	7875206	66.673	78.545	66.581	51824.859	66.751	90.933
Malayalam	5.133	7151096	5.133	5.359	5.133	182008.484	5.137	7.905
Norwegian	14.425	4223085	14.429	15.35	14.425	79399.109	14.434	23.621
Persian	6.509	2233196	6.507	6.782	6.509	107342.734	6.511	10.236
Polish	12.629	6547834.5	12.631	13.36	12.629	88912.945	12.632	21.372
Portuguese	15.318	6249820	15.319	15.927	15.318	78851.766	15.324	22.608
Romanian	10.893	5251915.5	10.895	11.526	10.893	71228.375	10.899	19.21
Russian	12.062	17596800	12.061	12.067	12.062	102639.602	12.066	18.504
Spanish	17.079	8220832.5	17.084	18.029	17.079	96575.547	17.084	24.007
Swahili	75.908	7875009	75.845	83.963	75.908	77765.133	75.8	131.709
Swedish	14.714	4712167.5	14.716	15.534	14.714	81574.734	14.717	22.628
Tamil	4.162	20660974	4.162	4.265	4.162	173728.312	4.164	5.881
Turkish	11.214	4489915	11.214	11.882	11.214	58347.055	11.218	19.76
Ukrainian	9.409	11689088	9.409	9.811	9.409	90008.312	9.414	14.807
Vietnamese	5.824	2235468	5.825	6.018	5.824	54187.02	5.825	9.015

Table 17: LLaMA-2-7B perplexity on 30 languages with 1% and 5% removal ratio. '100K' means that the region is selected from 100,000 samples. Here, we change the removal ratio from 3% to 1% and 5%.