

## A APPENDIX

### A.1 WMT MODEL AND TRAINING DETAILS

For our experiments, we use the Transformer Base model in (Chen et al., 2018). The sole difference is that we use a 64k vocabulary: our model therefore contains 142M parameters. For multilingual models, we share all parameters across language pairs including softmax layer in input/output word embeddings.

We use a 64k token vocabulary formed using a Sentence Piece Model (Kudo & Richardson, 2018). The vocabulary is shared on both the encoder and decoder side. To learn a joint SPM model given our imbalanced dataset, we followed the temperature based sampling strategy with a temperature of  $T = 5$ .

Finally, our models are optimized using the Adafactor optimizer (Shazeer & Stern, 2018) with momentum factorization and a per-parameter norm clipping threshold of 1.0. We followed a learning rate of 3.0, with 40K warm-up steps for the schedule, which is decayed with the inverse square root of the number of training steps after warm-up. BLEU scores presented in this paper are calculated using SacreBLEU Post (2018) on the WMT test sets.<sup>3</sup>

### A.2 WMT DATASET DETAILS

In Table 2 we provide the training set details for the WMT<sup>4</sup> setup we use (Siddhant et al., 2020). We provide the data sizes and WMT years of the Train, Dev and Test sets we use.

### A.3 INDIVIDUAL WMT BLEU SCORES

**Bilingual baselines:** We first train Transformer Base and Big models on each language pair. The results are in Table 3.

In Tables 5 and 6 we provide individual BLEU scores of the models discussed in Table 1.

### A.4 DETAILED BREAKDOWN OF PARAMETER COUNTS ON WMT

Table 6 describes the parameter counts of different parts of the Transformers compared in Table 1.

### A.5 DETAILED BREAKDOWN OF PARAMETER COUNTS

In Table 7 we describe the parameter counts of different parts of the Transformers discussed in Section 4.3.

### A.6 RESULTS ON LARGE MOE MODEL

In Table 8 we provide aggregate BLEU scores for the results in Figure 2.

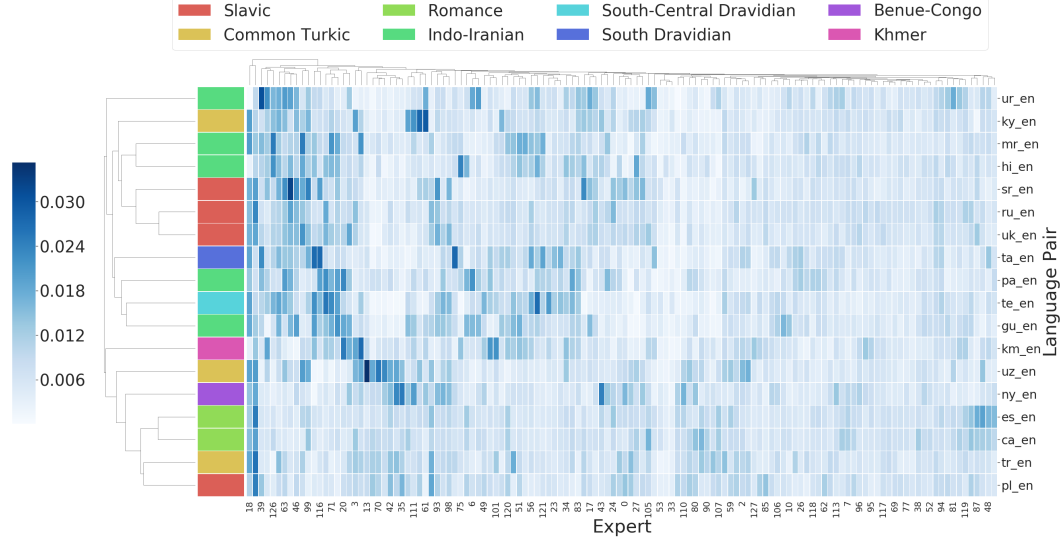
### A.7 GATING DECISIONS FOR TASK-LEVEL AND TOKEN-LEVEL MOES

In this section, we show the top expert distributions of different layers of the position-wise MoE model discussed in Section 4.3.4 in Figures 4, 5, 6 and 7.

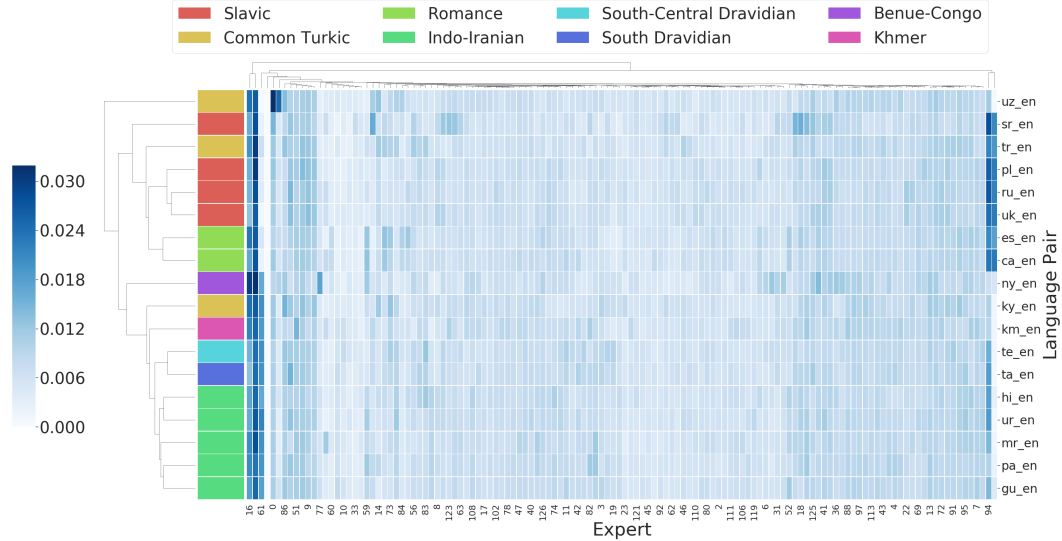
We also show expert distributions on MoE model routing by target language from EnX that was introduced in Section 4.3.2 in Figures 8 and 9. We omit results on XEn language pairs because they belong to the same task in the context of this model.

<sup>3</sup> BLEU+case.mixed+lang.<sl>-<tl>+numrefs.1+smooth.exp+tok.<tok>+version.1.3.0, where *sl* is the source language, *tl* is the target language and *tok* = *zh* if *tl* = *zh* and *intl* otherwise.

<sup>4</sup><http://www.statmt.org/wmt20/>

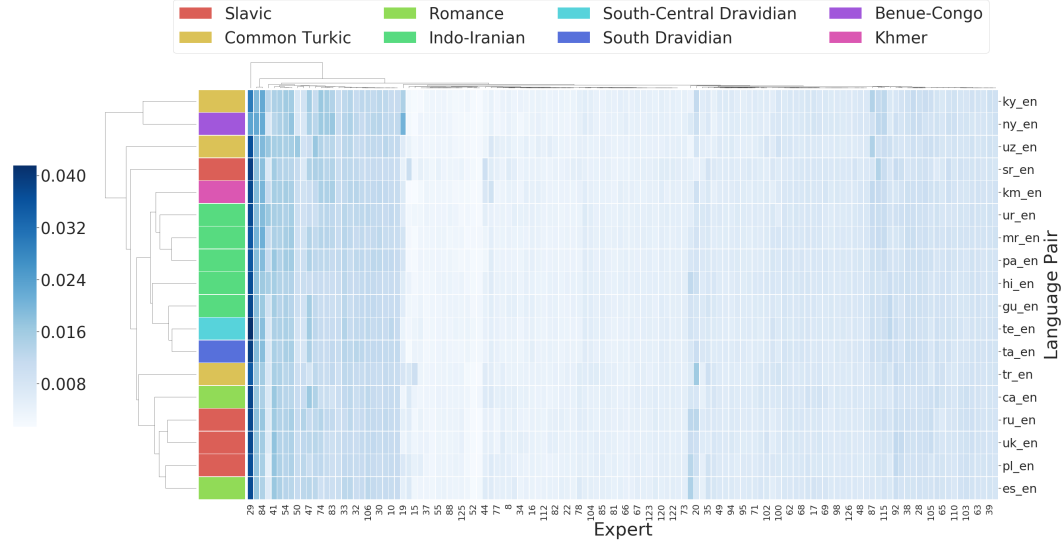


(a) Gating decisions of the first layer of the encoder for Xx-En language pairs.

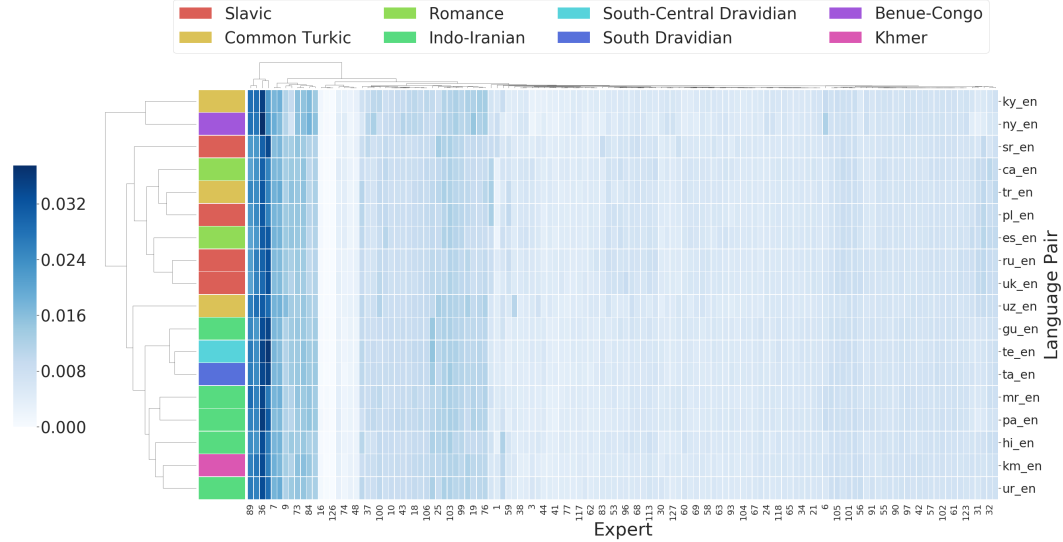


(b) Gating decisions of the last layer of the encoder for Xx-En language pairs.

Figure 5: Gating decisions of the encoder of the position-wise MoE model on Xx-En language pairs, trained on internal data on a multiway parallel dataset. In this diagram, the darker a cell, corresponding to, say en-sr and the 37th expert, the more the expert is used. In both the last layer of the encoder and decoder, the tokens from each language are fairly well distributed across experts. In (a) the first layer of the encoder, there does not seem to be any major pattern in the expert distribution whereas in (b) the last layer of the encoder, tokens from all tasks (*Xx-En*) seem to prefer the same set of few experts slightly over the others.

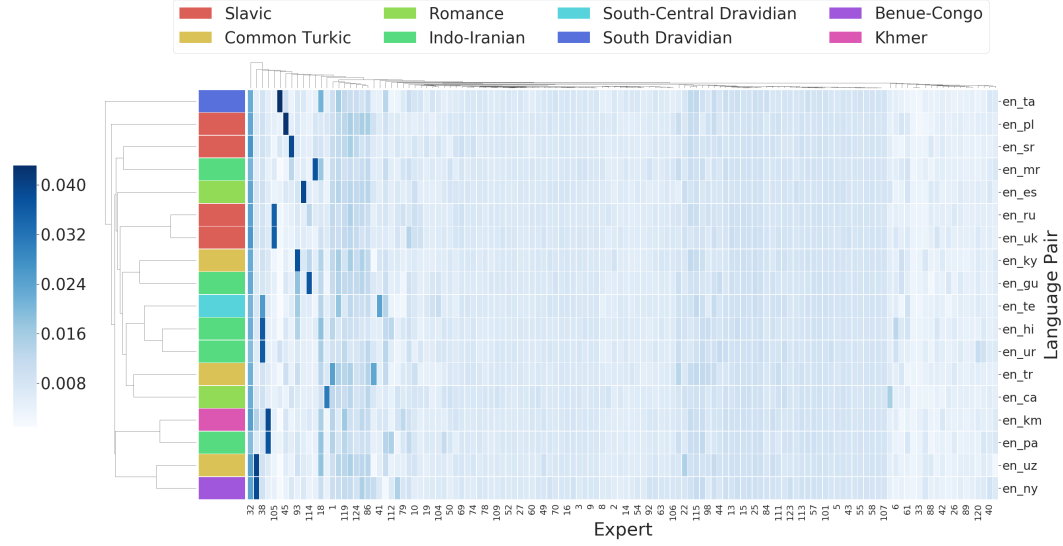


(a) Gating decisions of the first layer of the decoder for Xx-En language pairs.

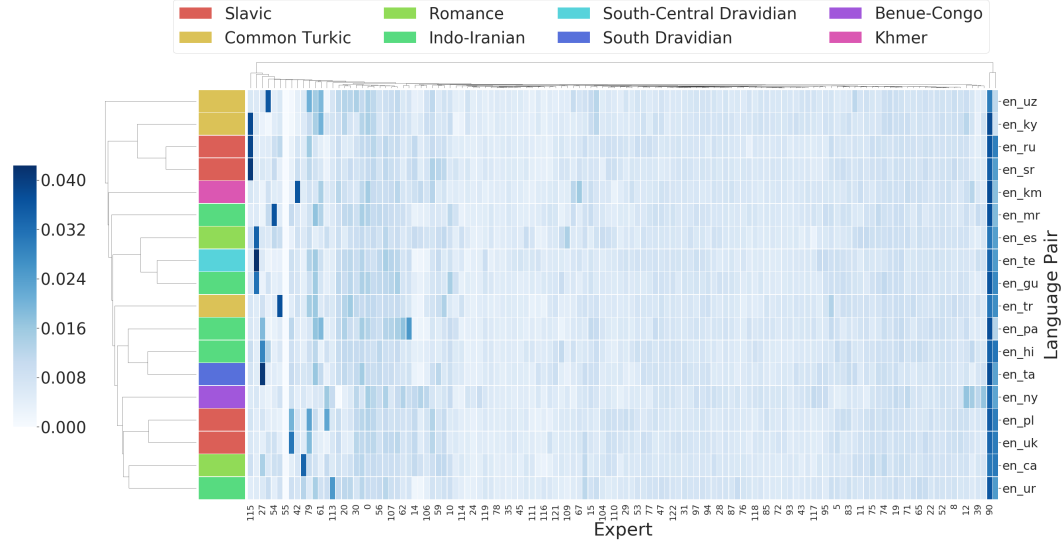


(b) Gating decisions of the last layer of the decoder for Xx-En language pairs.

Figure 6: Gating decisions of the decoder of the position-wise MoE model on Xx-En language pairs, trained on internal data on a multiway parallel dataset. In this diagram, the darker a cell, corresponding to, say en-sr and the 37th expert, the more the expert is used. In both the first and last layer of the decoder, the tokens from each language are fairly well distributed across experts. In fact, tokens from all tasks (*Xx-En*) seem to prefer the same set of few experts slightly over the others.

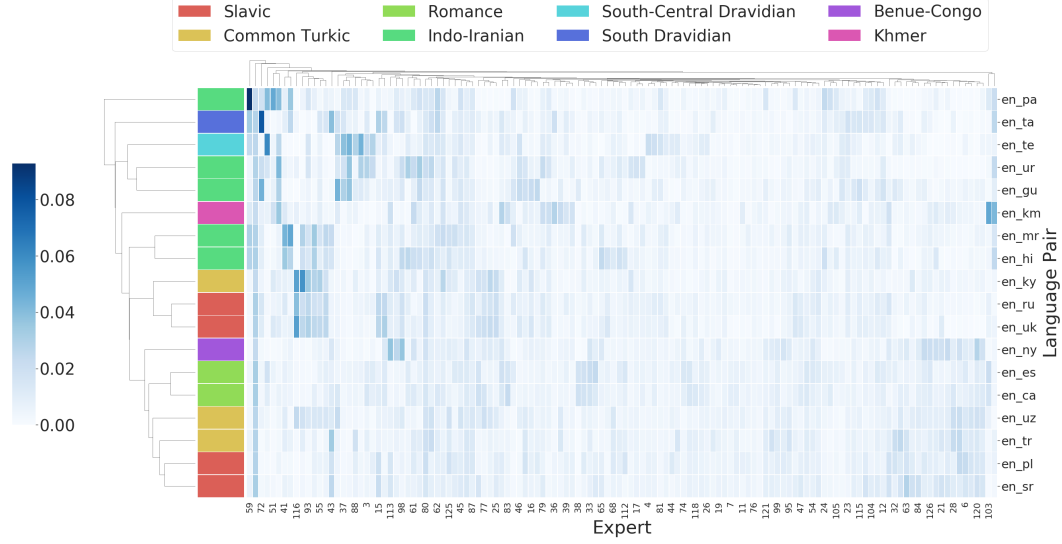


(a) Gating decisions of the first layer of the encoder for En-Xx language pairs.

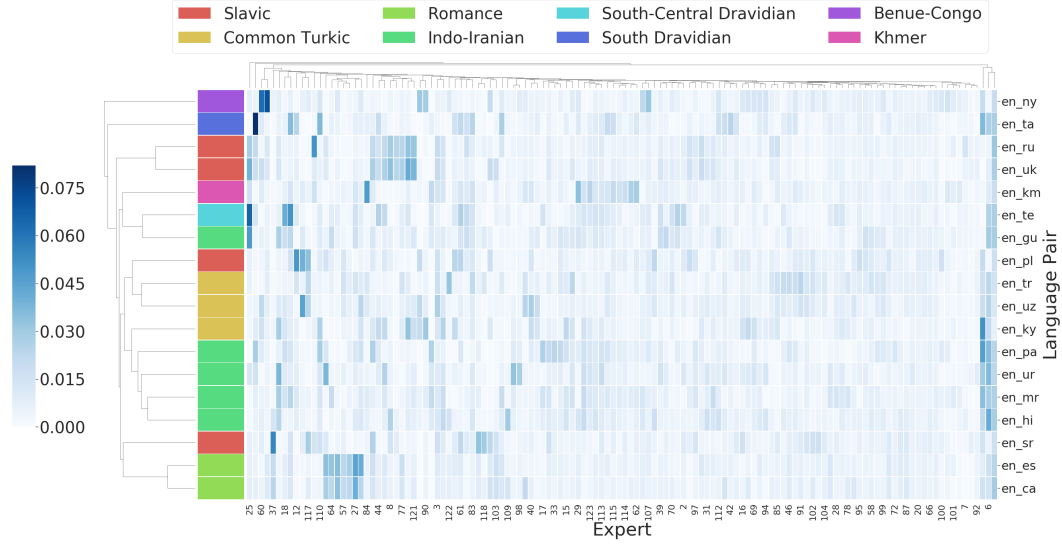


(b) Gating decisions of the last layer of the encoder for En-Xx language pairs.

Figure 7: Gating decisions of the encoder of the position-wise MoE model on En-Xx language pairs, trained on internal data on a multiway parallel dataset. In this diagram, the darker a cell, corresponding to, say en-sr and the 37th expert, the more the expert is used. In both the first and last layer of the encoder, the tokens from each language are fairly well distributed across experts. Each task (*En-Xx*) seems to slightly prefer a few experts over the other.

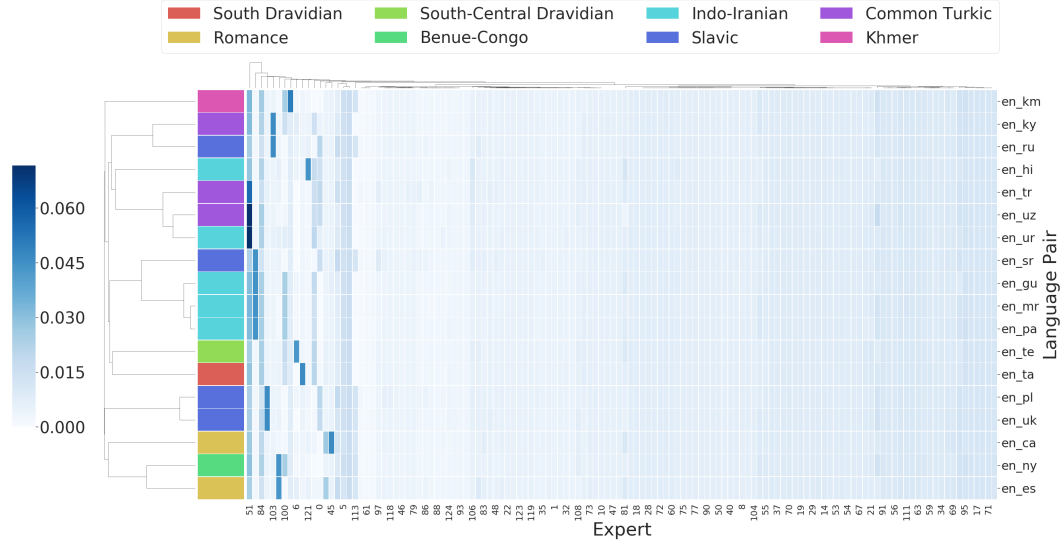


(a) Gating decisions of the first layer of the decoder for En-Xx language pairs.

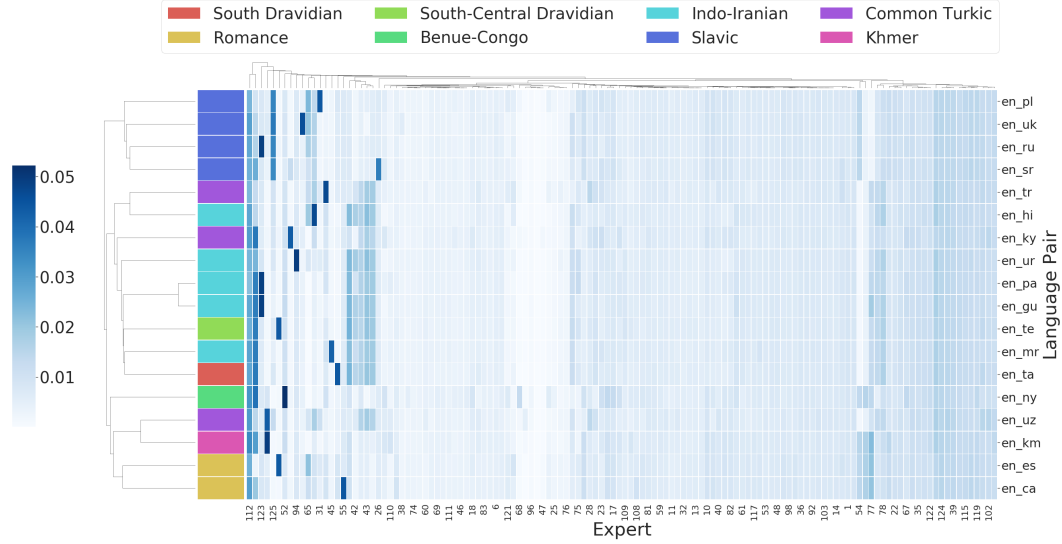


(b) Gating decisions of the last layer of the decoder for En-Xx language pairs.

Figure 8: Gating decisions of the decoder of the position-wise MoE model on En-Xx language pairs, trained on internal data on a multiway parallel dataset. In this diagram, the darker a cell, corresponding to, say en-sr and the 37th expert, the more the expert is used. In both the first and last layer of the decoder, the tokens from each language are fairly well distributed across experts. Each task (*En-Xx*) seems to slightly prefer a few experts over the other. Moreover, the set of experts appears to be similar for related languages. For example, English-Spanish and English-Catalan (two Romance Languages) have similar expert distributions and so do English-Russian and English-Ukrainian (two Slavic Languages).

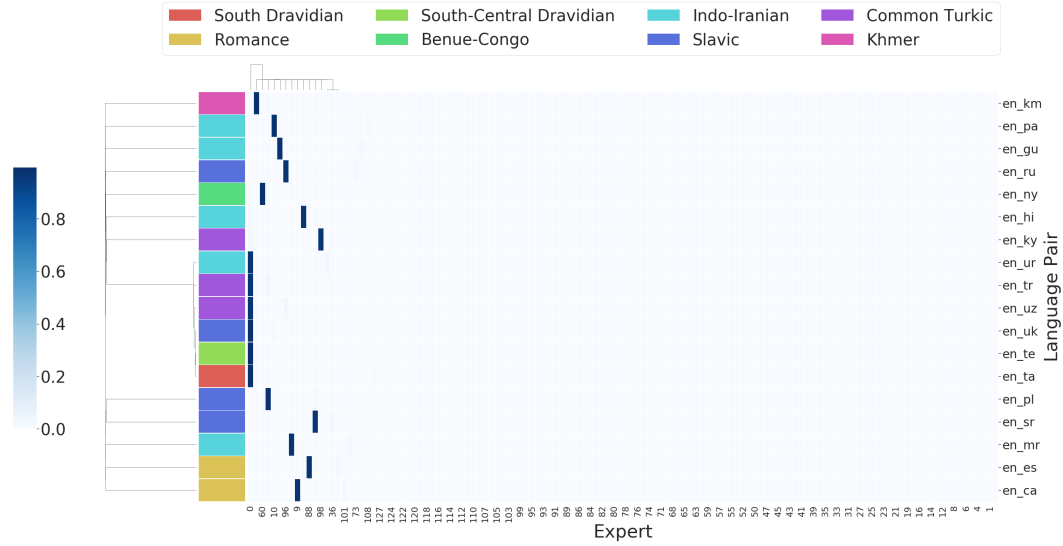


(a) Gating decisions of the first layer of the encoder for En-Xx language pairs.

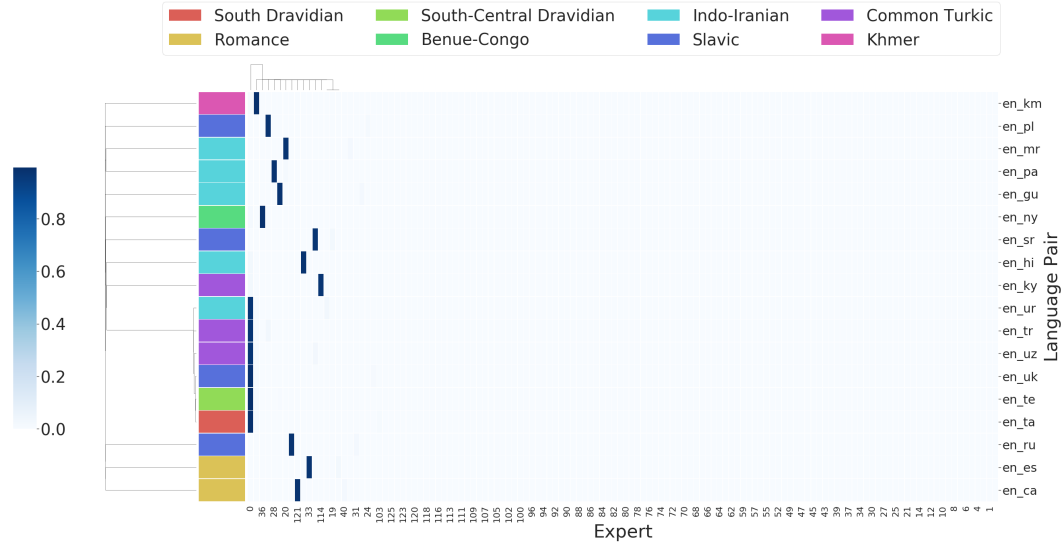


(b) Gating decisions of the last layer of the encoder for En-Xx language pairs.

Figure 9: Gating decisions of the encoder of the target language-wise MoE model on En-Xx language pairs, trained on internal data on a multiway parallel dataset. In this diagram, the darker a cell, corresponding to, say en-sr and the 37th expert, the more the expert is used. The encoder behaves similarly to that of the position-wise model: in both the first and last layer of the encoder, the tokens from each language are fairly well distributed across experts. Each task (*En-Xx*) seems to slightly prefer a few experts over the other.



(a) Gating decisions of the first layer of the decoder for En-Xx language pairs.



(b) Gating decisions of the last layer of the decoder for En-Xx language pairs.

Figure 10: Gating decisions of the decoder of the target language-wise MoE model on En-Xx language pairs, trained on internal data on a multiway parallel dataset. In this diagram, the darker a cell, corresponding to, say en-sr and the 37th expert, the more the expert is used. There seems to be some amount of expert sharing on a linguistic basis: en-ur, en-te and en-ta (two Dravidian Languages and an Indo-Iranian language) and en-tr, en-uz and en-uk (two Turkic languages and a Slavic language) share an expert. On the other hand, en-es and en-ca (two Romance languages) have different experts.

Language Pair	Data Sources			# Samples		
	Train	Dev	Test	Train	Dev	Test
cs→en	WMT'19	WMT'17	WMT'18	64336053	3005	2983
fr→en	WMT'15	WMT'13	WMT'14	40449146	3000	3003
ru→en	WMT'19	WMT'18	WMT'19	38492126	3000	2000
zh→en	WMT'19	WMT'18	WMT'19	25986436	3981	2000
es→en	WMT'13	WMT'13	WMT'13	15182374	3004	3000
fi→en	WMT'19	WMT'18	WMT'19	6587448	3000	1996
de→en	WMT'14	WMT'13	WMT'14	4508785	3000	3003
et→en	WMT'18	WMT'18	WMT'18	2175873	2000	2000
lv→en	WMT'17	WMT'17	WMT'17	637599	2003	2001
lt→en	WMT'19	WMT'19	WMT'19	635146	2000	1000
ro→en	WMT'16	WMT'16	WMT'16	610320	1999	1999
hi→en	WMT'14	WMT'14	WMT'14	313748	520	2507
kk→en	WMT'19	WMT'19	WMT'19	222424	2066	1000
tr→en	WMT'18	WMT'17	WMT'18	205756	3007	3000
gu→en	WMT'19	WMT'19	WMT'19	155798	1998	1016
en→cs	WMT'19	WMT'17	WMT'18	64336053	3005	2983
en→fr	WMT'15	WMT'13	WMT'14	40449146	3000	3003
en→ru	WMT'19	WMT'18	WMT'19	38492126	3000	2000
en→zh	WMT'19	WMT'18	WMT'19	25986436	3981	2000
en→es	WMT'13	WMT'13	WMT'13	15182374	3004	3000
en→fi	WMT'19	WMT'18	WMT'19	6587448	3000	1996
en→de	WMT'14	WMT'13	WMT'14	4508785	3000	3003
en→et	WMT'18	WMT'18	WMT'18	2175873	2000	2000
en→lv	WMT'17	WMT'17	WMT'17	637599	2003	2001
en→lt	WMT'19	WMT'19	WMT'19	635146	2000	1000
en→ro	WMT'16	WMT'16	WMT'16	610320	1999	1999
en→hi	WMT'14	WMT'14	WMT'14	313748	520	2507
en→kk	WMT'19	WMT'19	WMT'19	222424	2066	1000
en→tr	WMT'18	WMT'17	WMT'18	205756	3007	3000
en→gu	WMT'19	WMT'19	WMT'19	155798	1998	1016
fr→de	WMT'19	WMT'13	WMT'13	9824476	1512	1701
de→fr	WMT'19	WMT'13	WMT'13	9824476	1512	1701

Table 2: Data sources and number of samples for the parallel data in our corpus. Please note that we don't use parallel data in Fr-De for any of the experiments in the paper.

xx	cs	fr	ru	zh	es	fi	de	et	lv	lt	ro	hi	kk	tr	gu
Any-to-English (xx→en)	31.3	37.2	36.0	21.7	32.7	27.3	31.7	23.1	15.0	21.3	30.1	8.5	11.5	15.9	1.0
English-to-Any (en→xx)	23.8	41.3	26.4	31.3	31.1	18.1	29.9	18.2	14.2	11.5	23.4	4.5	1.9	13.6	0.6

Table 3: Bilingual baselines. xx refers to language in the column header. (Siddhant et al., 2020)



System		Routing Granularity		BLEU															
		AVG	xx2en	en2xx	HRL	LRL	cs_en	en_cs	fr_en	en_fr	ru_en	en_ru	zh_en	en_zh	es_en	en_es	de_fr	fr_de	
Multilingual Transformer-Base	-	20.03	23.69	17.5	23.25	15.88	27.2	18.1	34.1	36.1	31.7	21.1	18.9	17.2	31.3	29.2	17.4	5.5	
	-	23.84	26.10	22.03	27.69	18.89	31.03	23.24	37.75	40.43	35.2	25.09	20.02	25.99	33.45	32.27	20.07	20.98	
	Sentence	19.88	24.05	16.83	22.56	14.14	27.6	18.7	34.4	36.5	32.7	15.1	20.4	7.2	31.3	30.1	13.6	9.1	
	Token	22.58	24.91	20.35	27.49	16.28	29.8	21.8	36.4	40.1	34.6	25.7	19.9	23.7	33.9	32.8	23.9	19.9	
Task-level MoE – 32 experts	Language Pair	22.04	25.43	19.5	25.57	17.5	26.8	21.7	35.4	39.2	33	21	22.1	17.9	32.4	32.1	12.2	19.1	
	Target	22.88	25.63	20.19	27.21	17.3	29.1	21.7	36.1	40.2	33.8	24.7	21.9	24.8	32.6	33.1	25.8	18.8	
	Language Pair	22.45	25.58	20.34	26.85	16.79	30.3	21.5	36.7	40.3	34.8	25.1	21	25.9	33.6	32.4	12.9	16.6	
	Target	22.33	24.47	20.44	26.82	16.55	29.4	22	35.3	39.7	33.8	25.2	21	26.2	32.4	32.7	22.2	18.6	
	Language Pair	23.03	26.16	20.28	27.23	17.62	30.1	23.2	37.5	39.5	35.5	21.9	21.7	15.7	34.5	33.5	20.1	20.1	
	Token	23.62	25.95	21.09	28.48	17.37	30.5	22.5	37.1	39.9	35.4	25.6	21.4	27	34.3	33.5	27.7	22.4	

Table 4: Part 1 of the table with individual BLEU scores for Table1

System		Routing Granularity		BLEU																	
		fi.en	en.fi	de.en	en.de	et.en	en.et	lv.en	en.lv	lt.en	en.lt	ro.en	en.ro	hi.en	en.hi	kk.en	en.kk	tr.en	en.tr	gu.en	en.gu
Multilingual Transformer-Base	-	23.9	17	28.6	22	23.1	16.1	17.2	14.9	24.6	11.4	33.4	23.9	19.2	10.4	13.5	2.5	20.9	17.5	7.8	5.1
	-	27.89	20.83	30.72	27.37	28.49	17.59	20.32	17.76	26.1	26.1	35.84	26.83	20.87	14.61	10.4	5.23	22.69	19.44	10.68	7.67
	Sentence	Sentence 23.5	17.2	29.4	21.8	22	15.4	17.9	14.7	24.6	11.6	33.6	24.8	20.5	12.2	14	2.9	21.4	17.9	7.4	6.3
Token-level MoE – 32 experts	Token	27.3	20.2	31.2	26.7	27	19.9	18.7	17	23.7	13.9	33.7	26.5	19.8	11.5	8.5	2.4	20.3	18	8.8	5.1
	Language Pair	25.2	20.1	31.3	26.9	24.7	19.2	18.4	16.3	25.1	13.6	34.8	25.7	22.5	13.1	15	2.4	23.4	18.2	11.4	5.1
	Target	25.6	19.5	30.7	26.8	24.8	19.8	18.4	15.7	25.9	13.6	34.9	25.8	21.7	12.3	15.5	2.4	22.5	17.7	11	4.8
Task-level MoE – 32 experts	Language Pair	26.7	20	32.2	26.9	26.8	19.6	18.9	16.3	25.1	13.3	34.2	25.8	21.1	12.6	12.6	2.3	21.7	18.4	8	4.7
	Target	23.7	19.8	30.7	26.1	24.1	19.9	18	16.5	24.4	13.6	33.1	26.1	20	12.7	12.7	2.9	21.1	18.2	7.4	5
	Language Pair	27.8	21.1	32.3	27	27.6	21	19.8	17.2	26	14.6	36.4	26.8	20.4	14.2	12.3	3.3	21.5	19.4	9	5.8
Token	Token	27.9	20.5	32	27.1	27.3	20.5	19.4	17.6	25.9	14.4	36.2	26.6	20.1	13.3	11.6	3	21.2	19.2	9	5.7

Table 5: Part 2 of the table with individual BLEU scores for Table1

System	Routing Granularity		No. of Parameters					Effective n(params) at inference time		
	Encoder	Decoder	Vocabulary	Encoder	Decoder	Softmax	Total	Encoder	Decoder	Total
Multilingual Transformer-Base	-	-	33M	19M	25M	65M	142M	19M	25M	142M
Token-level MoE – 32 experts	Token	Token						214M	221M	533M
Sentence-level MoE – 32 expert	Sentence	Sentence						214M	221M	533M
	Language Pair	Language Pair						25M	32M	155M
	Target	Target						25M	32M	155M
Task-level MoE – 32 experts	Language Pair	Token	33M	214M	221M	65M	533M	214M	25M	338M
	Target	Token						214M	25M	338M
	Token	Language Pair						19M	221M	338M
	Token	Target						19M	221M	338M

Table 6: We break down the parameter counts of the models we compare in Section 4.2 by components.

System	Routing Granularity		No. of Parameters				Effective n(params) at inference time			
	Encoder	Decoder	Vocabulary	Encoder	Decoder	Softmax	Total	Encoder	Decoder	Total
Multilingual Transformer-Big	-	-		126M	151M		473M	126M	151M	473M
Token-level MoE – 128 experts	Token	Token	65M	6.5B	6.5B	131M	13B	6.5B	6.5B	13.3B
Task-level MoE – 128 experts	Token	Language		6.5B	6.5B		13B	6.5B	201M	6.9B
Task-level MoE – 128 experts	Token	Target		6.5B	6.5B		13B	6.5B	201M	6.9B

Table 7: We break down the parameter counts of the models we compare in Section 4.3.2 by components.

System	Routing Granularity		BLEU								
	Encoder	Decoder	AVG	En-X	X-En	High-25 (EnX)	Mid 52 (EnX)	Low 25 (Enx)	High-25 (XEn)	Mid 52 (XEn)	Low 25 (XEn)
Multilingual Transformer-Big	-	-	24.49	18.61	30.37	28.03	16.9	12.75	33.84	30.23	26.96
Token-level MoE – 128 experts	Token	Token	<b>28.37</b>	20.51	<b>36.26</b>	30.99	18.94	13.33	40.14	36.74	31.03
Task-level MoE – 128 experts	Token	Language	28.09	20.66	35.52	31.21	19.17	13.28	39.69	36.42	29.16
Task-level MoE – 128 experts	Token	Target	27.83	<b>20.76</b>	34.90	31.05	19.23	13.68	38.88	35.28	29.93

Table 8: We summarize the results in Figure 2 on scaled up 128 expert MoE models. Here, *High-25* means the average BLEU of the 25 highest resource languages, *Low-25* means the average BLEU of the 25 lowest resource languages while *Mid-52* is the average BLEU of the remaining 52 languages.