A COMPUTE RESOURCES USED

Here we provide details on the compute resources used in our experiments. All models were trained on 48GB A6000s, except for the Ensemble method, which was trained on 80GB A100s. The training time for each T5-GLUE experiment was approximately 108 hours while each ResNet-DomainNet experiment required approximately 11 hours of training.

B EXPERIMENT DETAILS

In this section, we provide details on the experimental setup and hyperparameter choices for the T5-GLUE and ResNet-DomainNet experiments described in the main text. We implemented Adamix Wang et al. (2022b) and ran with the hyperparameters listed in the below subsections. In Latent Skills Ponti et al. (2022), we use Adapters consistent with all other experiments and chose learning rate ratio of 10 for skill matrix, which we found to be best after sweeping for $\{1, 10, 100\}$ in both the settings. For exact implementation details of above methods, we refer the reader to each of the respective works.

It is also important to mention that SMEAR's memory footprint is comparable to that of other methods. Since each expert's size is moderate, we compute expert outputs by preparing an expert for each example and applying them to the examples in parallel. Thus, none of the methods need an extra weight-sized tensor. Training on T5-GLUE requires around 30GB of memory for SMEAR and other methods, with no significant differences observed in ResNet-DomainNet.

B.1 T5-GLUE

GLUE consists of nine datasets (SST-2 (Socher et al., 2013), CoLA (Warstadt et al., 2019)), MNLI (Williams et al., 2017), RTE (Bentivogli et al., 2009), QQP (Shankar et al., 2017), MRPC (Dolan & Brockett, 2005), STS-B (Cer et al., 2017), QNLI (Rajpurkar et al., 2016), and WNLI (Levesque et al., 2012)) that cover a wide range of natural language processing tasks. Following convention, we exclude WNLI and use the remaining eight datasets. We use the prompted form of these datasets available in PromptSource (Bach et al., 2022), which maps each example into a natural language request-and-response form. During training, we randomly select a prompt templates for each example; during evaluation, we evaluate each example using all of its dataset's templates. In the T5-GLUE experiments in this paper, we concatenated all 8 datasets of GLUE and perform multitask training. T5 models were trained for $600k$ steps using a learning rate of $3e^{-4}$, with $2k$ warmup steps, and batch size of 128. The AdamW optimizer was used with its default settings. We ran the ST-Gumbel estimator with a τ value of 10 and an anneal rate of $1e^{-6}$ by sweeping τ in the range of $\{1, 10\}$ and the anneal rate in the range of $\{1e^{-4}, 1e^{-5}, 1e^{-6}\}$. For the REINFORCE estimator in Equation 1, we used the same values as in (Clark et al., 2022), $\alpha = 1e^{-2}$, $\beta = 5e^{-4}$, and $\gamma = 1e^{-2}$. The adapters here use swish non-linearity in between.

B.2 RESNET-DOMAINNET

In the ResNet-DomainNet experiments, all the domains from DomainNet were concatenated to perform multitask training similar to T5-GLUE. ResNet models were trained for $100k$ steps with batch size of 128 and a learning rate of $1e^{-3}$, with no warm up, using Adam optimizer. We used τ value of 10 and anneal rate of $1e^{-4}$ for the ST-Gumbel estimator by sweeping τ in the range of $\{1, 10\}$ and the anneal rate in the range of $\{1e^{-4}, 1e^{-5}, 1e^{-6}\}$. The values of α, β, and γ for the REINFORCE estimators in Equation 1 are same as in T5-GLUE experiments. The hyperparameter that weighs entropy regularization in Dselect-k is chosen as 0.1 which worked best among $\{0.01, 0.1, 1\}$. The adapters also used a swish non-linearity in between.

C EXPERT DROPOUT

Table 1 illustrates the impact of expert dropout on methods that do not make use of metadata but learn adaptive routing, namely methods that are trained through gradient estimation and SMEAR. We conduct this ablation on a single seed to limit the amount of computation. The results show that

SMEAR benefits from an improvement of 2.9% on T5-GLUE, and Top-k achieves a 1.2% and 0.2% improvement on T5-GLUE and ResNet-DomainNet respectively. As a result, we include expert dropout for these two methods when discussed in the main text. However, expert dropout has a negative impact on the performance of ST-Gumbel and REINFORCE methods, and thus, we exclude it for these two methods in the main text.

Routing	T5-GLUE	ResNet-DomainNet
Top-k	77.2	59.8
w/ Expert dropout 0.1	78.4 (+1.2)	60.0 (+0.2)
ST-Gumbel	78.3	58.3
w/ Expert dropout 0.1	77.2 (-1.1)	57.9 (-0.4)
REINFORCE	79.8	59.8
w/ Expert dropout 0.1	78.2 (-1.6)	59.8 (+0.0)
SMEAR	80.2	62.0
w/ Expert dropout 0.1	83.1 (+2.9)	62.0 (+0.0)

Table 1: Performance comparision of different adaptive routing methods w and w/o dropout on a single seed. The results indicate that SMEAR and Top-k method benefit from the expert dropout, while ST-Gumbel and REINFORCE are negatively affected.

D LICENSE

T5 is licensed under Apache 2.0. The ResNet model we used is licensed under BSD 3-Clause License. QNLI uses CC BY-SA 4.0 license. MultiNLI uses data sources of multiple different licenses(Williams et al., 2017). CoLA, SST-2, RTE, MRPC, STS-B, QQP, and DomainNet allow non-commercial research use cases. Our code for T5-GLUE is based on HyperformerMahabadi et al. (2021), which is shared under Apache 2.0. The code for ResNet-DomainNet is developed by us.

E ETHICS STATEMENT

We are not aware of any negative ethical implications of our work. Our work does not involve human subjects and is primarily focused on diagnosing issues with an efficient class of neural networks. While conditional computation has been used to design extremely large neural networks (Shazeer et al., 2017; Fedus et al., 2021; Du et al., 2022) that have high computational costs (and, correspondingly, energy usage), our work primarily focuses on smaller-scale models.

F FULL RESULTS ON T5-GLUE AND RESNET-DOMAINNET

We show the full results of T5-GLUE in table 2 and ResNet-DomainNet in table 3.

G ROUTING DISTRIBUTION IN ALL ROUTING BLOCKS

Here we put the routing distribution in all routing blocks in both T5-GLUE and ResNet-DomainNet learnt by SMEAR in fig. 4, fig. 5, fig. 6, and fig. 7.

Routing	RTE acc	SST-2 acc	MRPC f1	MRPC acc	STS-B pearson	STS-B spearman	QQP f1	QQP acc	MNLI acc	QNLI acc	CoLA mcc	Average
SMEAR	$69.9_{2.6}$	$90.9_{0.8}$	$90.5_{1.5}$	$86.9_{2.2}$	$87.0_{0.7}$	$86.6_{0.8}$	$86.9_{0.3}$	$90.1_{0.2}$	$84.9_{0.5}$	$90.2_{0.6}$	$33.8_{6.4}$	$81.6_{1.0}$
$1\times$ parameters	$72.3_{2.1}$	$92.1_{0.5}$	$89.9_{0.5}$	$86.0_{0.8}$	$85.5_{0.8}$	$85.3_{0.9}$	$87.0_{0.3}$	$90.2_{0.2}$	$84.1_{0.5}$	$89.9_{0.7}$	$20.1_{8.1}$	$80.2_{0.8}$
$1\times$ compute	$67.3_{3.3}$	$91.9_{0.2}$	$89.2_{2.7}$	$85.5_{3.4}$	$87.4_{0.8}$	$87.3_{0.7}$	$85.6_{0.3}$	$89.2_{0.2}$	$84.5_{0.7}$	$89.9_{0.5}$	$5.1_{3.7}$	$78.4_{1.1}$
Adamix	$70.2_{3.2}$	$92.4_{0.6}$	$87.4_{1.4}$	$83.3_{1.2}$	$86.7_{0.6}$	$86.6_{0.7}$	$85.7_{0.2}$	$89.4_{0.1}$	$85.4_{0.3}$	$90.5_{0.4}$	$5.1_{1.1}$	$78.4_{0.4}$
Hash	$58.8_{2.7}$	$85.6_{1.3}$	$77.7_{2.6}$	$68.5_{3.2}$	$65.4_{2.4}$	$65.2_{1.8}$	$76.8_{0.2}$	$82.8_{0.3}$	$72.0_{0.9}$	$80.0_{0.5}$	$2.9_{2.5}$	$66.9_{0.9}$
Tag	$71.7_{2.9}$	$90.3_{0.5}$	$85.4_{0.7}$	$79.5_{0.8}$	$82.2_{1.1}$	$81.5_{1.2}$	$86.2_{0.4}$	$89.5_{0.3}$	$84.4_{0.8}$	$87.9_{0.9}$	$25.1_{8.3}$	$78.5_{1.2}$
Latent Skills	$70.4_{4.6}$	$90.8_{0.8}$	$90.0_{1.0}$	$85.8_{2.1}$	$86.6_{1.5}$	$86.3_{1.4}$	$86.4_{0.4}$	$89.8_{0.3}$	$84.9_{0.9}$	$89.3_{1.4}$	$30.8_{5.5}$	$81.0_{1.6}$
Top-k	$68.2_{2.3}$	$92.5_{0.4}$	$88.6_{0.7}$	$84.7_{1.7}$	$87.7_{1.6}$	$87.4_{1.7}$	$85.3_{0.4}$	$89.0_{0.5}$	$84.9_{0.9}$	$90.1_{0.9}$	$2.0_{2.0}$	$78.2_{0.9}$
ST-Gumbel	$67.6_{2.3}$	$92.1_{0.7}$	$88.8_{1.0}$	$84.8_{1.7}$	$86.9_{1.0}$	$86.8_{0.8}$	$85.7_{0.1}$	$89.2_{0.2}$	$84.5_{0.3}$	$89.1_{0.5}$	$1.3_{1.7}$	$77.9_{0.4}$
REINFORCE	$70.9_{3.3}$	$92.6_{0.5}$	$89.8_{1.6}$	$86.0_{2.1}$	$87.4_{0.6}$	$87.2_{0.5}$	$86.1_{0.3}$	$89.5_{0.2}$	$85.8_{0.4}$	$90.8_{0.7}$	$14.1_{6.9}$	$80.0_{0.8}$
Ensemble	$72.9_{1.6}$	$91.5_{0.4}$	$90.9_{1.4}$	$87.7_{1.4}$	$85.7_{1.5}$	$85.1_{1.6}$	$86.8_{0.3}$	$90.1_{0.2}$	$84.7_{0.5}$	$89.8_{0.6}$	$33.7_{6.1}$	$81.7_{1.0}$
SMEAR $2\times$	$70.9_{3.1}$	$90.9_{0.7}$	$89.5_{1.1}$	$85.8_{1.0}$	$86.9_{0.9}$	$86.5_{1.0}$	$86.8_{0.4}$	$90.1_{0.3}$	$84.4_{0.5}$	$89.6_{0.8}$	$33.1_{6.5}$	$81.3_{1.1}$

Table 2: Full T5-GLUE results.

Routing	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Final Accuracy
SMEAR	$64.2_{0.1}$	$31.2_{0.3}$	$57.8_{0.3}$	$62.3_{0.1}$	$74.3_{0.1}$	$56.0_{0.2}$	$62.0_{0.1}$
$1\times$ parameters	$63.3_{0.3}$	$29.8_{0.3}$	$56.4_{0.3}$	$61.5_{0.1}$	$72.9_{0.1}$	$54.9_{0.4}$	$60.8_{0.1}$
$1\times$ compute	$60.2_{0.3}$	$27.9_{0.3}$	$54.8_{0.1}$	$59.0_{0.2}$	$72.3_{0.1}$	$52.6_{0.2}$	$59.0_{0.1}$
Adamix	$58.9_{0.2}$	$27.0_{0.2}$	$54.1_{0.2}$	$57.2_{0.3}$	$72.1_{0.1}$	$51.2_{0.2}$	$58.0_{0.2}$
Hash	$53.5_{0.3}$	$23.4_{0.3}$	$49.8_{0.4}$	$48.6_{0.3}$	$68.5_{0.1}$	$45.7_{0.2}$	$52.4_{0.1}$
Tag	$62.8_{0.4}$	$30.2_{0.3}$	$58.0_{0.2}$	$61.7_{0.2}$	$74.1_{0.1}$	$55.1_{0.3}$	$61.4_{0.1}$
Latent Skills	$64.5_{0.4}$	$31.2_{0.4}$	$58.9_{0.1}$	$61.6_{0.3}$	$74.2_{0.1}$	$56.3_{0.2}$	$61.9_{0.2}$
Top-k	$61.6_{0.2}$	$29.6_{0.2}$	$55.8_{0.4}$	$60.2_{0.3}$	$73.0_{0.2}$	$53.5_{0.1}$	$60.0_{0.1}$
ST-Gumbel	$59.9_{0.3}$	$27.6_{0.4}$	$54.5_{0.3}$	$58.1_{0.5}$	$72.1_{0.2}$	$51.9_{0.4}$	$58.5_{0.2}$
REINFORCE	$61.3_{0.3}$	$29.1_{0.2}$	$55.9_{0.2}$	$60.4_{0.3}$	$72.8_{0.1}$	$53.6_{0.2}$	$60.0_{0.1}$
DSelect-k	$60.6_{0.2}$	$28.4_{0.3}$	$55.1_{0.3}$	$59.5_{0.3}$	$72.5_{0.2}$	$52.5_{0.4}$	$59.3_{0.2}$
Soft MoE	$62.7_{0.2}$	$29.3_{0.3}$	$56.5_{0.2}$	$60.3_{0.2}$	$73.3_{0.1}$	$54.9_{0.1}$	$60.5_{0.1}$
Ensemble	$65.7_{0.1}$	$32.3_{0.0}$	$58.5_{0.3}$	$63.7_{0.2}$	$74.6_{0.1}$	$57.6_{0.2}$	$62.9_{0.1}$
SMEAR $2\times$	$65.3_{0.1}$	$31.8_{0.1}$	$58.4_{0.5}$	$63.3_{0.3}$	$74.9_{0.2}$	$57.3_{0.1}$	$62.8_{0.1}$

Table 3: Full ResNet-DomainNet results.

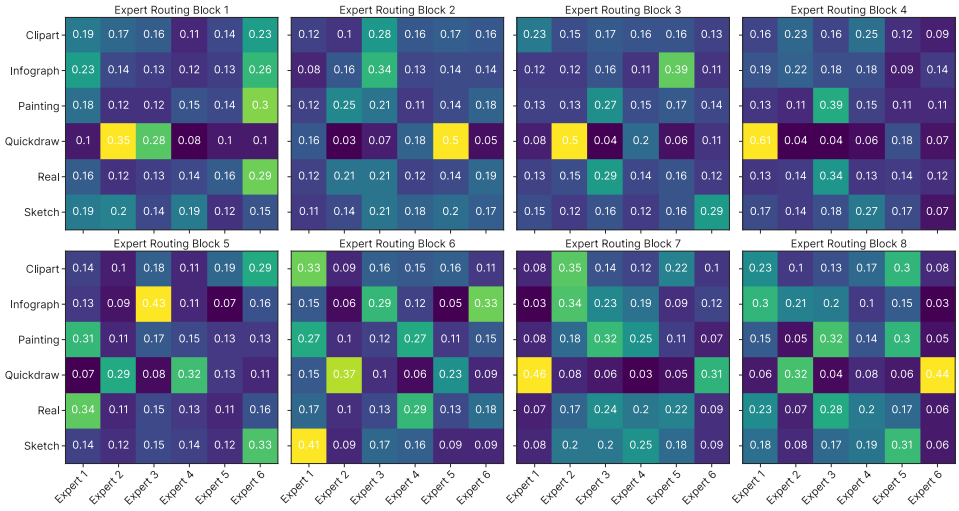

Figure 4: Routing distribution learnt by SMEAR in all routing blocks of ResNet-DomainNet

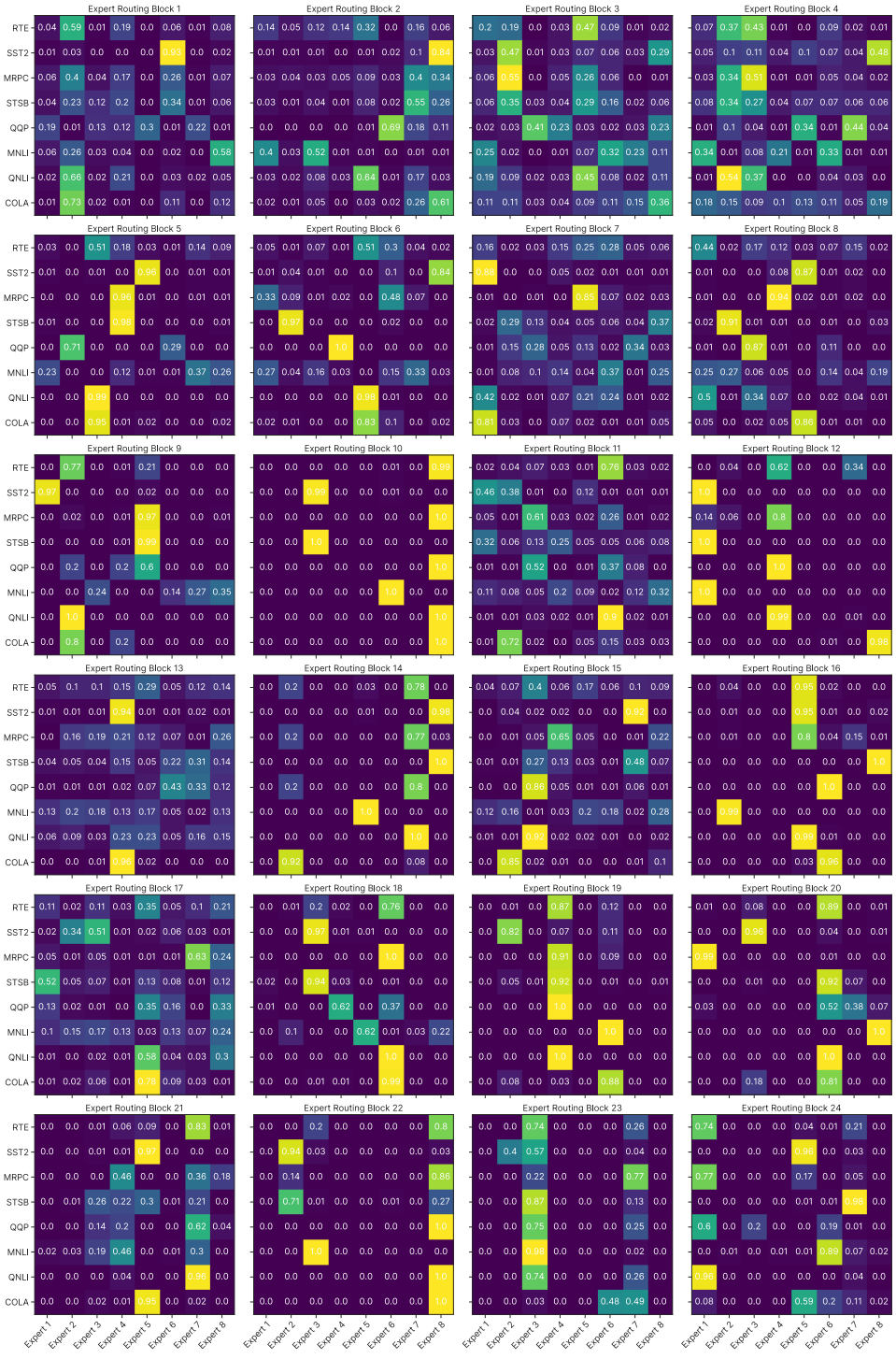

Figure 5: Routing distribution learnt by SMEAR in the encoder routing blocks (1-24) of T5-GLUE

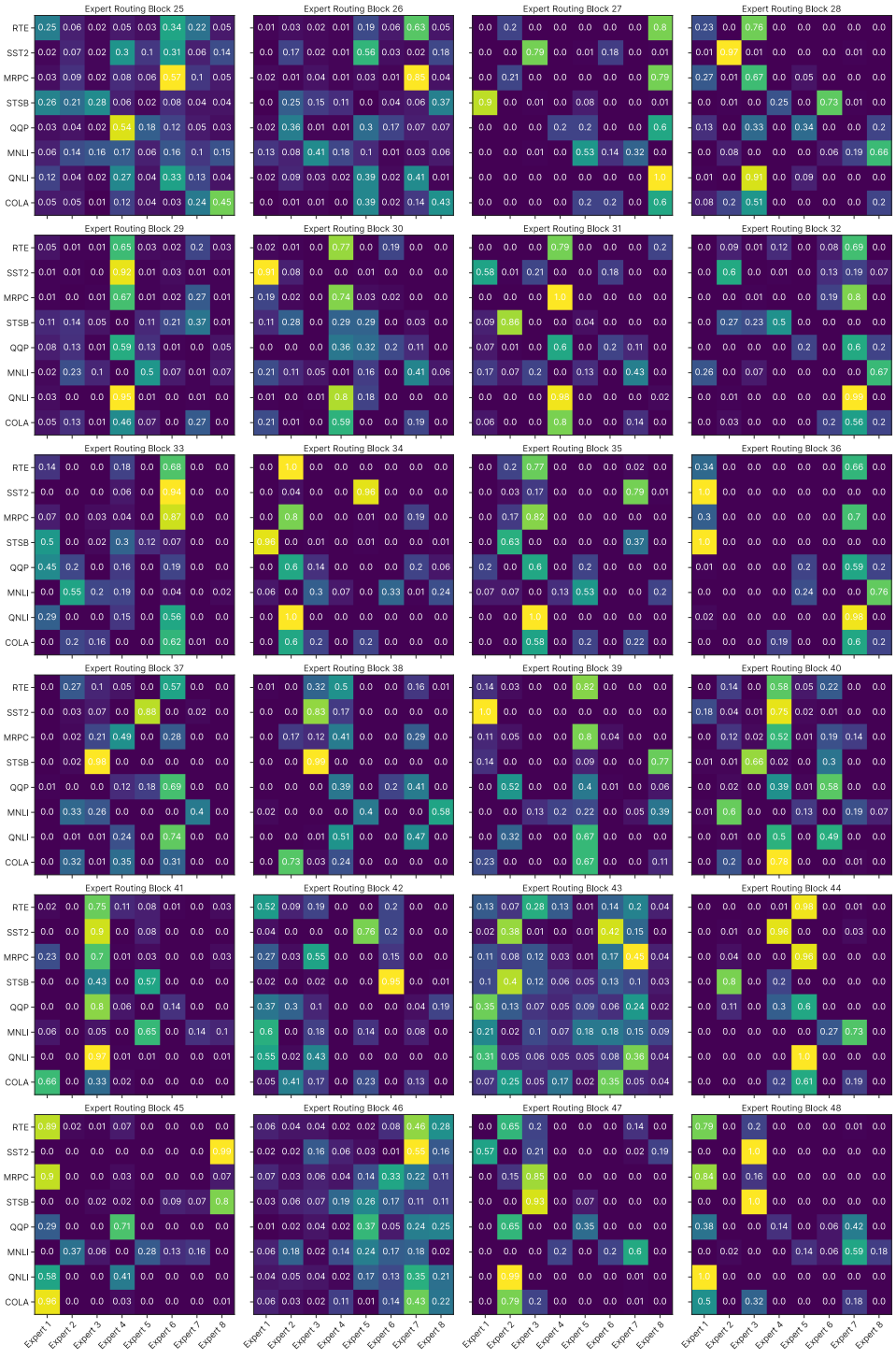

Figure 6: Routing distribution learnt by SMEAR in the decoder routing blocks (25-48) of T5-GLUE

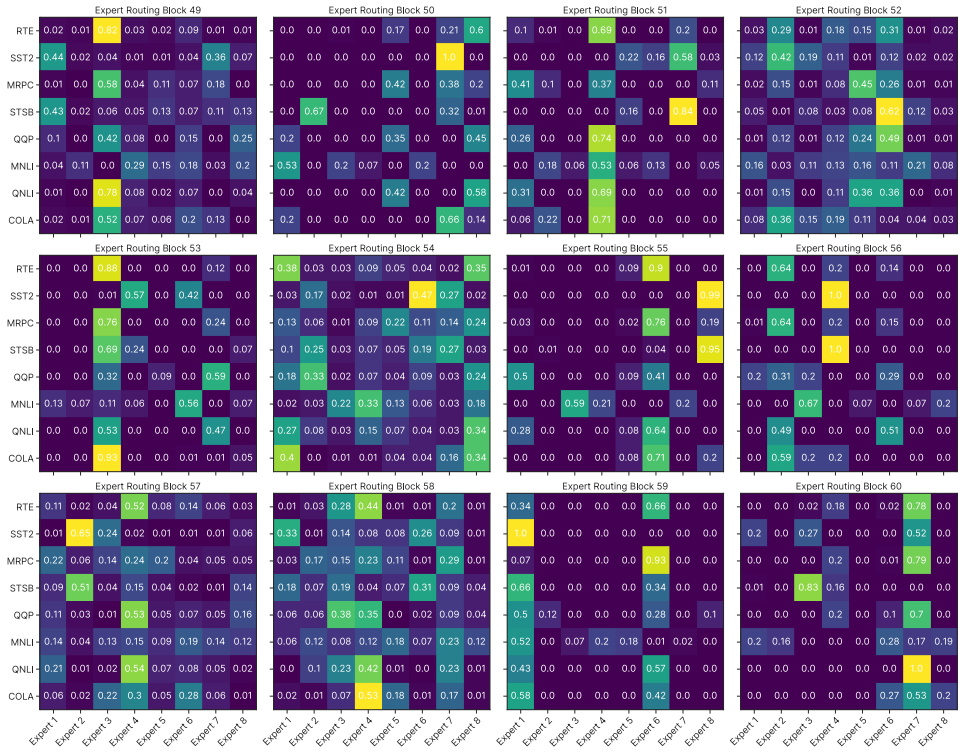

Figure 7: Routing distribution learnt by SMEAR in the decoder routing blocks (49-60) of T5-GLUE