A Scaling Law Study Details

A.1 Dataset

A concise view of evaluation datasets used for scaling laws (Section 4.2) is shown in Table. The ten OOD evaluation datsets span four categories (i) Web Domain (ii) News Domain (iii) Wikipedia (iv) Patents. They are either "source-original" or "target-original". There are two source-original and one target-original dataset in Web Domain, one source-original each in Wikipedia and Patents domain. We use publicly available WMT newstest2019 [Barrault et al., 2019] and WMT newstest2021 [Akhbardeh et al., 2021] for News Domain. Within this domain, we have five datasets: source-original, target-original, source-original-paraphrased [Freitag et al., 2020] and source-original-high-quality [Freitag et al., 2020] from WMT newstest2019 [Barrault et al., 2019], and wmt-reference-C from WMT newstest2021 [Akhbardeh et al., 2021].

Dataset Name	Domain	Type	Source
Train Subset	Web	mixed	In-house
Patents	Patents	mixed	In-house
Web domain 1	Web	source-original	In-house
Web domain 2	Web	source-original	In-house
Web domain 3	Web	target-original	In-house
Wikipedia	Wikipedia	source-original	In-house
wmt-high-quality	News	source-original	WMT newstest2019 [Freitag et al., 2020]
wmt-refC	News	source-original	WMT newstest2021 Ref-C [Akhbardeh et al., 2021]
wmt-paraphrased	News	source-original	WMT newstest2019 [Freitag et al., 2020]
wmt-src-orig	News	source-original	WMT newstest2019 [Barrault et al., 2019]
wmt-tgt-orig	News	target-original	WMT newstest2019 [Barrault et al., 2019]

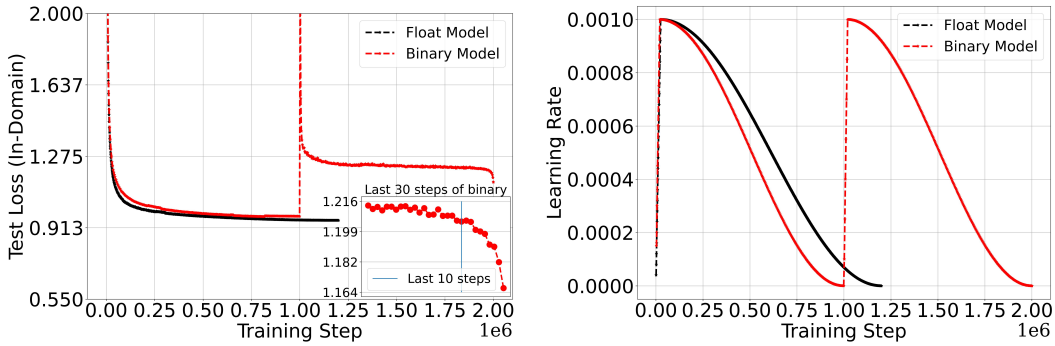
Table 2: Evaluation datasets used in Section 4.2.

A.2 Model & Training Details

All the models in Section 4.2 have an embedding dimension of 512, a hidden projection dimension of 2048, and 8 attention heads. The embedding parameters are shared on the source and the target side. The same embedding matrix (transposed) is also used for the linear readout (softmax) parameters on the decoder side. All models are trained with Adam optimizer [Kingma and Ba, 2014] and use cosine learning rate schedule. Due to the sufficiency in training data, we did not use label smoothing during training. In our experiments, enabling label smoothing resulted in poor development set performance across all the models. Training and Learning rate profiles of one model (6 encoder, 8 decoder layers) are shown in Figure 6. Float models are trained for 5 epochs, and binary models are trained for 9 epochs in two stages: float stage and a binarization stage. An independent but identical learning rate schedules are used (with warmup) in both the stages of the binary model training. We note that a significant amount of training (i.e. loss reduction) for binary models happens in the final 10 steps when the learning rate is extremely small. Raw values of last 15 steps of learning rates are [5.0e-7, 3.1e-7, 1.6e-7, 6.4e-7, 1.0e-8, {2.5e-15}x10]. We also tune binary models with a constant learning rate of values in {1e-8, 1e-11, 1e-15} for the last epoch (overriding the original schedule), however we observe degradation in the quality (loss plateaus). This phenomenon of significant learning in the final stages of binary models' training at extremely small learning rates is also observed by Liu et al. [2021], Zhang et al. [2022]. We leave further investigation of this behavior to future work.

A.3 Scaling Law Fit

Scaling law fit on all ten OOD evaluation datasets is shown in Figure 7. The slopes p_e and p_d are shown in Figure 8 and Table 3.



(a) Loss

(b) Learning Rate

Figure 6: Test loss and learning rate profiles of a 6L8L float and binary model as the training progresses.

Dataset	Float models		Binary models	
	p_e	p_d	p_e	p_d
Train Subset	0.18	0.31	0.16	0.28
Patents	0.20	0.30	0.19	0.32
Web Domain 1	0.14	0.25	0.14	0.27
Web Domain 2	0.19	0.37	0.16	0.30
Web Domain 3	0.12	0.18	0.14	0.23
Wikipedia	0.13	0.25	0.12	0.25
wmt-high-quality	0.20	0.31	0.18	0.30
wmt-refC	0.24	0.34	0.17	0.27
wmt-paraphrased	0.14	0.36	0.12	0.31
wmt-src-orig	0.22	0.37	0.23	0.36
wmt-tgt-orig	0.15	0.22	0.12	0.20

Table 3: Tabular representation of the same data (p_e & p_d) as shown in Figure 8.

B Generation Quality

Generation quality for decoder-scaling models is shown in Figure 9. We observe similar behavior as seen for encoder-scaling models in Section 4.3. BLEU scores for binary models are 2-3 BLEU points worse than the respective float models at the same model depth. MBR-BLEURT based decoding quality increases consistently by increasing the sample size.

B.1 Translation Samples

We generate several En-De translation samples as follows from both float and binary 6L10L encoder-decoder models. Inputs are taken from the WMT2017 dataset.

Example 1

- **Source:** The notice for the Nottingham East Labour meeting on Friday stated that "we want the meetings to be inclusive and productive."
- **Reference:** In der Mitteilung für das East Labour in Nottingham Treffen am Freitag heißt es: "Wir wollen, dass die Treffen integrativ und produktiv sind".
- **Float output:** In der Mitteilung für das Nottingham East Labour-Treffen am Freitag heißt es: "Wir wollen, dass die Treffen inklusive und produktiv sind."
- **Binary output:** Die Bekanntmachung für das Nottingham East Labour-Treffen am Freitag erklärte: "Wir wollen, dass die Sitzungen inklusive und produktiv sind."

Example 2

- **Source:** The Government of Wales Act 2017 gave the Welsh assembly the power to change its name.
- **Reference:** Mit dem Government of Wales Act 2017 erhielt das walisische Parlament die Möglichkeit, seinen Namen zu ändern.
- **Float output:** Der Government of Wales Act 2017 gab der walisischen Versammlung die Befugnis, ihren Namen zu ändern.
- **Binary output:** Das Government of Wales Act 2017 gab der walisischen Versammlung die Befugnis, ihren Namen zu ändern.

Example 3

- **Source:** Residents were seen returning to their destroyed homes, picking through waterlogged belongings, trying to salvage anything they could find.
- **Reference:** Anwohner wurden dabei beobachtet, wie sie in ihre zerstörten Häuser zurückkehrten, völlig durchnässte persönliche Gegenstände mitnahmen und versuchten zu retten, was zu retten ist.
- **Float output:** Die Bewohner wurden gesehen, wie sie in ihre zerstörten Häuser zurückkehrten, durch verstopfte Habseligkeiten pflückten und versuchten, alles zu retten, was sie finden konnten.
- **Binary output:** Die Bewohner wurden gesehen, wie sie in ihre zerstörten Häuser zurückkehrten, indem sie mit Wasser gefüllte Gegenstände pflückten und versuchten, alles zu retten, was sie finden konnten.

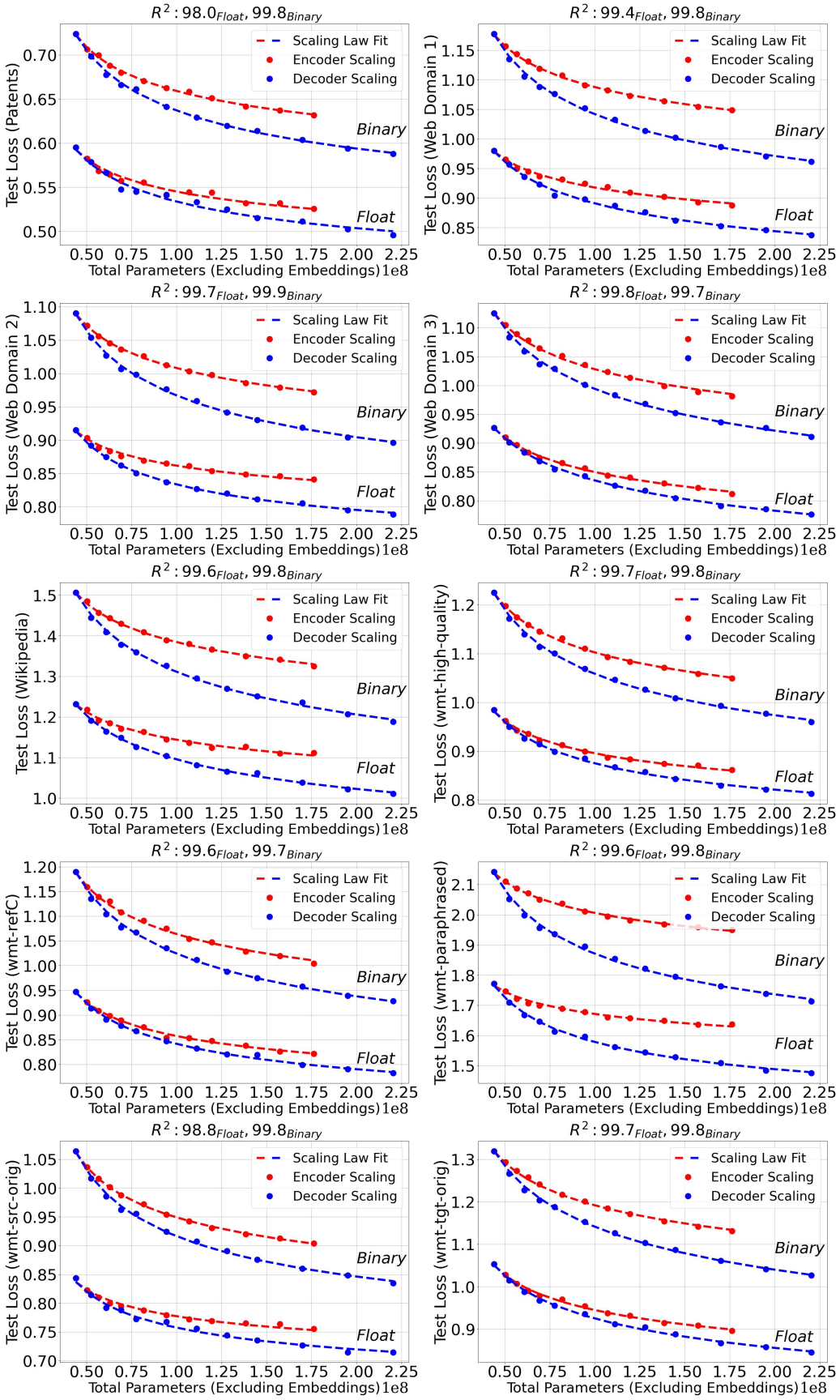


Figure 7: Scaling law studies on evaluation datasets defined in Section 4.2

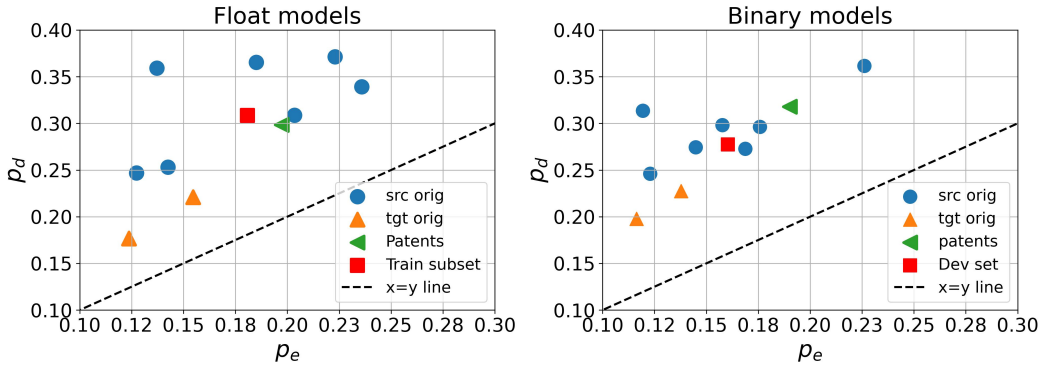
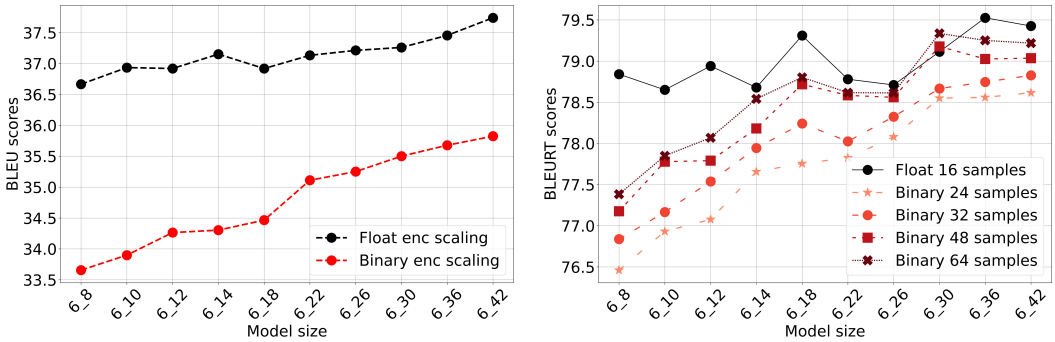


Figure 8: Encoder and Decoder scaling slopes (i.e. p_e & p_d) as per the scaling law defined in Section 4.2. Raw values are shown in Table 3.



(a) Beam Search BLEU

(b) MBR-BLEURT

Figure 9: Comparison on translation qualities between binarized and bfloat16 models for decoder-scaling.