A Training Algorithm

Algorithm 1 Simplified description of the training algorithm.

```
Load 4-bit Quantized Base Model
Load Training Hierarchy
for Level in Training Hierarchy do
    for Group in Level do
        Load data of all languages in Group
        if Level not 0 then
            Load relevant LoRAs from previous Levels
            Merge LoRAs with model
        end if
        Add new LoRA
        Train LoRA on relevant data
        Save current Group LoRA
    end for
end for
```

B Training Hyper Parameters

Table 5 shows all the hyperparameters used while training PTHQL.

Hyperparameter	Value
Dataset	AMR3.0 + Flores200
Max sequence length	256 tokens
Batch Size	8
Unique Instances per Language	30 997
Total Unique Instances	371 964
Epochs per Level (L)	1
Instances (L0)	371 964
Instances (L1)	185 982
Instances (L2)	61 994
Instances (L3)	30 997
Total Seen Instances	650 937
Real Total Instances	1 487 856
Checkpoints	Every 500 batches
Optimizer	Adafactor
Scheduler	Linear
Learning Rate	5e-5
Base Model	google/mt5-large
Base Model Parameters	1.2B
LoRA Parameters	293M
LoRA Rank	256
LoRA Alpha	256
LoRA Dropout	0.05
LoRA Scailing	Rank-Stabilized
LoRA Targets	All linear layer
Quantization	BnB 4-bit

Table 5: Hyper Parameters of the HQL training experiments.

C Other Metrics

C.1 LDC2020T07

Model	DEU	ENG	SPA	ITA
Ribeiro	49.4	—	57.2	54.0
Martinez	**55.8**	**73.4**	**64.0**	**60.7**
MonoQL	45.4	71.1	61.8	50.9
MultiQL	47.7	69.7	60.1	54.7
Gen&Trans*	54.9	71.1	63.7	59.6
DLHQL	49.0	70.5	62.3	56.9
PTHQL	50.6	70.1	62.3	57.1

Table 6: ChrF++ score on LDC2020T07 test data. *English Gen&Trans is simply the result of MonoQL.

Model	DEU	ENG	SPA	ITA
MonoQL	60.7	78.1	68.1	63.1
MultiQL	57.7	77.6	66.6	65.9
Gen&Trans*	**69.4**	78.1	**71.9**	**73.4**
DLHQL	62.3	78.9	71.3	71.0
PTHQL	63.7	**79.2**	70.7	71.4

Table 7: BLEURT-20 score on LDC2020T07 test data. *English Gen&Trans is simply the result of MonoQL.

C.2 FLORES-200

Model	DEU	LTZ	ENG	TPI	NLD	LIM	SPA	AST	ITA	SCN	FRA	HAT
MonoQL	39.2	37.1	58.5	**38.4**	36.8	30.8	37.1	34.3	37.0	32.5	42.2	37.6
MultiQL	39.8	37.0	60.7	36.6	38.4	30.2	39.1	32.8	37.9	32.8	42.3	37.8
Gen&Trans*	**44.0**	**39.5**	58.5	35.0	**41.8**	31.9	**41.3**	**47.4**	**42.3**	30.5	**49.2**	39.5
DLHQL	42.7	39.1	**64.4**	35.7	40.8	**32.2**	**41.3**	37.0	41.0	**35.9**	47.1	**39.9**
PTHQL	43.1	39.4	64.4	35.7	41.5	32.0	40.9	37.6	40.6	35.7	47.6	**39.7**

Table 8: ChrF++ score on our sub set of FLORES-200 test data. *English Gen&Trans is simply the result of MonoQL.

Model	DEU	LTZ	ENG	TPI†	NLD	LIM†	SPA	AST†	ITA	SCN†	FRA	HAT
MonoQL	57.8	34.9	68.5	59.3	63.2	33.7	47.5	28.9	50.4	16.4	40.3	45.7
MultiQL	52.4	36.0	70.8	60.0	62.2	34.6	51.8	30.1	51.6	17.7	40.5	45.8
Gen&Trans	**64.9**	**43.2**	68.5*	59.7	**64.7**	37.6	**59.3**	**38.2**	**63.2**	21.8	**56.8**	49.6
DLHQL	61.2	42.0	**74.5**	**60.6**	55.5	37.6	58.5	37.4	60.7	**22.3**	53.3	51.4
PTHQL	61.4	42.2	74.3	59.9	53.5	**38.0**	58.8	38.1	61.1	22.2	54.3	**51.4**

Table 9: BLEURT-20 score on our sub set of FLORES-200 test data. *English Gen&Trans is simply the result of MonoQL. †Languages not included in BLEURT pretraininig.

D Statistical Significance

For BLEU and ChrF++ which are corpus level metrics we performed paired bootstrap resampling (Koehn, 2004) to evaluate the statistical significance of the different models. We used our proposed PTHQL as the baseline for the pairwise comparison. For BLEURT, which is a sentence level metric, we used the Wilcoxon signed-rank test (Wilcoxon, 1945), since the scores of the test set do not follow a normal distribution.

D.1 LDC2020T07

Model	DEU	ENG	SPA	ITA
MonoQL	**0.00**	**0.00**	**0.01**	**0.00**
MultiQL	**0.00**	0.10	**0.00**	**0.00**
Gen&Trans	**0.00**	**0.00**	**0.00**	**0.00**
DLHQL	**0.00**	**0.02**	0.20	**0.05**

Table 10: P-value of the paired bootstrap resampling on the BLEU score on LDC2020T07 test data when compared to PTHQL. On bold the cases where $p \leq 0.05$.

Model	DEU	ENG	SPA	ITA
MonoQL	**0.00**	**0.02**	0.08	**0.00**
MultiQL	**0.00**	**0.05**	**0.00**	**0.00**
Gen&Trans	**0.00**	**0.02**	**0.00**	**0.00**
DLHQL	**0.00**	**0.04**	0.36	0.11

Table 11: P-value of the paired bootstrap resampling on the ChrF++ score on LDC2020T07 test data when compared to PTHQL. On bold the cases where $p \leq 0.05$.

Model	DEU	ENG	SPA	ITA
MonoQL	**0.00**	**0.01**	**0.00**	**0.00**
MultiQL	**0.00**	**0.00**	**0.00**	**0.00**
Gen&Trans	**0.00**	**0.01**	**0.01**	**0.00**
DLHQL	**0.00**	0.58	0.19	0.31

Table 12: P-value of the paired bootstrap resampling on the BLEURT-20 score on AMR3.0 test data when compared to PTHQL. On bold the cases where $p \leq 0.05$.

D.2 FLORES-200

Model	DEU	LTZ	ENG	TPI	NLD	LIM	SPA	AST	ITA	SCN	FRA	HAT
MonoQL	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.12	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
MultiQL	**0.00**	**0.00**	**0.00**	0.07	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gen&Trans	**0.01**	**0.05**	**0.00**	0.08	0.12	0.20	**0.05**	0.28	**0.03**	**0.00**	**0.00**	**0.02**
DLHQL	**0.03**	**0.04**	0.17	0.25	0.27	0.34	0.08	0.23	0.38	0.16	0.20	0.22

Table 13: P-value of the paired bootstrap resampling on the BLEU score on our sub set of FLORES-200 test data when compared to PTHQL. On bold the cases where $p \leq 0.05$.

Model	DEU	LTZ	ENG	TPI	NLD	LIM	SPA	AST	ITA	SCN	FRA	HAT
MonoQL	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
MultiQL	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gen&Trans	**0.03**	0.29	**0.00**	0.07	0.21	0.28	0.13	0.34	**0.00**	**0.00**	**0.00**	0.25
DLHQL	0.06	0.13	0.33	0.41	**0.01**	0.18	**0.04**	0.10	0.09	0.15	0.06	0.18

Table 14: P-value of the paired bootstrap resampling on the ChrF++ score on our sub set of FLORES-200 test data when compared to PTHQL. On bold the cases where $p \leq 0.05$.

Model	DEU	LTZ	ENG	TPI	NLD	LIM	SPA	AST	ITA	SCN	FRA	HAT
MonoQL	**0.00**	**0.00**	**0.00**	0.17	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
MultiQL	**0.00**	**0.00**	**0.00**	0.30	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gen&Trans	**0.00**	0.19	**0.00**	0.25	**0.01**	0.17	0.78	0.95	**0.00**	0.51	**0.00**	**0.00**
DLHQL	0.44	0.60	0.16	0.12	**0.01**	0.31	0.28	0.22	0.44	0.37	**0.04**	0.31

Table 15: P-value of the paired bootstrap resampling on the BLEURT-20 score on our sub set of FLORES-200 test data when compared to PTHQL. On bold the cases where $p \leq 0.05$.