SWE HW7

1. SETUP

vocab_size: 81

data_len: 99423

Train/val split: 90% / 10%

batch_size: 16

Optimizer: AdamW, lr=1e-3

Evaluation: average validation loss over 50 batches every 100 steps.

Training budget: 2000 steps for every configuration (same evaluation schedule).

2. CONFIGURATIONS

Only **n_layers**, **d_model**, and **block_size** changed. Everything else stayed the same.

Config	n_layers	d_model	block_size	Final val loss (step 2000)	Best val loss (step)
Config A	4	128	64	1.622	1.622 (step 2000)
Config B	6	128	64	1.701	1.636 (step 1600)
Config C	4	192	64	1.690	1.631 (step 1200)
Config D	4	128	128	1.715	1.616 (step 1400)

3. RESULTS AND ANALYSIS

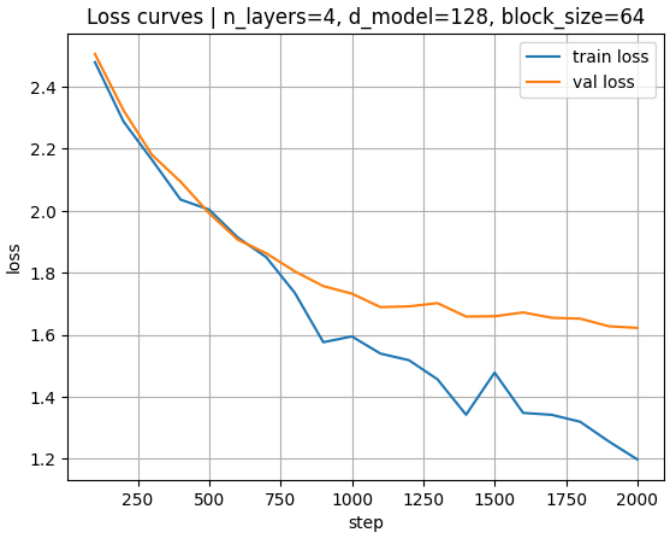
With a fixed 2000-step budget, Config A achieved the lowest validation loss (1.622). The lowest validation loss at any point across runs was Config D (1.616 at step 1400), but it overfit later and ended worse by step 2000.

Across all runs, training loss kept dropping while validation loss flattened and sometimes rose. That gap is a sign of overfitting at this training length. Early stopping would likely help, especially for the larger models.

3.1 CONFIG A (N_LAYERS=4, D_MODEL=128, BLOCK_SIZE=64)

Final losses at step 2000: train 1.198, val 1.622. Best validation loss: 1.622 at step 2000.

This run was the most stable overall. Validation loss kept improving slowly all the way to the end, and the curve did not drift upward. The generation sample still has a lot of misspellings, but it stays on topic more consistently than the other runs.



Sample generation: Rome was built rack his out an architecture charactering and dixircus upper Latin literary censusses individual briling. Althought military preceded system, their differing herist Afr, contain life, to had betwer their to geownered own clappor tradition (indepted from 2 Byzanting inscripted until stributions, the…

3.2 CONFIG B (N_LAYERS=6, D_MODEL=128, BLOCK_SIZE=64)

Final losses at step 2000: **train** 1.032, **val** 1.701. Best validation loss: 1.636 at step 1600.

Increasing depth lowered training loss faster, but validation loss stopped improving after the mid-run. After the best point, validation loss started to creep up while training kept improving. That looks like overfitting at 2000 steps with this depth.

Sample generation: Rome was built the Imperial mess to Jupiter has been Italian was libreteral provinced and at public games. The moss fulres Latin whal bessed monument model of the mos a mictuonism, a marriegate or routes. Sourned family and commancified with population (as emperor's to anstone and obs. The militarchy of Augustabl...

3.3 CONFIG C (N_LAYERS=4, D_MODEL=192, BLOCK_SIZE=64)

Final losses at step 2000: train 1.036, val 1.690. Best validation loss: 1.631 at step 1200.

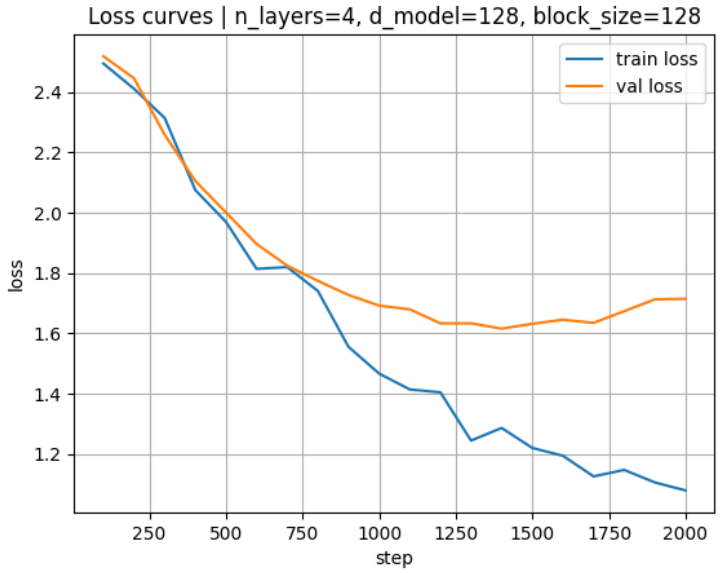
Increasing width also lowered training loss, and the best validation loss happened earlier than the end. Validation loss became less stable later in training. This model likely needs either stronger regularization or fewer steps to avoid overfitting.

Sample generation: Rome was built classical significance. Generally such as play and leas. Diocletian orate who had social or classinated no ose warlords—led stone was pirately state was not began to only dining and under Aurelium, but women seated at social enditiers and militaratar, massed no stracte. The cavalred by the Cruption…

3.4 CONFIG D (N_LAYERS=4, D_MODEL=128, BLOCK_SIZE=128)

Final losses at step 2000: train 1.079, val 1.715. Best validation loss: 1.616 at step 1400.

Increasing block_size gave the best validation loss during training, which suggests longer context helps on this data. However, the validation curve moved upward after the best point. With early stopping around step 1400, this would be the strongest option.



Sample generation: Rome was built circulate subsce oritions of the labsienden until the valuous shortud. The architecturing or refinding of military polive (in call be consider to provided state in most staged to (collegia to constitution (sauner) with the race daily stast equestrian order of the floors shad centralized to and impe…

4. CONCLUSION

Best final validation loss (step 2000): Config A with val loss 1.622.
Best validation loss during training: Config D with val loss 1.616 at step 1400.