

StaticPruning

November 26, 2025

```
[1]: import torch
from transformers import AutoModelForCausalLM, AutoTokenizer
from datasets import load_dataset
import pandas as pd
import random
import re
from tqdm import tqdm
import copy

# =====
# 1. Load Model
# =====
model_name = "Qwen/Qwen3-0.6B"

print("Loading model...")
tokenizer = AutoTokenizer.from_pretrained(model_name)
base_model = AutoModelForCausalLM.from_pretrained(
    model_name,
    torch_dtype=torch.float16,
    device_map="auto"
)

# =====
# 2. Load and prepare datasets (10%)
# =====
print("Loading datasets...")

mmlu = load_dataset("cais/mmlu", "all", split="test")
gsm8k = load_dataset("gsm8k", "main", split="test")

mmlu_10 = mmlu.shuffle(seed=42).select(range(int(0.1 * len(mmlu))))
gsm8k_10 = gsm8k.shuffle(seed=42).select(range(int(0.1 * len(gsm8k))))

# =====
# 3. MMLU Evaluation (MCQ)
# =====
def eval_mmlu(model, tokenizer, dataset):
```

```

correct = 0

for row in tqdm(dataset, desc="Evaluating MMLU"):
    question = row["question"]
    options = row["choices"]
    answer = row["answer"]

    prompt = "Answer the following multiple-choice question.\n\n"
    prompt += f"Question: {question}\n"
    for i, c in enumerate(options):
        prompt += f"{chr(65+i)}. {c}\n"
    prompt += "\nAnswer:\n"

    inputs = tokenizer(prompt, return_tensors="pt").to(model.device)

    output = model.generate(
        **inputs,
        max_new_tokens=2,
        do_sample=False
    )

    response = tokenizer.decode(output[0], skip_special_tokens=True)
    pred = response.strip()[-1:].upper()

    if pred == answer:
        correct += 1

return correct / len(dataset)

# =====
# 4. GSM8K Evaluation (open-ended math)
# =====

def extract_last_number(text):
    nums = re.findall(r"-?\d+\.\d*", text)
    return nums[-1] if nums else None

def eval_gsm8k(model, tokenizer, dataset):
    correct = 0

    for row in tqdm(dataset, desc="Evaluating GSM8K"):
        question = row["question"]
        gold_answer = row["answer"]

        prompt = (
            "Solve the following math word problem. "
            "Give only the final numeric answer.\n\n"
        )

```

```

        f"\n{question}\n\nAnswer:\n"
    )

inputs = tokenizer(prompt, return_tensors="pt").to(model.device)

output = model.generate(
    **inputs,
    max_new_tokens=200,
    do_sample=False
)

response = tokenizer.decode(output[0], skip_special_tokens=True)

pred = extract_last_number(response)
gold = extract_last_number(gold_answer)

if pred == gold:
    correct += 1

return correct / len(dataset)

# =====
# 5. STATIC PRUNING LOOP
# =====
layers = base_model.model.layers
num_layers = len(layers)

print(f"\nModel has {num_layers} transformer layers.\n")

results = []

for prune_layer in range(num_layers):
    print(f"\n===== TESTING WITH LAYER {prune_layer} REMOVED =====")

    # clone the original model each time
    model = AutoModelForCausalLM.from_pretrained(
        model_name,
        torch_dtype=torch.float16,
        device_map="auto"
    )

    # === REMOVE THE TARGET LAYER ===
    pruned_layers = torch.nn.ModuleList(
        [layer for i, layer in enumerate(model.model.layers) if i !=
         prune_layer]
    )

```

```

model.model.layers = pruned_layers
print(f"Layer {prune_layer} removed. New layer count: {len(model.model.
˓→layers)}")

# === Evaluate ===
mmlu_score = eval_mmlu(model, tokenizer, mmlu_10)
gsm8k_score = eval_gsm8k(model, tokenizer, gsm8k_10)

print(f"Layer {prune_layer}: MMLU={mmlu_score:.4f}, GSM8K={gsm8k_score:.
˓→4f}")

results.append({
    "removed_layer": prune_layer,
    "mmlu_accuracy": mmlu_score,
    "gsm8k_accuracy": gsm8k_score
})

# =====
# 6. Save results
# =====
df = pd.DataFrame(results)
df.to_csv("layer_pruning_results.csv", index=False)

print("\nSaved layer pruning results to: layer_pruning_results.csv")

```

Loading model...

```

tokenizer_config.json: 0.00B [00:00, ?B/s]
vocab.json: 0.00B [00:00, ?B/s]
merges.txt: 0.00B [00:00, ?B/s]
tokenizer.json: 0% | 0.00/11.4M [00:00<?, ?B/s]
config.json: 0% | 0.00/726 [00:00<?, ?B/s]
model.safetensors: 0% | 0.00/1.50G [00:00<?, ?B/s]
generation_config.json: 0% | 0.00/239 [00:00<?, ?B/s]

Loading datasets...
README.md: 0.00B [00:00, ?B/s]
dataset_infos.json: 0.00B [00:00, ?B/s]
all/test-00000-of-00001.parquet: 0% | 0.00/3.50M [00:00<?, ?B/s]
all/validation-00000-of-00001.parquet: 0% | 0.00/408k [00:00<?, ?B/s]
all/dev-00000-of-00001.parquet: 0% | 0.00/76.5k [00:00<?, ?B/s]

```

```

all/auxiliary_train-00000-of-00001.parquet(...): 0% | 0.00/47.5M [00:00<?
˓→, ?B/s]

Generating test split: 0% | 0/14042 [00:00<?, ? examples/s]

Generating validation split: 0% | 0/1531 [00:00<?, ? examples/s]

Generating dev split: 0% | 0/285 [00:00<?, ? examples/s]

Generating auxiliary_train split: 0% | 0/99842 [00:00<?, ? examples/
˓→s]

README.md: 0.00B [00:00, ?B/s]

main/train-00000-of-00001.parquet: 0% | 0.00/2.31M [00:00<?, ?B/s]

main/test-00000-of-00001.parquet: 0% | 0.00/419k [00:00<?, ?B/s]

Generating train split: 0% | 0/7473 [00:00<?, ? examples/s]

Generating test split: 0% | 0/1319 [00:00<?, ? examples/s]

```

Model has 28 transformer layers.

```

===== TESTING WITH LAYER 0 REMOVED =====
Layer 0 removed. New layer count: 27

Evaluating MMLU: 0% | 0/1404 [00:00<?,
?it/s]/root/miniconda3/envs/py3.10/lib/python3.10/site-
packages/transformers/generation/configuration_utils.py:631: UserWarning:
`do_sample` is set to `False`. However, `temperature` is set to `0.6` -- this
flag is only used in sample-based generation modes. You should set
`do_sample=True` or unset `temperature`.

    warnings.warn(
/root/miniconda3/envs/py3.10/lib/python3.10/site-
packages/transformers/generation/configuration_utils.py:636: UserWarning:
`do_sample` is set to `False`. However, `top_p` is set to `0.95` -- this flag is
only used in sample-based generation modes. You should set `do_sample=True` or
unset `top_p`.

    warnings.warn(
/root/miniconda3/envs/py3.10/lib/python3.10/site-
packages/transformers/generation/configuration_utils.py:653: UserWarning:
`do_sample` is set to `False`. However, `top_k` is set to `20` -- this flag is
only used in sample-based generation modes. You should set `do_sample=True` or
unset `top_k`.

    warnings.warn(
Evaluating MMLU: 100% | 1404/1404 [01:39<00:00, 14.10it/s]
Evaluating GSM8K: 100% | 131/131 [12:33<00:00, 5.75s/it]

Layer 0: MMLU=0.0000, GSM8K=0.0000

```

===== TESTING WITH LAYER 1 REMOVED =====

Layer 1 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:38<00:00, 14.27it/s]

Evaluating GSM8K: 100% | 131/131 [14:18<00:00, 6.56s/it]

Layer 1: MMLU=0.0000, GSM8K=0.0076

===== TESTING WITH LAYER 2 REMOVED =====

Layer 2 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.54it/s]

Evaluating GSM8K: 100% | 131/131 [13:41<00:00, 6.27s/it]

Layer 2: MMLU=0.0000, GSM8K=0.0076

===== TESTING WITH LAYER 3 REMOVED =====

Layer 3 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:38<00:00, 14.24it/s]

Evaluating GSM8K: 100% | 131/131 [14:33<00:00, 6.67s/it]

Layer 3: MMLU=0.0000, GSM8K=0.0763

===== TESTING WITH LAYER 4 REMOVED =====

Layer 4 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:39<00:00, 14.08it/s]

Evaluating GSM8K: 100% | 131/131 [14:34<00:00, 6.68s/it]

Layer 4: MMLU=0.0000, GSM8K=0.0534

===== TESTING WITH LAYER 5 REMOVED =====

Layer 5 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:38<00:00, 14.24it/s]

Evaluating GSM8K: 100% | 131/131 [14:40<00:00, 6.72s/it]

Layer 5: MMLU=0.0000, GSM8K=0.1527

===== TESTING WITH LAYER 6 REMOVED =====

Layer 6 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:38<00:00, 14.30it/s]

Evaluating GSM8K: 100% | 131/131 [14:29<00:00, 6.64s/it]

Layer 6: MMLU=0.0000, GSM8K=0.0611

===== TESTING WITH LAYER 7 REMOVED =====

Layer 7 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:38<00:00, 14.30it/s]

Evaluating GSM8K: 100% | 131/131 [14:20<00:00, 6.57s/it]

Layer 7: MMLU=0.0000, GSM8K=0.0840

===== TESTING WITH LAYER 8 REMOVED =====

Layer 8 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.59it/s]
Evaluating GSM8K: 100% | 131/131 [14:21<00:00, 6.58s/it]

Layer 8: MMLU=0.0000, GSM8K=0.1145

===== TESTING WITH LAYER 9 REMOVED =====

Layer 9 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:38<00:00, 14.28it/s]
Evaluating GSM8K: 100% | 131/131 [14:17<00:00, 6.55s/it]

Layer 9: MMLU=0.0000, GSM8K=0.0992

===== TESTING WITH LAYER 10 REMOVED =====

Layer 10 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.60it/s]
Evaluating GSM8K: 100% | 131/131 [14:19<00:00, 6.56s/it]

Layer 10: MMLU=0.0000, GSM8K=0.1145

===== TESTING WITH LAYER 11 REMOVED =====

Layer 11 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:37<00:00, 14.45it/s]
Evaluating GSM8K: 100% | 131/131 [14:14<00:00, 6.52s/it]

Layer 11: MMLU=0.0000, GSM8K=0.0305

===== TESTING WITH LAYER 12 REMOVED =====

Layer 12 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:38<00:00, 14.31it/s]
Evaluating GSM8K: 100% | 131/131 [14:22<00:00, 6.59s/it]

Layer 12: MMLU=0.0000, GSM8K=0.0763

===== TESTING WITH LAYER 13 REMOVED =====

Layer 13 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:37<00:00, 14.33it/s]
Evaluating GSM8K: 100% | 131/131 [14:19<00:00, 6.56s/it]

Layer 13: MMLU=0.0000, GSM8K=0.1069

===== TESTING WITH LAYER 14 REMOVED =====

Layer 14 removed. New layer count: 27

```
Evaluating MMLU: 100% | 1404/1404 [01:37<00:00, 14.47it/s]
Evaluating GSM8K: 100% | 131/131 [14:08<00:00, 6.48s/it]

Layer 14: MMLU=0.0000, GSM8K=0.0534

===== TESTING WITH LAYER 15 REMOVED =====
Layer 15 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:37<00:00, 14.40it/s]
Evaluating GSM8K: 100% | 131/131 [14:19<00:00, 6.56s/it]

Layer 15: MMLU=0.0000, GSM8K=0.0534

===== TESTING WITH LAYER 16 REMOVED =====
Layer 16 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:39<00:00, 14.12it/s]
Evaluating GSM8K: 100% | 131/131 [14:26<00:00, 6.62s/it]

Layer 16: MMLU=0.0000, GSM8K=0.0611

===== TESTING WITH LAYER 17 REMOVED =====
Layer 17 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.48it/s]
Evaluating GSM8K: 100% | 131/131 [14:12<00:00, 6.50s/it]

Layer 17: MMLU=0.0000, GSM8K=0.0687

===== TESTING WITH LAYER 18 REMOVED =====
Layer 18 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.56it/s]
Evaluating GSM8K: 100% | 131/131 [14:20<00:00, 6.57s/it]

Layer 18: MMLU=0.0000, GSM8K=0.0534

===== TESTING WITH LAYER 19 REMOVED =====
Layer 19 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:39<00:00, 14.13it/s]
Evaluating GSM8K: 100% | 131/131 [14:27<00:00, 6.62s/it]

Layer 19: MMLU=0.0000, GSM8K=0.0076

===== TESTING WITH LAYER 20 REMOVED =====
Layer 20 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:38<00:00, 14.30it/s]
Evaluating GSM8K: 100% | 131/131 [14:19<00:00, 6.56s/it]

Layer 20: MMLU=0.0000, GSM8K=0.0153
```

===== TESTING WITH LAYER 21 REMOVED =====

Layer 21 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.48it/s]
Evaluating GSM8K: 100% | 131/131 [14:18<00:00, 6.55s/it]

Layer 21: MMLU=0.0000, GSM8K=0.0153

===== TESTING WITH LAYER 22 REMOVED =====

Layer 22 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:38<00:00, 14.25it/s]
Evaluating GSM8K: 100% | 131/131 [14:17<00:00, 6.55s/it]

Layer 22: MMLU=0.0000, GSM8K=0.0382

===== TESTING WITH LAYER 23 REMOVED =====

Layer 23 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:35<00:00, 14.67it/s]
Evaluating GSM8K: 100% | 131/131 [14:05<00:00, 6.46s/it]

Layer 23: MMLU=0.0000, GSM8K=0.0763

===== TESTING WITH LAYER 24 REMOVED =====

Layer 24 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.55it/s]
Evaluating GSM8K: 100% | 131/131 [14:18<00:00, 6.55s/it]

Layer 24: MMLU=0.0000, GSM8K=0.0840

===== TESTING WITH LAYER 25 REMOVED =====

Layer 25 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.51it/s]
Evaluating GSM8K: 100% | 131/131 [14:17<00:00, 6.54s/it]

Layer 25: MMLU=0.0000, GSM8K=0.0840

===== TESTING WITH LAYER 26 REMOVED =====

Layer 26 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.52it/s]
Evaluating GSM8K: 100% | 131/131 [14:12<00:00, 6.51s/it]

Layer 26: MMLU=0.0000, GSM8K=0.0763

===== TESTING WITH LAYER 27 REMOVED =====

Layer 27 removed. New layer count: 27

Evaluating MMLU: 100% | 1404/1404 [01:36<00:00, 14.54it/s]
Evaluating GSM8K: 100% | 131/131 [14:10<00:00, 6.49s/it]

Layer 27: MMLU=0.0000, GSM8K=0.0840

Saved layer pruning results to: layer_pruning_results.csv

[]: