

Model Summary

TaoNet-mini-A2 is a 0.5B local-first language model intended for text generation experiments, lightweight instruction following, and research on efficient custom architectures.

This release is organized as a standard Hugging Face model package, while keeping the underlying TaoNet implementation in the repository for transparent loading and export.

Model Details

Model Specifications

Specification	Value
Model name	TaoNet-mini-A2
Model type	Causal language model
Architecture	TaoNetForCausalLM
Vocabulary size	8,192
Hidden size	1,024
Number of layers	16
Number of attention heads	8
Head dimension	128
Latent KV dimension	768
Feed-forward dimension	3,072
Maximum sequence length	1,024 tokens
Dropout	0.02
Embedding type	Factorized embedding
Rope scale	40.0
Tokenizer	SentencePiece
Special tokens	<UNK>, <BOS>, <EOS>, <PAD>

Hardware

Hardware

- GPU: 1 x RTX 5090

Software

- Training framework: TaoTrain

Quick Start

```
import torch
from transformers import AutoModelForCausalLM, AutoTokenizer

MODEL_NAME = "TaoTern/TaoNet-mini-A2"

device = "cuda" if torch.cuda.is_available() else "cpu"
dtype = torch.bfloat16 if device == "cuda" else torch.float32

tokenizer = AutoTokenizer.from_pretrained(MODEL_NAME, trust_remote_code=True)
model = AutoModelForCausalLM.from_pretrained(
    MODEL_NAME,
    trust_remote_code=True,
    torch_dtype=dtype,
).to(device)

prompt = "Fruit is now expensive so we should"
inputs = tokenizer(prompt, return_tensors="pt").to(device)

with torch.inference_mode():
    output_ids = model.generate(
        **inputs,
        max_new_tokens=64,
        temperature=0.7,
        top_p=0.85,
        repetition_penalty=1.2,
        do_sample=True,
        pad_token_id=tokenizer.pad_token_id,
        eos_token_id=tokenizer.eos_token_id,
    )

completion = tokenizer.decode(
    output_ids[0][inputs["input_ids"].shape[1]:],
    skip_special_tokens=True,
)
print(completion)
```

Benchmarks

The following scores were reported for TaoNet-mini-A2:

Benchmark	Score
MMLU	0.2412
HellaSwag	0.3162
ARC-Easy	0.4331
ARC-Challenge	0.2560

Benchmark	Score
PIQA	0.6137
WinoGrande	0.5083

These numbers should be treated as a snapshot of the current checkpoint, not as a universal capability guarantee.

Limitations

- This is a relatively small model, so it will not match larger frontier models on broad reasoning or long-horizon planning
- It may hallucinate or produce incorrect answers, especially on ambiguous prompts or tasks that require deep domain knowledge
- Outputs can be sensitive to prompt wording and generation parameters
- The model is not intended for safety-critical, legal, medical, or high-stakes decision-making without human review
- The reported benchmark scores are limited to the tasks listed above and do not describe full real-world quality

Citation

If you use TaoNet-mini-A2 in your research or product work, please cite:

```
@software{taonet_mini_a2_2026,  
  title={TaoNet-mini-A2},  
  author={Felix Thian},  
  year={2026},  
  url={https://huggingface.co/TaoTern/TaoNet-mini-A2}  
}
```

License

This repository is released under the MIT License.

Acknowledgments

- Hugging Face Transformers for the model-loading interface
- SentencePiece for tokenizer support
- The TaoTrain export pipeline used to package the checkpoint