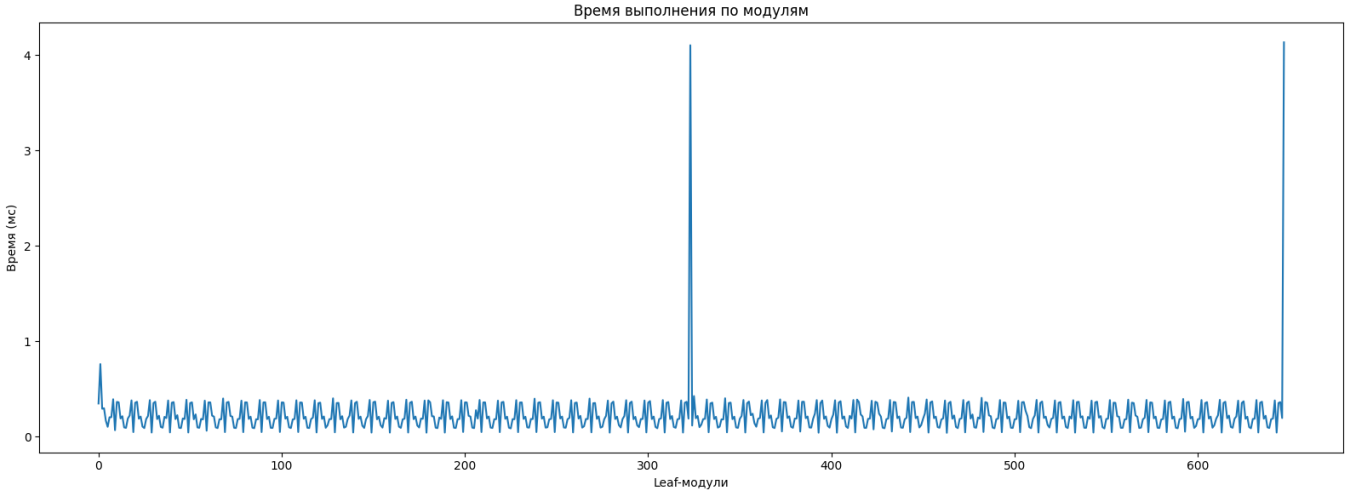
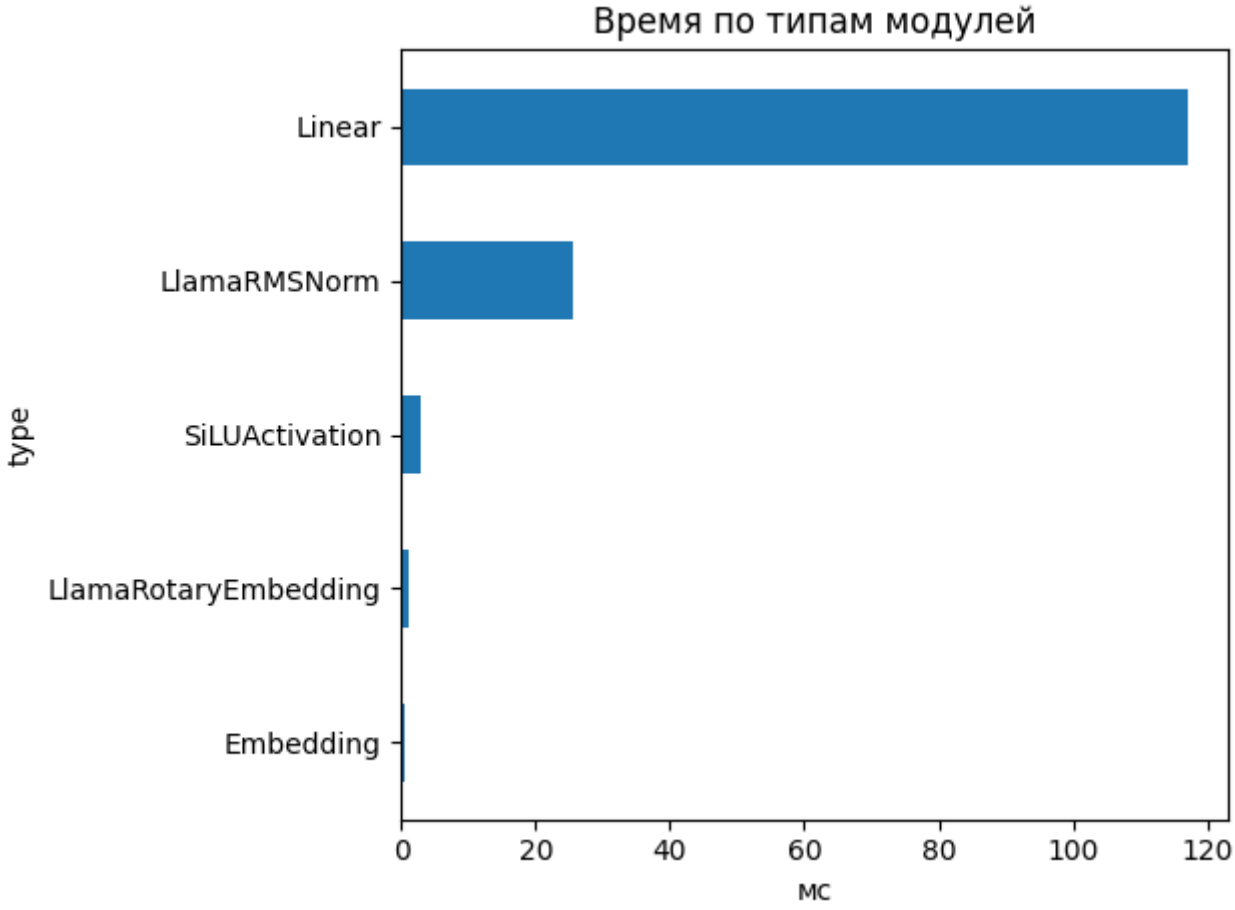
Llama-3.1-Nemotron-Nano-4B-v1.1

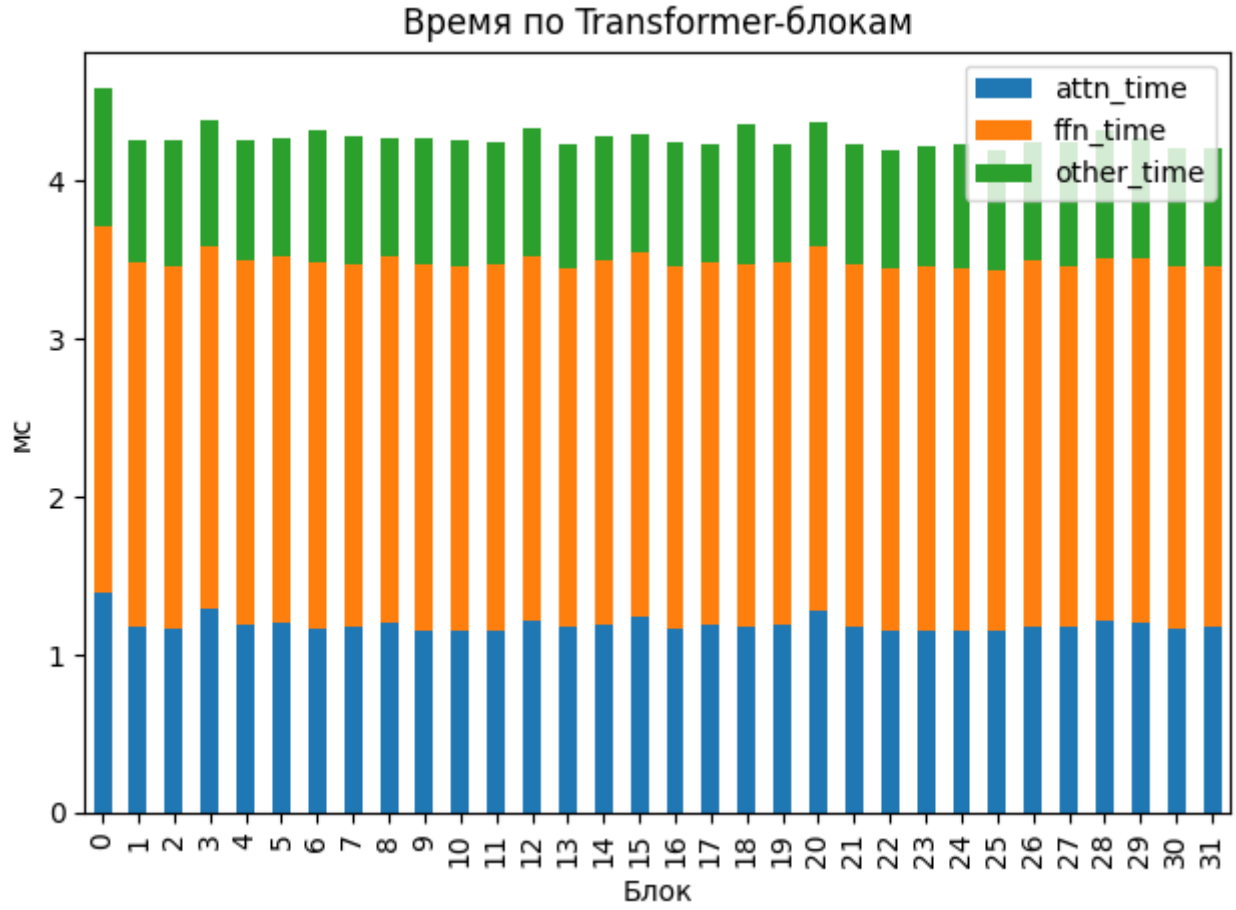
Общие параметры

- Время forward-pass: 106.08 ms
- Размер скрытого пространства: 3072
- Длина входной последовательности: 211
- Количество Transformer-блоков: 32
- Количество параметров: 4 118 544 384
- FLOPs / forward: 1170.30 GFLOPs
- Эффективная производительность: 11.03 TFLOPs

Графики







Transformer-блоки

- Размер скрытого пространства: 3072
- Размер внутреннего пространства FFN: 9216
- Отношение `ffn_dim / hidden_size`: 3.0
- Количество голов внимания: 32
- Количество K/V голов: 8
- Размер головы: 128
- Тип внимания: GQA
- Количество параметров в блоке: 116 391 936
- FLOPs attention: 12.677 GF
- FLOPs FFN: 23.895 GF

Эффективность по блокам

Номер блока	Эффективность (TFLOPs)	Номер блока	Эффективность (TFLOPs)
0	7.97	1	8.59
2	8.58	3	8.33
4	8.59	5	8.57
6	8.46	7	8.53
8	8.56	9	8.57
10	8.59	11	8.62
12	8.44	13	8.63
14	8.55	15	8.52
16	8.62	17	8.64
18	8.38	19	8.63
20	8.37	21	8.64
22	8.72	23	8.66
24	8.63	25	8.71
26	8.62	27	8.63
28	8.46	29	8.56
30	8.70	31	8.70

Сводная таблица времени по типам модулей

Тип	Кол-во	Суммарное время (мс)	Среднее (мс)
Linear	450	117.067	0.2601
LlamaRMSNorm	130	25.437	0.1957
SiLUActivation	64	2.940	0.0459

Тип	Кол-во	Суммарное время (мс)	Среднее (мс)
LlamaRotaryEmbedding	2	1.186	0.5929
Embedding	2	0.463	0.2314

Самые медленные модули (20)

- 4.135 ms — `lm_head` (Linear)
- 4.105 ms — `lm_head` (Linear)
- 0.761 ms — `model.rotary_emb` (LlamaRotaryEmbedding)
- 0.425 ms — `model.rotary_emb` (LlamaRotaryEmbedding)
- 0.410 ms — `model.layers.11.mlp.gate_proj` (Linear)
- 0.406 ms — `model.layers.15.mlp.gate_proj` (Linear)
- 0.402 ms — `model.layers.1.mlp.gate_proj` (Linear)
- 0.401 ms — `model.layers.12.mlp.gate_proj` (Linear)
- 0.400 ms — `model.layers.6.mlp.gate_proj` (Linear)
- 0.399 ms — `model.layers.26.mlp.gate_proj` (Linear)
- 0.398 ms — `model.layers.23.mlp.gate_proj` (Linear)
- 0.395 ms — `model.layers.26.mlp.gate_proj` (Linear)
- 0.392 ms — `model.layers.0.mlp.gate_proj` (Linear)
- 0.390 ms — `model.layers.4.mlp.gate_proj` (Linear)
- 0.389 ms — `model.layers.0.mlp.gate_proj` (Linear)
- 0.389 ms — `model.layers.12.mlp.gate_proj` (Linear)
- 0.388 ms — `model.layers.8.mlp.up_proj` (Linear)
- 0.387 ms — `model.layers.23.mlp.gate_proj` (Linear)
- 0.387 ms — `model.layers.16.mlp.gate_proj` (Linear)
- 0.386 ms — `model.layers.4.mlp.gate_proj` (Linear)