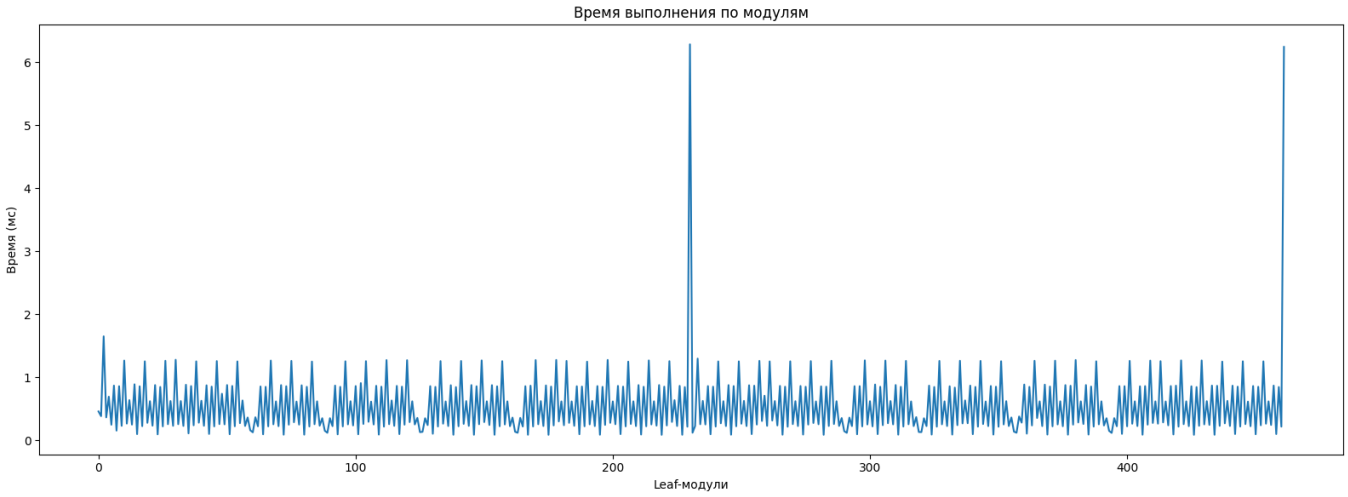
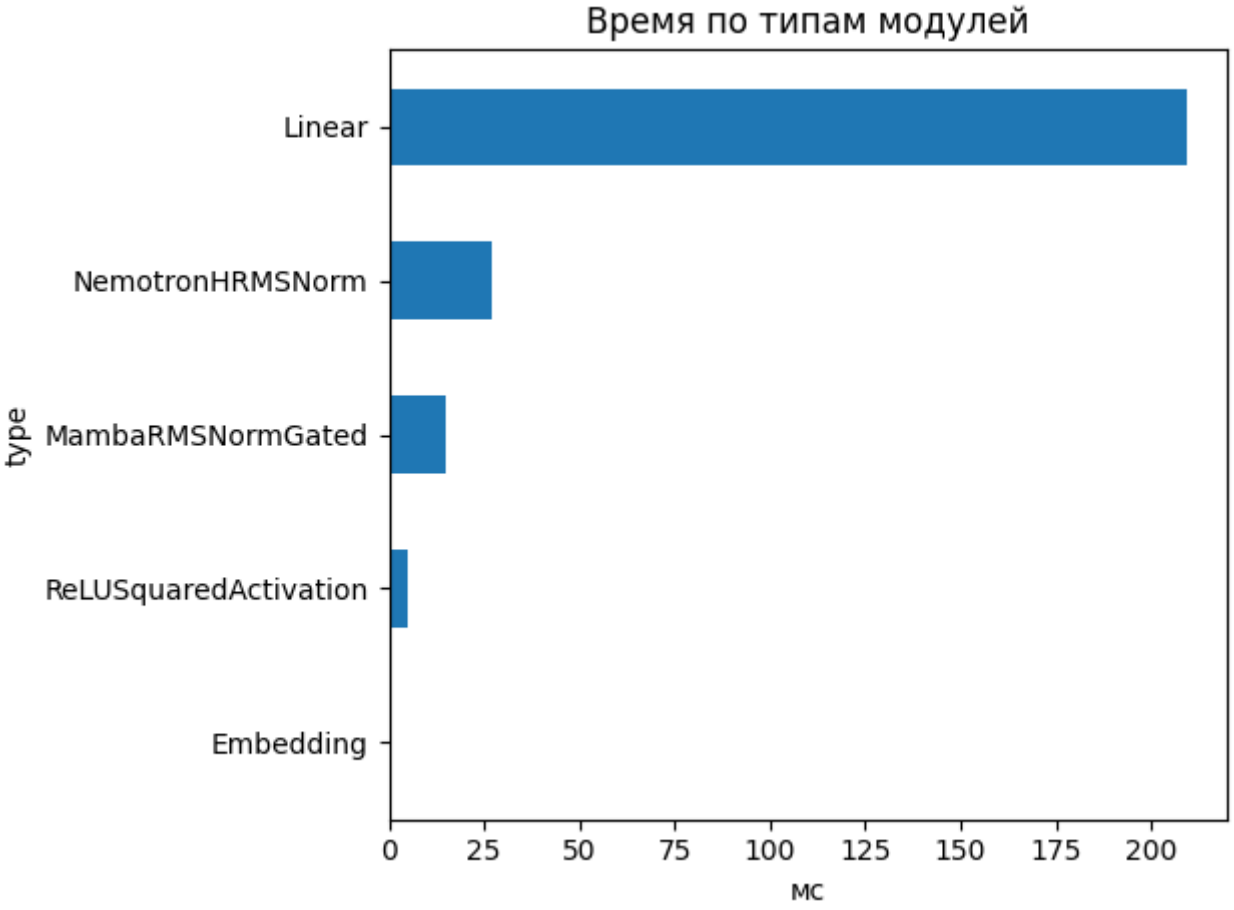
Llama-3.1-Nemotron-Nano-4B-v1.1

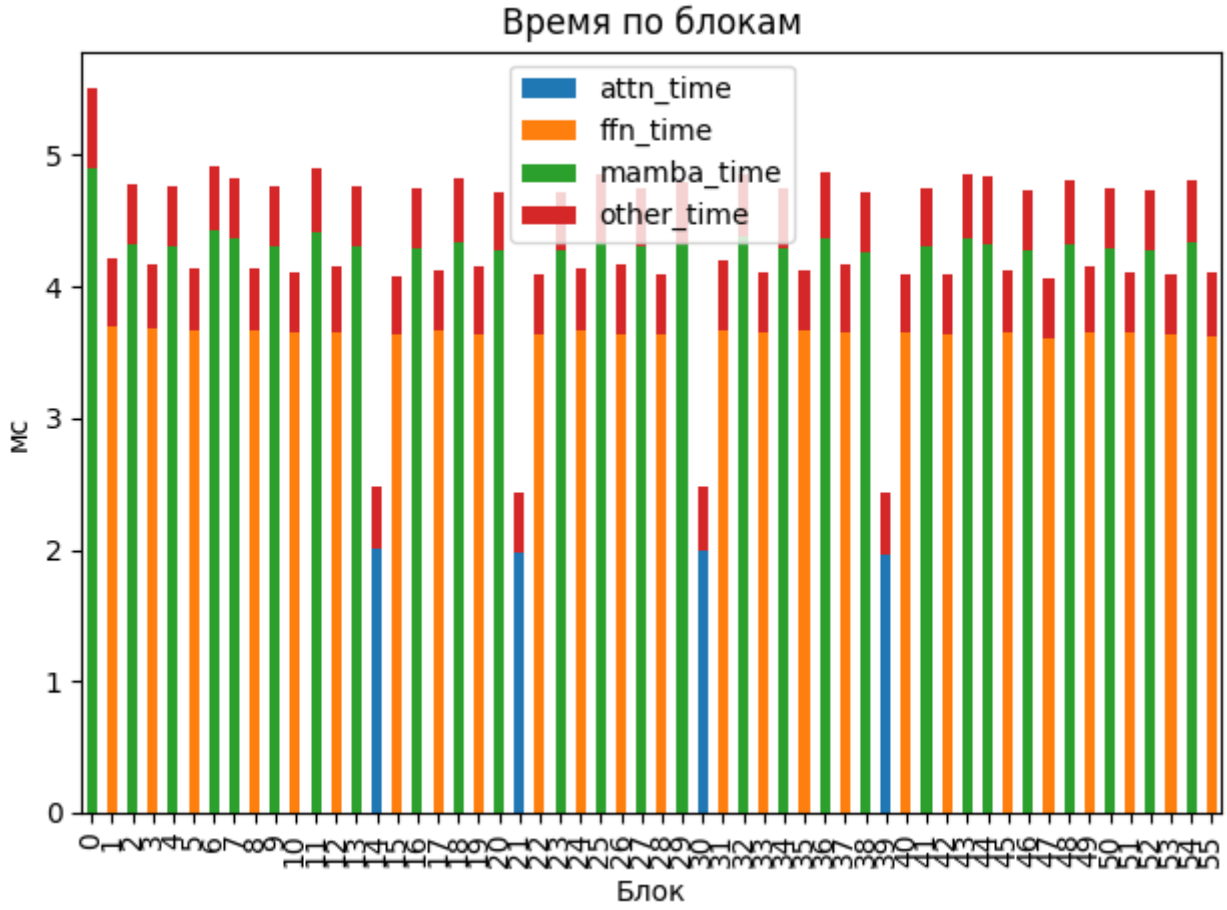
Общие параметры

- Время forward-pass: 183.69 ms
- Размер скрытого пространства: 4480
- Длина входной последовательности: 215
- Количество Transformer-блоков: 56
- Количество параметров: 8 298 823 680
- FLOPs / forward: 3221.39 GFLOPs
- Эффективная производительность: 17.54 TFLOPs

Графики







Transformer-блоки

- Размер скрытого пространства: 4480
- Количество параметров в блоке: 147 374 080

Эффективность по блокам

Номер блока	Эффективность (TFLOPs)	Номер блока	Эффективность (TFLOPs)
0	11.49	1	14.35
2	13.26	3	14.50
4	13.31	5	14.61
6	12.91	7	13.13
8	14.59	9	13.30
10	14.71	11	12.95
12	14.52	13	13.29
14	0.00	15	14.80
16	13.37	17	14.65
18	13.12	19	14.55
20	13.44	21	0.00
22	14.78	23	13.43
24	14.61	25	13.07
26	14.51	27	13.34
28	14.76	29	13.22
30	0.00	31	14.41
32	13.07	33	14.71
34	13.36	35	14.64
36	13.00	37	14.50
38	13.45	39	0.00
40	14.74	41	13.33
42	14.77	43	13.05
44	13.11	45	14.67
46	13.40	47	14.88
48	13.18	49	14.54
50	13.36	51	14.70
52	13.41	53	14.73

Номер блока	Эффективность (TFLOPs)	Номер блока	Эффективность (TFLOPs)
54	13.16	55	14.73

Сводная таблица времени по типам модулей

Тип	Кол-во	Суммарное время (мс)	Среднее (мс)
Linear	242	209.404	0.8653
NemotronHRMSNorm	114	26.961	0.2365
MambaRMSNormGated	54	14.775	0.2736
ReLUSquaredActivation	50	4.890	0.0978
Embedding	2	0.584	0.2921

Самые медленные модули (20)

- 6.282 ms — `lm_head` (Linear)
- 6.241 ms — `lm_head` (Linear)
- 1.652 ms — `backbone.layers.0.mixer.in_proj` (Linear)
- 1.299 ms — `backbone.layers.0.mixer.in_proj` (Linear)
- 1.280 ms — `backbone.layers.7.mixer.in_proj` (Linear)
- 1.277 ms — `backbone.layers.48.mixer.in_proj` (Linear)
- 1.277 ms — `backbone.layers.43.mixer.in_proj` (Linear)
- 1.276 ms — `backbone.layers.36.mixer.in_proj` (Linear)
- 1.275 ms — `backbone.layers.41.mixer.in_proj` (Linear)
- 1.275 ms — `backbone.layers.27.mixer.in_proj` (Linear)
- 1.274 ms — `backbone.layers.29.mixer.in_proj` (Linear)
- 1.270 ms — `backbone.layers.16.mixer.in_proj` (Linear)
- 1.270 ms — `backbone.layers.36.mixer.in_proj` (Linear)
- 1.270 ms — `backbone.layers.46.mixer.in_proj` (Linear)
- 1.270 ms — `backbone.layers.52.mixer.in_proj` (Linear)
- 1.269 ms — `backbone.layers.48.mixer.in_proj` (Linear)
- 1.267 ms — `backbone.layers.16.mixer.in_proj` (Linear)
- 1.267 ms — `backbone.layers.2.mixer.in_proj` (Linear)
- 1.267 ms — `backbone.layers.18.mixer.in_proj` (Linear)
- 1.267 ms — `backbone.layers.43.mixer.in_proj` (Linear)