

Clustering Report: Layer 1 Attention Block

SPD run: **s-275c8f21** (Pile LlamaSimpleMLP 4L) | Clustering: **c-93b8052d** (ensemble e-1ba0b590, run_idx=1) | Iteration: **4000**

1253

L1 Attn Components

846

In Multi-Member Clusters

20

Clusters (from top 50)

0.9675

Max Mean CI (L1 Attn)

Top 50 Layer 1 Attention Components by Mean CI

#	COMPONENT	MODULE	MEAN CI		CLUSTER	INT
1	h. 1.attn.q_proj: 279	L1.q_proj	0.9675	<div></div>	#151	unc
2	h. 1.attn.k_proj: 177	L1.k_proj	0.9380	<div></div>	#151	unc
3	h. 1.attn.o_proj: 796	L1.o_proj	0.5179	<div></div>	#151	tech aca <div>med</div>
4		L1.o_proj	0.3597	<div></div>	#4493	

	h. 1.attn.o_proj: 342					logi and high
5	h. 1.attn.k_proj: 204	L1.k proj	0.3297	<div></div>	#4008	tech sym iden
6	h. 1.attn.o_proj: 691	L1.o proj	0.2223	<div></div>	#4557	tech latin
7	h. 1.attn.o_proj: 678	L1.o proj	0.2205	<div></div>	singleton	tech stru frag med
8	h. 1.attn.v_proj: 202	L1.v proj	0.2067	<div></div>	#4829	dom tech term
9	h. 1.attn.o_proj: 580	L1.o proj	0.1622	<div></div>	#4536	prop citat
10	h. 1.attn.o_proj: 501	L1.o proj	0.1478	<div></div>	singleton	mat deli
11	h. 1.attn.o_proj: 799	L1.o proj	0.1462	<div></div>	singleton	form term med
12	h. 1.attn.k_proj: 137	L1.k proj	0.1451	<div></div>	singleton	form aca med
13	h. 1.attn.o_proj: 413	L1.o proj	0.1352	<div></div>	#4505	sec con phra
14	h. 1.attn.o_proj: 425	L1.o proj	0.1295	<div></div>	singleton	tech che med
15	h. 1.attn.k_proj: 231	L1.k proj	0.1260	<div></div>	singleton	pun synt high
16	h. 1.attn.o_proj: 253	L1.o proj	0.1249	<div></div>	singleton	tech scie med
17	h. 1.attn.o_proj: 639	L1.o proj	0.1200	<div></div>	#4546	all-c snal form
18		L1.o proj	0.1157	<div></div>	singleton	

	h. 1.attn.o_proj: 354						mat tech term
19	h. 1.attn.o_proj: 902	L1.o_proj	0.1137	<div></div>	#347		cod pun
20	h. 1.attn.o_proj: 406	L1.o_proj	0.1048	<div></div>	#4504		nou tech med
21	h. 1.attn.o_proj: 720	L1.o_proj	0.1029	<div></div>	singleton		tech form
22	h. 1.attn.o_proj: 44	L1.o_proj	0.1023	<div></div>	#4442		mat tabl
23	h. 1.attn.o_proj: 836	L1.o_proj	0.0977	<div></div>	singleton		tech boil
24	h. 1.attn.v_proj: 715	L1.v_proj	0.0964	<div></div>	singleton		unc
25	h. 1.attn.o_proj: 263	L1.o_proj	0.0928	<div></div>	singleton		sub- frag pun
26	h. 1.attn.v_proj: 767	L1.v_proj	0.0898	<div></div>	#4919		sent pun
27	h. 1.attn.o_proj: 892	L1.o_proj	0.0883	<div></div>	#4584		late nota
28	h. 1.attn.o_proj: 200	L1.o_proj	0.0873	<div></div>	#4467		prep phra india
29	h. 1.attn.o_proj: 597	L1.o_proj	0.0833	<div></div>	#454		new form high
30	h. 1.attn.o_proj: 507	L1.o_proj	0.0832	<div></div>	singleton		stru tech form med
31	h. 1.attn.v_proj: 416	L1.v_proj	0.0817	<div></div>	#4863		cap nou
32	h. 1.attn.o_proj: 560	L1.o_proj	0.0804	<div></div>	singleton		form deli

							trans med
33	h. 1.attn.v_proj : 566	L1.v_proj	0.0796	<div></div>	singleton	prop frag	
34	h. 1.attn.o_proj : 874	L1.o_proj	0.0780	<div></div>	singleton	tem spat prep med	
35	h. 1.attn.o_proj : 924	L1.o_proj	0.0770	<div></div>	singleton	tech aca iden	
36	h. 1.attn.v_proj : 209	L1.v_proj	0.0751	<div></div>	singleton	prep rela high	
37	h. 1.attn.o_proj : 917	L1.o_proj	0.0748	<div></div>	singleton	scie tech med	
38	h. 1.attn.o_proj : 82	L1.o_proj	0.0728	<div></div>	singleton	form and mar	
39	h. 1.attn.o_proj : 196	L1.o_proj	0.0722	<div></div>	singleton	med clini term	
40	h. 1.attn.v_proj : 904	L1.v_proj	0.0718	<div></div>	singleton	subj emp mod	
41	h. 1.attn.v_proj : 138	L1.v_proj	0.0696	<div></div>	#2807	wor and high	
42	h. 1.attn.o_proj : 932	L1.o_proj	0.0684	<div></div>	singleton	stru deli cod	
43	h. 1.attn.v_proj : 926	L1.v_proj	0.0676	<div></div>	#4192	mat cod high	
44	h. 1.attn.v_proj : 625	L1.v_proj	0.0671	<div></div>	#3703	shor vari abb high	
45	h. 1.attn.o_proj : 744	L1.o_proj	0.0670	<div></div>	singleton	boile lega	
46		L1.o_proj	0.0651	<div></div>	singleton		

	h. 1.attn.o_proj: 87						tech stru med
47	h. 1.attn.o_proj: 534	L1.o proj	0.0628	●	singleton		stat scie tern
48	h. 1.attn.o_proj: 530	L1.o proj	0.0607	●	singleton		ordi frac mod
49	h. 1.attn.v_proj: 814	L1.v proj	0.0598	●	#4173		mat cod high
50	h. 1.attn.o_proj: 995	L1.o proj	0.0596	●	singleton		unc

Cluster Details (20 clusters containing top L1 Attn components)

Cluster #151 (20 members: 3 L1-attn, 17 other)

max CI:
0.9737

Modules: L0.k proj L0.o proj L0.q proj L0.mlp.c_fc L1.k proj L1.o proj L1.q proj L1.mlp.c_fc L1.mlp.down proj L2.k proj L2.o proj L2.q proj L2.mlp.down proj L3.k proj L3.o proj L3.q proj L3.mlp.c_fc L3.mlp.down proj

COMPONENT	MODULE	MEAN CI		
h.3.attn.o_proj: 912	L3.o proj	0.9737	<div></div>	L
h.1.attn.q_proj: 279	L1.q proj	0.9675	<div></div>	L
h.0.attn.o_proj: 947	L0.o proj	0.9521	<div></div>	L
h.1.attn.k_proj: 177	L1.k proj	0.9380	<div></div>	L
h.3.mlp.c_fc: 2000	L3.mlp.c_fc	0.8724	<div></div>	t t
h.2.attn.o_proj: 585	L2.o proj	0.8483	<div></div>	t s G
h.2.attn.q_proj: 236	L2.q proj	0.8046	<div></div>	L

h. 1.mlp.down_proj: 1968	L1.mlp.down_proj	0.7772	<div></div>	L
h.3.attn.k_proj: 169	L3.k_proj	0.7758	<div></div>	L
h. 3.mlp.down_proj: 1999	L3.mlp.down_proj	0.7737	<div></div>	L
h.3.attn.q_proj: 31	L3.q_proj	0.7339	<div></div>	L
h.2.attn.k_proj: 193	L2.k_proj	0.7107	<div></div>	L
h.0.attn.q_proj: 1	L0.q_proj	0.6187	<div></div>	L
h. 3.mlp.down_proj: 2531	L3.mlp.down_proj	0.5759	<div></div>	f t Y
h.0.mlp.c_fc:444	L0.mlp.c_fc	0.5614	<div></div>	L
h.1.attn.o_proj: 796	L1.o_proj	0.5179	<div></div>	t a Y
h.1.mlp.c_fc: 3911	L1.mlp.c_fc	0.4981	<div></div>	c t R
h.0.attn.k_proj: 155	L0.k_proj	0.4942	<div></div>	L
h. 2.mlp.down_proj: 1802	L2.mlp.down_proj	0.4443	<div></div>	L
h.3.attn.k_proj: 206	L3.k_proj	0.4051	<div></div>	L

Cluster #4493 (2 members: 1 L1-attn, 1 other)

max CI:
0.3597

Modules: L1.o_proj L2.k_proj

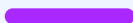


COMPONENT	MODULE	MEAN CI		INTERPRETATION
h. 1.attn.o_proj: 342	L1.o_proj	0.3597	<div></div>	logical connectors and transitions high

h.
2.attn.k_proj:8 L2.k_proj **0.3315**  unclear low

Cluster #4008 (3 members: 1 L1-attn, 2 other)

max CI:
0.3297

Modules: L0.mlp.down_proj L1.k_proj L3.v_proj

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.1.attn.k_proj:204	L1.k_proj	0.3297		technical and symbolic identifiers low
h.3.attn.v_proj:701	L3.v_proj	0.3061		unclear low
h.0.mlp.down_proj:2241	L0.mlp.down_proj	0.2886		unclear low

Cluster #4557 (3 members: 1 L1-attn, 2 other)

max CI:
0.2785

Modules: L1.o_proj L2.o_proj L2.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.2.mlp.c_fc:2266	L2.mlp.c_fc	0.2785		technical and symbolic formatting medium
h.1.attn.o_proj:691	L1.o_proj	0.2223		technical and non-latin text high
h.2.attn.o_proj:144	L2.o_proj	0.1699		grammatical function words and delimiters medium

Cluster #4829 (2 members: 1 L1-attn, 1 other)

max CI:
0.2067

Modules: L1.v_proj L2.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.1.attn.v_proj:202	L1.v_proj	0.2067		domain-specific technical terminology low

h.2.mlp.c_fc:
4042

L2.mlp.c_fc

0.1714



specialized vocabulary
suffixes medium

Cluster #4536 (2 members: 1 L1-attn, 1 other)

max CI:
0.1622

Modules: L1.o_proj L1.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.1.attn.o_proj: 580	L1.o_proj	0.1622		proper nouns and citations medium
h.1.mlp.c_fc: 2535	L1.mlp.c_fc	0.1123		academic and legal citations high

Cluster #4505 (3 members: 1 L1-attn, 2 other)

max CI:
0.1464

Modules: L1.o_proj L3.o_proj L3.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.3.mlp.c_fc: 1463	L3.mlp.c_fc	0.1464		unclear low
h. 1.attn.o_proj: 413	L1.o_proj	0.1352		second-person conversational phrases medium
h. 3.attn.o_proj: 660	L3.o_proj	0.1214		informal or profane speech high

Cluster #4546 (2 members: 1 L1-attn, 1 other)

max CI:
0.1200

Modules: L1.o_proj L1.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h. 1.attn.o_proj: 639	L1.o_proj	0.1200		all-caps and snake_case formatting high
h.1.mlp.c_fc: 2334	L1.mlp.c_fc	0.0726		uppercase and technical suffix continuation medium

Cluster #347 (3 members: 1 L1-attn, 2 other)

max CI:
0.1251

Modules: L0.o_proj L0.mlp.c_fc L1.o_proj

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h. 0.attn.o_proj: 266	L0.o_proj	0.1251	<div></div>	programming syntax and identifiers high
h.0.mlp.c_fc: 1014	L0.mlp.c_fc	0.1241	<div></div>	code and structured syntax high
h. 1.attn.o_proj: 902	L1.o_proj	0.1137	<div></div>	code syntax and punctuation high

Cluster #4504 (3 members: 1 L1-attn, 2 other)

max CI:
0.1048

Modules: L1.o_proj L1.mlp.c_fc L2.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h. 1.attn.o_proj: 406	L1.o_proj	0.1048	<div></div>	noun phrases and technical terms medium
h.2.mlp.c_fc: 1932	L2.mlp.c_fc	0.0745	<div></div>	technical and formal nouns medium
h.1.mlp.c_fc: 3342	L1.mlp.c_fc	0.0700	<div></div>	noun phrase heads medium

Cluster #4442 (2 members: 1 L1-attn, 1 other)

max CI:
0.1023

Modules: L1.o_proj L1.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h. 1.attn.o_proj: 44	L1.o_proj	0.1023	<div></div>	mathematical and table data high
h.1.mlp.c_fc: 2743	L1.mlp.c_fc	0.0774	<div></div>	mathematical and technical delimiters high

Cluster #4919 (2 members: 1 L1-attn, 1 other)

max CI:
0.0898

Modules: L1.v proj L3.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.1.attn.v_proj: 767	L1.v proj	0.0898	—	sentence-ending punctuation high
h.3.mlp.c_fc: 1420	L3.mlp.c_fc	0.0890	—	block-level structural delimiters high

Cluster #4584 (2 members: 1 L1-attn, 1 other)

max CI:
0.0883

Modules: L1.o proj L1.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.1.attn.o_proj: 892	L1.o proj	0.0883	—	latex and technical notation high
h.1.mlp.c_fc: 2082	L1.mlp.c_fc	0.0675	—	mathematical symbols and subscripts high

Cluster #4467 (2 members: 1 L1-attn, 1 other)

max CI:
0.0873

Modules: L1.o proj L1.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.1.attn.o_proj: 200	L1.o proj	0.0873	—	prepositional phrases and indicators medium
h.1.mlp.c_fc: 3194	L1.mlp.c_fc	0.0499	•	definite and indefinite articles medium

Cluster #454 (2 members: 1 L1-attn, 1 other)

max CI:
0.0833

Modules: L0.o proj L1.o proj

COMPONENT	MODULE	MEAN CI		INTERPRETATION
-----------	--------	---------	--	----------------

h.
1.attn.o_proj: **L1.o_proj** **0.0833**  newlines in formatted code **high**

h.
0.attn.o_proj: **L0.o_proj** **0.0538**  structural markers and code boilerplate **medium**

Cluster #4863 (2 members: 1 L1-attn, 1 other)

max CI:
0.0817



Modules: **L1.v_proj** **L2.q_proj**

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.1.attn.v_proj: 416	L1.v_proj	0.0817		capitalized proper noun stems high
h.2.attn.q_proj: 88	L2.q_proj	0.0492		technical and academic acronyms high

Cluster #2807 (2 members: 1 L1-attn, 1 other)

max CI:
0.1155



Modules: **L0.mlp.c_fc** **L1.v_proj**

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.0.mlp.c_fc: 3188	L0.mlp.c_fc	0.1155		technical and non-english suffixes medium
h.1.attn.v_proj: 138	L1.v_proj	0.0696		word fragments and sub-tokens high

Cluster #4192 (2 members: 1 L1-attn, 1 other)

max CI:
0.0676

Modules: **L0.mlp.down_proj** **L1.v_proj**

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h.1.attn.v_proj: 926	L1.v_proj	0.0676		mathematical and code punctuation high
h.0.mlp.down_proj: 2695	L0.mlp.down_proj	0.0596		programming and technical delimiters high

Cluster #3703 (3 members: 1 L1-attn, 2 other)

max CI:
0.0905

Modules: L0.mlp.down_proj L1.v_proj L1.mlp.c_fc

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h. 0.mlp.down_proj: 1410	L0.mlp.down_proj	0.0905	■	isolated single characters and abbreviations medium
h.1.attn.v_proj: 625	L1.v_proj	0.0671	●	short technical variables and abbreviations high
h.1.mlp.c_fc: 3161	L1.mlp.c_fc	0.0520	●	single characters and abbreviations high

Cluster #4173 (2 members: 1 L1-attn, 1 other)

max CI:
0.0608

Modules: L0.mlp.down_proj L1.v_proj

COMPONENT	MODULE	MEAN CI		INTERPRETATION
h. 0.mlp.down_proj: 2655	L0.mlp.down_proj	0.0608	●	mathematical and code delimiters high
h.1.attn.v_proj: 814	L1.v_proj	0.0598	●	mathematical and code delimiters high