Clustering Report: Layer 1 Attention Block

SPD run: **s-275c8f21** (Pile LlamaSimpleMLP 4L) | Clustering: **c-93b8052d** (ensemble e-1ba0b590, run_idx=1) | Iteration: **3100**

1253
L1 Attn Components

658
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	#152	uncl
2	h.1.attn.k_proj:177	L1.k_proj	0.9380	#152	uncl
3	h.1.attn.o_proj:796	L1.o_proj	0.5179	#152	tech aca mec
4		L1.o_proj	0.3597	#4757	

	h. 1.attn.o_proj: 342					logi and high
5	h. 1.attn.k_proj: 204	L1.k_proj	0.3297		#4217	tech sym iden
6	h. 1.attn.o_proj: 691	L1.o_proj	0.2223		#4826	tech latin
7	h. 1.attn.o_proj: 678	L1.o_proj	0.2205		singleton	tech stru frag mec
8	h. 1.attn.v_proj: 202	L1.v_proj	0.2067		#5180	dom tech tern
9	h. 1.attn.o_proj: 580	L1.o_proj	0.1622		#4803	prop cita
10	h. 1.attn.o_proj: 501	L1.o_proj	0.1478		singleton	mat deli
11	h. 1.attn.o_proj: 799	L1.o_proj	0.1462		singleton	form tern mec
12	h. 1.attn.k_proj: 137	L1.k_proj	0.1451		singleton	form acad mec
13	h. 1.attn.o_proj: 413	L1.o_proj	0.1352		#4770	seco conv phra
14	h. 1.attn.o_proj: 425	L1.o_proj	0.1295		singleton	tech chem mec
15	h. 1.attn.k_proj: 231	L1.k_proj	0.1260		singleton	pun synt high
16	h. 1.attn.o_proj: 253	L1.o_proj	0.1249		singleton	tech scie mec
17	h. 1.attn.o_proj: 639	L1.o_proj	0.1200		#4814	all-c snal form
18		L1.o_proj	0.1157		singleton	

	h.1.attn.o_proj:354					mat tech tern
19	h.1.attn.o_proj:902	L1.o_proj	0.1137		#351	cod pun
20	h.1.attn.o_proj:406	L1.o_proj	0.1048		#4768	nou tech mec
21	h.1.attn.o_proj:720	L1.o_proj	0.1029		singleton	tech form
22	h.1.attn.o_proj:44	L1.o_proj	0.1023		#4701	mat tabl
23	h.1.attn.o_proj:836	L1.o_proj	0.0977		singleton	tech boil
24	h.1.attn.v_proj:715	L1.v_proj	0.0964		singleton	unc
25	h.1.attn.o_proj:263	L1.o_proj	0.0928		singleton	sub- frag pun
26	h.1.attn.v_proj:767	L1.v_proj	0.0898		#5283	sent pun
27	h.1.attn.o_proj:892	L1.o_proj	0.0883		#4856	late nota
28	h.1.attn.o_proj:200	L1.o_proj	0.0873		#4728	prep phra indi
29	h.1.attn.o_proj:597	L1.o_proj	0.0833		#460	new form high
30	h.1.attn.o_proj:507	L1.o_proj	0.0832		singleton	stru tech form mec
31	h.1.attn.v_proj:416	L1.v_proj	0.0817		#5219	capi nou
32	h.1.attn.o_proj:560	L1.o_proj	0.0804		singleton	form deli

						tran med
33	h. 1.attn.v_proj: 566	L1.v_proj	**0.0796**		singleton	prop frag
34	h. 1.attn.o_proj: 874	L1.o_proj	**0.0780**		singleton	tem spat prep med
35	h. 1.attn.o_proj: 924	L1.o_proj	**0.0770**		singleton	tech aca ider
36	h. 1.attn.v_proj: 209	L1.v_proj	**0.0751**		singleton	prep rela high
37	h. 1.attn.o_proj: 917	L1.o_proj	**0.0748**		singleton	scie tech med
38	h. 1.attn.o_proj: 82	L1.o_proj	**0.0728**		singleton	form and mar
39	h. 1.attn.o_proj: 196	L1.o_proj	**0.0722**		singleton	med clini tern
40	h. 1.attn.v_proj: 904	L1.v_proj	**0.0718**		singleton	subj emp moc
41	h. 1.attn.v_proj: 138	L1.v_proj	**0.0696**		#2966	wor and high
42	h. 1.attn.o_proj: 932	L1.o_proj	**0.0684**		singleton	stru deli cod
43	h. 1.attn.v_proj: 926	L1.v_proj	**0.0676**		#4405	mat cod high
44	h. 1.attn.v_proj: 625	L1.v_proj	**0.0671**		#3902	sho vari abb high
45	h. 1.attn.o_proj: 744	L1.o_proj	**0.0670**		singleton	boil lega
46		L1.o_proj	**0.0651**		singleton	

	h. 1.attn.o_proj: 87					tech stru mec
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 moc
49	h. 1.attn.v_proj: 814	L1.v_proj	0.0598		#4386	mat cod high
50	h. 1.attn.o_proj: 995	L1.o_proj	0.0596		singleton	uncl

Cluster Details (20 clusters containing top L1 Attn components)

Cluster #152 (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	I
h.3.attn.o_proj: 912	L3.o_proj	0.9737	u
h.1.attn.q_proj: 279	L1.q_proj	0.9675	u
h.0.attn.o_proj: 947	L0.o_proj	0.9521	u
h.1.attn.k_proj: 177	L1.k_proj	0.9380	u
h.3.mlp.c_fc: 2000	L3.mlp.c_fc	0.8724	t t
h.2.attn.o_proj: 585	L2.o_proj	0.8483	t s
h.2.attn.q_proj: 236	L2.q_proj	0.8046	u

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

Cluster #4757 (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		logical connectors and transitions high

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

Cluster #4217 (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 #4826 (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 #5180 (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 #4803 (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 #4770 (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 #4814 (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 #351 (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**		programming syntax and identifiers high
h.0.mlp.c_fc:1014	`L0.mlp.c_fc`	**0.1241**		code and structured syntax high
h.1.attn.o_proj:902	`L1.o_proj`	**0.1137**		code syntax and punctuation high

Cluster #4768 (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**		noun phrases and technical terms medium
h.2.mlp.c_fc:1932	`L2.mlp.c_fc`	**0.0745**		technical and formal nouns medium
h.1.mlp.c_fc:3342	`L1.mlp.c_fc`	**0.0700**		noun phrase heads medium

Cluster #4701 (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**		mathematical and table data high
h.1.mlp.c_fc:2743	`L1.mlp.c_fc`	**0.0774**		mathematical and technical delimiters high

Cluster #5283 (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 #4856 (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 #4728 (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 #460 (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:597	L1.o_proj	0.0833		newlines in formatted code high
h.0.attn.o_proj:519	L0.o_proj	0.0538		structural markers and code boilerplate medium

Cluster #5219 (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 #2966 (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 #4405 (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 #3902 (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 #4386 (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