

Supplement C: Additional S3AE result in the other large language models

In this supplementary section, we provide the full results about S3AE in the 11 LLMs not reported in the main manuscript. In each figure, alphabetic labels and symbols denote the following:

- (A) Similarity between the vectors from the same layer.
- (B) Similarity between the vectors from different layers.
- The column index labels are abbreviations of the row index labels.

Also, each table is structured in the following way:

- Top-table: Increase in reconstruction loss (ratio) when the representational state vector (column) was masked in reconstructing the LLM activations.
- Middle-table: Proportion of samples having reconstruction loss increase when the representational state vector (column) was masked in reconstructing the LLM activations.
- Bottom-table: Thought classification performance (F1).
- The column index labels are abbreviations of the row index labels.

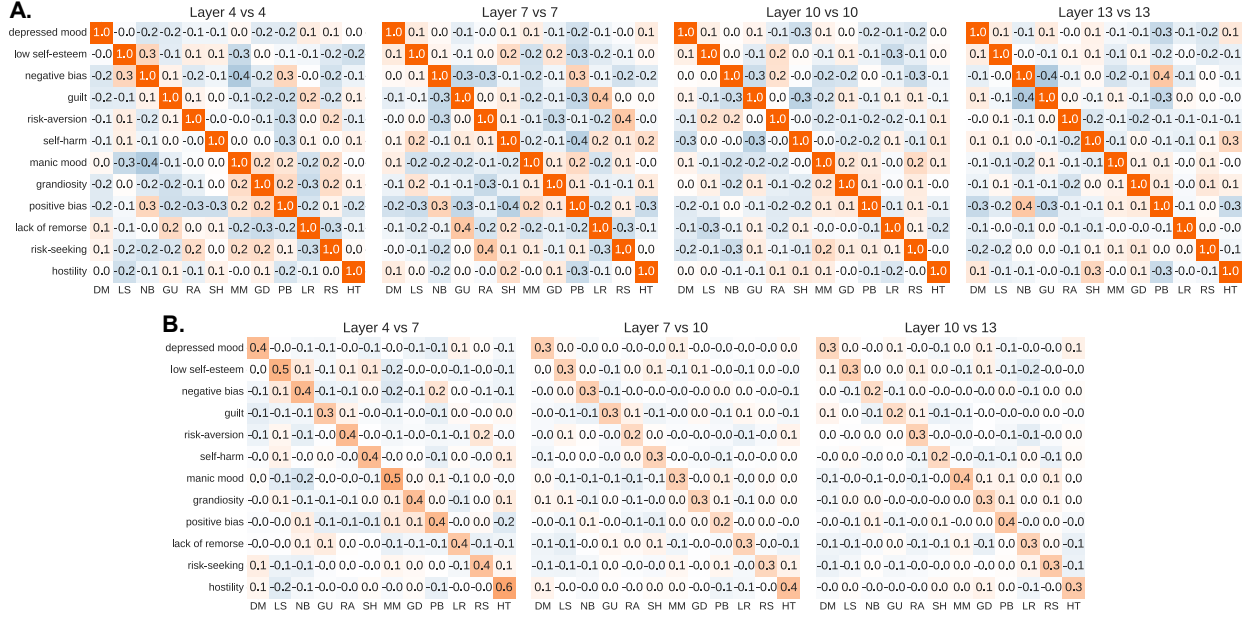


Figure 1: Cosine similarity between S3AE-learned representational state vectors of Gemma-3-270M.

Table 1: S3AE evaluation: Gemma-3-270M.

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
4	1.052	1.039	1.025	1.055	1.043	1.039	1.092	1.049	1.037	1.029	1.048	1.062
7	1.067	1.044	1.032	1.067	1.071	1.059	1.104	1.068	1.048	1.049	1.054	1.066
10	1.040	1.033	1.022	1.049	1.040	1.042	1.083	1.040	1.027	1.033	1.045	1.058
13	1.027	1.024	1.024	1.039	1.042	1.029	1.058	1.044	1.029	1.032	1.027	1.054
4	0.641	0.574	0.456	0.693	0.627	0.580	0.835	0.671	0.495	0.524	0.613	0.625
7	0.755	0.638	0.576	0.810	0.753	0.657	0.920	0.727	0.607	0.635	0.688	0.742
10	0.756	0.629	0.557	0.764	0.718	0.685	0.900	0.655	0.587	0.624	0.722	0.719
13	0.724	0.610	0.574	0.753	0.706	0.659	0.890	0.695	0.602	0.626	0.675	0.716
4	0.790	0.728	0.616	0.816	0.775	0.727	0.915	0.807	0.657	0.687	0.751	0.785
7	0.850	0.773	0.698	0.873	0.834	0.779	0.946	0.845	0.728	0.761	0.796	0.832
10	0.869	0.778	0.731	0.882	0.844	0.809	0.946	0.842	0.740	0.743	0.824	0.839
13	0.869	0.765	0.712	0.880	0.837	0.804	0.944	0.844	0.739	0.765	0.815	0.827

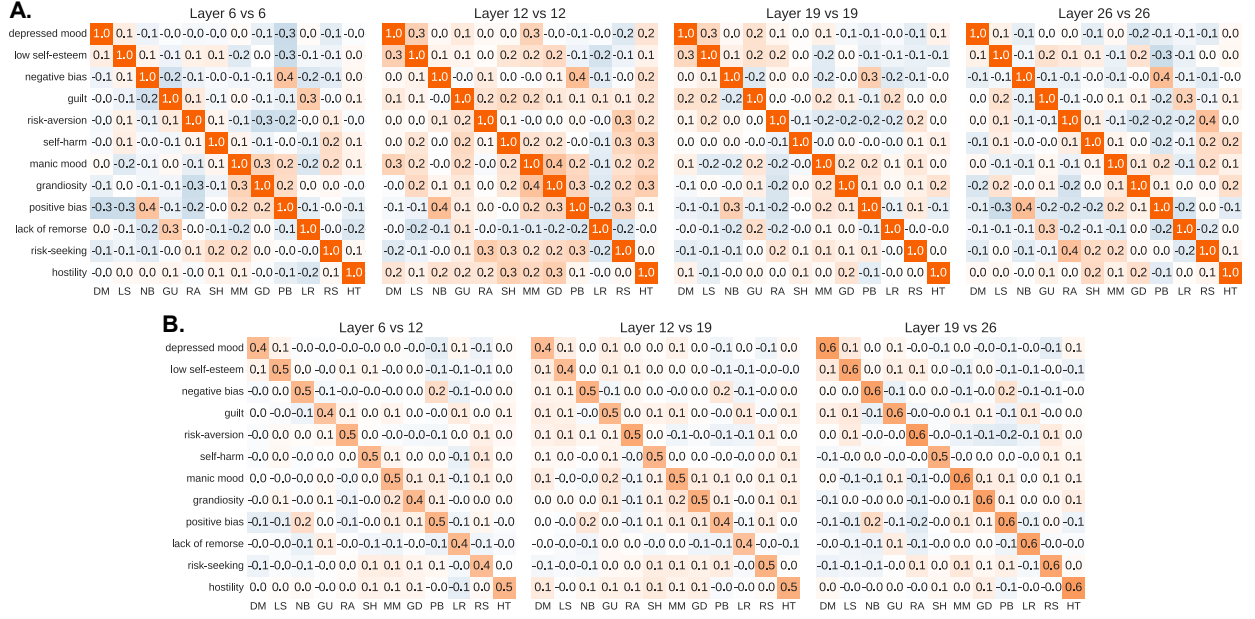


Figure 2: Cosine similarity between S3AE-learned representational state vectors of Gemma-3-4B.

Table 2: S3AE evaluation: Gemma-3-4B.

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
6	1.050	1.042	1.023	1.062	1.048	1.058	1.114	1.068	1.047	1.040	1.040	1.071
12	1.115	1.077	1.043	1.151	1.089	1.096	1.183	1.154	1.070	1.076	1.079	1.127
19	1.125	1.062	1.042	1.117	1.094	1.075	1.182	1.095	1.059	1.068	1.073	1.124
26	1.134	1.059	1.043	1.119	1.091	1.098	1.218	1.119	1.092	1.078	1.102	1.118
6	0.816	0.716	0.641	0.843	0.749	0.773	0.944	0.790	0.662	0.709	0.755	0.822
12	0.879	0.778	0.758	0.881	0.833	0.860	0.938	0.844	0.759	0.824	0.837	0.843
19	0.905	0.804	0.811	0.906	0.882	0.911	0.969	0.873	0.816	0.883	0.883	0.899
26	0.901	0.791	0.786	0.910	0.885	0.909	0.978	0.885	0.825	0.891	0.893	0.909
6	0.899	0.814	0.759	0.905	0.851	0.866	0.964	0.877	0.791	0.820	0.834	0.886
12	0.925	0.850	0.843	0.943	0.904	0.919	0.967	0.909	0.857	0.887	0.900	0.921
19	0.941	0.868	0.870	0.957	0.922	0.944	0.981	0.921	0.883	0.920	0.919	0.935
26	0.945	0.870	0.862	0.955	0.922	0.942	0.986	0.924	0.886	0.920	0.927	0.938



Figure 3: Cosine similarity between S3AE-learned representational state vectors of Gemma-3-12B.

Table 3: S3AE evaluation: Gemma-3-12B.

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
8	1.047	1.051	1.032	1.068	1.049	1.069	1.138	1.090	1.050	1.061	1.053	1.077
18	1.155	1.110	1.047	1.201	1.111	1.156	1.174	1.177	1.076	1.122	1.113	1.146
27	1.155	1.083	1.053	1.169	1.122	1.121	1.242	1.178	1.080	1.103	1.130	1.158
37	1.152	1.068	1.039	1.110	1.074	1.093	1.191	1.134	1.079	1.080	1.085	1.112
8	0.783	0.709	0.665	0.858	0.748	0.785	0.946	0.827	0.684	0.750	0.772	0.819
18	0.911	0.834	0.853	0.954	0.884	0.960	0.978	0.904	0.846	0.948	0.887	0.930
27	0.911	0.824	0.851	0.954	0.881	0.939	0.980	0.911	0.840	0.957	0.902	0.934
37	0.919	0.814	0.814	0.930	0.891	0.952	0.992	0.906	0.849	0.961	0.926	0.922
8	0.896	0.816	0.773	0.912	0.854	0.872	0.966	0.885	0.802	0.839	0.852	0.892
18	0.958	0.892	0.905	0.973	0.933	0.970	0.989	0.936	0.914	0.966	0.942	0.956
27	0.955	0.887	0.892	0.972	0.932	0.962	0.988	0.940	0.906	0.964	0.938	0.953
37	0.960	0.889	0.883	0.974	0.935	0.968	0.995	0.941	0.910	0.961	0.945	0.955

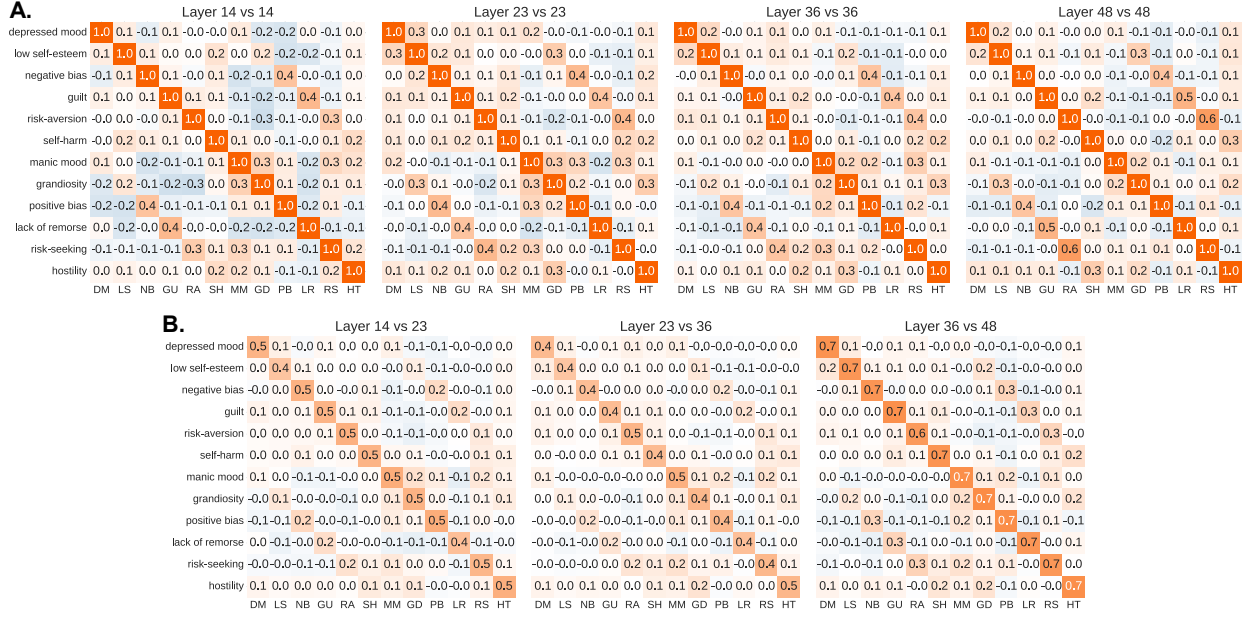


Figure 4: Cosine similarity between S3AE-learned representational state vectors of Gemma-3-27B.

Table 4: S3AE evaluation: Gemma-3-27B

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
14	1.068	1.056	1.036	1.105	1.062	1.070	1.135	1.104	1.056	1.065	1.055	1.072
23	1.136	1.085	1.050	1.182	1.106	1.132	1.232	1.188	1.077	1.097	1.126	1.154
36	1.144	1.076	1.049	1.168	1.088	1.110	1.211	1.144	1.075	1.112	1.111	1.143
48	1.140	1.068	1.043	1.171	1.110	1.112	1.197	1.133	1.076	1.121	1.139	1.120
14	0.866	0.782	0.766	0.924	0.812	0.864	0.970	0.854	0.767	0.862	0.833	0.865
23	0.934	0.853	0.866	0.961	0.898	0.952	0.989	0.901	0.871	0.949	0.918	0.922
36	0.926	0.874	0.887	0.961	0.911	0.972	0.991	0.910	0.902	0.980	0.938	0.956
48	0.924	0.867	0.827	0.967	0.901	0.970	0.997	0.918	0.918	0.977	0.915	0.951
14	0.943	0.867	0.843	0.955	0.897	0.931	0.984	0.918	0.870	0.915	0.900	0.925
23	0.960	0.899	0.903	0.975	0.934	0.967	0.993	0.941	0.923	0.963	0.944	0.956
36	0.966	0.910	0.918	0.980	0.944	0.980	0.994	0.951	0.939	0.976	0.956	0.969
48	0.969	0.916	0.901	0.983	0.945	0.983	0.999	0.954	0.949	0.984	0.953	0.973

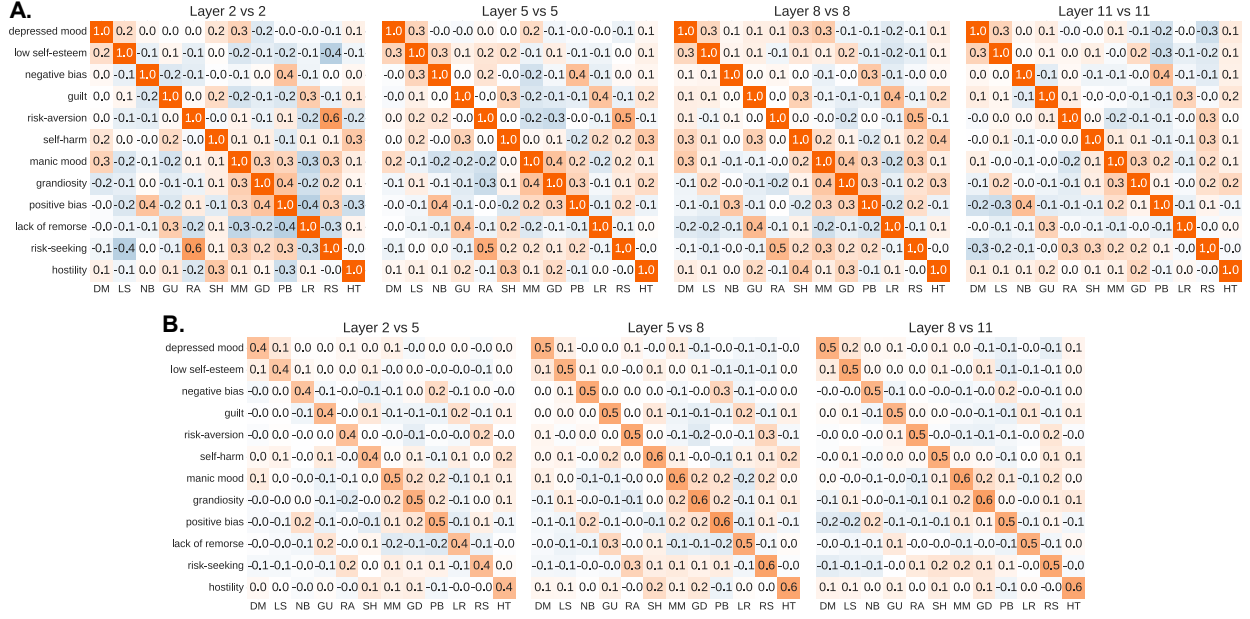


Figure 5: Cosine similarity between S3AE-learned representational state vectors of Llama-3.2-1B.

Table 5: S3AE evaluation: Llama-3.2-1B.

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
2	1.076	1.066	1.035	1.073	1.080	1.062	1.113	1.088	1.058	1.051	1.093	1.064
5	1.104	1.085	1.051	1.163	1.113	1.111	1.202	1.143	1.085	1.094	1.118	1.110
8	1.150	1.071	1.056	1.176	1.106	1.154	1.249	1.165	1.097	1.093	1.140	1.127
11	1.111	1.058	1.033	1.092	1.085	1.064	1.191	1.120	1.067	1.066	1.088	1.091
2	0.797	0.706	0.647	0.870	0.775	0.826	0.955	0.820	0.723	0.789	0.783	0.828
5	0.865	0.785	0.806	0.942	0.867	0.918	0.968	0.862	0.824	0.904	0.866	0.903
8	0.893	0.809	0.826	0.943	0.886	0.950	0.989	0.890	0.869	0.934	0.905	0.920
11	0.883	0.784	0.769	0.921	0.854	0.886	0.986	0.856	0.813	0.900	0.876	0.883
2	0.909	0.834	0.796	0.931	0.866	0.900	0.977	0.899	0.834	0.877	0.870	0.899
5	0.945	0.871	0.867	0.962	0.911	0.950	0.990	0.925	0.900	0.940	0.920	0.937
8	0.952	0.882	0.884	0.967	0.926	0.965	0.995	0.935	0.913	0.959	0.939	0.946
11	0.943	0.865	0.851	0.959	0.915	0.945	0.992	0.923	0.885	0.938	0.923	0.935

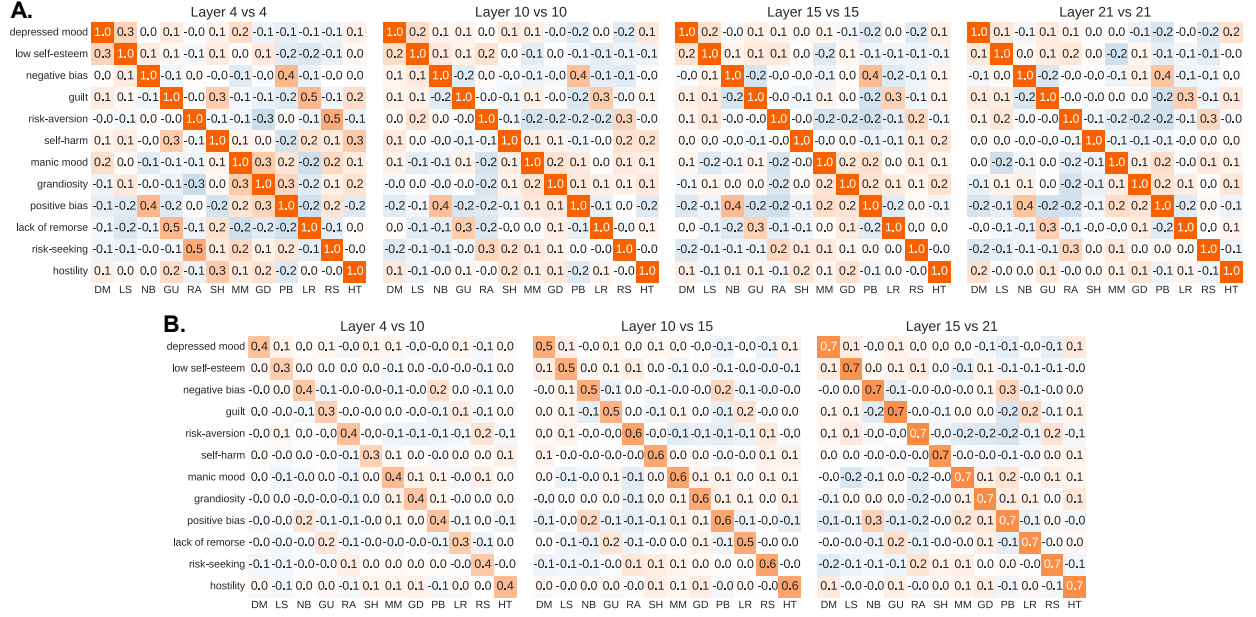


Figure 6: Cosine similarity between S3AE-learned representational state vectors of Llama-3.2-3B.

Table 6: S3AE evaluation: Llama-3.2-3B

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
4	1.093	1.070	1.045	1.122	1.077	1.093	1.141	1.120	1.077	1.082	1.095	1.089
10	1.101	1.061	1.042	1.105	1.089	1.077	1.145	1.110	1.067	1.080	1.075	1.127
15	1.125	1.059	1.044	1.103	1.089	1.072	1.173	1.101	1.067	1.081	1.075	1.114
21	1.120	1.050	1.029	1.084	1.078	1.059	1.166	1.085	1.051	1.064	1.072	1.090
4	0.877	0.806	0.794	0.945	0.854	0.901	0.989	0.877	0.822	0.904	0.847	0.899
10	0.921	0.854	0.867	0.954	0.902	0.972	0.999	0.906	0.908	0.972	0.925	0.935
15	0.911	0.850	0.852	0.950	0.913	0.975	0.999	0.914	0.905	0.982	0.930	0.931
21	0.918	0.829	0.824	0.956	0.904	0.963	0.998	0.906	0.880	0.976	0.911	0.915
4	0.949	0.877	0.861	0.961	0.910	0.946	0.993	0.929	0.890	0.936	0.911	0.941
10	0.969	0.909	0.915	0.983	0.943	0.983	0.999	0.948	0.941	0.983	0.953	0.966
15	0.974	0.913	0.922	0.985	0.948	0.987	1.000	0.952	0.946	0.990	0.960	0.970
21	0.974	0.908	0.911	0.984	0.946	0.985	1.000	0.952	0.942	0.987	0.956	0.967

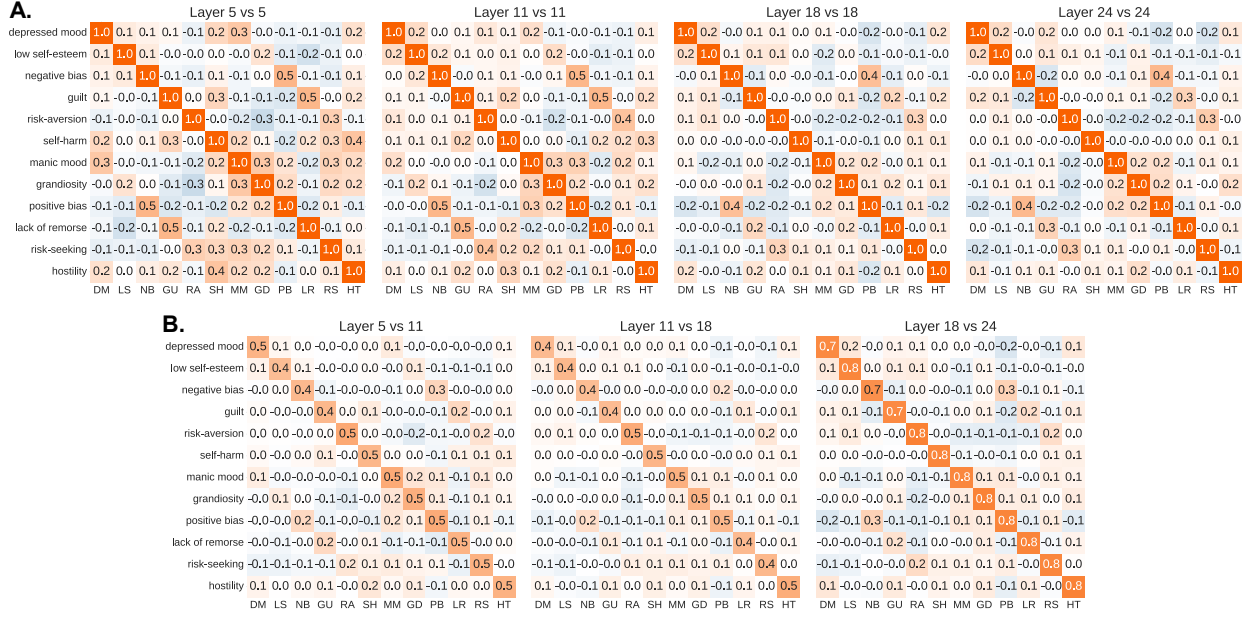


Figure 7: Cosine similarity between S3AE-learned representational state vectors of Llama-3.1-8B.

Table 7: S3AE evaluation: Llama-3.1-8B.

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
5	1.118	1.057	1.046	1.122	1.074	1.125	1.199	1.114	1.071	1.089	1.068	1.096
11	1.136	1.089	1.056	1.173	1.104	1.131	1.219	1.175	1.086	1.102	1.105	1.152
18	1.118	1.057	1.034	1.090	1.086	1.067	1.170	1.090	1.062	1.075	1.072	1.117
24	1.126	1.051	1.034	1.082	1.078	1.058	1.182	1.088	1.054	1.065	1.072	1.108
5	0.913	0.822	0.850	0.962	0.861	0.961	0.996	0.896	0.875	0.945	0.887	0.940
11	0.940	0.883	0.910	0.976	0.921	0.988	1.000	0.928	0.930	0.989	0.942	0.953
18	0.907	0.858	0.860	0.952	0.917	0.989	0.999	0.914	0.922	0.987	0.935	0.928
24	0.929	0.846	0.857	0.956	0.910	0.965	0.998	0.917	0.904	0.982	0.929	0.942
5	0.967	0.904	0.903	0.978	0.935	0.979	0.999	0.949	0.930	0.971	0.944	0.964
11	0.980	0.925	0.936	0.989	0.953	0.994	1.000	0.961	0.959	0.995	0.968	0.979
18	0.981	0.926	0.934	0.990	0.956	0.996	1.000	0.965	0.964	0.997	0.970	0.981
24	0.982	0.926	0.931	0.992	0.955	0.996	1.000	0.965	0.960	0.997	0.970	0.979

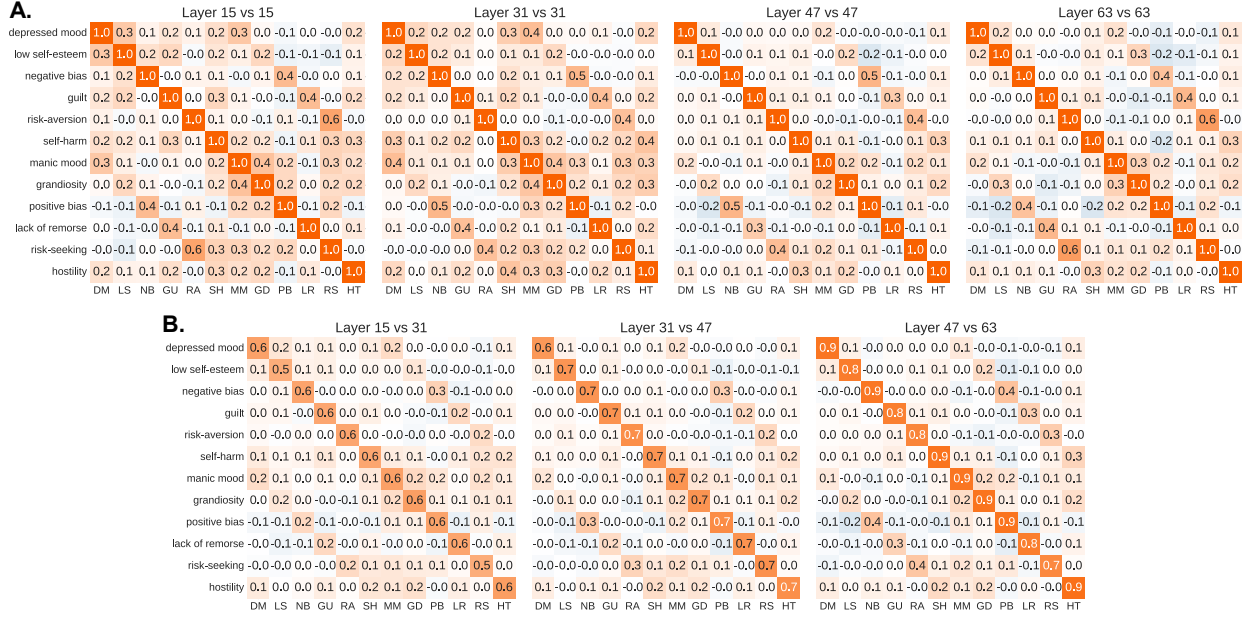


Figure 8: Cosine similarity between S3AE-learned representational state vectors of Llama-3.3-70B.

Table 8: S3AE evaluation: Llama-3.3-70B

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
15	1.048	1.032	1.030	1.061	1.070	1.043	1.088	1.039	1.043	1.017	1.073	1.038
31	1.100	1.057	1.037	1.110	1.055	1.102	1.192	1.094	1.054	1.072	1.060	1.081
47	1.100	1.045	1.032	1.108	1.065	1.083	1.167	1.084	1.055	1.065	1.065	1.087
63	1.102	1.049	1.036	1.105	1.087	1.088	1.178	1.084	1.046	1.064	1.104	1.074
15	0.866	0.835	0.903	0.944	0.911	0.959	0.964	0.777	0.905	0.792	0.913	0.787
31	0.914	0.909	0.879	0.982	0.844	0.987	0.995	0.883	0.909	0.969	0.932	0.925
47	0.920	0.870	0.794	0.978	0.908	0.980	0.998	0.891	0.890	0.964	0.919	0.935
63	0.931	0.877	0.866	0.972	0.910	0.980	0.997	0.874	0.816	0.931	0.909	0.922
15	0.956	0.915	0.923	0.977	0.945	0.985	0.996	0.948	0.944	0.967	0.948	0.951
31	0.976	0.931	0.939	0.994	0.938	0.994	1.000	0.962	0.953	0.985	0.977	0.966
47	0.981	0.935	0.889	0.996	0.964	0.994	1.000	0.974	0.946	0.982	0.978	0.971
63	0.981	0.926	0.933	0.993	0.958	0.991	1.000	0.948	0.904	0.964	0.964	0.965

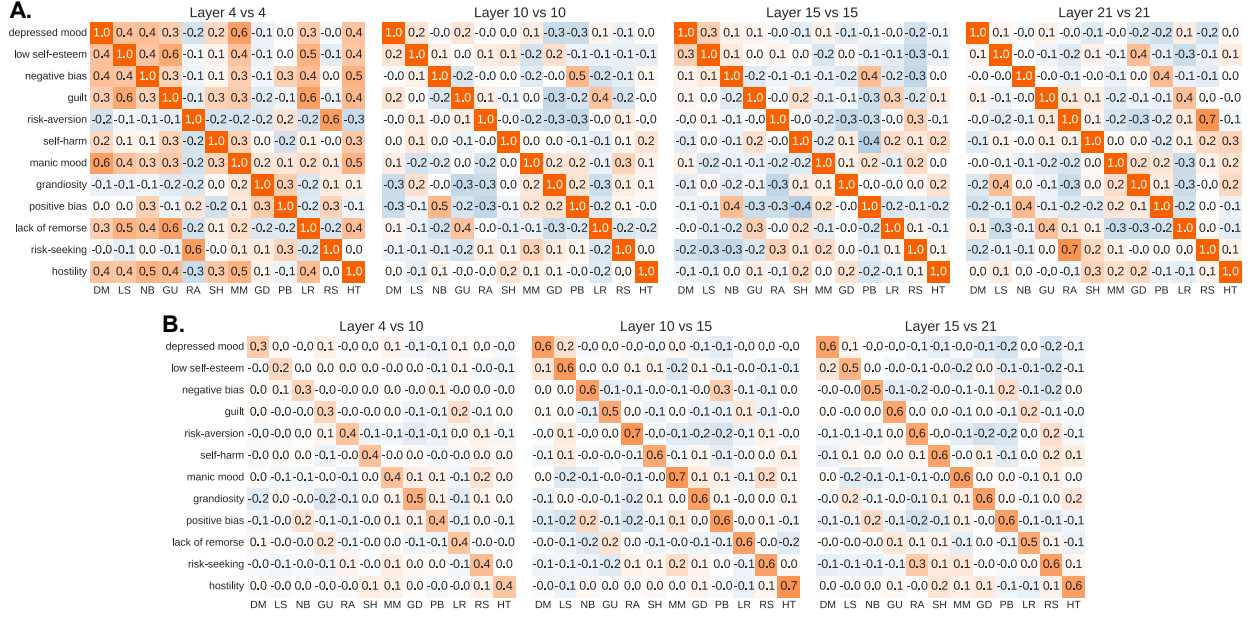


Figure 9: Cosine similarity between S3AE-learned representational state vectors of Qwen3-0.6B.

Table 9: S3AE evaluation: Qwen3-0.6B.

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
4	1.092	1.144	1.045	1.105	1.085	1.071	1.181	1.081	1.048	1.081	1.081	1.074
10	1.061	1.050	1.030	1.087	1.064	1.042	1.111	1.075	1.053	1.055	1.045	1.087
15	1.079	1.048	1.034	1.128	1.089	1.080	1.129	1.079	1.066	1.054	1.084	1.076
21	1.071	1.070	1.031	1.106	1.142	1.070	1.152	1.117	1.066	1.061	1.159	1.084
4	0.744	0.641	0.497	0.753	0.697	0.709	0.905	0.732	0.590	0.586	0.702	0.681
10	0.797	0.716	0.671	0.863	0.781	0.768	0.942	0.802	0.715	0.754	0.780	0.809
15	0.848	0.719	0.704	0.889	0.837	0.825	0.952	0.807	0.787	0.793	0.826	0.830
21	0.854	0.732	0.732	0.919	0.848	0.848	0.970	0.813	0.779	0.826	0.820	0.838
4	0.843	0.764	0.679	0.866	0.816	0.813	0.944	0.842	0.726	0.768	0.789	0.825
10	0.894	0.808	0.775	0.919	0.868	0.863	0.961	0.875	0.808	0.843	0.850	0.874
15	0.911	0.827	0.814	0.932	0.895	0.893	0.971	0.893	0.854	0.864	0.887	0.888
21	0.915	0.831	0.810	0.935	0.870	0.908	0.981	0.892	0.853	0.893	0.874	0.896

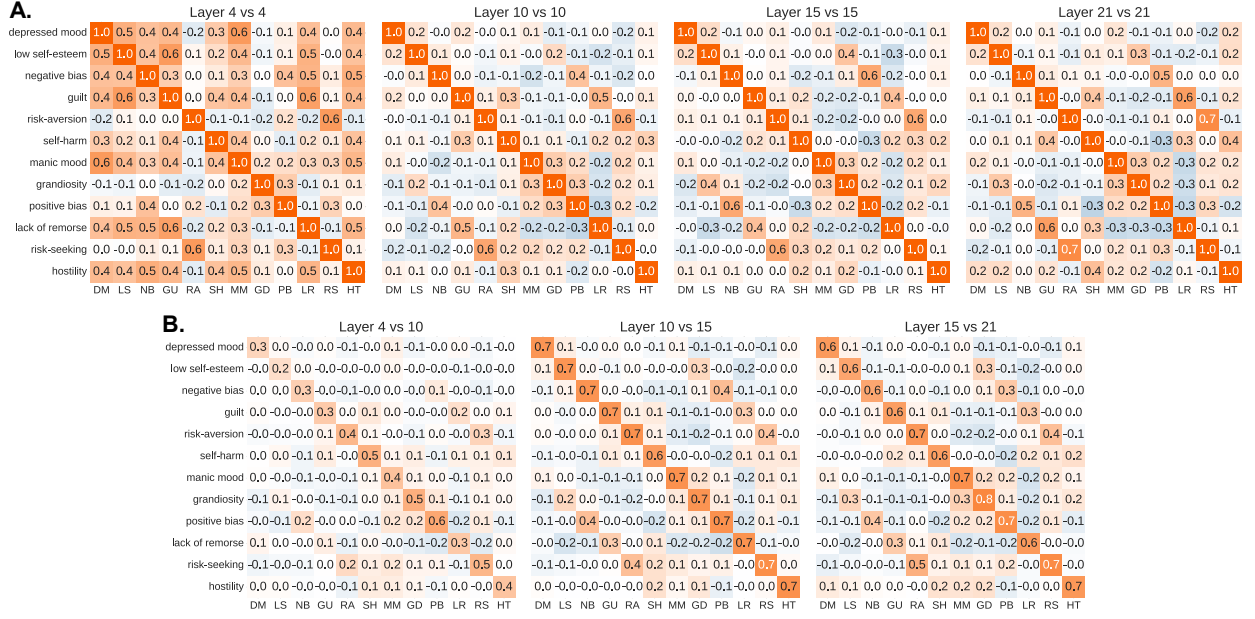


Figure 10: Cosine similarity between S3AE-learned representational state vectors of Qwen3-1.7B.

Table 10: S3AE evaluation: Qwen3-1.7B.

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
4	1.099	1.163	1.046	1.117	1.069	1.068	1.159	1.065	1.038	1.076	1.069	1.079
10	1.072	1.058	1.045	1.168	1.104	1.126	1.161	1.132	1.090	1.078	1.123	1.109
15	1.099	1.080	1.057	1.157	1.131	1.098	1.180	1.158	1.097	1.075	1.170	1.117
21	1.099	1.084	1.042	1.202	1.173	1.128	1.183	1.133	1.097	1.105	1.211	1.135
4	0.809	0.692	0.583	0.789	0.747	0.790	0.922	0.764	0.648	0.669	0.760	0.756
10	0.876	0.784	0.814	0.954	0.880	0.926	0.982	0.847	0.835	0.917	0.872	0.905
15	0.885	0.806	0.838	0.958	0.898	0.945	0.995	0.892	0.886	0.935	0.905	0.895
21	0.899	0.800	0.785	0.960	0.872	0.939	0.987	0.898	0.866	0.928	0.901	0.924
4	0.888	0.808	0.754	0.909	0.853	0.872	0.961	0.877	0.787	0.836	0.840	0.872
10	0.953	0.884	0.874	0.971	0.920	0.959	0.997	0.931	0.908	0.953	0.930	0.944
15	0.955	0.887	0.895	0.970	0.930	0.964	0.997	0.932	0.924	0.961	0.940	0.945
21	0.933	0.833	0.848	0.895	0.821	0.933	0.987	0.908	0.891	0.934	0.877	0.928

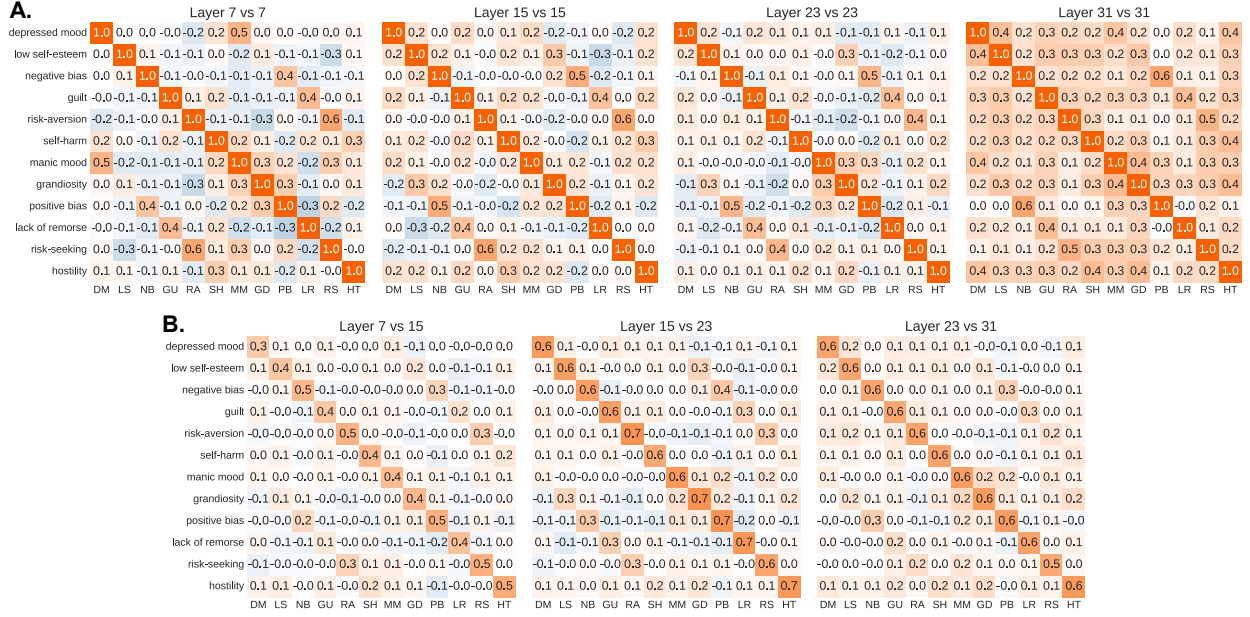


Figure 11: Cosine similarity between S3AE-learned representational state vectors of Qwen3-14B.

Table 11: S3AE evaluation: Qwen3-14B.

Layer	DM	LS	NB	GU	RA	SH	MM	GD	PB	LR	RS	HT
7	1.068	1.048	1.025	1.072	1.069	1.070	1.132	1.077	1.046	1.053	1.073	1.055
15	1.126	1.087	1.051	1.139	1.116	1.116	1.176	1.123	1.084	1.081	1.125	1.153
23	1.121	1.085	1.045	1.146	1.116	1.098	1.183	1.155	1.077	1.089	1.106	1.142
31	1.171	1.068	1.045	1.166	1.099	1.108	1.259	1.131	1.084	1.088	1.105	1.170
7	0.823	0.762	0.699	0.905	0.812	0.843	0.970	0.819	0.761	0.831	0.814	0.828
15	0.931	0.858	0.870	0.965	0.903	0.966	0.998	0.905	0.905	0.972	0.918	0.933
23	0.917	0.862	0.879	0.969	0.920	0.978	0.997	0.904	0.933	0.993	0.935	0.948
31	0.926	0.836	0.869	0.960	0.911	0.978	0.998	0.910	0.912	0.980	0.932	0.951
7	0.928	0.853	0.815	0.944	0.890	0.927	0.986	0.909	0.851	0.901	0.893	0.912
15	0.973	0.912	0.915	0.982	0.942	0.986	1.000	0.953	0.943	0.987	0.957	0.969
23	0.979	0.925	0.937	0.990	0.957	0.996	1.000	0.961	0.966	0.998	0.972	0.978
31	0.971	0.914	0.920	0.981	0.947	0.990	1.000	0.954	0.954	0.992	0.963	0.972