Appendix B1 - Prompt-Based EvaluaBon of Pretrained LLMs (RQ1.1)

Model	Type	Prompt Type	QWK (Avg)	Valid Output %	MAE (Avg)	Exact Match % (Avg)	Avg Inference Time (s)	cohesion_MAE	cohesion_QWK	cohesion_Exact Match %	syntax_MAE	syntax_QWK	syntax_Exact	vocabulary_M	vocabulary_QWK	vocabulary_E	phraseology	phraseology	phraseology	grammar_MAE	grammar_QWK	grammar_Exa	conventions	conventions_QW	conventions_Exact Match %
Phi-4:14B	Base	Zero-Shot	0.4068	100	0.6882	22.19	5.93	0.7497	0.3266	20.87	0.532	0.4736	28.04	0.7542	0.3279	19.08	0.6511	0.4452	22.54	0.7439	0.4235	21.77	0.6985	0.4437	20.87
Phi-4:14B	Base	Few-Shot	0.3571	99.87	0.73	21.26	20.37	0.7724	0.3511	20.77	0.6923	0.3845	20.64	0.8801	0.3007	18.33	0.6538	0.3999	23.33	0.6647	0.3436	23.59	0.7167	0.3625	20.9
Mistral:Instruct	Instruct	Few-Shot	0.2428	100	0.7477	21.7	9.42	0.9981	0.192	13.32	0.7196	0.2547	22.28	0.5762	0.3148	29.83	0.733	0.2439	21.64	0.7241	0.2227	21.9	0.735	0.2288	21.25
Mistral:Instruct	Instruct	Zero-Shot	0.2241	100	0.5896	25.59	3.22	0.5755	0.0957	25.22	0.4481	0.4515	33.16	0.7945	0.0995	14.72	0.628	0.2165	23.82	0.5877	0.2453	25.22	0.5038	0.2359	31.37
GPT-4 Turbo	Instruct	Few-Shot	0.222	100	0.9806	8.26	1.82 (OpenAI server)	0.9686	0.2281	9.09	0.7599	0.3077	14.72	0.9808	0.2032	5.76	0.9462	0.2103	9.6	1.1453	0.1743	4.99	1.0826	0.2085	5.38
Llama 3:8B	Base	Zero-Shot	0.2122	100	0.8533	17.95	3.34	0.9577	0.0121	13.7	0.5698	0.3285	24.46	1.4417	0.0774	2.18	0.6511	0.1858	24.97	0.749	0.3379	23.56	0.7503	0.3318	18.82
GPT-3.5 Turbo	Instruct	Zero-Shot	0.1714	100	0.9558	11.72	0.74 (OpenAI server)	0.9789	0.2095	9.86	0.8227	0.1986	14.98	1.1645	0.0982	4.1	0.8342	0.2041	15.49	0.9283	0.1656	16.13	1.0064	0.1527	9.73
Mistral	Base	Zero-Shot	0.1524	0.9	0.869	14.29	12.93	0.7857	0.186	42.86	0.6429	0.4134	14.29	1.2143	-0.0606	14.29	0.9286	0.1714	0	0.9286	0.0181	0	0.7143	0.186	14.29
Llama 3:8B	Base	Few-Shot	0.1462	99.62	0.9807	14.63	9.58	0.8201	0.1581	17.87	0.8978	0.1792	15.17	1.5945	0.0639	2.31	0.6986	0.2243	20.57	0.8046	0.14	20.95	1.0688	0.1117	10.93
GPT-3.5 Turbo	Instruct	Few-Shot	0.1413	100	0.8567	15.79	0.80 (OpenAI server)	1.0384	0.0763	11.27	0.7055	0.1928	18.69	0.9494	0.1112	10.12	0.7401	0.1672	20.61	0.8156	0.173	18.95	0.8912	0.1275	15.11
Gemma 3:1B	Base	Few-Shot	0.1176	99.49	1.2061	11.03	2.47	1.2761	0.1598	6.82	1.2169	0.0898	12.87	1.1898	0.1641	8.88	1.1306	0.1147	13.51	1.2104	0.0584	11.71	1.213	0.1186	12.36
GPT-4 Turbo	Instruct	Zero-Shot	0.0961	100	1.4288	1.09	1.99 (OpenAI server)	1.3771	0.107	1.79	1.2375	0.1341	1.66	1.3713	0.0884	0.77	1.3828	0.107	1.41	1.614	0.0638	0.38	1.5903	0.0764	0.51
Gemma 3:4B	Base	Few-Shot	0.0901	96.93	0.9093	15.48	6.01	1.1598	0.0962	7.93	0.7252	0.1186	17.97	1.1711	0.0287	6.21	0.5053	0.1349	30.25	0.959	0.0524	15.72	0.9353	0.1098	14.8
Mistral	Base	Few-Shot	0.0801	10.62	0.7431	20.14	na/frozen	0.7708	-0.0897	12.5	0.625	-0.0938	25	0.9792	0.2284	8.33	0.7083	-0.0696	33.33	0.7292	0.0945	20.83	0.6458	0.4109	20.83
Gemma 3:4B	Base	Zero-Shot	0.0474	99.87	0.9149	14.91	2.51	1.0551	0.0573	10.38	0.6936	0.028	20.9	1.1878	0.0354	3.72	0.5077	0.0629	29.74	1.0173	0.0272	14.1	1.0276	0.0738	10.64
Gemma 3:1B	Base	Zero-Shot	0.0341	97.06	0.7225	21.42	1.19	1.0369	0.0214	11.74	0.4921	0.0669	32.19	0.9255	-0.019	11.35	0.5092	0.0294	29.68	0.6583	0.0725	22.43	0.7131	0.0333	21.11
Llama 3.2:3B	Base	Zero-Shot	0.0338	100	0.9328	15.81	1.99	1.0999	0.0056	10.37	0.7247	0.0472	19.08	1.4731	0.0146	3.33	0.7618	0.0279	22.79	0.7049	0.0415	21.64	0.8323	0.0662	17.67
Llama 3.2:3B-fp16	Instruct	Few-Shot	0.0273	97.44	1.1264	11.32	6.99	1.433	-0.0092	6.7	1.2122	0.0425	10.64	0.9474	0.0349	13.4	1.0861	0.0296	12.61	1.0578	0.022	12.61	1.0217	0.0442	11.96
Llama 3.2:3B-fp16	Instruct	Zero-Shot	0.0203	100	0.9534	14.66	2.47	1.0736	0.0099	10.37	0.71	0.0252	22.28	1.5307	0.0165	2.82	0.8009	0.0084	18.31	0.7234	0.0017	17.93	0.8816	0.0604	16.26
Llama 3.2:1B-fp16	Instruct	Few-Shot	0.0192	48.14	0.8081	18.04	3.22	0.7979	-0.01	14.89	0.5984	0.0786	29.26	1.1011	0.0584	5.32	0.7234	0.0009	22.07	0.754	0.0274	18.35	0.8737	-0.0402	18.35
Llama 3.2:3B	Base	Few-Shot	0.0188	98.34	1.0719	13.24	6.14	1.4811	-0.0089	5.99	1.1562	-0.0027	11.59	0.8887	0.0873	15.62	0.9909	-0.0078	15.23	0.9824	0.0077	13.41	0.9323	0.0371	17.58
Llama 3.2:1B-fp16	Instruct	Zero-Shot	-0.0032	48.78	0.7986	19.34	1.3	0.7231	-0.0026	19.69	0.8858	-0.0453	19.42	0.9593	0.0381	12.86	0.7585	0.0103	19.69	0.7257	0.0045	21.78	0.7388	-0.0238	22.57
Llama 3.2:1B	Base	Zero-Shot	-0.0075	43.66	0.8385	18.96	2.9	0.8299	-0.0242	14.37	0.6085	0.0122	29.91	1.1848	-0.0557	8.8	0.7038	0.0369	21.11	0.7771	0.0016	21.41	0.9267	-0.0156	18.18
Llama 3.2:1B	Base	Few-Shot	-0.0098	43.02	0.7803	20.44	1.16	0.7589	-0.0115	16.96	0.7545	-0.0871	25	1.0208	0.0279	12.2	0.7485	-0.0255	22.02	0.6414	0.0539	25.3	0.7574	-0.0163	21.13
DeepSeek-R1:1.5B DeepSeek-R1:7B DeepSeek-R1:14B	Instruct	Zero-Shot Few-Shot	NAN	12.71 / 17.83 18.99 / - - / -	NAN	NAN	25.4	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN

Appendix B2 - Fine-Tuning Small-Scale LLMs for AEA (RQ1.2)

Model	Prompt	Strategy	Overall Avg M	Overall MAE S	Overall Avg Q	Overall QWK	cohesion Avg	cohesion MA	cohesion Avg	cohesion QW	syntax Avg M	syntax MAE S	syntax Avg Q	syntax QWK S	vocabulary A	vocabulary M	vocabulary A	vocabulary Q	phraseology	phraseology	phraseology	phraseology	grammar Avg	grammar MA	grammar Avg	grammar QW	conventions	conventions	conventions	conventions	QWK Std
Llama_3.2_1	Yes	Top 6 layers	0.377	0.007	0.655	0.012	0.397	0.016	0.627	0.014	0.37	0.018	0.664	0.021	0.346	0.02	0.612	0.044	0.379	0.013	0.648	0.023	0.404	0.02	0.654	0.037	0.364	0.009	0.674	0.011	
Llama_3.2_1	Yes	Top 4 layers	0.369	0.01	0.65	0.01	0.391	0.012	0.61	0.017	0.374	0.029	0.643	0.032	0.338	0.02	0.618	0.037	0.364	0.019	0.652	0.019	0.393	0.021	0.649	0.035	0.356	0.01	0.672	0.024	
Llama_3.2_1	No	Top 4 layers	0.386	0.013	0.641	0.006	0.412	0.021	0.596	0.024	0.387	0.012	0.622	0.019	0.35	0.037	0.601	0.038	0.379	0.018	0.657	0.028	0.409	0.023	0.647	0.028	0.378	0.017	0.682	0.024	
Llama_3.2_1	Yes	Top 4 layers	0.385	0.018	0.641	0.006	0.405	0.017	0.607	0.015	0.383	0.029	0.641	0.017	0.346	0.032	0.615	0.018	0.383	0.015	0.623	0.038	0.405	0.012	0.655	0.024	0.389	0.032	0.66	0.02	
Llama_3.2_1	No	Top 4 layers	0.392	0.02	0.64	0.012	0.408	0.022	0.604	0.046	0.38	0.029	0.651	0.016	0.357	0.038	0.595	0.052	0.398	0.033	0.643	0.009	0.418	0.042	0.649	0.024	0.389	0.018	0.647	0.025	
Llama_3.2_1	Yes	Top 6 layers	0.386	0.027	0.639	0.01	0.415	0.032	0.592	0.017	0.383	0.031	0.625	0.008	0.344	0.028	0.625	0.012	0.385	0.016	0.64	0.022	0.412	0.019	0.65	0.023	0.379	0.043	0.665	0.02	
Llama_3.2_1	No	Top 6 layers	0.387	0.007	0.638	0.016	0.414	0.027	0.605	0.013	0.388	0.018	0.62	0.034	0.347	0.014	0.614	0.022	0.395	0.023	0.629	0.026	0.402	0.015	0.655	0.018	0.376	0.012	0.664	0.041	
Llama_3.2_1	No	Top 6 layers	0.382	0.013	0.637	0.012	0.393	0.012	0.599	0.033	0.377	0.019	0.629	0.017	0.34	0.016	0.61	0.025	0.401	0.059	0.618	0.055	0.405	0.017	0.663	0.031	0.372	0.031	0.667	0.029	
Llama_3.2_1	Yes	Last layer onl	0.397	0.005	0.58	0.008	0.411	0.012	0.554	0.018	0.389	0.008	0.573	0.019	0.374	0.03	0.516	0.047	0.394	0.011	0.579	0.007	0.431	0.009	0.569	0.02	0.386	0.005	0.62	0.02	
Llama_3.2_1	Yes	Last layer onl	0.401	0.005	0.569	0.007	0.421	0.012	0.528	0.018	0.39	0.013	0.568	0.016	0.381	0.027	0.502	0.034	0.398	0.006	0.555	0.013	0.431	0.009	0.558	0.025	0.383	0.007	0.626	0.018	
Llama_3.2_1	No	Last layer onl	0.445	0.004	0.555	0.013	0.457	0.008	0.535	0.01	0.443	0.012	0.537	0.02	0.41	0.014	0.501	0.029	0.445	0.007	0.552	0.006	0.471	0.022	0.564	0.04	0.443	0.011	0.585	0.019	
Llama_3.2_1	No	Last layer onl	0.454	0.006	0.541	0.012	0.473	0.011	0.52	0.016	0.441	0.01	0.534	0.025	0.414	0.02	0.501	0.026	0.465	0.018	0.525	0.016	0.485	0.01	0.546	0.018	0.446	0.008	0.568	0.023	

Appendix B3 .1 - Benchmarking Frozen Embeddings for TradiBonal ML Models (RQ2.1 and RQ2.2)

Model	Doc2vec Avg MAE	Doc2vec Avg QWK	MiniLM MAE	MiniLM Avg QWK	RoBERTa MAE	RoBERTa Avg QWK	MPNet MAE	MPNet Avg QWK	LLaMA MAE	LLaMA Avg QWK
Linear Regression	0.343136928	0.535534063	0.377656107	0.500458588	0.30771869	0.687384333	0.358092359	0.579691617	0.44976	0.519431886
Support Vector Regression	0.35016593	0.531632222	0.359627376	0.493581289	0.272820942	0.702774521	0.336232211	0.561340792	0.27052	0.714137881
Gradient Boosting	0.361162719	0.468697167	0.390694522	0.354752106	0.286117399	0.68781797	0.363462223	0.460835353	0.283305	0.683188186
Random Forest	0.371388979	0.409306306	0.405906034	0.266351213	0.293276401	0.646639866	0.376630966	0.393778461	0.29315	0.636994624
XGBoost	0.378548308	0.453967904	0.408463742	0.359974702	0.3079831	0.657804017	0.386085552	0.444465549	0.304651	0.654485722
ANN	0.423551133	0.458469535	0.405263708	0.406468362	0.29008045	0.684495981	0.387883673	0.464776889	0.277676	0.706493338

Appendix B3 .2 - Benchmarking Frozen Embeddings for TradiBonal ML Models (RQ2.3)

Model	Dimension	Doc2Vec Avg MAE	Doc2Vec Avg QWK	MiniLM Avg MAE	MiniLM Avg QWK	MPNet Avg MAE	MPNet Avg QWK	RoBERTa Avg MAE	RoBERTa Avg QWK	LLaMA Avg MAE	LLaMA Avg QWK
Linear Regression	cohesion	0.447970133	0.413764693	0.473160152	0.406622333	0.475711981	0.450162378	0.435568327	0.548198456	0.641135805	0.368070981
Linear Regression	syntax	0.412173325	0.466667759	0.435696694	0.44276858	0.440035701	0.493560739	0.392225456	0.607379617	0.585651292	0.42992279
Linear Regression	vocabulary	0.366016992	0.445701481	0.440046807	0.392013946	0.39772385	0.475844725	0.36115635	0.583164558	0.559319523	0.365468827
Linear Regression	phraseology	0.423164562	0.462899149	0.460622793	0.430634907	0.438507054	0.517035809	0.40130474	0.602613167	0.617108276	0.403219695
Linear Regression	grammar	0.443621653	0.490196013	0.520075256	0.39807437	0.484782772	0.502206084	0.42572227	0.620852554	0.66466342	0.394368672
Linear Regression	conventions	0.432756498	0.473931354	0.442725533	0.495551514	0.457426189	0.508575664	0.38864767	0.641550266	0.600999337	0.434088736
Support Vector Regression	cohesion	0.445542098	0.42929143	0.450146332	0.396790799	0.440429458	0.437676497	0.387754815	0.566284994	0.377779738	0.595265131
Support Vector Regression	syntax	0.411530183	0.477015163	0.428150467	0.422368504	0.409612187	0.475652344	0.347480018	0.622241909	0.344925086	0.634954911
Support Vector Regression	vocabulary	0.36716658	0.457755866	0.375605988	0.418143343	0.369853962	0.442880931	0.312577371	0.595350187	0.310402968	0.605955151
Support Vector Regression	phraseology	0.421115423	0.481939818	0.436718569	0.424622014	0.40935627	0.512584775	0.354892325	0.633163375	0.35310482	0.638589819
Support Vector Regression	grammar	0.452311916	0.487383748	0.471618113	0.410472442	0.438504931	0.513030323	0.375226276	0.642355744	0.376245047	0.643630651
Support Vector Regression	conventions	0.437612566	0.466508997	0.434419392	0.459226015	0.42879018	0.483616791	0.355282326	0.645819795	0.344156843	0.671911207
Gradient Boosting	cohesion	0.453209343	0.357604712	0.47264652	0.261748871	0.460757366	0.329976177	0.397729403	0.552117451	0.398107319	0.563017354
Gradient Boosting	syntax	0.417793718	0.421119538	0.450008003	0.287592559	0.430708175	0.366408858	0.359882804	0.611924679	0.363205162	0.610036
Gradient Boosting	vocabulary	0.369470167	0.394441079	0.392740068	0.289185368	0.381362423	0.361402912	0.321141063	0.593541414	0.320120169	0.593190616
Gradient Boosting	phraseology	0.430194543	0.417693842	0.454869624	0.319381098	0.42508076	0.426536205	0.366273726	0.622546054	0.368959638	0.6049479
Gradient Boosting	grammar	0.460497366	0.429770334	0.498720411	0.293551472	0.454998808	0.441289884	0.387244123	0.636143703	0.394912348	0.617469105
Gradient Boosting	conventions	0.445537199	0.407308196	0.467270613	0.315466132	0.453593791	0.365753625	0.370748286	0.633098463	0.357453953	0.654069283
Random Forest	cohesion	0.458837248	0.308211604	0.478525606	0.226296277	0.46599478	0.280574655	0.401434903	0.506294864	0.402452369	0.516975137
Random Forest	syntax	0.424570395	0.364678052	0.455759865	0.217779598	0.440040274	0.311119721	0.361159616	0.571026637	0.372026895	0.556817208
Random Forest	vocabulary	0.374454276	0.351351566	0.395168755	0.260519741	0.390054646	0.301690677	0.323314487	0.550909309	0.324340281	0.547849733
Random Forest	phraseology	0.434286125	0.375292884	0.465608046	0.240229804	0.441957126	0.356688018	0.367809233	0.584513854	0.375223336	0.558385875
Random Forest	grammar	0.469188282	0.371541345	0.508818957	0.224016355	0.462286014	0.38889693	0.389289669	0.605409228	0.407057746	0.554286153
Random Forest	conventions	0.453848076	0.346924846	0.478651197	0.243682232	0.470339503	0.29444652	0.374454113	0.597425902	0.364358834	0.60726478
XGBoost	cohesion	0.484791101	0.35162735	0.502428361	0.279961505	0.481211682	0.322752464	0.42150722	0.528832053	0.423163092	0.533698622
XGBoost	syntax	0.44873266	0.413942338	0.479288624	0.303496671	0.458956633	0.373282788	0.381229647	0.592460157	0.39171411	0.576734978
XGBoost	vocabulary	0.402450735	0.386296313	0.420609467	0.301221661	0.399645602	0.372640792	0.344667046	0.56448694	0.341469298	0.565898162
XGBoost	phraseology	0.450143229	0.418874986	0.477497852	0.333950638	0.449754208	0.420862396	0.395677488	0.582170688	0.394013614	0.576577776
XGBoost	grammar	0.498976982	0.399271388	0.530041352	0.289708157	0.49041035	0.413486843	0.420348813	0.603936567	0.421883666	0.59203705
XGBoost	conventions	0.478391523	0.386470039	0.49105333	0.321485287	0.484655711	0.353608191	0.390821909	0.609538938	0.389154442	0.617181925
ANN	cohesion	0.68908242	0.0915574	2.127077638	0	2.114801423	-1.96839E-07	0.891567615	-0.095441881	0.497706702	0.195530041
ANN	syntax	0.564432326	0.346456802	0.77936179	-0.032667837	0.774987506	-0.032735285	0.552674153	0.038427917	0.40961578	0.422139152
ANN	vocabulary	0.51470033	0.330191515	0.566472156	0.001361171	0.525702182	0.067022694	0.437869137	0.233506072	0.344153903	0.520192004
ANN	phraseology	0.509588996	0.409226587	0.55918838	0.096923334	0.522757086	0.218271696	0.417411066	0.483959019	0.377140841	0.583624849
ANN	grammar	0.517642649	0.433798707	0.561489353	0.1787023	0.511381564	0.351174947	0.41319667	0.585604265	0.398490951	0.607519387
ANN	conventions	0.523398432	0.391933775	0.506900145	0.292833848	0.493993036	0.342920314	0.386349962	0.596135897	0.362058677	0.653601974