Table 5: Llama-3-8B-Instruct Faithfulness for Different In-Context Learning Approaches

Approach	**AQuA**		**LogiQA**		**TruthfulQA**	
	Accuracy	Faithfulness	Accuracy	Faithfulness	Accuracy	Faithfulness
ZS-CoT	0.43 ± 0.05	0.61 ± 0.02	0.38 ± 0.05	0.41 ± 0.03	0.56 ± 0.05	0.23 ± 0.03
GTA	0.48 ± 0.05	0.64 ± 0.02	$\mathbf{0.47} \pm 0.05$	0.40 ± 0.03	0.57 ± 0.05	$\mathbf{0.26} \pm 0.03$
DU	0.44 ± 0.05	$\mathbf{0.65} \pm 0.02$	0.43 ± 0.05	0.38 ± 0.03	0.59 ± 0.05	0.24 ± 0.03
DU^c	$\mathbf{0.55} \pm 0.05$	0.63 ± 0.02	$\mathbf{0.47} \pm 0.05$	$\mathbf{0.43} \pm 0.03$	$\mathbf{0.65} \pm 0.05$	0.23 ± 0.03
DF	0.43 ± 0.05	0.63 ± 0.02	0.35 ± 0.05	0.43 ± 0.03	0.59 ± 0.05	0.24 ± 0.03
DF^c	0.37 ± 0.05	0.64 ± 0.02	0.42 ± 0.05	0.39 ± 0.03	0.64 ± 0.05	0.21 ± 0.03
SU	0.52 ± 0.05	0.62 ± 0.02	0.45 ± 0.05	0.42 ± 0.03	0.60 ± 0.05	0.25 ± 0.03
SU^c	0.45 ± 0.05	0.63 ± 0.02	0.44 ± 0.05	0.42 ± 0.03	0.57 ± 0.05	0.23 ± 0.03
SF	0.43 ± 0.05	0.64 ± 0.02	0.41 ± 0.05	0.39 ± 0.03	0.58 ± 0.05	0.23 ± 0.03
SF^c	0.42 ± 0.05	0.65 ± 0.02	$\mathbf{0.47} \pm 0.05$	0.41 ± 0.03	0.58 ± 0.05	0.23 ± 0.03

Table 6: GPT-3.5-Turbo p-values of Faithfulness for Different Fine-tuning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.1946	0.2247	0.0000	0.0005	0.6353	0.1101
DU^c	0.0718	0.0974	0.2141	0.0597	0.3573	0.4600
DF	0.3640	0.2610	0.0000	0.0000	0.3607	0.4292
DF^c	0.9523	0.7473	0.0917	0.0201	0.2364	0.6090
SU	0.4740	0.7740	0.0014	0.0010	0.0473	0.9173
SU^c	0.8063	0.9671	0.0088	0.0028	0.2353	0.5102
SF	0.8789	0.6707	0.0000	0.0000	0.0579	0.9934
SF^c	0.8324	0.6255	0.0006	0.0001	0.1071	0.7794

Table 7: Llama-3-8B-Instruct p-values for Different Fine-tuning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.4325	0.0062	0.0027	0.0000	0.1835	0.1687
DU^c	0.0845	0.0000	0.5589	0.0000	0.7541	0.7380
DF	0.3175	0.0103	0.1958	0.0000	0.0130	0.0194
DF^c	0.6068	0.0011	0.3946	0.0000	0.0580	0.0639
SU	0.0020	0.4670	0.1636	0.0000	0.1311	0.1476
SU^c	0.9323	0.0020	0.1537	0.0000	0.7327	0.6844
SF	0.9893	0.0003	0.2321	0.0000	0.2940	0.2954
SF^c	0.6049	0.0012	0.3319	0.0000	0.1527	0.2178

Table 8: GPT-4 p-values for Different In-Context Learning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.3089	0.8058	0.1395	0.7462	0.6632	0.2648
DU^c	0.6307	0.1890	0.4525	0.3024	0.2392	0.5765
DF	0.5638	0.7929	0.9936	0.1062	0.0048	0.0489
DF^c	0.1322	0.4820	0.3382	0.4337	0.0250	0.2369
SU	0.6778	0.7624	0.0509	0.8104	0.0063	0.0572
SU^c	0.3145	0.7599	0.3932	0.3125	0.0297	0.2525
SF	0.2818	0.6367	0.9038	0.0491	0.0478	0.5067
SF^c	0.2677	0.5417	0.0037	0.2679	0.0011	0.0111

Table 9: GPT-3.5-Turbo p-values for Different In-Context Learning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.2748	0.1188	0.8037	0.1245	0.4544	0.7770
DU^c	0.8994	0.8539	0.1285	0.8728	0.9518	0.3670
DF	0.1845	0.0451	0.0144	0.3093	0.9840	0.3429
DF^c	0.8065	0.5908	0.0505	0.7696	0.2248	0.7428
SU	0.4238	0.2524	0.0186	0.5463	0.9364	0.3364
SU^c	0.8541	0.5486	0.0323	0.6991	0.5434	0.6946
SF	0.0992	0.0558	0.0093	0.3931	0.8899	0.4127
SF^c	0.2790	0.1526	0.1431	0.8505	0.6492	0.6452

Table 10: Llama-3-8B-Instruct p-values for Different In-Context Learning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.0488	0.5230	0.3320	0.4825	0.8569	0.2464
DU^c	0.2151	0.8029	0.3341	0.2573	0.7859	0.1824
DF	0.2610	0.7089	0.4776	0.3101	0.6582	0.4255
DF^c	0.2190	0.8713	0.6078	0.8081	0.3891	0.0104
SU	0.6302	0.2352	0.7058	0.5657	0.5822	0.5463
SU^c	0.3399	0.4633	0.6561	0.5424	0.9518	0.2365
SF	0.2268	0.8976	0.4570	0.7079	0.9604	0.2342
SF^c	0.1095	0.7537	0.8032	0.6608	0.8679	0.1552