Color coding:
Relative difference between test runs and original experiment

	MODEL	Indicator	Source of info	dataset SST PAPER	SST delta rel%	SST TEST RU AVG	dataset imdb PAPER	imdb delta rel%	imdb TEST RU AVG	dataset Yelp PAPER	Yelp delta rel%	Yelp TEST RU AVG	dataset Amazon PAPER	Amazon delta rel%	Amazon TEST RU AVG	dataset 20News PAPER	20News delta rel%	20News TEST RU AVG	dataset Tweets PAPER	Tweets delta rel%	Tweets TEST RU AVG
TABLE 2	**vanilla lstm**	test accuracy bes	evaluate.json	0,818	-2%	0,803	0,895	0%	0,893	0,956	-1%	0,949	0,937	0%	0,937	0,936	-3%	0,908	0,870	-4%	0,833
Benchmark data		conicity_mean	evaluate.json	0,680	5%	0,713	0,690	-13%	0,602	0,530	1%	0,536	0,500	-9%	0,457	0,770	-1%	0,761	0,770	1%	0,776
		conicity_std	evaluate.json	n/avail	-	0,173	n/avail	-	0,135	n/avail	-	0,121	n/avail	-	0,084	n/avail	-	0,189		-	0,159
	ortho lstm	test accuracy bes	evaluate.json	0,801	-3%	0,776	0,887	0%	0,883	0,960	-2%	0,945	0,930	1%	0,936	0,922	0%	0,919	0,832	1%	0,839
		conicity_mean	evaluate.json	0,280	1%	0,283	0,180	-9%	0,163	0,180	3%	0,186	0,160	7%	0,171	0,230	2%	0,235	0,270	-4%	0,260
		conicity_std	evaluate.json	n/avail	-	0,058	n/avail	-	0,041	n/avail	-	0,044	n/avail	-	0,032	n/avail	-	0,068		-	0,062
	diversity lstm	test accuracy bes	evaluate.json	0,800	0%	0,800	0,885	-1%	0,878	0,954	-2%	0,938	0,929	0%	0,932	0,910	0%	0,908	0,870	-2%	0,854
		conicity_mean	evaluate.json	0,200	-6%	0,188	0,080	11%	0,089	0,060	478%	0,347	0,050	-2%	0,049	0,150	-10%	0,135	0,240	-4%	0,231
		conicity_std	evaluate.json	n/avail	-	0,050	n/avail	-	0,021	n/avail	-	0,014	n/avail	-	0,024	n/avail	-	0,046		-	0,067
FIGURE 3		Indicator	Source of info																		
Fraction of hid	**vanilla lstm - ATT**	median	importance_ranking_		-	0,820	0,75	-27%	0,550	0,90	-6%	0,85	0,90	-	-	0,60	-38%	0,370			1,000
Box plots		1st quartile	importance_ranking_		-	0,480	0,18	-33%	0,120	0,90	-6%	0,85	0,65	-	-	0,60	-38%	0,370			1,000
visual inspection paper		3rd quartile	importance_ranking_		-	1,000	1,00	0%	1,000	1,00	0%	1,00	1,00	-	-	1,00	0%	1,000			1,000
	RANDOM	median	importance_ranking_		-	0,850	0,98	-1%	0,970	0,93	-3%	0,90	1,00	-	-	0,93	5%	0,980			0,930
		1st quartile	importance_ranking_		-	0,850	0,88	-6%	0,830	0,93	-3%	0,90	0,93	-	-	0,93	5%	0,980			0,930
		3rd quartile	importance_ranking_		-	1,000	1,00	0%	1,000	1,00	0%	1,00	1,00	-	-	1,00	0%	1,000			1,000
	orth lstm - ATTN	median	importance_ranking_		-	0,220	0,05	-20%	0,040	0,25	-28%	0,180	0,30	-	-	0,15	-20%	0,120			0,530
		1st quartile	importance_ranking_		-	0,100	0,02	0%	0,020	0,13	-8%	0,120	0,20	-	-	0,07	-43%	0,040			0,190
		3rd quartile	importance_ranking_		-	0,350	0,12	-17%	0,100	0,33	-18%	0,270	0,40	-	-	0,22	45%	0,320			0,750
	RANDOM	median	importance_ranking_		-	0,780	0,92	-5%	0,870	0,93	-1%	0,920	0,90	-	-	0,93	-1%	0,920			0,930
		1st quartile	importance_ranking_		-	0,570	0,73	-15%	0,620	0,83	-10%	0,750	0,80	-	-	0,80	-8%	0,740			0,750
		3rd quartile	importance_ranking_		-	1,000	0,98	-1%	0,970	1,00	-2%	0,980	0,97	-	-	0,99	-1%	0,980			1,000
	diversity lstm - A	median	importance_ranking_		-	0,180	0,07	-29%	0,050	0,13	246%	0,450	0,17	-	-	0,10	-20%	0,080			0,370
		1st quartile	importance_ranking_		-	0,080	0,02	0%	0,020	0,08	125%	0,180	0,10	-	-	0,04	-25%	0,030			0,200
		3rd quartile	importance_ranking_		-	0,390	0,14	-29%	0,100	0,18	383%	0,870	0,25	-	-	0,23	13%	0,260			0,530
	RANDOM	median	importance_ranking_		-	0,770	0,88	0%	0,880	0,90	6%	0,950	0,90	-	-	0,93	-3%	0,900			0,880
		1st quartile	importance_ranking_		-	0,500	0,62	5%	0,650	0,73	12%	0,820	0,70	-	-	0,75	-1%	0,740			0,680
		3rd quartile	importance_ranking_		-	1,000	0,98	-1%	0,970	0,95	5%	1,000	1,00	-	-	1,00	-3%	0,970			1,000
FIGURE 4		Indicator	Source of info																		
Comparison	**vanilla lstm - [0.0**	median	Permutation.png		-	0,050	0,16	13%	0,180	0,03	67%	0,05	0,05	-	-	0,05	-60%	0,020		-	0,030
Violin plots	[0.25-0.50]	median	Permutation.png		-	0,080	0,17	18%	0,200	0,05	40%	0,07	0,07	-	-	0,15	-47%	0,080		-	0,050
visual inspecti	[0.50-0.75]	median	Permutation.png		-	0,230	n/appl	-	-	0,20	-25%	0,15	0,15	-	-	0,23	-	-		-	0,100
	[0.75-1.00]	median	Permutation.png		-	-	n/appl	-	-	n/appl	-	-	0,15	-	-	0,28	-	-		-	-
	ortho lstm - [0.0	median	Permutation.png		-	0,120	0,42	7%	0,450	0,37	16%	0,430	0,37	-	-	0,49	-12%	0,430		-	0,180
	[0.25-0.50]	median	Permutation.png		-	0,230	0,43	12%	0,480	0,33	21%	0,400	0,33	-	-	0,49	-6%	0,460		-	0,200
	[0.50-0.75]	median	Permutation.png		-	0,330	0,43	21%	0,520	0,35	-6%	0,330	0,40	-	-	0,48	-10%	0,430		-	0,250
	[0.75-1.00]	median	Permutation.png		-	0,360	0,37	54%	0,570	0,47	-36%	0,300	n/appl	-	-	0,50	-6%	0,470		-	0,300
	diversity lstm - [	median	Permutation.png		-	0,100	0,39	10%	0,430	0,43	-30%	0,30	0,3	-	-	0,46	2%	0,470		-	0,200
	[0.25-0.50]	median	Permutation.png		-	0,230	0,44	7%	0,470	0,47	-38%	0,29	0,4	-	-	0,48	-2%	0,470		-	0,250
	[0.50-0.75]	median	Permutation.png		-	0,350	0,48	6%	0,510	0,53	-45%	0,29	0,5	-	-	0,46	4%	0,480		-	0,300
	[0.75-1.00]	median	Permutation.png		-	0,420	n/appl	-	-	0,55	-27%	0,40	n/appl	-	-	0,49	-4%	0,470		-	0,350
TABLE 3		Indicator	Source of info																		
Mean attentio	**vanilla lstm**	Rationale attenti	rationale_summary_te	0,348	113%	0,742	0,472	105%	0,968	0,44	-2%	0,430	0,35	-	-	0,63	-2%	0,617	0,28	189%	0,820
		Rationale length	rationale_summary_te	0,240	200%	0,719	0,217	324%	0,920	0,17	120%	0,380	0,16	-	-	0,22	173%	0,588	0,23	260%	0,809
	ortho lstm	Rationale attenti	rationale_summary_te		-	0,354		-	0,921		-	0,535		-	-			0,860		-	
		Rationale length	rationale_summary_te		-	0,102		-	0,269		-	0,114		-	-			0,240		-	
	diversity lstm	Rationale attenti	rationale_summary_te	0,624	-12%	0,550	0,76	20%	0,912	0,57	-54%	0,266	0,40	-	-	0,88	6%	0,940	0,76	-23%	0,591
		Rationale length	rationale_summary_te	0,175	3%	0,180	0,17	33%	0,224	0,16	19%	0,191	0,24	-	-	0,17	53%	0,265	0,31	4%	0,319
TABLE 4		Indicator	Source of info																		
Comparison tc	**vanilla lstm**	Overall mean Pea	Attn_Gradient_X_val_	0,71	-13%	0,619	0,80	8%	0,863	0,55	25%	0,690	0,43	-	-	0,72	-29%	0,513	0,65	-7%	0,605
Correlatons		Std Pearson corr	Attn_Gradient_X_val_	0,21			0,07			0,16			0,19	-		0,28			0,24		
We use std.dev in paper to judg		Overall mean Pea	Attn_Integrated_Grad	0,62	-19%	0,503	0,68	17%	0,793	0,40	15%	0,458	0,43	-	-	0,65	-40%	0,390	0,56	-14%	0,483
		Std Pearson corr	Attn_Integrated_Grad	0,24			0,09			0,19			0,19	-		0,32			0,25		
		Overall mean JS (	Attn_Gradient_X_val_	0,10	-10%	0,090	0,09	-11%	0,080	0,15	-19%	0,121	0,17	-	-	0,15	21%	0,182	0,08	-6%	0,075
		Std JS div	Attn_Gradient_X_val_	0,04			0,02			0,04			0,04	-		0,07			0,03		
		Overall mean JS (	Attn_Integrated_Grad	0,12	-18%	0,099	0,13	-13%	0,114	0,19	-7%	0,177	0,21	-	-	0,21	6%	0,222	0,08	-5%	0,076
		Std JS div	Attn_Integrated_Grad	0,05			0,02			0,05			0,06			0,06			0,04		
	ortho lstm	Overall mean Pea	Attn_Gradient_X_val_		-	0,726	n/avail	-	0,903	n/avail	-	0,686	n/avail	-	-		-	0,878	n/avail	-	0,747
		Std Pearson corr	Attn_Gradient_X_val_																		
		Overall mean Pea	Attn_Integrated_Grad		-	0,689	n/avail	-	0,714	n/avail	-	0,668	n/avail	-	-		-	0,858	n/avail	-	0,708
		Std Pearson corr	Attn_Integrated_Grad																		
		Overall mean JS (	Attn_Gradient_X_val_		-	0,143	n/avail	-	0,126	n/avail	-	0,163	n/avail	-	-		-	0,174	n/avail	-	0,184
		Std JS div	Attn_Gradient_X_val_																		
		Overall mean JS (	Attn_Integrated_Grad		-	0,148	n/avail	-	0,183	n/avail	-	0,170	n/avail	-	-		-	0,152	n/avail	-	0,186
		Std JS div	Attn_Integrated_Grad																		
	diversity lstm	Overall mean Pea	Attn_Gradient_X_val_	0,83	2%	0,849	0,89	3%	0,917	0,79	-18%	0,645	0,77	-	-	0,96	-5%	0,914	0,80	-3%	0,780
		Std Pearson corr	Attn_Gradient_X_val_	0,19			0,04			0,12			0,14	-		0,08			0,21		
		Overall mean Pea	Attn_Integrated_Grad	0,79	3%	0,815	0,78	1%	0,789	0,79	-29%	0,559	0,77	-	-	0,67	29%	0,867	0,74	-5%	0,707
		Std Pearson corr	Attn_Integrated_Grad	0,22			0,07			0,14			0,14	-		0,11			0,22		
		Overall mean JS (	Attn_Gradient_X_val_	0,08	18%	0,094	0,09	17%	0,105	0,13	33%	0,173	0,12	-	-	0,06	180%	0,168	0,12	-24%	0,091
		Std JS div	Attn_Gradient_X_val_	0,05			0,01			0,04			0,04	-		0,04			0,07		
		Overall mean JS (	Attn_Integrated_Grad	0,09	12%	0,101	0,13	13%	0,147	0,19	-2%	0,187	0,12	-	-	0,07	91%	0,134	0,15	-31%	0,103
		Std JS div	Attn_Integrated_Grad	0,05			0,02			0,05			0,04			0,05			0,06		
FIGURE 5		Indicator	Source of info																		
Distribution of	**vanilla lstm**	Order of POS pre	quant_pos_attn.png	[NOUN,		[NOUN,				[PUNC,!		[NOUN,				[NOUN,		[NOUN,	[NOUN,		[NOUN,
		%attention PUNC		0%		0%			0%	28%		0%				2%		3%	10%		11%
		%attention ADJ		20%		17%			28%	17%		17%				11%		8%	12%		10%
	ortho lstm	Order of POS pre	quant_pos_attn.png									[NOUN,						[NOUN,			[NOUN,
		%attention PUNC				0%			0%			-						0%			3%
		%attention ADJ				25%			35%			-						10%			14%
	diversity lstm	Order of POS pre	quant_pos_attn.png	[NOUN,		[NOUN,				[ADJ,NO		[NOUN,				[NOUN,	[NOUN,	[NOUN,	[NOUN,		[NOUN,
		%attention PUNC		0		0%			0%	3%		0%				0%	0%	0%	3%		5%
		%attention ADJ		28%		25%			35%	30%		17%				10%	8%	13%			12%
	%Change	%attention PUNC																			
		%attention ADJ		40%		47%			25%	76%		0%				-9%	0%	8%			20%
TABLE 5		Indicator	Source of info																		
Preference given to vanilla vs diversity models (human annotators)																					