Color coding:
Relative difference between test runs and original experiment

	MODEL	Indicator	Source of info	dataset SNLI PAPER	SNLI delta rel%	SNLI TEST RU AVG	dataset QQP PAPER	QQP delta rel%	QQP TEST RU AVG	dataset Babi-1 PAPER	Babi-1 delta rel%	Babi-1 TEST RU AVG	dataset Babi-2 PAPER	Babi-2 delta rel%	Babi-2 TEST RU AVG	dataset Babi-3 PAPER	Babi-3 delta rel%	Babi-3 TEST RU AVG	dataset CNN PAPER	CNN delta rel%	CNN TEST RU AVG
TABLE 2 Benchmark data	**vanilla lstm**	test accuracy bes	evaluate.json	0,782	-1%	0,773	0,787	0%	0,784	0,991	1%	1,000	0,401	36%	0,544	0,477	-56%	0,211	0,631	-6%	0,595
		conicity_mean	evaluate.json	0,560	5%	0,589	0,590	-1%	0,583	0,560	37%	0,766	0,480	-10%	0,432	0,430	116%	0,927	0,450	-12%	0,395
		conicity_std	evaluate.json	n/avail	-	0,079	n/avail	-	0,131	n/avail	-	0,023	n/avail	-	0,048	n/avail	-	0,044	n/avail	-	0,046
	ortho lstm	test accuracy bes	evaluate.json	0,765	0%	0,766	0,786	0%	0,786	0,999	0%	0,999	0,561	5%	0,590	0,512	13%	0,577	0,543	-1%	0,536
		conicity_mean	evaluate.json	0,270	13%	0,306	0,330	-4%	0,317	0,220	5%	0,230	0,210	-20%	0,169	0,120	12%	0,135	0,070	41%	0,099
		conicity_std	evaluate.json	n/avail	-	0,071	n/avail	-	0,055	n/avail	-	0,049	n/avail	-	0,030	n/avail	-	0,018	n/avail	-	-
	diversity lstm	test accuracy bes	evaluate.json	0,770	-4%	0,740	0,784	0%	0,782	1,000	0%	1,000	0,402	36%	0,546	0,509	11%	0,563	0,582	-20%	0,463
		conicity_mean	evaluate.json	0,120	-64%	0,043	0,040	-17%	0,033	0,070	5%	0,073	0,050	158%	0,129	0,100	9%	0,109	0,060	528%	0,377
		conicity_std	evaluate.json	n/avail	-	0,022	n/avail	-	0,023	n/avail	-	0,023	n/avail	-	0,030	n/avail	-	0,024	n/avail	-	-
FIGURE 3 Fraction of hid Box plots visual inspection paper		Indicator	Source of info																		
	vanilla lstm - ATT	median	importance_ranking_MAvDY_all.png	0,330			1,00	0%	1,000	0,06	-33%	0,040	[redacted]	-	0,020	[redacted]	-	0,030	[redacted]	-	0,020
		1st quartile	importance_ranking_MAvDY_all.png	0,100			0,66	-39%	0,400	0,05	-60%	0,020		-	0,010		-	0,010		-	0,010
		3rd quartile	importance_ranking_MAvDY_all.png	0,860			1,00	0%	1,000	0,10	-40%	0,060		-	0,030		-	0,050		-	0,030
	RANDOM	median	importance_ranking_MAvDY_all.png	0,800			1,00	0%	1,000	1,00	-50%	0,500		-	0,550		-	0,550		-	0,660
		1st quartile	importance_ranking_MAvDY_all.png	0,530			0,85	-18%	0,700	0,90	-67%	0,300		-	0,300		-	0,250		-	0,330
		3rd quartile	importance_ranking_MAvDY_all.png	1,000			1,00	0%	1,000	1,00	-30%	0,700		-	0,800		-	0,850		-	0,900
	orth lstm - ATTN	median	importance_ranking_MAvDY_all.png	0,300			0,60	17%	0,700	0,04	0%	0,040		-	0,020		-	0,010		-	0,050
		1st quartile	importance_ranking_MAvDY_all.png	0,100			0,20	25%	0,250	0,01	100%	0,020		-	0,010		-	0,000		-	0,010
		3rd quartile	importance_ranking_MAvDY_all.png	0,800			1,00	0%	1,000	0,07	-14%	0,060		-	0,030		-	0,020		-	0,200
	RANDOM	median	importance_ranking_MAvDY_all.png	0,750			1,00	0%	1,000	0,45	11%	0,500		-	0,550		-	0,500		-	0,830
		1st quartile	importance_ranking_MAvDY_all.png	0,470			0,66	-9%	0,600	0,25	20%	0,300		-	0,300		-	0,250		-	0,470
		3rd quartile	importance_ranking_MAvDY_all.png	1,000			1,00	0%	1,000	0,70	0%	0,700		-	0,800		-	0,750		-	0,950
	diversity lstm - A	median	importance_ranking_MAvDY_all.png	0,460			0,80	9%	0,870	0,04	0%	0,040		-	0,020		-	0,020		-	0,020
		1st quartile	importance_ranking_MAvDY_all.png	0,100			0,50	10%	0,550	0,01	100%	0,020		-	0,010		-	0,010		-	0,010
		3rd quartile	importance_ranking_MAvDY_all.png	0,820			1,00	0%	1,000	0,08	-25%	0,060		-	0,030		-	0,030		-	0,030
	RANDOM	median	importance_ranking_MAvDY_all.png	0,820			1,00	0%	1,000	0,70	-29%	0,500		-	0,550		-	0,500		-	0,700
		1st quartile	importance_ranking_MAvDY_all.png	0,500			0,80	-6%	0,750	0,35	-14%	0,300		-	0,300		-	0,250		-	0,330
		3rd quartile	importance_ranking_MAvDY_all.png	1,000			1,00	0%	1,000	1,00	-30%	0,700		-	0,800		-	0,750		-	0,900
FIGURE 4 Comparison Violin plots visual inspecti		Indicator	Source of info																		
	vanilla lstm - [0.(	median	Permutation.png	[redacted]	-	-	[redacted]	-	0,070	[redacted]	-	-	[redacted]	-	0,800	[redacted]	-	0,500	[redacted]	-	0,600
	[0.25-0.50]	median	Permutation.png		-	-		-	0,100		-	0,820		-	0,920		-	0,650		-	0,700
	[0.50-0.75]	median	Permutation.png		-	-		-	0,100		-	0,870		-	0,920		-	0,650		-	0,900
	[0.75-1.00]	median	Permutation.png		-	-		-	0,100		-	0,900		-	0,970		-	0,700		-	0,950
	ortho lstm - [0.0(	median	Permutation.png	[redacted]	-	-	[redacted]	-	0,070	[redacted]	-	0,400	[redacted]	-	0,750	[redacted]	-	0,700	[redacted]	-	0,600
	[0.25-0.50]	median	Permutation.png		-	-		-	0,100		-	0,750		-	0,900		-	0,750		-	0,800
	[0.50-0.75]	median	Permutation.png		-	-		-	0,100		-	0,980		-	0,900		-	0,850		-	0,950
	[0.75-1.00]	median	Permutation.png		-	-		-	0,150		-	0,980		-	0,950		-	0,950		-	-
	diversity lstm - [(	median	Permutation.png	[redacted]	-	-	[redacted]	-	0,070	[redacted]	-	-	[redacted]	-	0,820	[redacted]	-	0,750	[redacted]	-	0,550
	[0.25-0.50]	median	Permutation.png		-	-		-	0,170		-	0,850		-	0,950		-	0,850		-	0,700
	[0.50-0.75]	median	Permutation.png		-	-		-	0,250		-	0,870		-	0,970		-	0,900		-	0,900
	[0.75-1.00]	median	Permutation.png		-	-		-	0,200		-	0,870		-	0,970		-	0,950		-	0,800
TABLE 3 Mean attentio		Indicator	Source of info																		
	vanilla lstm	Rationale attenti	rationale_summary_test.txt	[redacted]	-	[redacted]	[redacted]	-	-	[redacted]	-	-	[redacted]	-	-	[redacted]	-	-	[redacted]	-	-
		Rationale length	rationale_summary_test.txt		-			-	-		-	-		-	-		-	-		-	-
	ortho lstm	Rationale attenti	rationale_summary_test.txt	-		[redacted]	[redacted]	-	-	[redacted]	-	-	[redacted]	-	-	[redacted]	-	-	[redacted]	-	-
		Rationale length	rationale_summary_test.txt	-				-	-		-	-		-	-		-	-		-	-
	diversity lstm	Rationale attenti	rationale_summary_test.txt	-		[redacted]	[redacted]	-	-	[redacted]	-	-	[redacted]	-	-	[redacted]	-	-	[redacted]	-	-
		Rationale length	rationale_summary_test.txt	-				-	-		-	-		-	-		-	-		-	-
TABLE 4 Comparison tc Correlatons We use std.dev in paper to judg		Indicator	Source of info																		
	vanilla lstm	Overall mean Pea	Attn_Gradient_X_val_	0,58	-4%	0,556	0,19	116%	0,410	0,56	71%	0,959	0,16	294%	0,630	0,39	21%	0,470	0,58	35%	0,782
		Std Pearson corr	Attn_Gradient_X_val_	0,33			0,34			0,34			0,23			0,24			0,25		
		Overall mean Pea	Attn_Integrated_Grad	0,38	1%	0,384	-0,06	-240%	0,084	0,33	138%	0,785	0,05	432%	0,266	-0,01	#####	0,362	0,45	42%	0,641
		Std Pearson corr	Attn_Integrated_Grad	0,40			0,34			0,37			0,22			0,08			0,28		
		Overall mean JS (	Attn_Gradient_X_val_	0,11	4%	0,114	0,15	-33%	0,101	0,33	-64%	0,120	0,53	-27%	0,389	0,46	-43%	0,263	0,22	-29%	0,157
		Std JS div	Attn_Gradient_X_val_	0,07			0,08			0,12			0,09			0,08			0,07		
		Overall mean JS (	Attn_Integrated_Grad	0,16	-10%	0,143	0,19	-22%	0,148	0,43	-43%	0,247	0,58	-12%	0,510	0,64	-46%	0,348	0,30	-25%	0,225
		Std JS div	Attn_Integrated_Grad	0,09			0,10			0,13			0,09			0,05			0,10		
	ortho lstm	Overall mean Pea	Attn_Gradient_X_val_	n/avail	-	0,506	n/avail	-	0,378	n/avail	-	0,867	n/avail	-	0,709	n/avail	-	0,665	n/avail	-	0,204
		Std Pearson corr	Attn_Gradient_X_val_pearsonr.csv																		
		Overall mean Pea	Attn_Integrated_Grad	n/avail	-	0,365	n/avail	-	0,238	n/avail	-	0,709	n/avail	-	0,227	n/avail	-	0,044	n/avail	-	0,099
		Std Pearson corr	Attn_Integrated_Gradient_X_val_pearsonr.csv																		
		Overall mean JS (	Attn_Gradient_X_val_	n/avail	-	0,120	n/avail	-	0,120	n/avail	-	0,212	n/avail	-	0,384	n/avail	-	0,426	n/avail	-	0,391
		Std JS div	Attn_Gradient_X_val_jsd.csv																		
		Overall mean JS (	Attn_Integrated_Grad	n/avail	-	0,146	n/avail	-	0,140	n/avail	-	0,282	n/avail	-	0,540	n/avail	-	0,639	n/avail	-	0,439
		Std JS div	Attn_Integrated_Gradient_X_val_jsd.csv																		
	diversity lstm	Overall mean Pea	Attn_Gradient_X_val_	0,51	-2%	0,498	0,58	-15%	0,494	0,91	-7%	0,850	0,70	-13%	0,612	0,67	10%	0,736	0,75	-50%	0,375
		Std Pearson corr	Attn_Gradient_X_val_	0,35			0,31			0,10			0,13			0,19			0,20		
		Overall mean Pea	Attn_Integrated_Grad	0,26	5%	0,274	0,21	-45%	0,116	0,91	-9%	0,824	0,75	3%	0,771	0,47	-44%	0,265	0,66	-87%	0,086
		Std Pearson corr	Attn_Integrated_Grad	0,39			0,36			0,10			0,10			0,25			0,23		
		Overall mean JS (	Attn_Gradient_X_val_	0,10	9%	0,109	0,10	10%	0,110	0,21	7%	0,226	0,23	72%	0,396	0,37	-3%	0,359	0,17	101%	0,342
		Std JS div	Attn_Gradient_X_val_	0,06			0,05			0,08			0,06			0,07			8,00		
		Overall mean JS (	Attn_Integrated_Grad	0,13	11%	0,144	0,15	-6%	0,141	0,24	-8%	0,221	0,19	206%	0,582	0,41	56%	0,641	0,21	143%	0,511
		Std JS div	Attn_Integrated_Grad	0,06			0,06			0,08			0,05			0,08			0,10		
FIGURE 5 Distribution of		Indicator	Source of info																		
	vanilla lstm	Order of POS pre	quant_pos_attn.png	[redacted]		[NOUN,	[NOUN,	[redacted]	[NOUN,	[redacted]		[NOUN]	[redacted]		[NOUN]	[redacted]		[NOUN,	[redacted]		[NOUN,
		%attention PUNC					23%		9%			0%			1%			8%			3%
		%attention ADJ					7%		9%			0%			0%			0%			9%
	ortho lstm	Order of POS pre	quant_pos_attn.png	[redacted]		[NOUN,	[redacted]		[NOUN,	[redacted]		[NOUN,	[redacted]		[NOUN]	[redacted]		[NOUN]	[redacted]		[PUNC,
		%attention PUNC							17%			0%			0%			0%			20%
		%attention ADJ							10%			0%			0%			0%			3%
	diversity lstm	Order of POS pre	quant_pos_attn.png	[redacted]		[NOUN,	[NOUN,		[NOUN,	[redacted]		[NOUN]	[redacted]		[NOUN]	[redacted]		[NOUN]	[redacted]		[NOUN,
		%attention PUNC					3%		5%			0%			0%			0%			3%
		%attention ADJ					10%		9%			0%			0%			0%			9%
	%Change	%attention PUNC																			
		%attention ADJ					43%		0%												
TABLE 5 Preference given to vanilla vs diversity models (human annotators)		Indicator	Source of info										[redacted]			[redacted]			[redacted]		