Appendix A: Detailed Results

BAGEL			SFHOTEL			SFREST		
Inf: 0.16*	Nat: 0.36*	Qua: 0.38*	Inf: 0.41*	Nat: 0.47*	Qua: 0.52*	Inf: 0.35*	Nat: 0.29*	Qua: 0.35*
TGEN: 0.42*	LOLS: 0.24*		RNNLG: 0.52*	LOLS: 0.45*		RNNLG: 0.28*	LOLS: 0.38*	
Total BAGEL: 0.31*			Total SFHOTEL: 0.50*			Total SFREST: 0.35*		
Total all data: 0.45*								

Table 7: Intra-class correlation coefficient (ICC) for human ratings across the three datasets. "*" denotes statistical significance ($p < 0.05$).

	BAGEL		SFHOTEL		SFREST	
	TGEN	LOLS	RNNLG	LOLS	RNNLG	LOLS
metric	*Avg / StDev*	*Avg / StDev*	*Avg / StDev*	*Avg / StDev*	*Avg / StDev*	*Avg / StDev*
TER	0.36/0.24	0.33/0.24	0.28*/0.27	0.65*/0.32	0.41*/0.35	0.65*/0.27
BLEU1	0.75*/0.21	0.81*/0.16	0.85*/0.18	0.66*/0.23	0.73*/0.24	0.59*/0.23
BLEU2	0.68/0.23	0.72/0.21	0.78*/0.25	0.54*/0.28	0.62*/0.31	0.45*/0.29
BLEU3	0.60/0.28	0.63/0.26	0.69*/0.32	0.42*/0.33	0.52*/0.37	0.34*/0.33
BLEU4	0.52/0.32	0.53/0.33	0.56*/0.40	0.28*/0.33	0.44*/0.41	0.24*/0.32
ROUGE	0.76/0.18	0.78/0.17	0.83*/0.18	0.64*/0.21	0.72*/0.24	0.58*/0.22
NIST	4.44*/2.05	4.91*/2.04	4.37*/2.19	3.49*/1.99	4.86*/2.55	4.01*/2.07
LEPOR	0.46*/0.22	0.50*/0.19	0.52*/0.23	0.30*/0.16	0.51*/0.25	0.30*/0.17
CIDEr	2.92/2.40	3.01/2.27	3.08*/2.05	1.66*/1.67	3.39*/2.53	2.09*/1.73
METEOR	0.50/0.22	0.53/0.23	0.62*/0.27	0.44*/0.20	0.54*/0.28	0.41*/0.19
SIM	0.66/0.09	0.65/0.12	0.76*/0.15	0.73*/0.14	0.76/0.13	0.77/0.14
RE	86.79/19.48	83.39/20.41	70.90/17.07	69.62/19.14	64.67/19.07	64.27/22.22
msp	0.04*/0.21	0.14*/0.37	0.68/0.78	0.69/0.77	0.78/0.82	0.85/0.89
prs	84.51*/25.78	93.30*/27.04	97.58*/32.58	107.90*/36.41	93.74/34.98	97.20/39.30
len	38.20*/14.22	42.54*/14.11	49.06*/15.77	51.69*/17.30	53.27*/19.50	50.92*/18.74
wps	10.08*/3.10	10.94*/3.19	11.43*/3.63	12.07*/4.17	11.15*/4.37	10.52*/4.21
sps	13.15*/4.98	14.61*/5.13	16.03*/4.88	17.02*/5.90	16.39*/6.17	15.41*/5.92
cpw	3.77/0.60	3.88/0.59	4.34/0.58	4.36/0.63	4.86*/0.64	4.94*/0.76
spw	1.30/0.22	1.33/0.23	1.43/0.23	1.43/0.26	1.50/0.26	1.50/0.29
pol	2.22/1.21	2.40/1.16	1.24/1.04	1.33/1.04	1.69/1.12	1.57/1.07
ppw	0.22/0.09	0.22/0.09	0.11/0.10	0.12/0.09	0.16/0.11	0.16/0.12
informativeness	4.77/1.09	4.91/1.23	5.47*/0.81	5.27/1.02	5.29*/0.94	5.16/1.07
naturalness	4.76/1.26	4.67/1.25	4.99*/1.13	4.62/1.28	4.86/1.13	4.74/1.23
quality	4.77/1.19	4.54/1.28	4.54/1.18	4.53/1.26	4.51/1.14	4.58/1.33

Table 8: The systems' performance for all datasets. *Avg* denotes a mean value, *StDev* stands for standard deviation, "*" denotes a statistically significant difference ($p < 0.05$) between the two systems on the given dataset.

	BAGEL						SFHOTEL						SFREST					
	TGEN			LOLS			RNNLG			LOLS			RNNLG			LOLS		
metric	*inf*	*nat*	*qual*	*inf*	*nat*	*qual*	*inf*	*nat*	*qual*	*inf*	*nat*	*qual*	*inf*	*nat*	*qual*	*inf*	*nat*	*qual*
TER	-0.21*	-0.19*	-0.16*	-0.16*	-0.19*	-0.16*	-0.03	-0.09	-0.08	-0.06	**-0.20***	-0.12*	0.02	-0.14*	-0.08	**-0.16***	-0.14*	-0.14*
BLEU1	**0.30***	0.15*	0.13	0.13	0.15*	0.13	0.09	0.09*	0.08	0.01	0.12*	0.06	0.02	0.12*	0.06	**0.19***	0.15*	0.13*
BLEU2	**0.30***	0.17*	0.14	0.12	0.14*	0.11	0.08	0.09*	0.07	0.00	0.12*	0.07	0.01	0.13*	0.07	**0.14***	0.10*	0.08*
BLEU3	**0.27***	0.17*	0.12	0.11	0.13	0.10	0.06	0.08	0.06	0.01	0.11*	0.08	0.02	0.13*	0.09*	0.12*	0.08	0.07
BLEU4	**0.23***	0.15*	0.11	0.11	0.13	0.10	0.06	0.05	0.07	0.00	0.02	0.03	0.03	0.12*	0.07	0.12*	0.04	0.05
ROUGE	0.20*	0.11	0.09	0.20*	0.17*	0.15*	0.07	0.09	0.08	-0.01	0.04	0.02	0.04	0.17*	0.09*	0.12*	0.11*	0.08
NIST	0.24*	0.07	0.02	0.16*	0.13	0.11	0.07	0.05	0.01	0.02	0.14*	0.11*	0.03	0.07	0.01	**0.15***	0.08	0.07
LEPOR	**0.17***	0.12	**0.07**	-0.07	0.02	-0.04	0.03	0.03	0.03	0.14*	**0.17***	0.10*	0.00	0.05	-0.02	**0.28***	0.17*	0.18*
CIDEr	**0.26***	0.14*	0.10	0.14*	0.19*	0.14*	0.07	0.07	0.00	0.03	0.13*	0.09	0.02	0.12*	0.03	0.10*	0.11*	0.08
METEOR	0.29*	0.09	0.09	0.20*	0.18*	0.16*	0.07	0.10*	0.10*	0.05	0.06	0.04	0.06	0.16*	0.09*	**0.23***	0.19*	0.17*
SIM	0.16*	0.04	0.06	0.14*	0.13	0.09	-0.05	-0.12*	-0.11*	0.03	-0.03	-0.08	0.13*	-0.06	-0.08*	0.19*	0.01	0.02
RE	-0.06	**0.09**	0.13	-0.09	-0.04	0.04	0.00	0.03	0.10*	-0.01	0.03	0.09	0.00	-0.05	0.02	0.04	**0.09***	0.08*
cpw	0.03	**-0.12**	**-0.19***	0.08	0.05	-0.03	0.02	-0.02	**-0.09***	**0.13***	**0.14***	0.06	0.02	0.11*	0.01	0.06	0.10*	0.09*
len	**0.25***	-0.25*	-0.21*	0.04	-0.19*	-0.24*	0.01	-0.17*	-0.09	0.12*	-0.08	-0.07	0.11*	-0.17*	-0.08	0.21*	-0.14*	-0.09*
wps	**0.33***	-0.17*	-0.12	-0.05	**-0.28***	**-0.29***	0.01	-0.15*	-0.05	0.08	-0.12*	-0.08	0.11*	-0.19*	-0.07	0.18*	-0.15*	-0.11*
sps	**0.25***	-0.20*	-0.17*	0.03	-0.17*	-0.23*	-0.02	-0.16*	-0.08	0.02	-0.18*	-0.16*	0.07	-0.17*	-0.08	0.12*	-0.21*	-0.16*
spw	0.01	-0.07	**-0.13**	0.10	**0.09**	0.02	-0.08	-0.02	-0.11*	-0.10*	-0.10*	-0.17*	-0.07	0.06	-0.03	-0.14*	**-0.10***	-0.11*
pol	0.16*	-0.06	-0.07	0.11	-0.03	-0.12	-0.07	-0.10*	-0.15*	0.01	-0.09	-0.14*	-0.04	-0.04	-0.03	-0.02	-0.13*	-0.11*
ppw	-0.02	0.06	0.00	**0.16***	0.15*	0.08	-0.09	-0.06	-0.16*	-0.02	-0.01	-0.09	-0.09*	**0.08**	0.00	-0.13*	-0.05	-0.07
msp	-0.02	-0.06	-0.11	0.02	-0.02	-0.10	-0.01	-0.10*	-0.08	0.05	-0.02	-0.03	0.05	0.02	-0.06	0.12*	0.01	0.07
prs	**-0.23***	0.18*	0.13	0.05	0.24*	0.31*	-0.02	0.13*	0.09	-0.13*	0.05	0.04	-0.11*	0.15*	0.11*	-0.16*	0.20*	0.16*

Table 9: Spearman correlation between metrics and human ratings for individual datasets and systems. "*" denotes statistically significant correlation ($p < 0.05$), bold font denotes significantly stronger correlation when comparing two systems on the same dataset.

	BAGEL			SFHOTEL			SFREST		
	inf	*nat*	*qual*	*inf*	*nat*	*qual*	*inf*	*nat*	*qual*
TER	-0.19*	-0.19*	-0.15*	-0.10*	-0.19*	-0.07*	-0.09*	-0.15*	-0.08*
BLEU1	0.23*	0.14*	0.11*	0.11*	0.18*	0.07*	0.11*	0.14*	0.07*
BLEU2	0.21*	0.15*	0.12*	0.10*	0.17*	0.07*	0.09*	0.13*	0.06*
BLEU3	0.19*	0.15*	0.11*	0.09*	0.16*	0.07*	0.08*	0.12*	0.06*
BLEU4	0.18*	0.14*	0.10*	0.08*	0.10*	0.06	0.09*	0.09*	0.05
ROUGE	0.20*	0.13*	0.11*	0.09*	0.15*	0.06	0.09*	0.15*	0.06*
NIST	0.21*	0.09	0.06	0.07*	0.13*	0.06	0.10*	0.08*	0.03
LEPOR	0.07	0.07	0.01	0.13*	0.15*	0.05	0.16*	0.12*	0.04
CIDEr	0.21*	0.16*	0.12*	0.10*	0.16*	0.05	0.08*	0.12*	0.04
METEOR	0.25*	0.13*	0.12*	0.11*	0.15*	0.08*	0.15*	0.18*	0.11*
SIM	0.15*	0.09	0.07	0.01	-0.04	-0.09*	0.15*	-0.02	-0.02
RE	-0.08	0.03	0.09	0.01	0.04	0.10*	0.02	0.02	0.06
cpw	0.05	-0.04	-0.12*	0.07*	0.05	-0.02	0.04	0.10*	0.06
len	0.14*	-0.22*	-0.24*	0.05	-0.14*	-0.07*	0.16*	-0.15*	-0.09*
wps	0.14*	-0.23*	-0.23*	0.03	-0.14*	-0.06	0.14*	-0.17*	-0.10*
sps	0.14*	-0.19*	-0.21*	-0.01	-0.18*	-0.12*	0.10*	-0.18*	-0.12*
spw	0.05	0.00	-0.06	-0.10*	-0.06	-0.14*	-0.11*	-0.02	-0.07*
pol	0.13*	-0.05	-0.10*	-0.04	-0.10*	-0.14*	-0.03	-0.08*	-0.08*
ppw	0.06	0.11*	0.04	-0.06	-0.04	-0.13*	-0.11*	0.01	-0.04
msp	0.02	-0.04	-0.11*	0.02	-0.06	-0.06	0.08*	0.01	0.01
prs	-0.10	0.22*	0.25*	-0.05	0.12*	0.07	-0.13*	0.18*	0.13*

Table 10: Spearman correlation between metrics and human ratings for each dataset. "*" denotes statistically significant correlation ($p < 0.05$).

	TGEN			LOLS			RNNLG		
	inf	*nat*	*qual*	*inf*	*nat*	*qual*	*inf*	*nat*	*qual*
TER	-0.21*	-0.19*	-0.16*	-0.07*	-0.15*	-0.11*	-0.02	-0.13*	-0.08*
BLEU1	0.30*	0.15*	0.13	0.08*	0.12*	0.08*	0.07*	0.13*	0.07*
BLEU2	0.30*	0.17*	0.14	0.05	0.11*	0.07*	0.06*	0.14*	0.08*
BLEU3	0.27*	0.17*	0.12	0.04	0.09*	0.07*	0.06	0.13*	0.08*
BLEU4	0.23*	0.15*	0.11	0.04	0.04	0.04	0.06	0.11*	0.08*
ROUGE	0.20*	0.11	0.09	0.05	0.09*	0.05	0.07*	0.15*	0.09*
NIST	0.25*	0.07	0.02	0.07*	0.11*	0.09*	0.04	0.06*	0.01
LEPOR	0.17*	0.12	0.07	0.13*	0.13*	0.11*	0.02	0.05	0.00
CIDEr	0.26*	0.14*	0.10	0.05	0.13*	0.09*	0.04	0.10*	0.02
METEOR	0.29*	0.09	0.09	0.14*	0.13*	0.12*	0.08*	0.15*	0.10*
SIM	0.16*	0.04	0.06	0.14*	0.02	0.00	0.05	-0.08*	-0.09*
RE	-0.06	0.09	0.13	-0.02	0.04	0.07*	0.02	-0.01	0.06*
cpw	0.03	-0.12	-0.19*	0.11*	0.11*	0.08*	-0.02	0.02	-0.05
len	0.25*	-0.25*	-0.21*	0.17*	-0.12*	-0.10*	0.06	-0.18*	-0.08*
wps	0.33*	-0.17*	-0.12	0.11*	-0.17*	-0.13*	0.07*	-0.17*	-0.06
sps	0.25*	-0.20*	-0.17*	0.09*	-0.19*	-0.17*	0.03	-0.17*	-0.08*
spw	0.01	-0.07	-0.13	-0.07*	-0.06*	-0.10*	-0.09*	0.01	-0.07*
pol	0.16*	-0.06	-0.07	-0.02	-0.09*	-0.11*	-0.08*	-0.08*	-0.09*
ppw	-0.02	0.06	0.00	-0.08*	0.00	-0.05	-0.11*	0.00	-0.07*
msp	-0.02	-0.06	-0.11	0.10*	0.00	0.02	0.02	-0.04	-0.07*
prs	-0.23*	0.18*	0.13	-0.12*	0.16*	0.15*	-0.07*	0.14*	0.10*

Table 11: Spearman correlation between metrics and human ratings for each system. "*" denotes statistical significance ($p < 0.05$).

Accuracy	rand	TER	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	NIST	LEPOR	CIDEr	METE	RE	SIM
BAGEL													
inform	37.13	45.05*	41.58*	41.58*	42.57*	42.08*	43.07*	43.07*	41.58*	43.07*	45.54*	37.13	41.09*
natural	42.08	47.03*	46.04*	45.54*	44.06*	45.05*	46.04*	44.55*	46.53*	45.05*	45.05*	42.08	43.07*
quality	33.17	45.54*	43.07*	40.10*	40.59*	43.56*	43.07*	41.09*	40.59*	42.08*	41.58*	37.62	42.57*
SFHOTEL													
inform	25.38	34.92*	35.68*	35.18*	35.68*	34.67*	36.43*	31.41	32.16*	33.92*	36.43*	34.92*	33.92*
natural	41.96	45.73	46.48	45.48	46.48	45.23	48.74	41.21	43.72	44.72	49.75*	37.19	46.98
quality	44.47	40.95	40.95	42.21	44.72	41.46	43.22	40.2	40.95	42.46	45.98	33.67	37.44
SFREST													
inform	33.68	36.27	35.41	34.02	34.72	36.96	33.16	35.58	36.27	32.47	34.72	38.34*	42.66*
natural	36.10	40.41	40.07	38.86	38.34	38.86	38.17	39.38	41.11*	36.79	39.38	39.38	38.00
quality	39.38	37.13	36.96	39.21	37.65	39.55	36.10	38.69	39.72	35.23	34.89	40.93	37.31
SFREST, quant.													
inform	31.95	35.75*	36.27*	34.37*	35.92*	34.54*	36.44*	39.55*	37.13*	36.27*	36.79*	38.17*	42.83*
quality	39.21	33.33	34.37	32.3	30.57	26.94	34.54	33.16	35.92	30.92	31.61	32.47	35.41
naturalness	37.13	37.82	38.69	36.1	35.75	32.3	36.96	39.21	38.86	35.23	38.34	34.2	36.1

Table 12: Accuracy of metrics predicting relative human ratings, with "*" denoting statistical significance ($p < 0.05$).

	informativeness		*naturalness*		*quality*	
	Bad	*Good and avg*	*Bad*	*Good and avg*	*Bad*	*Good and avg*
TER	**0.48***	0.07*	**0.31***	0.15*	0.08	0.06*
BLEU1	**0.45***	0.11*	**0.26***	0.13*	0.07	0.04
BLEU2	**0.49***	0.09*	**0.29***	0.13*	0.05	0.04*
BLEU3	**0.40***	0.08*	**0.25***	0.13*	0.01	0.05*
BLEU4	**0.41***	0.07*	**0.21***	0.08*	0.01	0.04
ROUGE	**0.50***	0.08*	**0.28***	0.13*	0.07	0.04*
NIST	**0.26**	0.08*	**0.23***	0.08*	0.08	0.03
LEPOR	**0.40***	0.09*	**0.23***	0.10*	0.03	0.01
CIDEr	**0.42***	0.09*	**0.21***	0.12*	0.02	0.04
METEOR	**0.45***	0.14*	0.24*	0.15*	0.03	0.08*
SIM	**0.37***	0.12*	**0.29***	-0.03	**0.21***	-0.08*

Table 13: Spearman correlation between WBM scores and human ratings for utterances from the *Bad* bin and utterances from the *Good* and *Average* bins. "*" denotes statistically significant correlation ($p < 0.05$), bold font denotes significantly stronger correlation for the *Bad* bin compared to the *Good* and *Average* bins.

	informativeness		*naturalness*		*quality*	
	Inform	*Not inform*	*Inform*	*Not inform*	*Inform*	*Not inform*
TER	-0.08*	-0.10	-0.17*	-0.18*	-0.09*	-0.11*
BLEU1	0.11*	0.09	0.14*	0.20*	0.07*	0.11*
BLEU2	0.09*	0.10	0.14*	0.20*	0.07*	0.13*
BLEU3	0.07*	0.11*	0.13*	0.20*	0.06*	0.14*
BLEU4	0.06*	0.11*	0.09*	0.18*	0.05*	0.14*
ROUGE	0.08*	0.12*	0.14*	0.22*	0.06*	0.16*
NIST	0.08*	0.05	0.10*	0.06	0.07*	-0.06
LEPOR	0.09*	0.16*	0.11*	0.16*	0.05*	0.04
CIDEr	0.10*	0.01	**0.16***	0.04	**0.07***	0.02
METEOR	0.14*	0.17*	0.15*	0.22*	0.09*	0.18*
SIM	0.15*	0.09	-0.01	-0.03	-0.05*	-0.10
cpw	**0.12***	-0.15*	**0.09***	-0.14*	**0.01**	-0.11*
len	0.17*	0.08	-0.15*	-0.12*	-0.12*	-0.05
wps	0.11*	0.19*	**-0.19***	-0.03	**-0.12***	0.01
sps	0.09*	0.18*	**-0.20***	-0.02	**-0.17***	0.02
spw	-0.06*	0.09	-0.03	0.01	**-0.12***	0.01
pol	**-0.08***	0.05	-0.10*	-0.03	-0.09*	-0.03
ppw	**-0.14***	-0.01	0.00	-0.03	-0.03	-0.05
msp	**0.11***	-0.03	0.00	-0.08	-0.03	-0.08
prs	-0.10*	-0.18*	**0.18***	0.04	**0.15***	0.02

Table 14: Spearman correlation between automatic metrics and human ratings for utterances of the *inform* MR type and utterances of other MR types. "*" denotes statistically significant correlation ($p < 0.05$), bold font denotes significantly stronger (absolute) correlation for *inform* MRs compared to non-*inform* MRs.