	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	81.45	11.73	48.57	30.57	35.97	49.43	81.21	15.05	64.55	44.67	77.06	53.08
APE	75.22	13.39	49.43	28.62	42.83	46.40	78.32	19.64	56.07	45.22	79.20	48.64
AVF	76.44	13.61	48.25	28.65	39.76	45.00	79.50	18.77	57.20	44.51	80.27	49.27
PNMA	74.10	10.60	43.87	24.51	35.60	38.24	74.23	15.19	55.29	38.30	75.43	42.95
Our	**83.83**	**16.27**	**57.28**	**33.09**	**43.69**	**51.26**	**82.16**	**24.94**	**74.91**	**46.40**	**82.39**	**55.43**

Content Preservation

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	**86.29**	76.54	73.85	**84.59**	83.01	77.49	76.03	**90.99**	79.89	**64.10**	76.63	74.89
APE	77.26	85.28	**78.18**	83.33	**89.48**	**83.52**	**77.28**	88.09	**81.72**	59.37	76.62	74.06
AVF	77.08	**85.73**	77.85	84.59	88.12	81.00	77.10	88.93	80.99	59.54	**78.05**	74.41
PNMA	77.01	85.12	76.27	83.67	87.77	82.13	76.98	88.06	79.52	57.90	75.28	72.91
Our	85.43	85.51	77.59	80.63	84.29	75.48	77.05	83.55	78.38	61.82	75.60	**75.79**

Fluency

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	87.93	86.58	112.55	190.29	86.41	63.86	101.40	90.29	196.47	135.93	172.68	122.16
APE	93.56	89.34	128.33	187.91	87.70	65.33	105.00	91.75	246.71	131.55	146.58	123.43
AVF	96.03	85.80	128.28	188.03	84.33	71.31	111.61	95.92	219.96	122.60	151.32	126.41
PNMA	103.47	90.65	131.33	190.10	91.86	77.24	108.24	99.59	256.96	132.28	154.48	126.53
Our	**87.16**	**76.93**	**80.75**	**171.37**	**81.08**	**62.28**	**100.91**	**81.45**	**146.72**	**113.43**	**140.09**	**107.88**

Table 10: **Main Results (70B model):** Style transfer accuracy (higher values are better; ↑), content preservation (↑) and fluency (↓) on 6 datasets across 12 transfer directions. Best results are highlighted in bold.

Content Preservation (Paraphrase Model)												
	Formality		**Toxicity**		**Politics**		**Politness**		**Authorship**		**Sentiment**	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	**85.95**	74.71	73.54	82.71	82.48	75.77	75.32	**89.14**	78.75	**62.28**	76.17	**74.47**
APE	76.72	85.06	**76.72**	83.00	**87.99**	**82.21**	76.80	87.89	80.07	57.61	**76.52**	73.53
AVF	75.21	84.53	76.63	**83.57**	86.92	80.68	**76.94**	87.32	**80.94**	58.98	76.15	73.95
PNMA	75.52	84.11	75.67	82.54	86.79	80.67	76.04	86.93	79.22	57.42	75.04	72.67
Our	85.84	**86.28**	75.85	80.10	82.32	74.96	75.65	82.47	77.19	60.92	75.25	74.21
Content Preservation (LaBSE model)												
	Formality		**Toxicity**		**Politics**		**Politness**		**Authorship**		**Sentiment**	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	0.75	**0.90**	0.77	0.86	0.86	0.79	0.72	**0.92**	0.86	0.70	0.76	0.76
APE	0.75	0.88	0.78	0.87	**0.90**	0.85	0.74	0.90	0.86	0.66	0.75	0.75
AVF	0.74	0.88	0.78	0.87	0.89	0.83	0.75	0.90	0.87	0.67	0.75	0.76
PNMA	**0.79**	0.90	**0.79**	**0.89**	0.89	**0.87**	**0.86**	0.90	**0.89**	**0.74**	**0.82**	**0.81**
Our	0.74	0.89	0.74	0.81	0.84	0.75	0.69	0.90	0.84	0.54	0.75	0.64
Content Preservation (BLEURT)												
	Formality		**Toxicity**		**Politics**		**Politness**		**Authorship**		**Sentiment**	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	**0.089**	**0.527**	0.132	0.345	0.304	0.039	**0.625**	0.040	0.307	0.320	0.136	0.084
APE	0.069	0.449	0.156	**0.376**	**0.488**	**0.262**	0.094	**0.535**	0.328	**0.461**	0.193	0.078
AVF	0.043	0.440	0.157	0.376	0.424	0.191	0.122	0.522	**0.344**	0.426	**0.207**	0.095
PNMA	0.002	0.433	0.139	0.360	0.399	0.181	0.074	0.513	0.334	0.417	0.197	0.085
Our	0.073	0.478	**0.157**	0.329	0.460	0.232	0.557	0.473	0.324	0.386	0.199	**0.133**

Table 11: Different content preservation metrics: sentence embedding model trained from paraphrase datasets, sentence embedding model from multilingual representation model and BLEURT metrics.

Style Transfer Accuracy												
	Formality		**Toxicity**		**Politics**		**Politeness**		**Authorship**		**Sentiment**	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
NS ($p = 0.95$)	79.36	11.95	50.55	27.84	36.17	50.19	76.10	21.70	72.79	43.89	74.34	50.96
CS	79.46	12.40	54.05	30.12	36.04	48.39	79.61	22.62	72.18	43.94	76.54	52.82
Our	**80.80**	**14.40**	**55.36**	**31.98**	**37.81**	**50.30**	**80.63**	**23.27**	**73.40**	**45.14**	**77.93**	**54.73**

Table 12: Comparison of three different decoding methods: nucleus sampling (NP; p=0.95), contrastive search (CS) and our decoding method.