category	Model Score	cohen kappa score	krippendorff score ordinal	dist normalized + std
Fluency	target	1	1	0 ± 0
	t5detox	0.6734	0.8062	0.04687 ± 0.1717
	t5formal	0.574	0.5992	0.03906 ± 0.1619
	t5large	0.6216	0.7360	0.0625 ± 0.1889
	flanlarge	0.4693	0.5688	0.05468 ± 0.2015
Corpify Level	target	0.3081	0.3005	0.052 ± 0.1989
	t5detox	0.3467	0.67695	0.1757 ± 0.2385
	t5formal	0.4291	0.769	0.1562 ± 0.2293
	t5large	0.3715	0.6786	0.1835 ± 0.2449
	flanlarge	0.105	0.256	0.3 ± 0.296
Content Intention	target	-0.0212	-0.016	0.0468 ± 0.213
	t5detox	0.2592	0.3654	0.3281 ± 0.39
	t5formal	0.4069	0.6387	0.2109 ± 0.306
	t5large	0.5492	0.6826	0.164 ± 0.3093
	flanlarge	0.6097	0.7631	0.1484 ± 0.2911
essential details	target	0.1978	0.1969	0.1328 ± 0.2708
	t5detox	0.3922	0.6319	0.2265 ± 0.2945
	t5formal	0.4394	0.618	0.2109 ± 0.306
	t5large	0.4491	0.6257	0.2031 ± 0.3177
	flanlarge	0.5582	0.7807	0.1562 ± 0.2653

Table 4: annotation agreement

Category’s Score Model	**Fluency (0-2)**	**’Corpy’ Level (-2 - +2)**	**Content Intention (0-2)**	**Essential Details (0-2)**
Target	2±0	1.9375 ± 0.2439	1.9765 ± 0.1065	1.8203 ± 0.3492
T5-detox	1.7968 ± 0.5249	0.7578125 ± 1.3944	1.125 ± 0.74	1.039 ± 0.7832
T5-formal	1.8671 ± 0.4475	0.6718 ± 1.3487	1.2734 ± 0.7762	1.1953 ± 0.7695
T5-large	1.76562 ± 0.5342	0.414 ± 1.4241	**1.3046** ± **0.8243**	**1.3125** ± **0.774**
Flan-Large	**1.8828** ± **0.3752**	**0.8515** ± **1.0603**	0.9609 ± 0.8877	0.9062 ± 0.8492

Table 5: Human annotation results. Each one of the 64 *test set*’s sentences was rated by 2 M.Sc. students, in 4 different categories (Fluency, "Corpy" level, preserving content intention, and preserving essential details). Each score in this table is the mean of all sentences of a given category and a given model. Scores scale given by raters is in brackets. full guidelines is available here