Appendix

Anonymous Author(s)*

A HUMAN EVALUATION

In order to eliminate the impact of showing reference summaries to the evaluators, we conduct another set of human evaluation experiments: displaying only code and generated summaries to participants, without reference summaries. The evaluators are asked to assign scores from 0 to 4 to measure the informativeness and naturalness of the generated summary. Note that similarity is not measured because in this setting there are no references shown to the evaluators. As shown in Table 1 and Table 2, the conclusion is consistent with the previous human evaluation (showing reference summaries): CoCoSUM outperforms Ast-attendgru, ASTNN, and Rencos.

Table 1: Results of human evaluation not showing references to participants (standard deviation in parentheses).

Model	Informativeness	Naturalness
CoCoSUM	**3.34** (1.20)	**3.10** (1.22)
Ast-attendgru	3.01 (1.14)	2.55 (1.24)
ASTNN	2.35 (1.26)	2.10 (0.99)
Rencos	2.56 (1.28)	2.70 (1.1)

Table 2: Statistics significance p-value of CoCoSUM over other methods in human evaluation not showing references to participants.

Model	Informativeness	Naturalness
Ast-attendgru	$8.25e^{-19}$	$1.04e^{-2}$
ASTNN	$6.29e^{-73}$	$1.36e^{-6}$
Rencos	$3.97e^{-52}$	$4.06e^{-2}$