INLG 2024

Automatic Metrics in NLG: A Survey of Current Evaluation Practices

Patrícia Schmidtová, Saad Mahamood, Simone Balloccu, Ondřej Dušek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondřej Plátek, Adarsa Sivaprasad

Introduction

- Automatic metrics are quick proxies, but...
    - Some have a poor correlation with human judgment
    - Many cannot capture factuality or faithfulness issues in text
    - Different implementations make results hard to interpret and reproduce
    - They can be over-reported without adding any informational value

How are automatic metrics used in NLG?

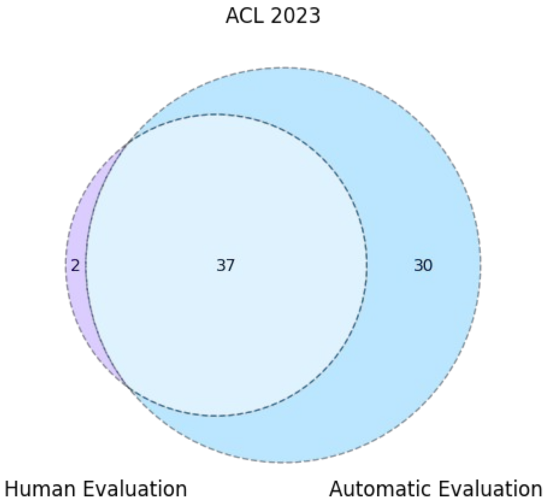
Method

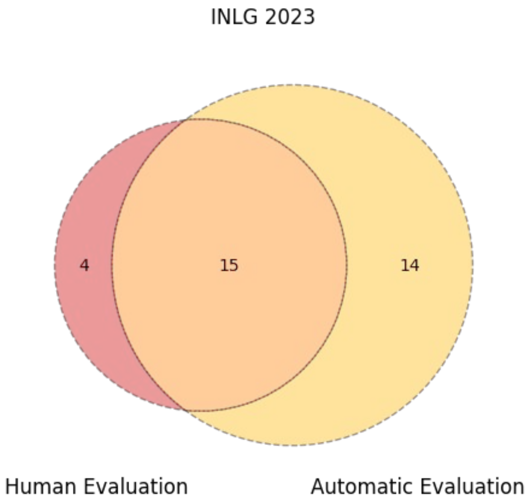
We collected papers from INLG 2023 and ACL 2023 Generation track and annotated the following information:

- **Name** of the evaluation method
- Was the method **newly introduced**?
- Which **task**(s) was this metric used to evaluate?
- Did the authors comment on any **correlation between automatic** and **human evaluation**?
- Did the authors provide **implementation details** for the metric?
- Was the metric **only** reported in the **Appendix**?
- Did the authors explain the **rationale** for the metric?

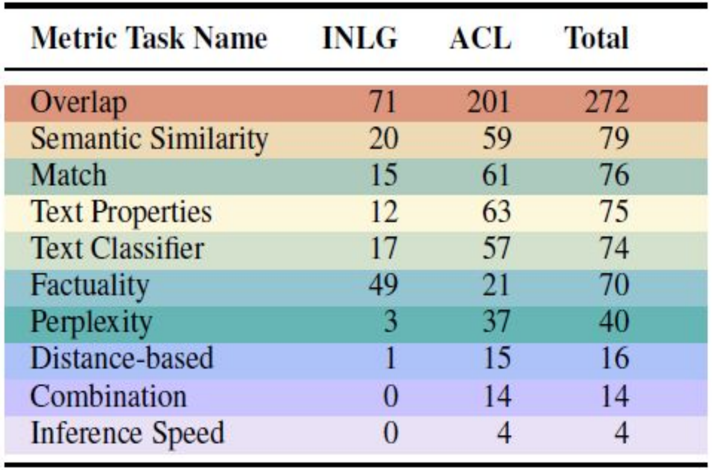
Overview of Results

- **110** papers total (**36** from **INLG** and **74** from **ACL**)
- **102** papers **included** any **evaluation**
- **57%** use **human** evaluation
- **94%** use **automatic** evaluation
- **51%** use **both**
- **634 counts** of automatic metrics (**283 unique**)

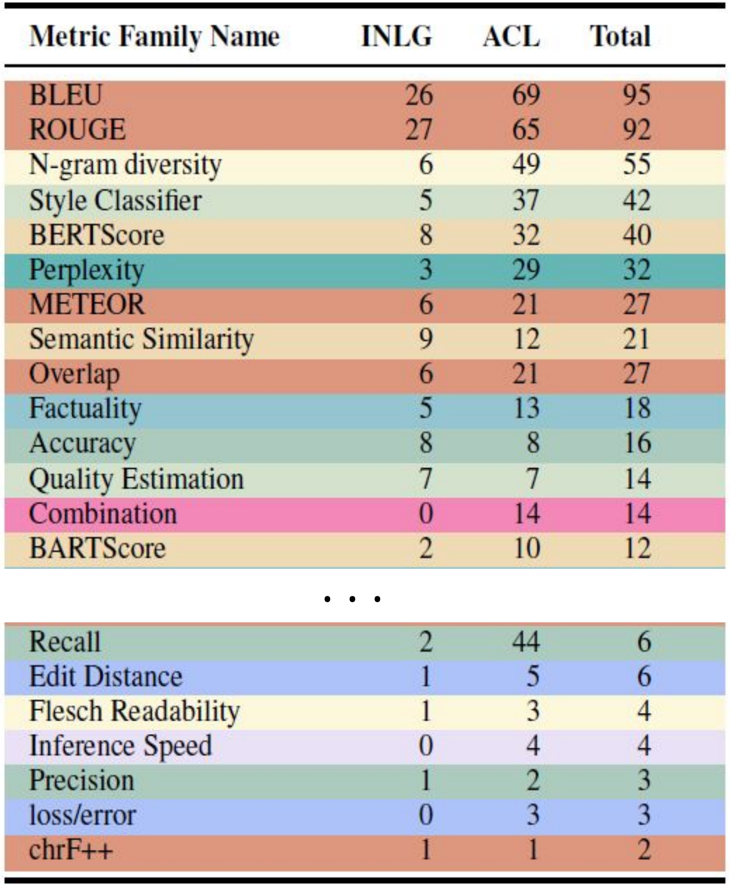




Metric Families & Categories

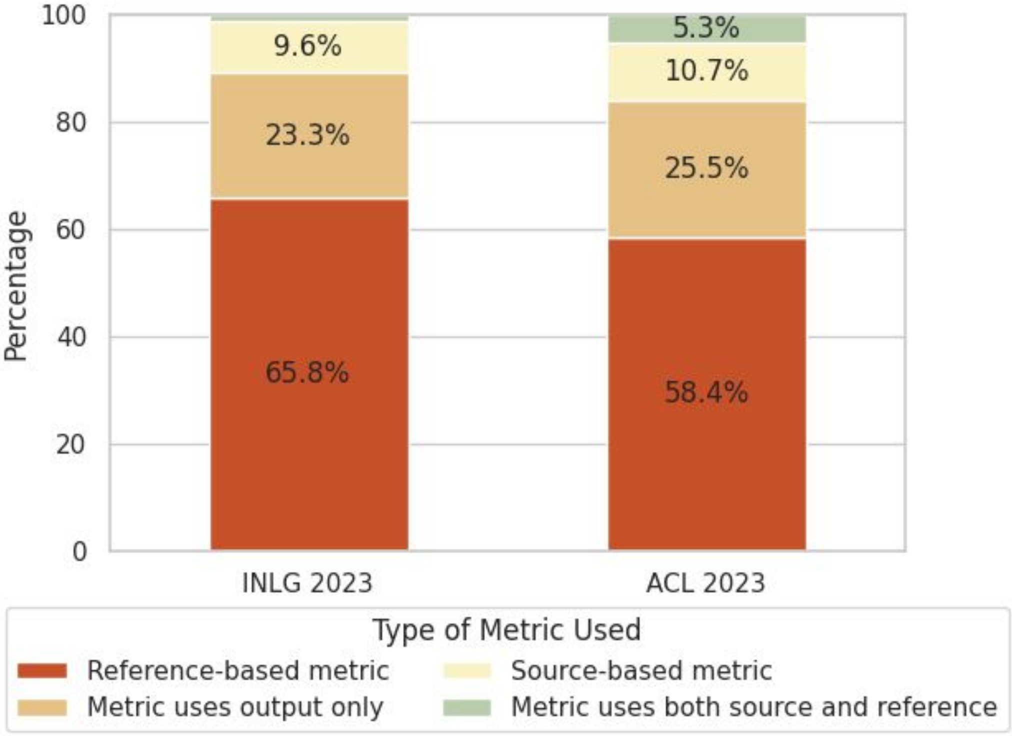


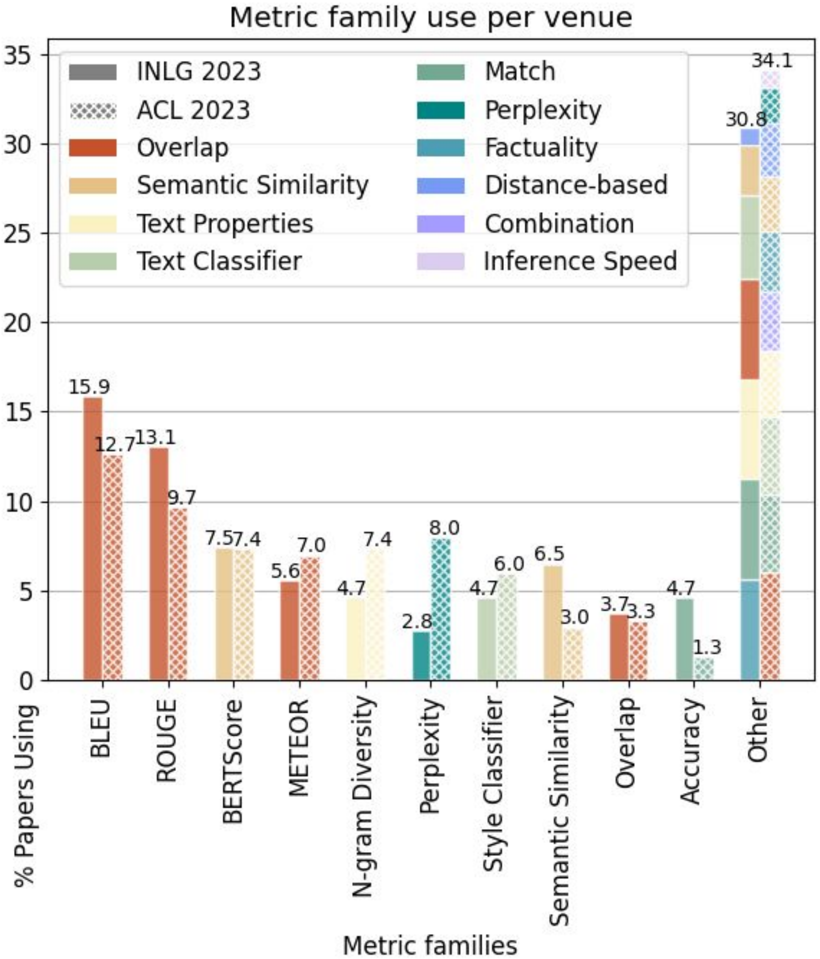
Metric Task Name	INLG	ACL	Total
Overlap	71	201	272
Semantic Similarity	20	59	79
Match	15	61	76
Text Properties	12	63	75
Text Classifier	17	57	74
Factuality	49	21	70
Perplexity	3	37	40
Distance-based	1	15	16
Combination	0	14	14
Inference Speed	0	4	4



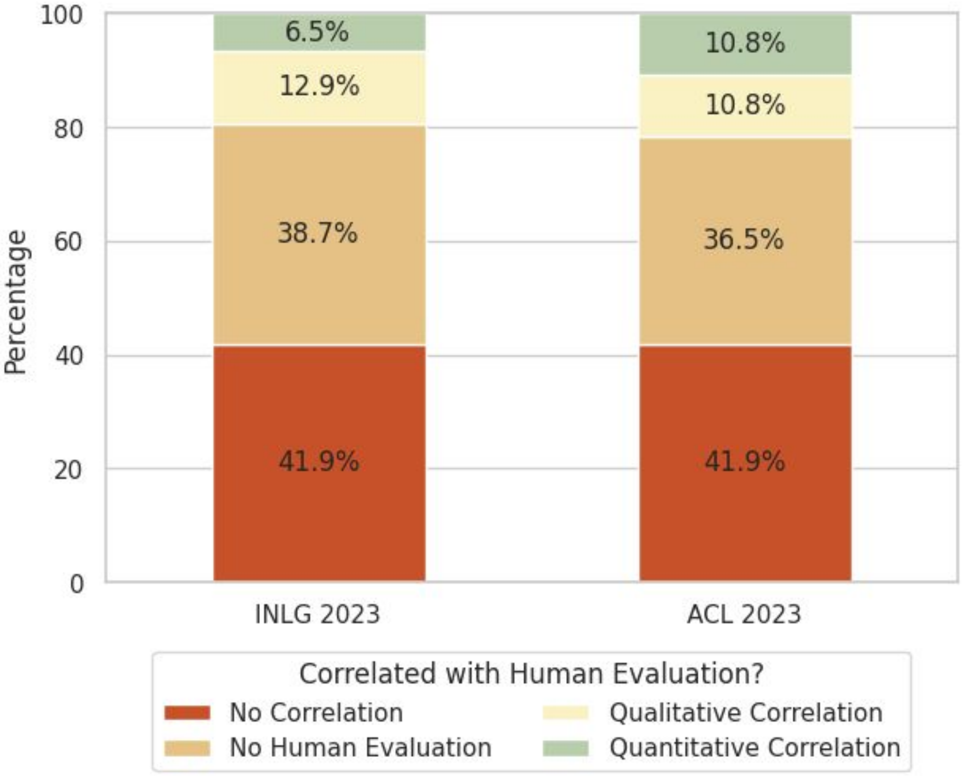
Metric Family Name	INLG	ACL	Total
BLEU	26	69	95
ROUGE	27	65	92
N-gram diversity	6	49	55
Style Classifier	5	37	42
BERTScore	8	32	40
Perplexity	3	29	32
METEOR	6	21	27
Semantic Similarity	9	12	21
Overlap	6	21	27
Factuality	5	13	18
Accuracy	8	8	16
Quality Estimation	7	7	14
Combination	0	14	14
BARTScore	2	10	12
. . .			
Recall	2	44	6
Edit Distance	1	5	6
Flesch Readability	1	3	4
Inference Speed	0	4	4
Precision	1	2	3
loss/error	0	3	3
chrF++	1	1	2

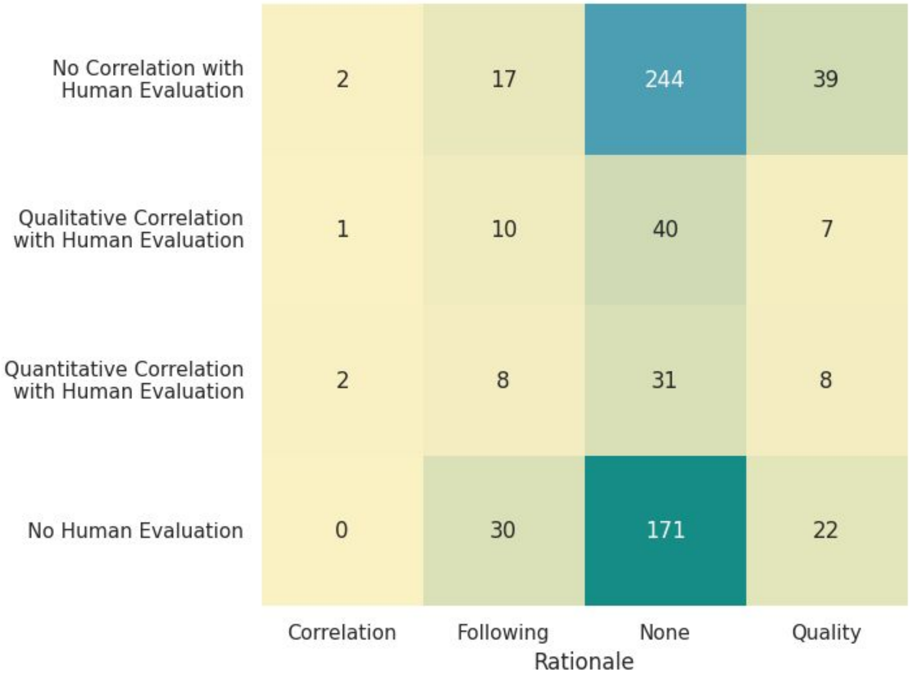
What kinds of metrics were used?





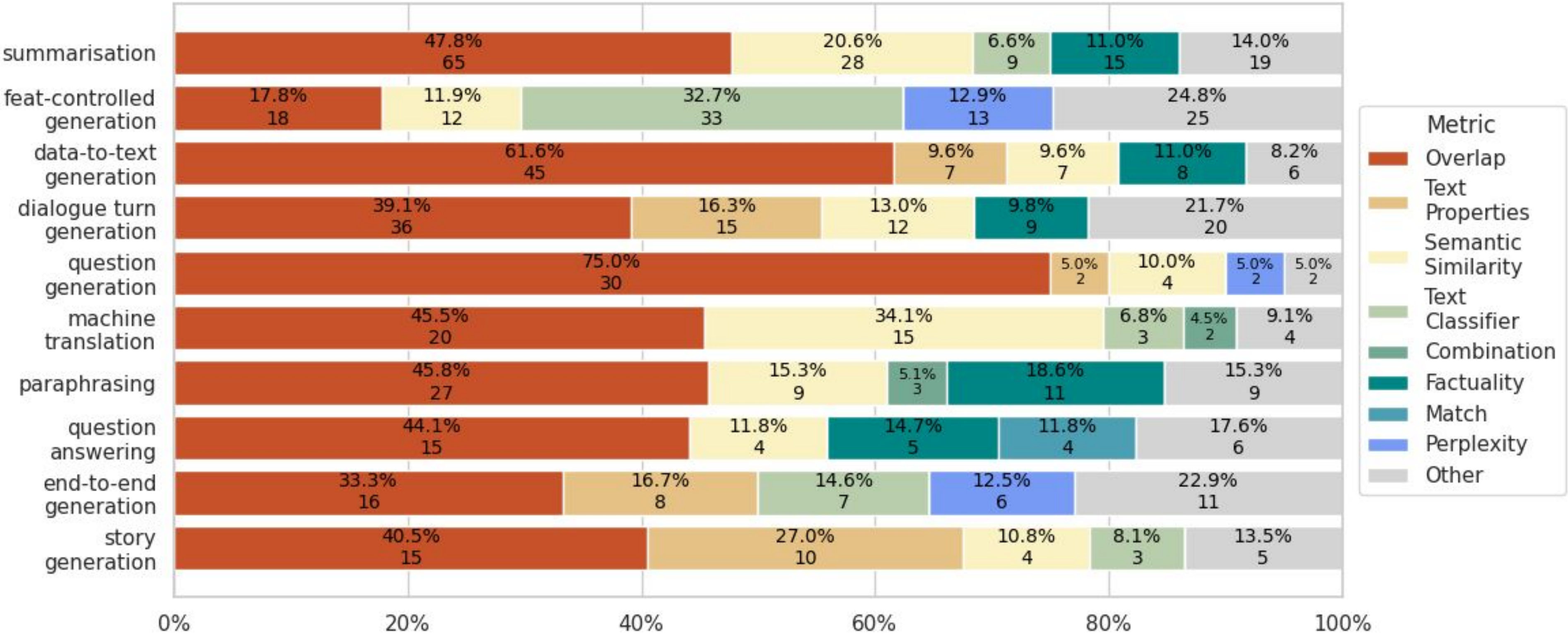
Correlation with Human Evaluation



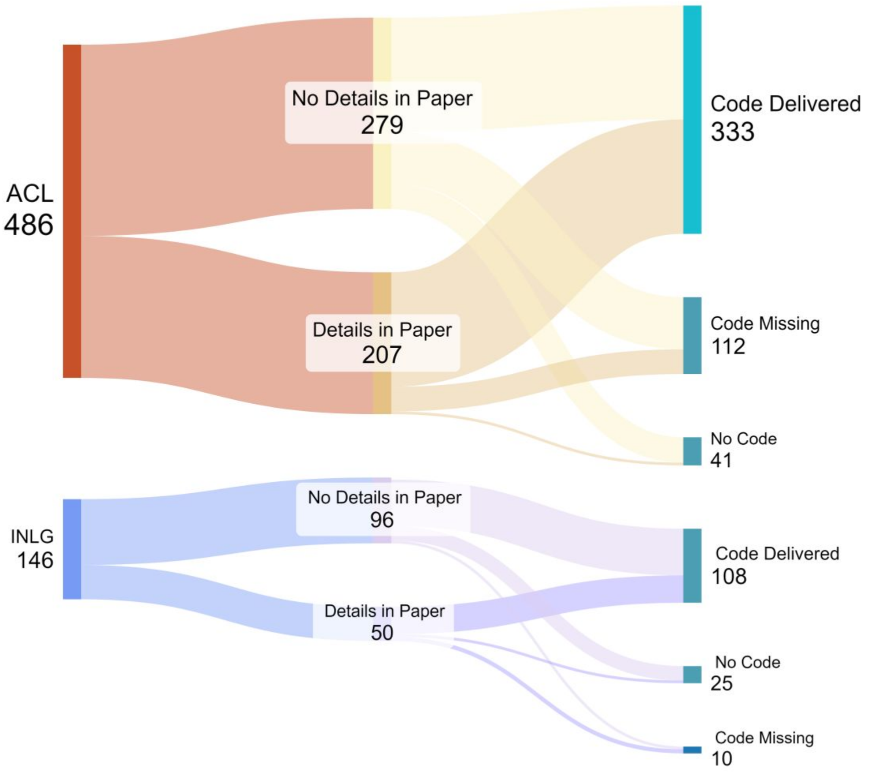


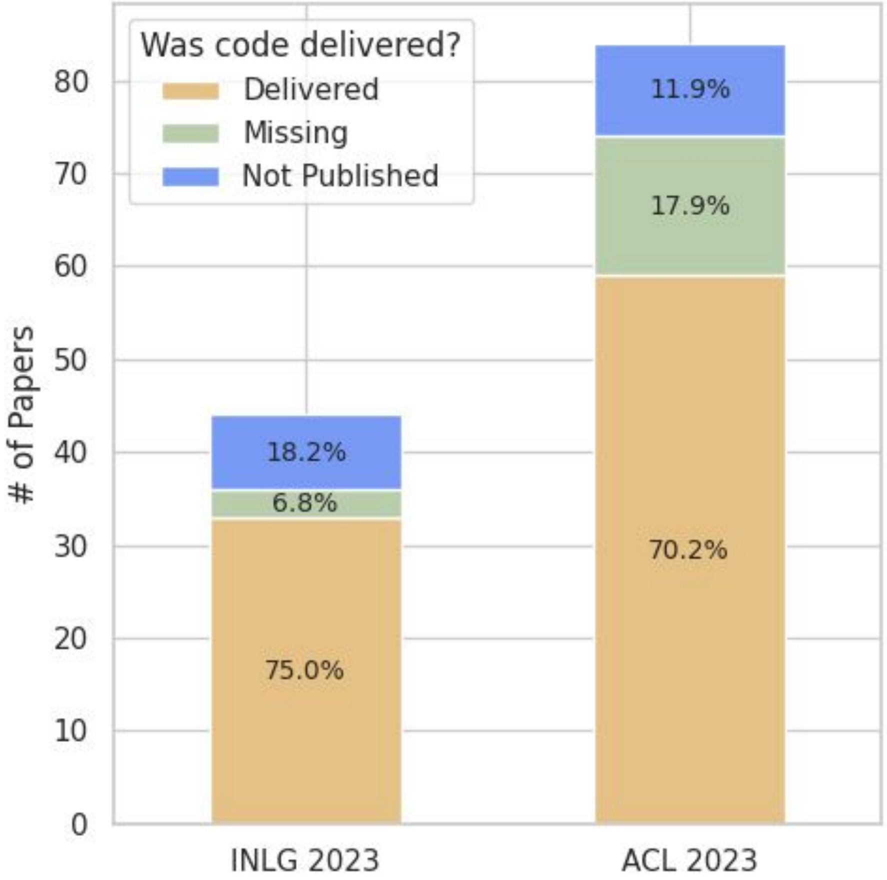
	Correlation	Following	None	Quality
No Correlation with Human Evaluation	2	17	244	39
Qualitative Correlation with Human Evaluation	1	10	40	7
Quantitative Correlation with Human Evaluation	2	8	31	8
No Human Evaluation	0	30	171	22

Results per Task



The Lack of Implementation Details





Recommendations - Evaluation Quality

- Rationalize your selection of metrics
- Comment on metric combinations
- Do not copy-paste widely used metrics
- Respect the intended use of metrics
- Discuss (dis)agreements between human and automatic evaluation

Recommendations - Evaluation Reproducibility

- Share evaluation details
- Share data samples
- Release code

Thank You! Questions?

Patricia

Saad

Simone

Ondřej

Albert

Dimitra

Dave

Ondřej

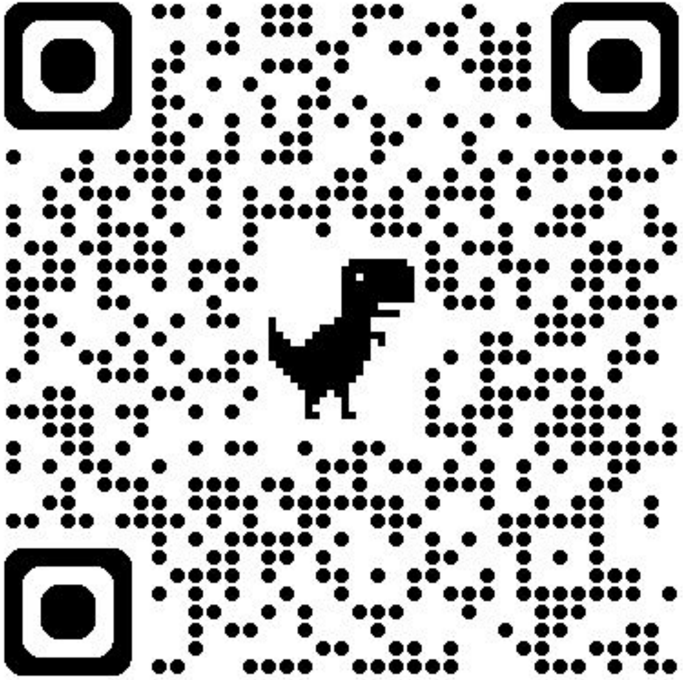
Adarsa

Correspondence to: schmidtova@ufal.mff.cuni.cz

Supported by the European Research Council (Grant agreement No. 101039303 NG-NLG) & Charles University Project SVV 260 698.
Using the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

Links

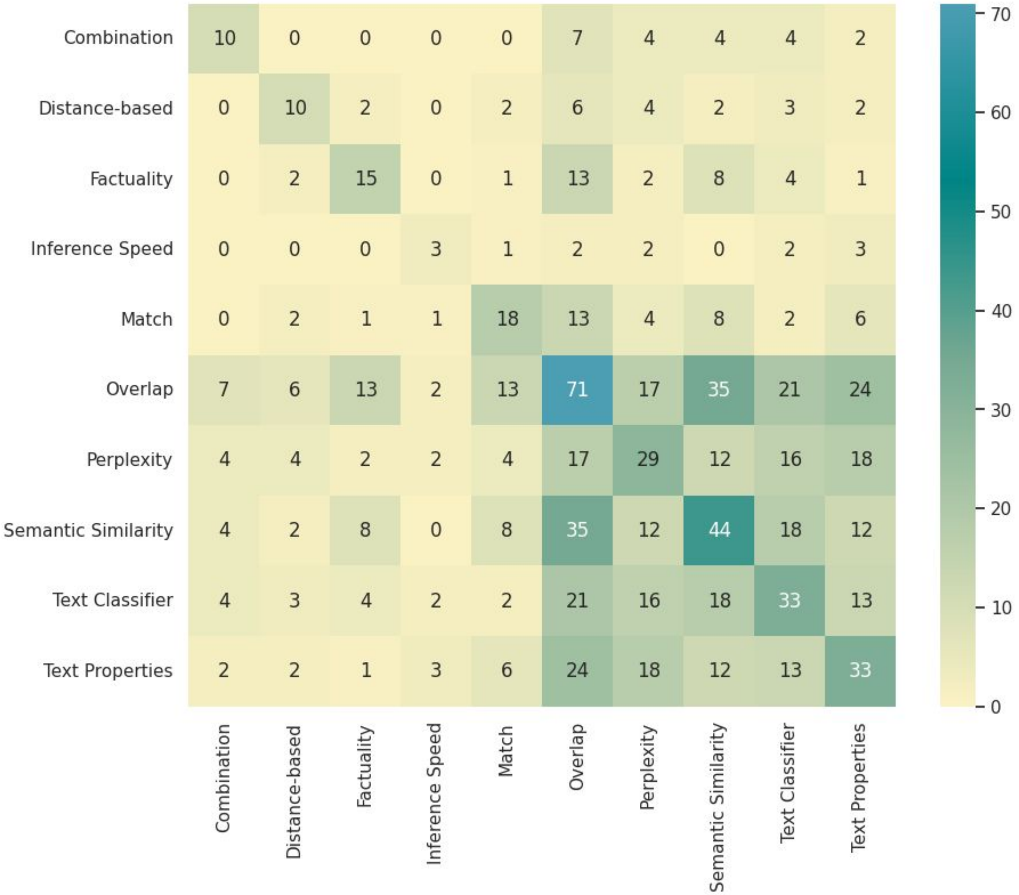
Link to the paper



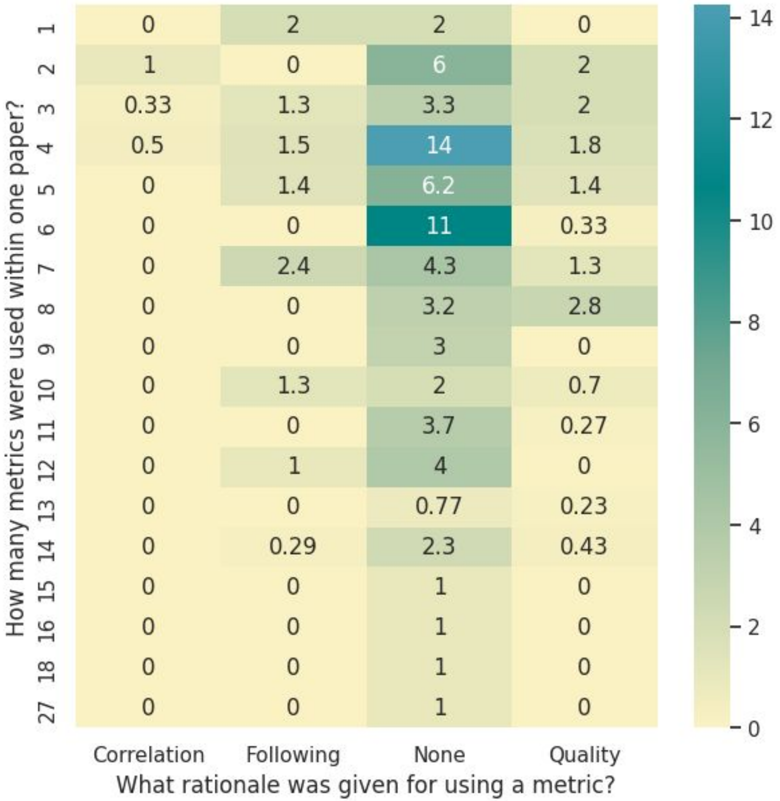
Link to GitHub

Backup Slides

Correlations

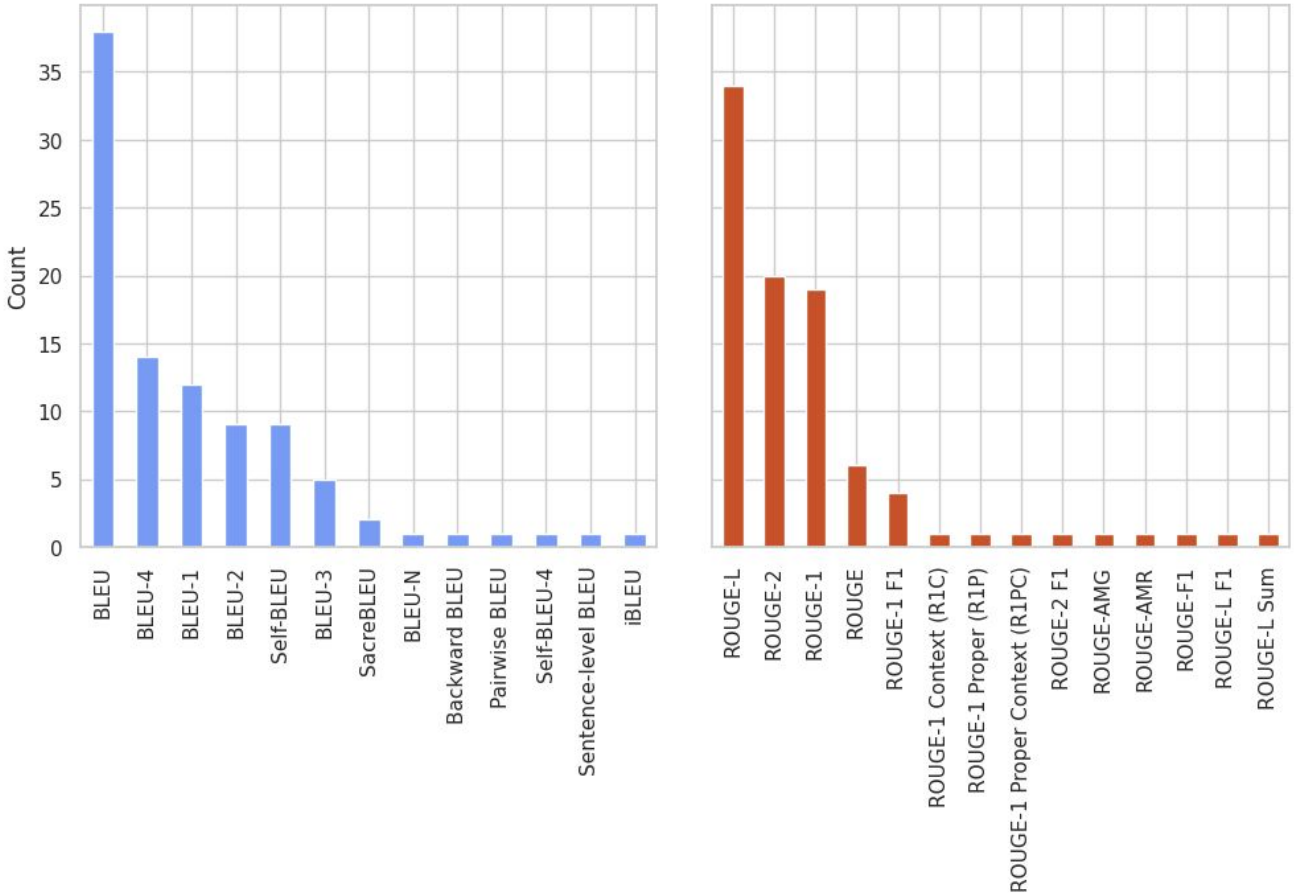


	Combination	Distance-based	Factuality	Inference Speed	Match	Overlap	Perplexity	Semantic Similarity	Text Classifier	Text Properties
Combination	10	0	0	0	0	7	4	4	4	2
Distance-based	0	10	2	0	2	6	4	2	3	2
Factuality	0	2	15	0	1	13	2	8	4	1
Inference Speed	0	0	0	3	1	2	2	0	2	3
Match	0	2	1	1	18	13	4	8	2	6
Overlap	7	6	13	2	13	71	17	35	21	24
Perplexity	4	4	2	2	4	17	29	12	16	18
Semantic Similarity	4	2	8	0	8	35	12	44	18	12
Text Classifier	4	3	4	2	2	21	16	18	33	13
Text Properties	2	2	1	3	6	24	18	12	13	33



	Correlation	Following	None	Quality
1	0	2	2	0
2	1	0	6	2
3	0.33	1.3	3.3	2
4	0.5	1.5	14	1.8
5	0	1.4	6.2	1.4
6	0	0	11	0.33
7	0	2.4	4.3	1.3
8	0	0	3.2	2.8
9	0	0	3	0
10	0	1.3	2	0.7
11	0	0	3.7	0.27
12	0	1	4	0
13	0	0	0.77	0.23
14	0	0.29	2.3	0.43
15	0	0	1	0
16	0	0	1	0
18	0	0	1	0
27	0	0	1	0

Variants of BLEU and ROUGE


Count
BLEU
BLEU-4
BLEU-1
BLEU-2
Self-BLEU
BLEU-3
SacreBLEU
BLEU-N
Backward BLEU
Pairwise BLEU
Self-BLEU-4
Sentence-level BLEU
iBLEU
ROUGE-L
ROUGE-2
ROUGE-1
ROUGE
ROUGE-1 F1
ROUGE-1 Context (R1C)
ROUGE-1 Proper (R1P)
ROUGE-1 Proper Context (R1PC)
ROUGE-2 F1
ROUGE-AMG
ROUGE-AMR
ROUGE-F1
ROUGE-L F1
ROUGE-L Sum