Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation

Cyril Chhun[1] **Pierre Colombo**[2*] **Fabian M. Suchanek**[1] **Chloé Clavel**[1]
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris
[2]Lab of Mathematics and Informatics (MICS), CentraleSupélec, Université Paris-Saclay
cyril.chhun@telecom-paris.fr

Abstract

Research on Automatic Story Generation (ASG) relies heavily on human and automatic evaluation. However, there is no consensus on which human evaluation criteria to use, and no analysis of how well automatic criteria correlate with them. In this paper, we propose to re-evaluate ASG evaluation. We introduce a set of 6 orthogonal and comprehensive human criteria, carefully motivated by the social sciences literature. We also present HANNA, an annotated dataset of 1,056 stories produced by 10 different ASG systems. HANNA allows us to quantitatively evaluate the correlations of 72 automatic metrics with human criteria. Our analysis highlights the weaknesses of current metrics for ASG and allows us to formulate practical recommendations for ASG evaluation.

1 Introduction

Storytelling is at the heart of human societies: skillful storytelling allows a narrator to connect more authentically with their audience and listeners, and to understand the essence of complex concepts better (Suzuki et al., 2018). Numerous applications could benefit from strong automatic story generation systems, including gaming (Hartsook et al., 2011), communication (Alhussain and Azmi, 2021), and education (Aylett et al., 2007). Several approaches have been explored to generate stories automatically or with minimum editing efforts (Alabdulkarim et al., 2021). *Automatic story generation* (ASG) takes as input a short sentence (a *prompt*) and aims at generating a narrative from it (Cavazza and Pizzi, 2006; Lebowitz, 1985). Advances in neural language models (Radford et al., 2018, 2019; Brown et al., 2020) have allowed substantial progress in ASG.

To further improve the quality of generated stories, it is indispensable to systematically evaluate ASG models. However, there is little work that specifically studies ASG evaluation. Most research works rely on human criteria such as coherence (Xu et al., 2018; Colombo et al., 2019; Jalalzai et al., 2020), relevance (Jhamtani and Berg-Kirkpatrick, 2020), overall quality (Brahman and Chaturvedi, 2020), narrative flow (Rashkin et al., 2020), and creativity (Pascual et al., 2021). However, taken individually, these criteria fail to encompass all aspects of the task, and there is no consensus on a set of criteria that would cover those aspects in a complete and non-redundant fashion. Due to the high cost of human annotation, system quality is also often evaluated using automatic metrics. However, it is not clear how these metrics correlate with human judgment in ASG, and thus how suitable they are at all for the evaluation of ASG.

Contributions. In this work, we revisit both human and automatic evaluation of ASG. We believe that this meta-evaluation is a missing piece in the ASG literature and a crucial step to strengthening the foundations of ASG. Formally, our contributions to the ASG field are:

1. **A comprehensive set of non-redundant human criteria for ASG evaluation.** Motivated by the social sciences literature (McCabe and Peterson, 1984; Dickman, 2003; Bae et al., 2021), we introduce six human criteria: relevance, coherence, empathy, surprise, engagement and complexity.

2. **HANNA**[1]**, a large annotated dataset of Human-ANnotated NArratives for ASG evaluation**, which contains 1,056 stories generated from 96 prompts. Each prompt is linked to a human story and stories generated by 10 different ASG generation systems. Each story was annotated by 3 different human raters along our 6 proposed human criteria.

*Previously from Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec, CNRS, Université Paris-Saclay.

[1]The HANNA dataset and corresponding code are available on https://github.com/dig-team/hanna-benchmark-asg.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5794–5836
October 12–17, 2022.

3. **A meta-evaluation of ASG with fine-grained recommendations.** Relying on HANNA, we perform an extensive study of the performance of the ASG systems and we analyze the correlations of 72 existing automatic metrics with our proposed human criteria. The obtained results demonstrate the limitations of current automatic evaluation methods and allow us to make recommendations on which metrics to use for ASG evaluation.

2 Related work

2.1 Human evaluation

van der Lee et al. (2019) advise to define separate and precise criteria for human evaluation to make it as accurate as possible. However, in ASG, there is no consensus on the criteria to be used: among others, we find a pairing task (Fan et al., 2018), fluency and coherence (Xu et al., 2018), creativity (Pascual et al., 2021), faithfulness (Peng et al., 2018; Wang et al., 2020), fidelity (Yao et al., 2019), grammar and logicality (Guan et al., 2019, 2020), overall quality and relevance (Jhamtani and Berg-Kirkpatrick, 2020; Goldfarb-Tarrant et al., 2020; Guan et al., 2021b), outline utilisation and narrative flow (Rashkin et al., 2020), emotion faithfulness (Witon et al., 2018), and content quality (Brahman and Chaturvedi, 2020). Many of these criteria are not specific to ASG (fluency, grammar, overall quality, content quality), overlap with one another (pairing task, faithfulness, and fidelity are variations of relevance; logicality and narrative flow, of coherence) or are ascribed to a specific setting (outline utilisation, emotion faithfulness). Furthermore, evaluation protocols mostly use only two or three criteria, which is not enough to grasp all aspects of a task as complex as ASG. They also do not associate Likert scales with explicit descriptions, even though such descriptions could reduce the subjectivity of the labelling process.

2.2 Automatic evaluation

Although most of the research work in ASG relies on BLEU and ROUGE, there exists a plethora of automatic metrics to evaluate ASG. These can be classified into two categories: *reference-based* (Ξ) metrics evaluate a candidate text by comparing it to a reference text (in our case, the human story), and *reference-free* (¤) metrics rely only on the candidate story (and, possibly, on the prompt). In both categories, we find *string-based* (§), *embedding-based* (ε) and *model-based* (Δ) metrics. String-based metrics evaluate the textual representation of the inputs; they cannot handle synonyms or paraphrases. By contrast, embedding-based metrics rely on word embeddings, *e.g.* word2vec (Mikolov et al., 2013a,b), or contextualized embeddings, *e.g.* obtained from BERT (Devlin et al., 2019). Finally, model-based metrics leverage regression or pretrained language models to return a score. A synoptic classification can be found in Tab. 1[2].

	Reference-based (Ξ)	Reference-free (¤)
String-based (§)	BLEU (Papineni et al., 2002) ROUGE (Lin, 2004) METEOR (Banerjee and Lavie, 2005) CHRF (Popović, 2015) CIDEr (Vedantam et al., 2015)	Coverage (Grusky et al., 2018) Density (Grusky et al., 2018) Compression (Grusky et al., 2018) Text length (Fabbri et al., 2021) Novelty (Fabbri et al., 2021) Repetition (Fabbri et al., 2021)
Embedding-based (ε)	ROUGE-WE (Ng and Abrecht, 2015) BERTScore (Zhang et al., 2020) MoverScore (Zhao et al., 2019) BaryScore (Colombo et al., 2021d) DepthScore (Staerman et al., 2021)	SUPERT (Gao et al., 2020)
Model-based (Δ)	S3 (Peyrard et al., 2017) SummaQA (Scialom et al., 2019) InfoLM (Colombo et al., 2022c)	BLANC (Vasilyev et al., 2020)
	BARTScore (Yuan et al., 2021)	BARTScore (Yuan et al., 2021)

Tab. 1: Classification of the automatic metrics considered in our study with symbols for easier identification.

2.3 Meta-evaluation

Several previous works have studied the relationship between automatic metrics and human judgment (Zhang et al., 2004; Ma et al., 2019), reporting weak correlation (Novikova et al., 2017; Stent et al., 2005; Mathur et al., 2020) and strong bias towards specific systems (Callison-Burch et al., 2006). Meta-evaluation has been done in image description (Elliott and Keller, 2014), dialogue response generation (Liu et al., 2016), question generation (Nema and Khapra, 2018), table-to-text generation (Dhingra et al., 2019), question answering (Chen et al., 2019), and summarization (Bhandari et al., 2020). In ASG, Guan et al. (2021b) introduced the OpenMEVA benchmark which compares the overall quality of human and generated stories; their work especially focused on the textual features of stories. We build upon it and perform a comprehensive analysis of the correlations between 72 automatic metrics and 6 human criteria specifically tailored for ASG.

[2] BARTScore was designed to be either reference-based or reference-free depending on the setting.

3 HANNA for ASG evaluation

3.1 ASG datasets

Story evaluation has been widely studied in different scenarii. ROCStories (Mostafazadeh et al., 2016), a corpus of 50k 5-sentence stories with titles, was designed for the Story Cloze Test: the prediction of the final sentence of a story given the four others. Huang et al. (2016) developed the VisualStorytelling dataset, which contains sequences of images with corresponding descriptions divided in three tiers of temporal context. More recently, Ammanabrolu et al. (2020) proposed the WorldGeneration dataset which adapts story generation to adventure games by guiding the generation process with location, character and object triplets. The WritingPrompts (WP) dataset (Fan et al., 2018) contains stories generated from short sentences called *prompts*. For our work, we chose the WP dataset, because it has been extensively used in previous literature for the design of ASG models (Rashkin et al., 2020; Goldfarb-Tarrant et al., 2020; Fang et al., 2021; Wilmot and Keller, 2021; Guan et al., 2021a). While ROCStories has also been used in several works, the shortness of the stories made it less suited for our evaluation. An example prompt and story from WP is shown in Tab. 2 (Fan et al., 2018).

3.2 Chosen setting

HANNA, our annotated dataset for ASG, contains outputs from 10 different systems aligned on 96 common prompts with human stories from the WP dataset, for 1,056 stories in total, with 3 human annotations per story (19,008 annotations in total) and automatic metric scores, allowing for an analysis of the correlations between these metrics (Sec. 4).

3.3 Chosen ASG sytems

We directly contacted the authors of articles that introduced ASG systems and asked for the outputs of their systems. We managed to collect the outputs of **3 ASG systems**[3] on the WP dataset: Fusion (Fan et al., 2018), HINT (Guan et al., 2021a), and TD-VAE (Wilmot and Keller, 2021). We extracted 96 stories aligned on common prompts. We then fine-tuned **7 pre-trained language models** for ASG on a causal language modeling task on WP to generate stories on the same 96 prompts, using the *Transformers* library (Wolf et al., 2020)[4]. We trained BertGeneration (Rothe et al., 2020), CTRL (Keskar et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019), and GPT-2 (tag), another instance of GPT-2 trained with <EOP> (End Of Prompt) tags, as inspired by Bai et al. (2021), who argued that such tags could improve generation.

3.4 Proposed human criteria

As mentioned in Ssec. 2.1, there is no consensus on human criteria for ASG evaluation. At the same time, work in social sciences has looked extensively at the features that make for a "good" story (McCabe and Peterson, 1984; Dickman, 2003; Bae et al., 2021). We condense them as follows into a new, comprehensive set of criteria:

1. **Relevance** (RE): how well the story matches its prompt, used in Jhamtani and Berg-Kirkpatrick (2020); Goldfarb-Tarrant et al. (2020);

2. **Coherence** (CH): how much the story makes sense, used in Xu et al. (2018); Peng et al. (2018); Yao et al. (2019); Pascual et al. (2021);

3. **Empathy** (EM): how well the reader understood the character's emotions, derived from the importance of emotional commentary (McCabe and Peterson, 1984), passion (Dickman, 2003), and empathy (Keen et al., 2007; Bae et al., 2021);

4. **Surprise** (SU): how surprising the end of the story was, derived from the importance of schema violation, or unexpectedness (Schank, 1978; Bae et al., 2021), postdictability (Behrooz et al., 2019), and novelty (Randall, 1999);

5. **Engagement** (EG): how much the reader engaged with the story; a more subjective criterion associated with projecting volitive modality (making the reader formulate a subjective judgment and express a desire to see something accomplished) (Toolan, 2012) and story outcome, which is an underlying cause of story liking (Iran-Nejad, 1987);

6. **Complexity** (CX): how elaborate the story is; derived from the importance of detailed descriptions and sophisticated problem-solving (McCabe and Peterson, 1984) and good world-building (Roine, 2016).

[3] We also collected outputs from two other systems (Goldfarb-Tarrant et al., 2020; Bai et al., 2021); unfortunately, these were not aligned with the others.

[4] https://github.com/huggingface/transformers

The four last criteria are an original contribution and were designed to evaluate story features that are different from the first two criteria (RE and CH), which are currently most used in the ASG literature. Examples of annotations w.r.t. those criteria are shown in Tab. 2.

3.5 Annotation Protocol

To evaluate our human criteria on the 1,056 stories of HANNA, we conducted an annotation campaign on Amazon Mechanical Turk. As advised by Karpinska et al. (2021), for each task, we provided the human story alongside the story to be annotated, so that the workers could calibrate their judgment. Each of the stories was evaluated by three workers on our six proposed criteria. For this evaluation, we chose a 5-point Likert scale rather than a rank-based comparison because we reckoned that it would be tedious to order the large number of evaluated systems. We estimated that a HIT should take between 90 and 120 seconds, so we set the remuneration at \$0.28 per HIT, or between \$8.40 and \$11.40 per hour. To ensure that annotators spoke fluent English, we restricted access to the experiment to workers located in the UK, the US, Canada, Australia and New Zealand. We further required them to have the Masters Qualification. To remove noisy annotations and ensure that the workers read the stories, we chose to reject judgments that were made in fewer than 30 seconds. We additionally asked workers to write down the name of the first-mentioned fictional character of the story. The detailed instructions of the experiment and the inter-annotator agreement analysis can be found in the appendix (see Ap. A and Ssec. 4.1). Finally, following the recommendations of Shapira et al. (2019), we obtained the human score of a story by averaging the results of the three workers.

3.6 Meta-evaluation strategies

Notations. Let y_i^j be the story generated by system $j \in \{1, \dots, S\}$ for prompt $i \in \{1, \dots, N\}$, and $m(y_i^j)$ the score associated to y_i^j by a (human or automatic) metric m. Given a correlation coefficient K (*e.g.* Pearson's r (**?**), Spearman's ρ (Melamed et al., 2003) or Kendall's τ (Kendall, 1938)), two meta-evaluation strategies are commonly used to evaluate metric quality.

Story-level correlation ($K^{\textbf{story}}_{m_1,m_2}$) measures how suited m_1 is w.r.t. m_2 if used as a loss or reward for a model. The correlation is applied to each story among all system outputs and the mean is taken. Formally:

$$K^{\text{story}}_{m_1,m_2} \triangleq \frac{1}{N}\sum_{i=1}^{N} K(\mathbf{C}^{\text{story}}_{m_1,i}, \mathbf{C}^{\text{story}}_{m_2,i}), \quad (1)$$

$$\text{where} \quad \mathbf{C}^{\text{story}}_{m,i} \triangleq \left[m(y_i^1), \cdots, m(y_i^S)\right].$$

System-level correlation ($K^{\textbf{sys}}_{m_1,m_2}$) measures how suited m_1 is w.r.t. m_2 if used to compare the performance of two systems. The correlation is applied to the mean values over all stories for all systems for both metrics. Formally:

$$K^{\text{sys}}_{m_1,m_2} \triangleq K\left(\frac{1}{N}\mathbf{C}^{\text{sys}}_{m_1}, \frac{1}{N}\mathbf{C}^{\text{sys}}_{m_2}\right), \quad (2)$$

$$\text{where} \quad \mathbf{C}^{\text{sys}}_{m} \triangleq \left[\sum_{i=1}^{N} m(y_i^1), \dots, \sum_{i=1}^{N} m(y_i^S)\right].$$

Statistical significance. Correlations computed for two automatic metrics on the same annotated dataset are not independent. We follow Graham and Baldwin (2014) and use the Williams test (Williams, 1959; Moon, 2019)[5] to evaluate the significance of an increase in dependent correlations (Steiger, 1980).

4 HANNA Analysis

In this section, we analyse the scores of HANNA in detail. Tab. 4 shows that human stories achieve significantly higher scores than generated stories. Following Mathur et al. (2020), who advise to remove outliers, we compute correlations with human stories removed[6].

4.1 Inter-annotator agreement

To estimate the reliability of the annotations, we computed an intra-class coefficient for each criterion. Among the annotators which took part in the experiment, three of them covered 2490 stories, *i.e.*, more than 78% of the dataset, but no annotator graded the same story twice. Since the reliability is to be estimated for the average of the three ratings, the ICC2k coefficient (ICC for *average random raters*) is the most relevant one, according to Hallgren (2012). In particular, it accounts for the systematic errors of raters and random residual errors. The results are shown in Tab. 3.

Coefficients are dispersed between 29% and 56% with relatively small confidence intervals (except for RE and CH), which can be considered between

[5] https://github.com/inmoonlight/nlp-williams

[6] The same applies for Sec. 5.

Prompt: When you die, the afterlife is an arena where you face every insect and animal you killed in your life. If you win you go to heaven, lose you go to hell. Your job was an exterminator on earth.

Human: 3,000 years have I been fighting. Every morning, the raccoons scratch at my eyes. Every evening, the skunks spray me while the opossums chew at my feet. I have never had any tools. I have only my hands. I don't remember the place I came from before this. [...]

Story #1: First of all, not everyone was entitled to be an exterminator. But the ones that were – maybe were, like, *genius*, because, yes, I had once belonged to a less fortunate class of people – had all the opportunity to work for the damn plant killer, and it's hard work. [...]

Story #2: It was hell. Not exactly a place of torture. There were no guards in prison and you couldn't just walk through it, either, because you would get killed regardless. hell was a young man, and he was lying on his floor. He was unconscious. [...]

Story	RE	CH	EM	SU	EG	CX
Human	5 2 4	5 2 4	1 3 3	3 2 2	4 2 4	1 3 4
Story #1	2 2 2	4 2 3	3 2 2	1 1 3	1 2 3	1 2 3
Story #2	5 3 3	5 2 4	3 3 3	3 2 4	3 2 4	2 3 3

Metric	Human	Story #1	Story #2
BLEU$^{\Xi§}$ (%)	1.00	0.01	0.01
ROUGE-1$^{\Xi§}$	1.00	0.24	0.33
chrF$^{\Xi§}$ (%)	1.00	0.32	0.39
BERTScore$^{\Xi\varepsilon}$	1.00	0.50	0.52
MoverScore$^{\Xi\varepsilon}$	1.00	0.51	0.51
BaryScore$^{\Xi\varepsilon}$	0.00	0.92	0.92
S3$^{\Xi\Delta}$	1.39	0.07	0.15
BARTScore$^{\Xi\Delta}$	-0.98	-3.97	-4.03
SUPERT$^{¤\varepsilon}$	0.94	0.37	0.36

Tab. 2: Example prompt, human and generated stories from HANNA with human annotations and metric scores

"fair" and "moderate" according to Landis and Koch (1977). These values are in tune with existing literature (Karpinska et al., 2021; Habernal and Gurevych, 2017; Spooren and Degand, 2010; Ritter et al., 2011; Graham et al., 2017) and show the difficulty of evaluating natural language generation. Therefore, we follow the methodology of Craggs and Wood (2005) and Artstein and Poesio (2008), who argue against setting a specific agreement threshold as long as there is a detailed reporting of the methodology (see Ssec. 3.5 and Tab. 7) and confidence intervals (Tab. 3).

Criterion	LB	ICC2k	UB
RE	0.18	0.48	0.65
CH	0.10	0.29	0.48
EM	0.25	0.34	0.41
SU	0.16	0.28	0.37
EG	0.34	0.46	0.55
CX	0.48	0.56	0.63

Tab. 3: Intra-class coefficient per criterion. LB and UB are the lower- and upper-bounds of the 95% confidence interval

4.2 Evaluating our human criteria

In this experiment, we study the relationship between the proposed human criteria. To compute story-level (Fig. 1) and system-level (Fig. 2) correlations, we average the human ratings.

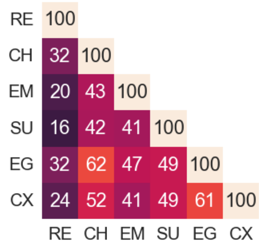

Fig. 1: Story-level Kendall correlations (%) between human criteria

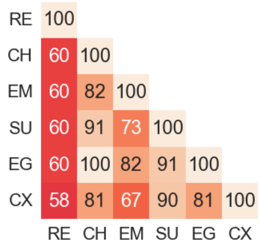

Fig. 2: System-level Kendall correlations (%) between human criteria

Story-level analysis (Fig. 1). Kendall correlations range from 16% (RE–SU) to 62% (CH–EG), averaging at 40.7%. In the appendix, we also show correlations with Spearman's ρ (Fig. 10) and Pearson's r (Fig. 12). EG correlates slightly more with CH and CX; this could confirm that coherent and intricate plots make readers more likely to connect with a story. In contrast, RE is poorly correlated to the other criteria, which makes sense: an excellent story in every other aspect can be completely unrelated to a prompt, and vice versa. Overall, moderate to weak correlations suggest that our criteria evaluate distinct aspects of storytelling which cannot be regrouped in fewer criteria.

System-level analysis (Fig. 2). Compared to story-level correlations, system level correlations are higher. Spearman (Fig. 11) and Pearson (Fig. 13)

Model	RE	CH	EM	SU	EG	CX	Average
Human	4.17 ±0.14	4.43 ±0.10	3.22 ±0.14	3.15 ±0.15	3.88 ±0.12	3.73 ±0.13	3.76 ±0.06
BertGeneration	2.46 ±0.16	3.14 ±0.16	2.28 ±0.13	**2.09** ±0.13	2.67 ±0.12	2.41 ±0.11	2.51 ±0.06
CTRL	2.54 ±0.16	2.93 ±0.16	2.26 ±0.13	1.93 ±0.12	2.53 ±0.12	2.23 ±0.10	2.40 ±0.06
GPT	2.40 ±0.16	**3.22** ±0.15	**2.37** ±0.12	**2.13** ±0.13	2.76 ±0.13	2.49 ±0.12	2.56 ±0.06
GPT-2	* **2.81** ±0.16	**3.29** ±0.14	* **2.47** ±0.12	**2.21** ±0.13	**2.86** ±0.12	2.68 ±0.10	**2.72** ±0.06
GPT-2 (tag)	**2.67** ±0.16	* **3.31** ±0.15	* **2.47** ±0.12	* **2.22** ±0.13	* **2.92** ±0.12	* **2.80** ±0.11	* **2.73** ±0.06
RoBERTa	2.54 ±0.16	3.22 ±0.16	2.27 ±0.12	**2.12** ±0.13	2.74 ±0.12	2.41 ±0.11	2.55 ±0.06
XLNet	2.39 ±0.17	2.88 ±0.16	2.10 ±0.12	1.95 ±0.12	2.46 ±0.13	2.36 ±0.11	2.36 ±0.06
Fusion	2.09 ±0.16	2.86 ±0.16	1.99 ±0.12	1.72 ±0.12	2.27 ±0.14	1.92 ±0.11	2.14 ±0.06
HINT	2.29 ±0.16	2.38 ±0.16	1.74 ±0.13	1.56 ±0.11	1.75 ±0.12	1.45 ±0.10	1.86 ±0.06
TD-VAE	2.51 ±0.16	2.99 ±0.15	2.07 ±0.11	**2.10** ±0.12	2.59 ±0.12	2.49 ±0.11	2.46 ±0.06

Tab. 4: Average system ratings per criterion with 95% confidence interval. Best value in **bold** marked with an asterisk (*), values in the confidence interval of the best value in **bold** without asterisk

correlations are also higher than their story-level counterparts. This suggests that a given system tends to be uniformly better or worse than other systems across all criteria.

4.3 Finding the best systems

On Tab. 4, we observe that generic fine-tuned models perform better than ASG systems according to human annotators. The best system is `GPT-2`, which scores better than all other systems on all criteria. The `GPT-2` variant trained with `<EOP>` tags shows marginal improvement compared to `GPT-2`, as reported in Bai et al. (2021). It is worth noting that all models are still noticeably below human performance, which emphasizes that current systems are still a long way from human-like narrative intelligence.

5 Evaluation of automatic metrics using HANNA

In this section we evaluate how suitable existing automatic metrics are for ASG evaluation, using the `SummEval` library (Fabbri et al., 2021)[7]. Due to space constraints, in each figure, we selected representative metrics from each of the categories introduced in Ssec. 2.2. Full figures can be found in the appendix.

5.1 Correlations with human judgement

Story-level analysis (Fig. 3). Most metrics have either a moderate (between 30% and 50%) or weak (below 30%) correlation with human criteria. `RE` is particularly elusive, except for the `SUPERT`$^{¤\varepsilon}$ metric, which is reference-free and compares the prompt and the story. This corroborates Novikova et al. (2017), who argue that automatic metrics do not accurately reflect human judgment when comparing instances despite performing reliably at the system level. We also observe vertical "color stripes": metric performance is consistent across criteria. A weak metric will correlate poorly with all criteria, whereas a more robust metric will be uniformly better.

System-level analysis (Fig. 4). Correlations are indeed higher at the system-level, hovering between 40% and 70% for most metrics. Therefore, while metrics are poor estimators of human criteria at the story level, they can be used to compare systems with reasonable accuracy.

Best metrics per criterion (Tab. 5). We observe that a few metrics are heavily represented in the top 3 for each level. Pretrained transformer-based metrics achieve strong results. The metrics that correlate best with human criteria at the system level are all reference-based: `ROUGE-S*`$^{\Xi\S}$, `BaryScore`$^{\Xi\varepsilon}$, `DepthScore`$^{\Xi\varepsilon}$, and `BARTScore`$_1^{\Xi\Delta}$. At the story level, `BARTScore`$_2^{¤\Delta}$, `Novelty-1`$^{¤\S}$ and `Repetition-3`$^{¤\S}$ are reference-free while `chrF`$^{\Xi\S}$, `BERTScore`$^{\Xi\varepsilon}$, `S3`$^{\Xi\Delta}$ are reference-based. As `Novelty-1` and `Repetition-3` are simple data statistics, their outperforming all metrics for `SU` and `CH` respectively highlights the shortcomings of current metrics.

5.2 Correlations between automatic metrics

Story-level analysis (Fig. 5). Reference-based metrics are moderately to highly correlated with one another. By contrast, embedding- and model-based reference-free metrics such as `SUPERT`$^{¤\varepsilon}$ and `BLANC`$^{¤\Delta}$ are almost independent from all other metrics, even other reference-free metrics.

System-level analysis (Fig. 6). Previous obser-

[7] https://github.com/Yale-LILY/SummEval

Fig. 3: Story-level absolute Kendall correlations (%) between metrics and criteria. Full version: Fig. 14.

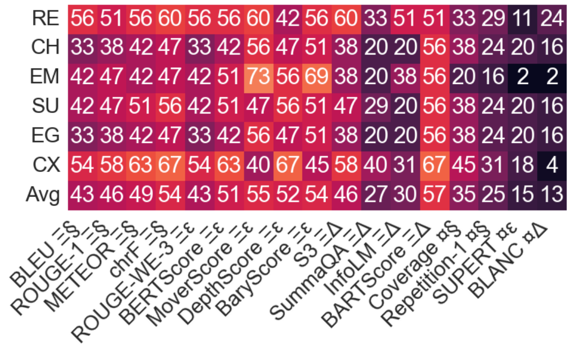

Fig. 4: System-level absolute Kendall correlations (%) between metrics and criteria. Full version: Fig. 15.

Level	Criterion	Metric #1	$\|r\|$ (%)	Metric #2	$\|r\|$ (%)	Metric #3	$\|r\|$ (%)
Story	RE	BARTScore$_2^{¤Δ}$	42.6	SUPERT$_1^{¤ε}$	41.2	SUPERT$_2^{¤ε}$	40.2
	CH	Repetition-3$^{¤§}$	38.1	BERTScore$_R^{Ξε}$	37.1	S3^{ΞΔ}	37.1
	EM	S3^{ΞΔ}	32.8	chrF^{Ξ§}	32.4	BERTScore$_R^{Ξε}$	32.1
	SU	Novelty-1$^{¤§}$	32.9	chrF^{Ξ§}	32.7	ROUGE-1^{Ξ§}	31.3
	EG	BERTScore$_R^{Ξε}$	43.0	Novelty-1$^{¤§}$	42.3	chrF^{Ξ§}	41.1
	CX	chrF^{Ξ§}	58.8	BERTScore$_R^{Ξε}$	55.8	ROUGE-1^{Ξ§}	55.0
System	RE	ROUGE-S*$_F^{Ξ§}$	80.4	ROUGE-SU*$_F^{Ξ§}$	80.3	ROUGE-S*$_R^{Ξ§}$	80.2
	CH	BaryScore$_1^{Ξε}$	88.2	BaryScore$_2^{Ξε}$	88.0	BERTScore$_F^{Ξε}$	87.9
	EM	BaryScore$_1^{Ξε}$	90.0	BaryScore$_2^{Ξε}$	90.0	BERTScore$_F^{Ξε}$	88.7
	SU	BARTScore$_1^{ΞΔ}$	92.7	BERTScore$_R^{Ξε}$	91.1	DepthScore^{Ξε}	90.7
	EG	DepthScore^{Ξε}	93.4	BARTScore$_1^{ΞΔ}$	92.4	SUPERT$_2^{¤ε}$	92.2
	CX	DepthScore^{Ξε}	95.6	BERTScore$_R^{Ξε}$	95.5	Compression$^{¤§}$	94.3

Tab. 5: Top 3 metrics per criterion per level (story or system) of absolute Pearson (r) correlation. Indices denote different variants.

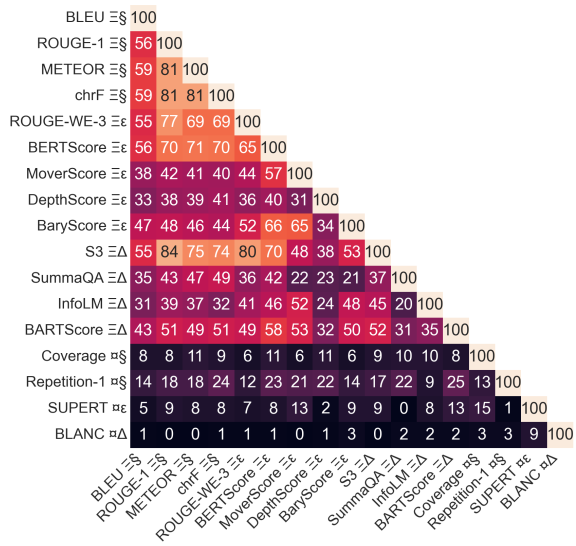

Fig. 5: Story-level absolute Kendall correlations (%) between metrics. Full version: Fig. 20.

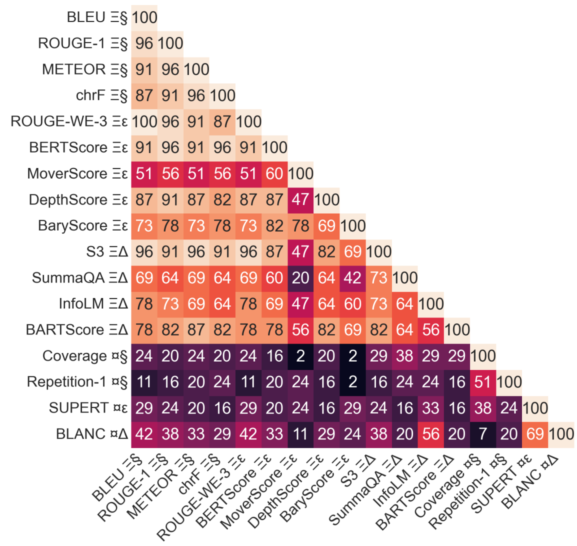

Fig. 6: System-level absolute Kendall correlations (%) between metrics. Full version: Fig. 21.

vations at the story level remain mostly valid, although correlations are overall higher. Reference-based metrics form a large group of very highly correlated metrics, with a majority of correlations surpassing 70%. Embedding- and model-based reference-free metrics remain weakly correlated to other metrics.

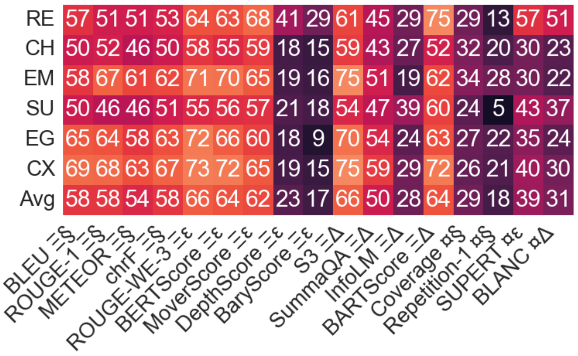

Fig. 7: Weighted macro F1-scores of paired bootstrap resampling. Full version: Fig. 26.

5.3 Fine-grained analysis

Top-k systems (Fig. 8). Here, we evaluate whether automatic metrics can reliably quantify differences between systems of competitive performances. For all criteria except RE and CX, correlations follow a convex curve between $k = 10$ and $k = 4$, suggesting that metrics should not be used to compare systems of high variance in quality unless there are enough of them. Indeed, removing a few systems causes correlations to worsen significantly, until the remaining systems are few enough and of competitive performance. RE correlations interestingly increase as k decreases, which indicates that system quantity is a lesser concern for RE.

Pairwise system comparison (Fig. 7). Here, we evaluate the pairwise discriminative power of automatic metrics. Following Bhandari et al. (2020), we take all system pairs (s_1, s_2) and compute their average ratings per criterion using paired bootstrap resampling (Koehn, 2004; Dror et al., 2018). We assign a label $y_{\text{true}} = 1$ if s_1 is better than s_2 with 95% confidence, $y_{\text{true}} = 2$ if s_2 is better, and $y_{\text{true}} = 0$ if confidence is below 95%. We then repeat the procedure for each metric m, getting $y^{(m)}_{\text{pred}}$ labels, and calculate the weighted macro F1-scores (Goutte and Gaussier, 2005) between y_{true} and $y^{(m)}_{\text{pred}}$ to evaluate if m is a good proxy for human criteria. We observe that reference-based metrics again perform better than reference-free metrics, with S3$^{\Xi\Delta}$ and ROUGE-WE-3$^{\Xi\varepsilon}$ at the top. DepthScore$^{\Xi\varepsilon}$ and BaryScore$^{\Xi\varepsilon}$ prove to be very unsuited for pairwise system comparisons, despite showing high system-level correlations (see Fig. 3). Finally, SU appears to be the most troublesome criterion for this task, suggesting that the surprise factor is especially difficult to evaluate.

Statistical significance. Using the Williams test (Ap. B), we found that increases in correlation with human criteria between top 3 metrics per criterion (Tab. 5) are not statistically significant, which suggests that best-scoring metrics are of similar performance. However, except for the RE criterion, we notably find that the increases in correlation of chrF$^{\Xi\S}$ and BERTScore$^{\Xi\varepsilon}$ compared to BLEU$^{\Xi\S}$ and ROUGE$^{\Xi\S}$ variants are statistically significant.

5.4 Aggregated rankings of metrics

To aggregate the scores obtained by the three correlation measures (Kendall, Pearson and Spearman), we use the work of Colombo et al. (2022a)[8], who rely on the Kemeny consensus (Kemeny, 1959; Myerson, 1996) and recommend to use the Borda Count (BC) as an efficient approximation (Sibony, 2014). They experimentally show that Kemeny consensus has more desirable properties than a ranking obtained through a mean-aggregation procedure. We report the results in Tab. 6. To compare system performance, model- or embedding-based metrics (*e.g.*. BARTScore$^{\Xi\varepsilon}$ or BERTScore$^{\Xi\varepsilon}$) seem most adapted. However, at the story level, chrF$^{\Xi\S}$ and BERTScore$^{\Xi\varepsilon}$ are among the best metrics, while BLEU$^{\Xi\S}$ is completely absent from the top spots. ROUGE$^{\Xi\S}$ does appear in the ranking, albeit below chrF$^{\Xi\S}$.

Level	Metric	BC
Story	chrF$^{\Xi\S}$	1237
	S3$_p^{\Xi\Delta}$	1198
	ROUGE-1$^{\Xi\S}$	1186
	S3$_r^{\Xi\Delta}$	1177
	BERTScore$_R^{\Xi\varepsilon}$	1158
System	BARTScore$^{\Xi\Delta}$	1120
	BaryScore$_5^{\Xi\varepsilon}$	1110
	BERTScore$_F^{\Xi\varepsilon}$	1095
	MoverScore$^{\Xi\varepsilon}$	1070
	DepthScore$^{\Xi\varepsilon}$	1069

Tab. 6: Top 5 metrics computed by one-level ranking per aggregation level, higher Borda count is better

6 Conclusions

Our analysis yields the following conclusions:

1. **Large pre-trained language models seem to produce the best results for ASG.** Our benchmark shows that GPT-2 performed better than systems specifically tailored for ASG despite being older

[8] https://github.com/PierreColombo/RankingNLPSystems

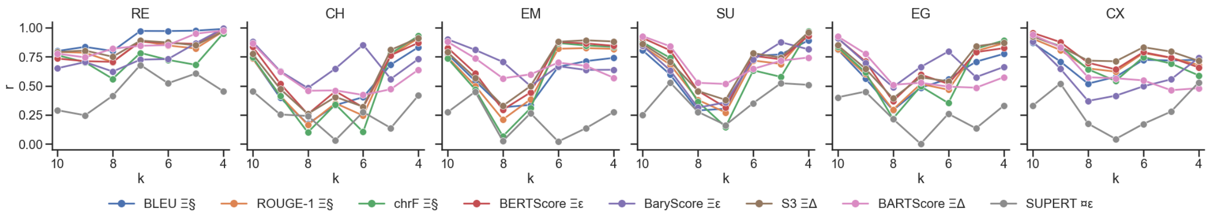

Fig. 8: System-level absolute Pearson correlations (%) between automatic metrics and our proposed human criteria on top-k systems

than some of them. Overall, all systems remain significantly inferior to human output, illustrating that ASG remains a challenging task for current language models.

2. **Stronger metrics, tailored explicitly for specific criteria of ASG, are desperately needed.** The weak correlations of automatic metrics with human criteria still leave much to be desired. Ideally, we would have automatic metrics which reflect each of our proposed criteria.

3. **Awaiting specific ASG metrics, researchers should use better metrics than `BLEU`$^{\Xi\S}$ and `ROUGE`$^{\Xi\S}$.** `chrF`$^{\Xi\S}$ and `BARTScore`$^{\Xi\varepsilon}$ are the best performers at the story- and system-level respectively. Given the overall weak results, however, we strongly advise to rely on human annotations for ASG evaluation.

4. **Our new set of human criteria allows for a standardized and extensive human evaluation.** The criteria are overall weakly correlated with one another, which shows that they are non-redundant, and produce coherent system rankings.

Future directions. Motivated by our search for human criteria from the social science literature, we reckon more collaboration between the NLP and social science communities may yield valuable insights into the question of how to computationally capture good indicators of story quality. In this spirit, we hope that `HANNA` will pave the way for further progress in the evaluation of ASG.

Acknowledgements

We thank Yejin Choi, Richard Bai, Le Fang, Jian Guan, Hannah Rashkin, David Wilmot and Eden Bensaid for answering to our requests for data. This work was granted access to the HPC resources of IDRIS under the allocation 2022-101838 made by GENCI and was partially funded by the grant ANR-20-CHIA-0012-01 (“NoRDF”).

References

Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic story generation: Challenges and attempts. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 72–83, Virtual. Association for Computational Linguistics.

Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: a survey of approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark O. Riedl. 2020. Bringing stories alive: Generating interactive fiction worlds. *CoRR*, abs/2001.10161.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Ruth Aylett, Marco Vala, Pedro Sequeira, and Ana Paiva. 2007. Fearnot!–an emergent narrative approach to virtual dramas for anti-bullying education. In *International Conference on Virtual Storytelling*, pages 202–205. Springer.

Byung-Chull Bae, Suji Jang, Youngjune Kim, and Seyoung Park. 2021. A preliminary survey on story interestingness: Focusing on cognitive and emotional interest. In *International Conference on Interactive Digital Storytelling*, pages 447–453. Springer.

He Bai, Peng Shi, Jimmy Lin, Luchen Tan, Kun Xiong, Wen Gao, Jie Liu, and Ming Li. 2021. Semantics of the unwritten: The effect of end of paragraph and sequence tokens on text generation with GPT2. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 148–162, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Morteza Behrooz, Justus Robertson, and Arnav Jhala. 2019. Story quality as a matter of perception: using word embeddings to estimate cognitive interest. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 3–9.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Marc Cavazza and David Pizzi. 2006. Narratology for interactive storytelling: A critical introduction. In *International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, pages 72–83. Springer.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, Institut polytechnique de Paris.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021a. Code-switched inspired losses for spoken dialog representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8320–8337.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloé Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7594–7601. AAAI Press.

Pierre Colombo, Chloé Clavel, Chouchang Yack, and Giovanna Varni. 2021b. Beam search with bidirectional strategies for neural response generation. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 139–146, Trento, Italy. Association for Computational Linguistics.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clemencon. 2022a. What are the best systems? new perspectives on nlp benchmarking. *arXiv preprint arXiv:2202.03799*.

Pierre Colombo, Pablo Piantanida, and Chloé Clavel. 2021c. A novel estimator of mutual information for learning to disentangle textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6539–6550, Online. Association for Computational Linguistics.

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021d. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022b. Learning disentangled textual representations via statistical measures of similarity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2614–2630, Dublin, Ireland. Association for Computational Linguistics.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long and Short Papers), pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. 2022c. Infolm: A new metric to evaluate summarization & data2text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10554–10562.

Richard Craggs and Mary McGee Wood. 2005. Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics*, 31(3):289–296.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Robert Dickman. 2003. The four elements of every successful story. *Reflections - Society for Organizational Learning*, 4(3):51–58.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Transformer-based conditional variational autoencoder for controllable story generation. *arXiv preprint arXiv:2101.00828*.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Alexandre Garcia, Pierre Colombo, Florence d'Alché Buc, Slim Essid, and Chloé Clavel. 2019. From the token to the review: A hierarchical multimodal approach to opinion mining. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5539–5548, Hong Kong, China. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021a. Long text generation by modeling sentence-level and discourse-level coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6379–6393, Online. Association for Computational Linguistics.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6473–6480. AAAI Press.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021b. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Ken Hartsook, Alexander Zook, Sauvik Das, and Mark O Riedl. 2011. Toward supporting stories with procedurally generated game worlds. In *2011 IEEE Conference on Computational Intelligence and Games (CIG'11)*, pages 297–304. IEEE.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Asghar Iran-Nejad. 1987. Cognitive and affective causes of interest and liking. *Journal of Educational Psychology*, 79(2):120.

Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Éric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2020. Narrative text generation with a latent discrete plan. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3637–3650, Online. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Suzanne Keen et al. 2007. *Empathy and the Novel*. Oxford University Press on Demand.

John G Kemeny. 1959. Mathematics without numbers. *Daedalus*, 88(4):577–591.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Michael Lebowitz. 1985. Story-telling as planning and learning. *Poetics*, 14(6):483–502.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Allyssa McCabe and Carole Peterson. 1984. What makes a good story. *Journal of Psycholinguistic Research*, 13(6):457–480.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 61–63.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Jihyung Moon. 2019. Significance test of increase in correlation for nlp evaluations in python. `https://github.com/inmoonlight/nlp-williams/`.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Roger B Myerson. 1996. Fundamentals of social choice theory. Technical report, Discussion Paper.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. 2022. A differential entropy estimator for training neural networks. In *International Conference on Machine Learning*, pages 17691–17715. PMLR.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

William Lowell Randall. 1999. Narrative intelligence and the novelty of our lives. *Journal of aging Studies*, 13(1):11–28.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Hanna-Riikka Roine. 2016. Imaginative, immersive and interactive engagements. the rhetoric of worldbuilding in contemporary speculative fiction.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Roger C Schank. 1978. Interestingness: Controlling inferences. Technical report, YALE UNIV NEW HAVEN CONN DEPT OF COMPUTER SCIENCE.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Sibony. 2014. Borda count approximation of kemeny's rule and pairwise voting inconsistencies. In *Proceedings of the NIPS 2014 Workshop on Analysis of Rank Data*.

Wilbert Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

Guillaume Staerman, Pavlo Mozharovskyi, Stéphan Clémençon, and Florence d'Alché Buc. 2021. A pseudo-metric between probability distributions based on depth-trimmed regions. *arXiv preprint arXiv:2103.12711*.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *international conference on intelligent text processing and computational linguistics*, pages 341–351. Springer.

Wendy A Suzuki, Mónica I Feliú-Mójer, Uri Hasson, Rachel Yehuda, and Jean Mary Zarate. 2018. Dialogues: The science and power of storytelling. *Journal of Neuroscience*, 38(44):9468–9470.

Michael Toolan. 2012. Engagement via emotional heightening in" passion": On the grammatical texture of emotionally-immersive passages in short fiction. *Narrative*, 20(2):210–225.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Su Wang, Greg Durrett, and Katrin Erk. 2020. Narrative interpolation for generating and understanding stories. *arXiv preprint arXiv:2008.07466*.

Evan J. Williams. 1959. *Regression Analysis*. Wiley, New York.

David Wilmot and Frank Keller. 2021. A temporal variational model for story generation. *arXiv preprint arXiv:2109.06807*.

Wojciech Witon, Pierre Colombo, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at IEST 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 248–253, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7378–7385.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Appendix

Table of Contents

A Amazon Mechanical Turk experiment details

To complement section 3.5, the details of the instructions we gave in our Amazon Mechanical Turk experiment can be found in Tab. 7 below.

B Names of metric variants

Here we define the names we give to some variants of the automatic metrics we used.
SUPERT and BLANC are summarization metrics which normally require a source document and a summary. In our setting, we have a prompt and a generated story. The suffix PS means we used the "Prompt as the Summary", and SS means the "Story as the Summary". The Golden suffix means we used the reference human story as the source document and the generated story as the summary.
Given a couple of texts (x, y), BARTScore computes a score based on the log probability of y given x. We used the suffixes SH for (Story, Human), HS for (Human, Story), SP for (Story, Prompt) and PS for (Prompt, Story).

All other names are defined in their respective papers.

Amazon Mechanical Turk example task	
Please read the prompt, the human story and the subject story (both stories might be the same). The story you will have to rate is the **subject story**. **Important**: we will reject HITs which were done in **fewer than 30 seconds** (unless both stories are exceptionally short). Please rest assured: if you take the work seriously, we have no reason to reject it. **Disclaimer**: some stories have been automatically generated and might contain explicit or offensive content. **Note**: some stories have been abruptly cut in the middle of a sentence. Please rate them as if they ended just before the unfinished sentence. **Note**: if the story is not relevant with respect to the prompt, it **only** affects the **Relevance** criterion! Do not rate 1 everywhere, or we will **reject**! Then, please write down the name of the **first character** that is mentioned in the **subject story**; if no name is mentioned, write "None". Only proper nouns count as names. Then, please rate the **subject story** on a scale from 1 (worst) to 5 (best) on the following criteria: relevance, coherence, empathy, surprise, engagement, and complexity.	
Prompt	When you die the afterlife is an arena where you face every insect and animal you killed in your life. If you win you go to heaven, lose you go to hell. Your job was an exterminator on earth.
Human story	3,000 years have I been fighting. Every morning, the raccoons scratch at my eyes. Every evening, the skunks spray me while the opossums chew at my feet. I have never had any tools. I have only my hands. I don't remember the place I came from before this. All I remember is the daily fight between me and these animals. No matter how many times I kill them, they come back the next day. [...]
Subject story	First of all, not everyone was entitled to be an exterminator. But the ones that were – maybe were, like, *genius*, because, yes, I had once belonged to a less fortunate class of people – had all the opportunity to work for the damn plant killer, and it's hard work. And the horrifying truth is, once you die, and the entire planet turns into a glade that contains a golden fish that would've been crushed by a million million goldfish just moments ago, you're not really good enough for heaven. Why? [...]
Name of the first mentioned character in the subject story	`[Area to fill]`
Relevance — measures how well the story matches its prompt	1 — The story has no relationship with the prompt at all. 2 — The story only has a weak relationship with the prompt. 3 — The story roughly matches the prompt. 4 — The story matches the prompt, except for one or two small aspects. 5 — The story matches the prompt exactly.

Coherence — measures whether the story makes sense	1 — The story does not make sense at all. For instance, the setting and/or characters keep changing, and/or there is no understandable plot. 2 — Most of the story does not make sense. 3 — The story mostly makes sense but has some incoherences. 4 — The story almost makes sense overall, except for one or two small incoherences. 5 — The story makes sense from beginning to end.
Empathy — measures how well you understood the characters' emotions (regardless of whether you agreed with them)	1 — The characters seemed apathetic to you. 2 — At least one character slightly related to you on an emotional level. 3 — You recognized specific, but not necessarily strong, emotions (eg sadness, joy, fear. . .) in at least one character. 4 — At least one character emotionally involved you, but minor details prevented you from completely relating to them. 5 — At least one character completely involved you on an emotional level.
Surprise — measures how surprising the end of the story was	1 — The ending seemed completely obvious from the start, or doesn't make any sense at all. 2 — The ending was easily predictable after a few sentences. 3 — The ending was predictable after half of the story. 4 — The ending surprised you, but would have been difficult to predict. 5 — The ending surprised you, and still seemed as if it could very reasonably have been predicted, ie, there were enough clues in the story.
Engagement — measures how much you engaged with the story	1 — You found the story boring and were glad it was over. 2 — You found one or two things interesting in the story, but no more. 3 — The story was mildly interesting. 4 — The story almost kept you engaged until the end. 5 — You were so engaged that you wished there was a sequel.
Complexity — measures how elaborate the story is	1 — The setting of the story is extremely simple; it only involves one or two characters or concepts. 2 — The setting of the story is simple; one or two characters, a simple plot, maybe an indication of time or location. 3 — The story is somewhat developed: it involves at least one of the following: complex concepts, realistic characters, an intricate plot, an underlying history or circumstances, precise descriptions. 4 — The story is developed: it involves at least two of the following: complex concepts, realistic characters, an intricate plot, an underlying history or circumstances, precise descriptions. 5 — The story is well thought-out: it involves at least three of the following: complex concepts, realistic characters, an intricate plot, an underlying history or circumstances, precise descriptions.

Tab. 7: Example task from our Amazon Mechanical Turk experiment

C Distributions of human annotations per system

Here we report the violin plots of the distributions of human annotations per system. Human output scores visibly better than language models. Note that for our generation, we do not use beam search (Colombo et al., 2021b, 2020; Pichler et al., 2022; Colombo, 2021; Colombo et al., 2022b; Garcia et al., 2019; Colombo et al., 2021c). To further improve the generation a domain pre-trained language model could be considered (Chapuis et al., 2020; Colombo et al., 2021a).

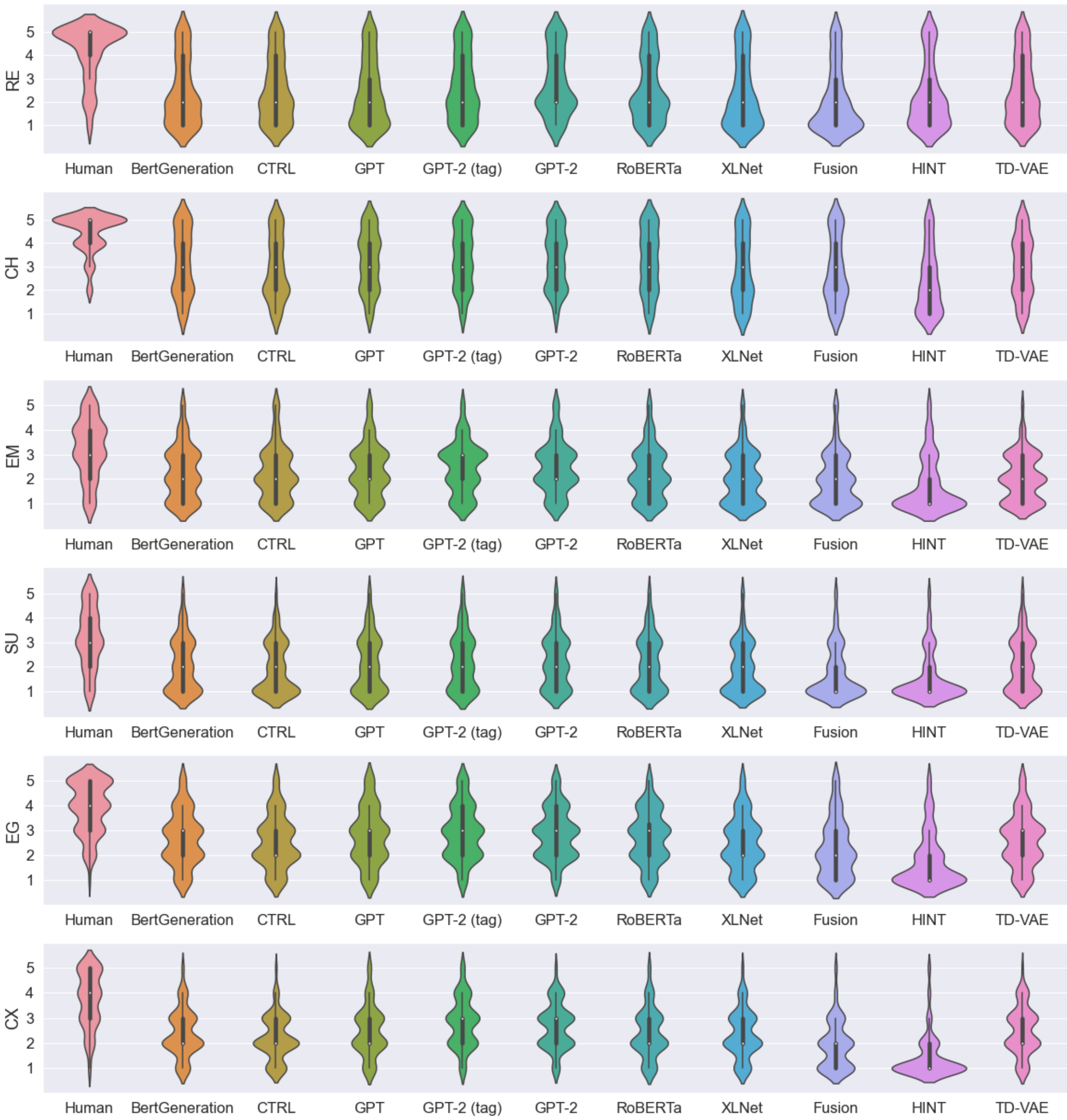

Fig. 9: Violin plots of the distributions of human annotations per system

D Correlations between human criteria

Here we report the story-level and system-level absolute correlations between human criteria with Spearman's ρ (Fig. 10 and Fig. 11) and Pearson's r (Fig. 12 and Fig. 13).

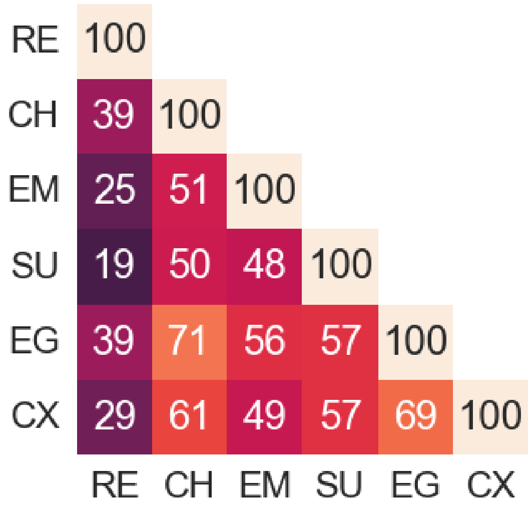

Fig. 10: Story-level Spearman correlations (%) between our proposed human criteria

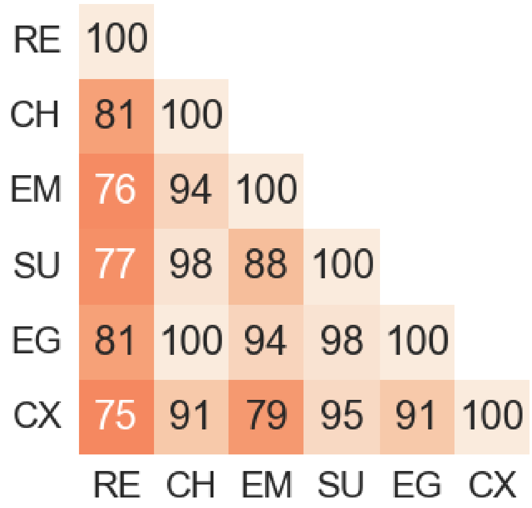

Fig. 11: System-level Spearman correlations (%) between our proposed human criteria

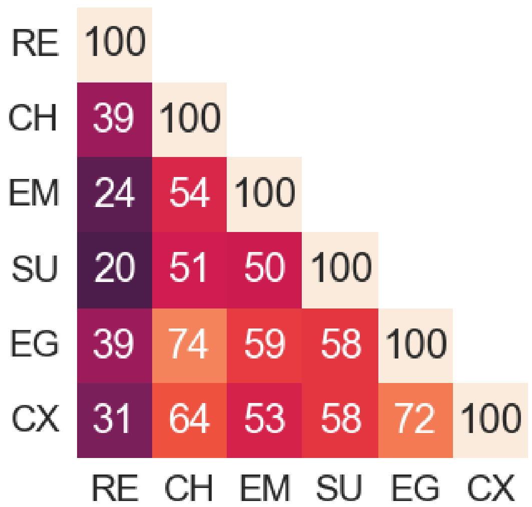

Fig. 12: Story-level Pearson correlations (%) between our proposed human criteria

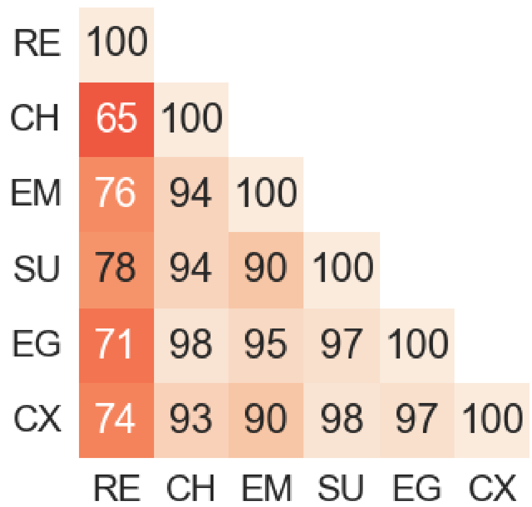

Fig. 13: System-level Pearson correlations (%) between our proposed human criteria

E Correlations between human criteria and automatic metrics

Here we report the full figures of story-level and system-level absolute correlations between human criteria and automatic metrics with all three correlation coefficients.

Fig. 14: Story-level absolute Kendall correlations (%) between automatic metrics and our proposed human criteria

Fig. 15: System-level absolute Kendall correlations (%) between automatic metrics and our proposed human criteria

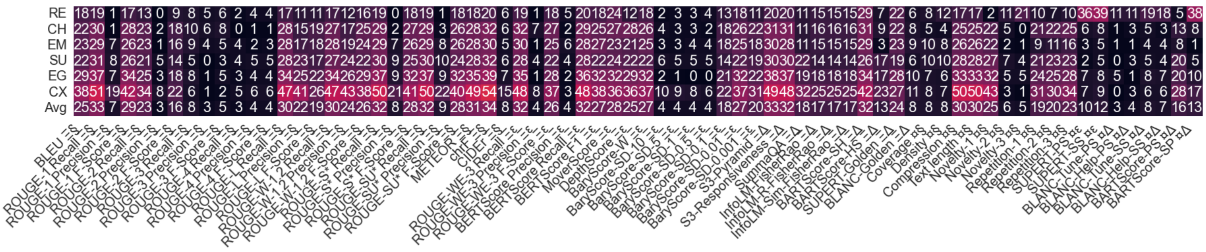

Fig. 16: Story-level absolute Spearman correlations (%) between automatic metrics and our proposed human criteria

Fig. 17: System-level absolute Spearman correlations (%) between automatic metrics and our proposed human criteria

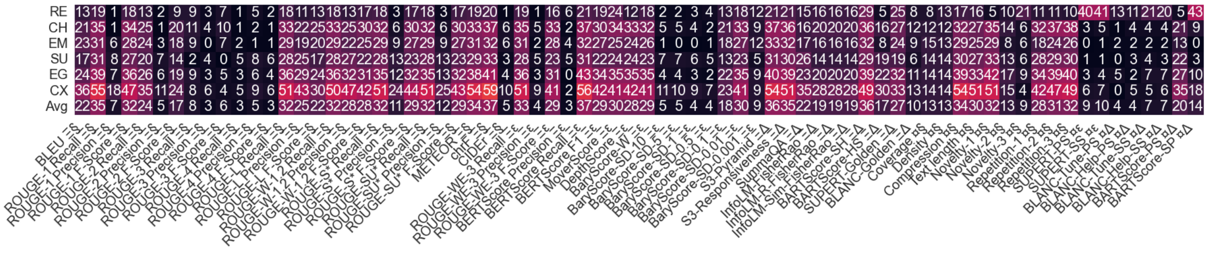

Fig. 18: Story-level absolute Pearson correlations (%) between automatic metrics and our proposed human criteria

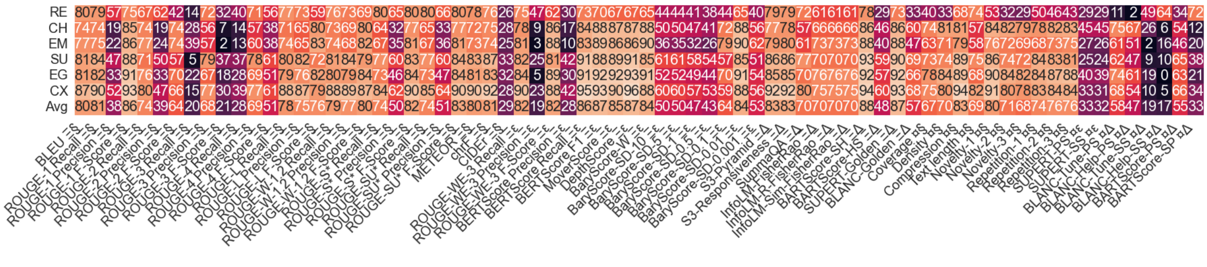

Fig. 19: System-level absolute Pearson correlations (%) between automatic metrics and our proposed human criteria

F Correlations between automatic metrics

Here we report the full figures of story-level and system-level absolute correlations between automatic metrics with all three correlation coefficients.

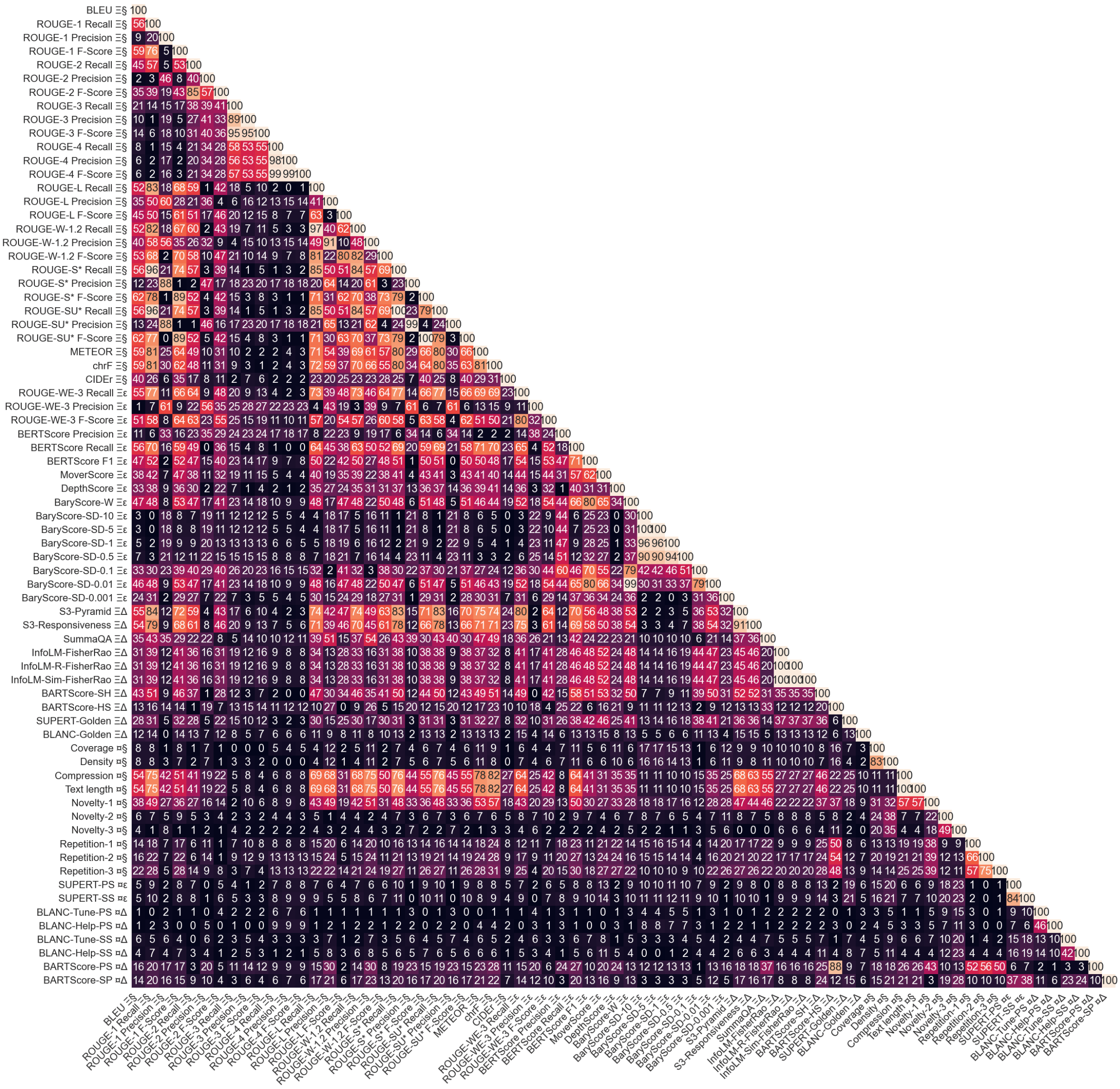

Fig. 20: Story-level absolute Kendall correlations (%) between automatic metrics

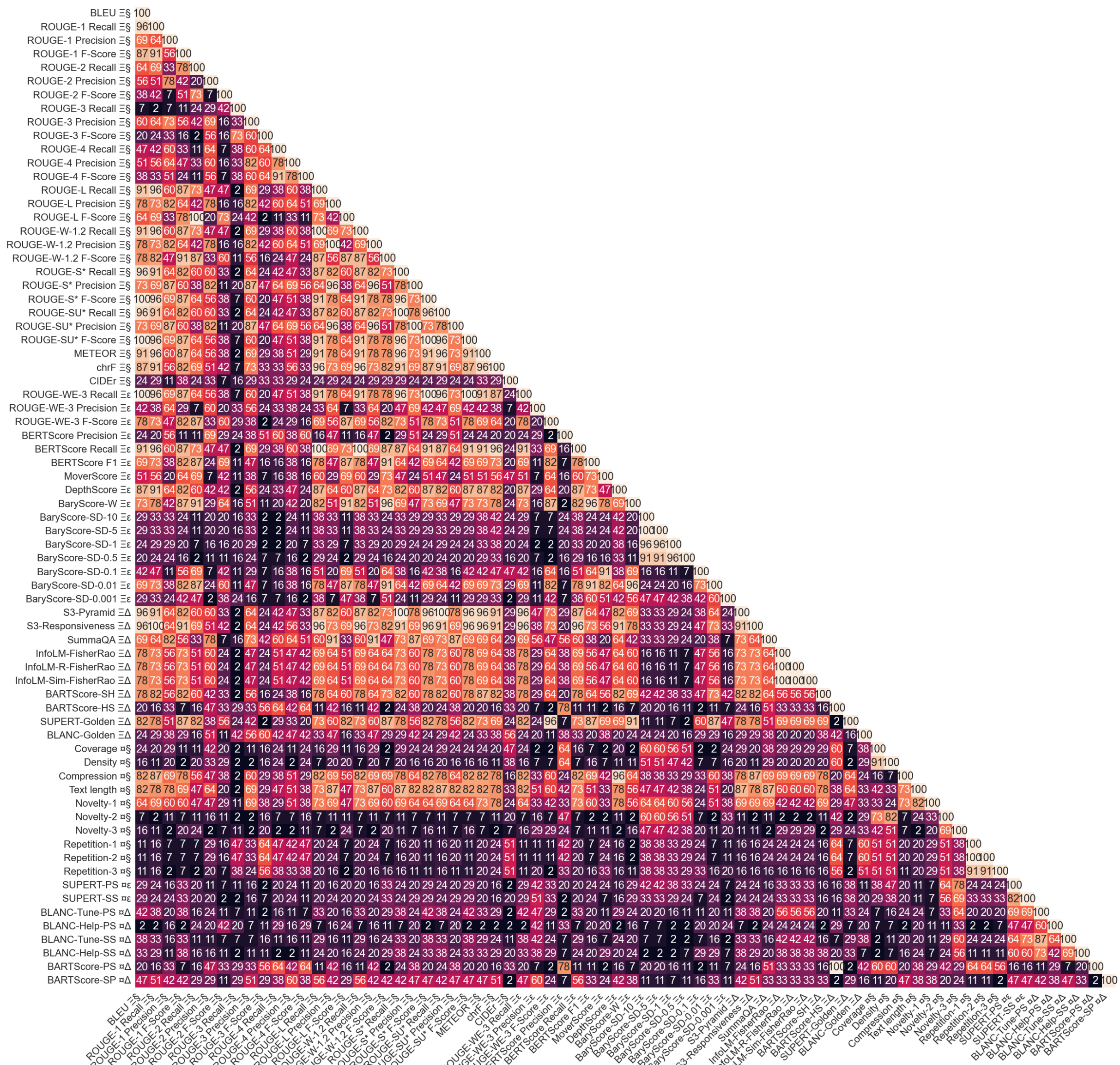

Fig. 21: System-level absolute Kendall correlations (%) between automatic metrics

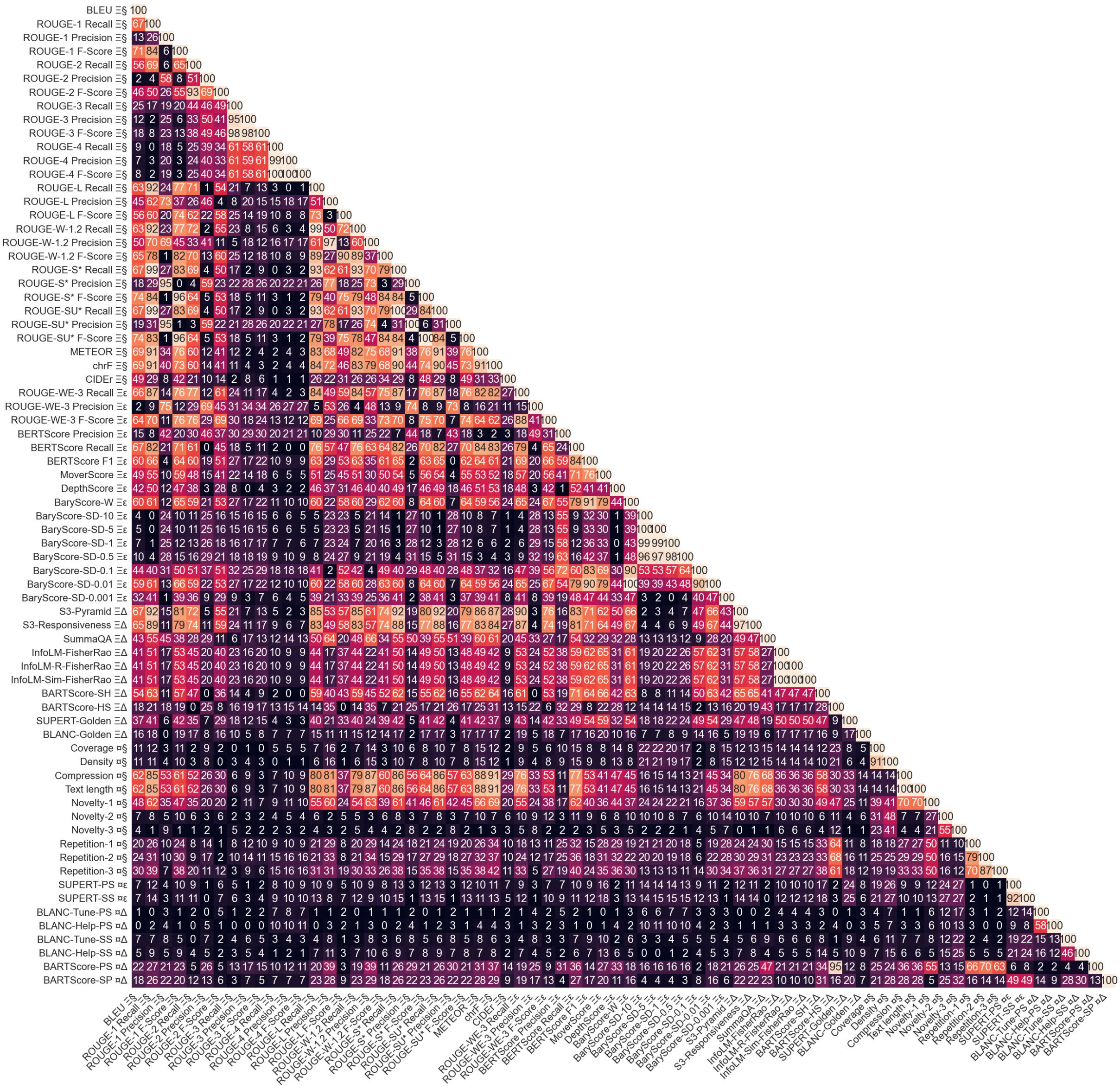

Fig. 22: Story-level absolute Spearman correlations (%) between automatic metrics

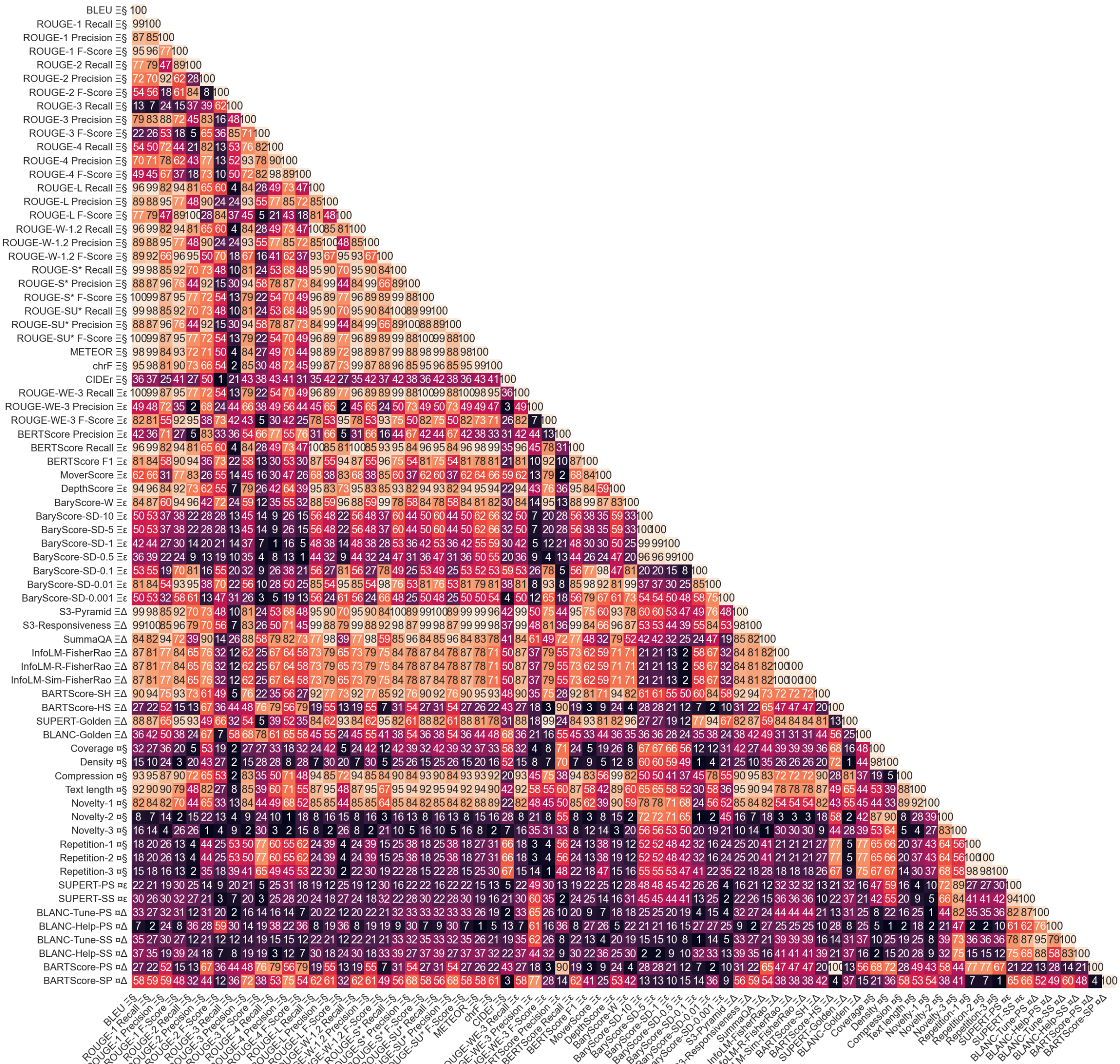

Fig. 23: System-level absolute Spearman correlations (%) between automatic metrics

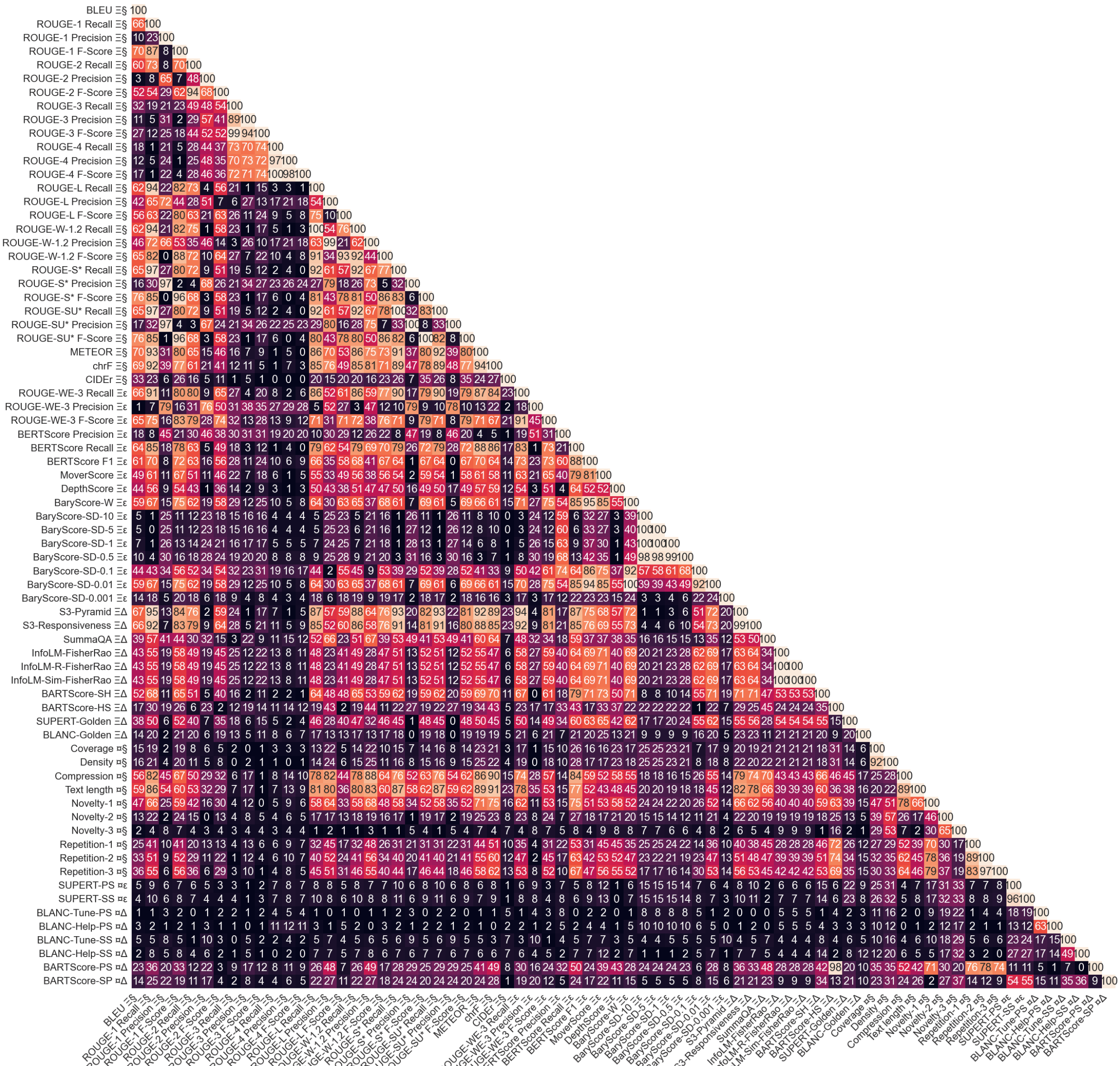

Fig. 24: Story-level absolute Pearson correlations (%) between automatic metrics

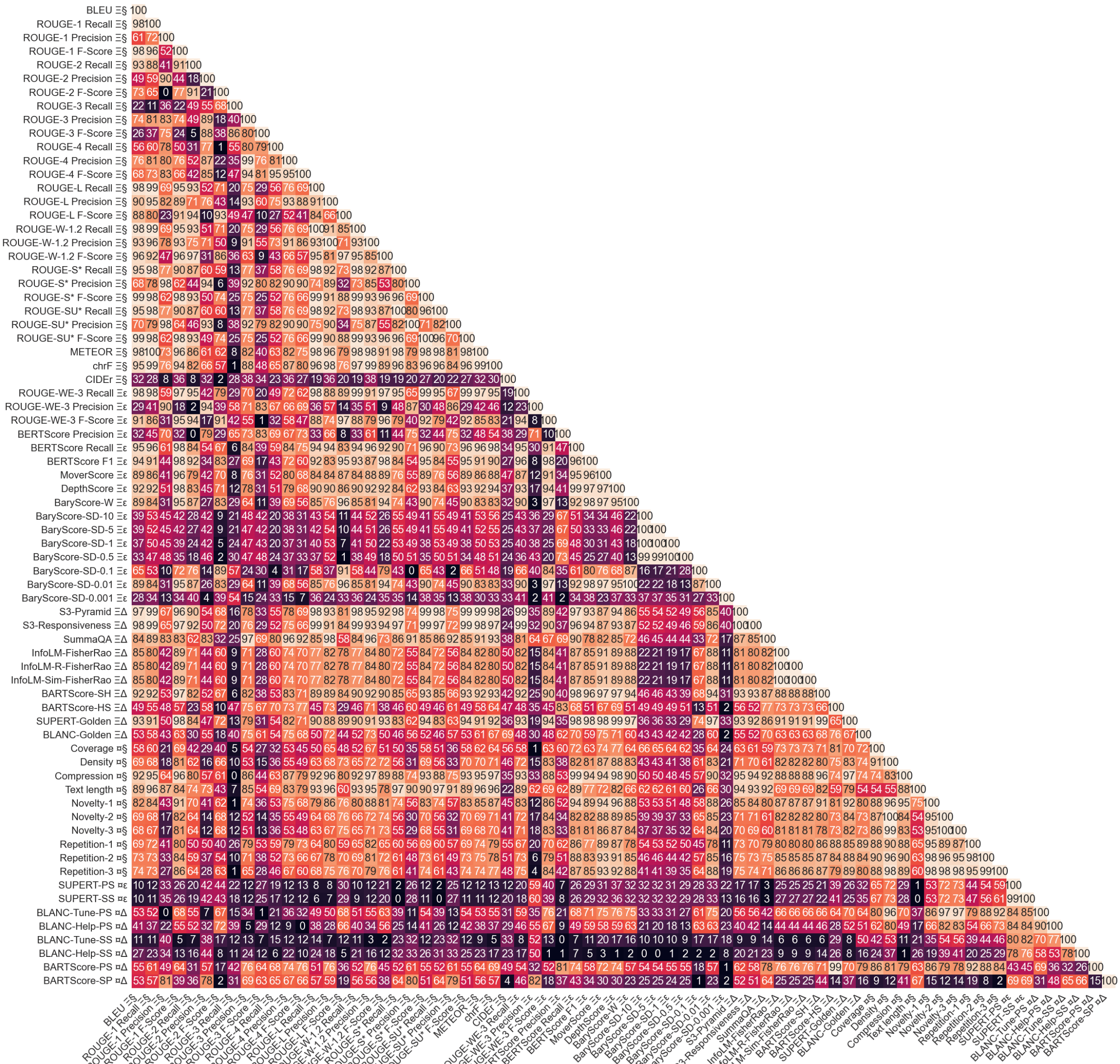

Fig. 25: System-level absolute Pearson correlations (%) between automatic metrics

G Best metrics per criterion per level of correlation per correlation coefficient

Here we report the top 5 metrics per criterion per story-level and system-level absolute correlation coefficient.

Criterion	$\lvert\tau\rvert$ (%)		$\lvert\rho\rvert$ (%)		$\lvert r\rvert$ (%)	
	SUPERT-SS$^{\Xi\varepsilon}$	29.95	SUPERT-SS$^{\Xi\varepsilon}$	38.58	BARTScore-SP$^{\text{¤}\triangle}$	42.55
	BARTScore-SP$^{\text{¤}\triangle}$	29.61	BARTScore-SP$^{\text{¤}\triangle}$	37.98	SUPERT-SS$^{\Xi\varepsilon}$	41.16
RE	SUPERT-PS$^{\Xi\varepsilon}$	28.59	SUPERT-PS$^{\Xi\varepsilon}$	36.40	SUPERT-PS$^{\Xi\varepsilon}$	40.15
	BARTScore-SH$^{\Xi\triangle}$	22.32	BARTScore-SH$^{\Xi\triangle}$	28.53	BARTScore-SH$^{\Xi\triangle}$	28.98
	MoverScore$^{\Xi\varepsilon}$	19.12	MoverScore$^{\Xi\varepsilon}$	23.67	SUPERT-Golden$^{\Xi\varepsilon}$	24.72
	ROUGE-WE-3 Recall$^{\Xi\varepsilon}$	25.29	ROUGE-WE-3 Recall$^{\Xi\varepsilon}$	32.22	Repetition-3$^{\text{¤}\S}$	38.12
	BARTScore-SH$^{\Xi\triangle}$	25.06	CHRF$^{\Xi\S}$	32.03	BERTScore Recall$^{\Xi\varepsilon}$	37.12
CH	CHRF$^{\Xi\S}$	24.61	BARTScore-SH$^{\Xi\triangle}$	31.38	S3-Pyramid$^{\Xi\triangle}$	37.05
	S3-Pyramid$^{\Xi\triangle}$	24.39	S3-Responsiveness$^{\Xi\triangle}$	31.31	CHRF$^{\Xi\S}$	36.99
	S3-Responsiveness$^{\Xi\triangle}$	24.28	S3-Pyramid$^{\Xi\triangle}$	31.14	Repetition-2$^{\text{¤}\S}$	36.54
	ROUGE-WE-3 Recall$^{\Xi\varepsilon}$	23.58	ROUGE-WE-3 Recall$^{\Xi\varepsilon}$	29.85	S3-Pyramid$^{\Xi\triangle}$	32.78
	CHRF$^{\Xi\S}$	23.33	CHRF$^{\Xi\S}$	29.81	CHRF$^{\Xi\S}$	32.43
EM	S3-Pyramid$^{\Xi\triangle}$	23.19	S3-Pyramid$^{\Xi\triangle}$	29.68	BERTScore Recall$^{\Xi\varepsilon}$	32.06
	ROUGE-SU* Recall$^{\Xi\S}$	23.13	ROUGE-SU* Recall$^{\Xi\S}$	29.38	S3-Responsiveness$^{\Xi\triangle}$	31.78
	ROUGE-S* Recall$^{\Xi\S}$	23.08	ROUGE-S* Recall$^{\Xi\S}$	29.32	BARTScore-SH$^{\Xi\triangle}$	31.66
	CHRF$^{\Xi\S}$	24.45	CHRF$^{\Xi\S}$	31.55	Novelty-1$^{\text{¤}\S}$	32.86
	ROUGE-1 Recall$^{\Xi\S}$	23.67	ROUGE-1 Recall$^{\Xi\S}$	30.86	CHRF$^{\Xi\S}$	32.65
SU	S3-Responsiveness$^{\Xi\triangle}$	23.35	S3-Responsiveness$^{\Xi\triangle}$	30.41	ROUGE-1 Recall$^{\Xi\S}$	31.32
	Novelty-1$^{\text{¤}\S}$	23.11	ROUGE-SU* Recall$^{\Xi\S}$	30.30	S3-Pyramid$^{\Xi\triangle}$	31.07
	ROUGE-SU* Recall$^{\Xi\S}$	22.85	ROUGE-S* Recall$^{\Xi\S}$	30.25	BERTScore Recall$^{\Xi\varepsilon}$	30.98
	CHRF$^{\Xi\S}$	30.77	CHRF$^{\Xi\S}$	39.03	BERTScore Recall$^{\Xi\varepsilon}$	42.95
	S3-Pyramid$^{\Xi\triangle}$	29.62	S3-Pyramid$^{\Xi\triangle}$	37.74	Novelty-1$^{\text{¤}\S}$	42.27
EG	ROUGE-1 Recall$^{\Xi\S}$	29.19	ROUGE-1 Recall$^{\Xi\S}$	37.02	CHRF$^{\Xi\S}$	41.07
	S3-Responsiveness$^{\Xi\triangle}$	29.01	S3-Responsiveness$^{\Xi\triangle}$	36.85	S3-Pyramid$^{\Xi\triangle}$	40.34
	BERTScore Recall$^{\Xi\varepsilon}$	28.93	ROUGE-S* Recall$^{\Xi\S}$	36.60	Repetition-3$^{\text{¤}\S}$	39.53
	CHRF$^{\Xi\S}$	43.31	CHRF$^{\Xi\S}$	54.11	CHRF$^{\Xi\S}$	58.76
	ROUGE-1 Recall$^{\Xi\S}$	40.65	ROUGE-1 Recall$^{\Xi\S}$	50.60	BERTScore Recall$^{\Xi\varepsilon}$	55.83
CX	ROUGE-SU* Recall$^{\Xi\S}$	39.83	Text length$^{\text{¤}\S}$	50.19	ROUGE-1 Recall$^{\Xi\S}$	55.01
	Text length$^{\text{¤}\S}$	39.82	Compression$^{\text{¤}\S}$	50.19	METEOR$^{\Xi\S}$	54.41
	Compression$^{\text{¤}\S}$	39.82	ROUGE-SU* Recall$^{\Xi\S}$	50.10	Compression$^{\text{¤}\S}$	54.38

Tab. 8: Top 5 metrics per criterion per story-level correlation coefficient

Criterion	$\|\tau\|$ (%)		$\|\rho\|$ (%)		$\|r\|$ (%)	
RE	S3-Pyramid$^{\Xi\triangle}$	60.00	MoverScore$^{\Xi\varepsilon}$	78.18	ROUGE-S* F-Score$^{\Xi\S}$	80.39
	CHRF$^{\Xi\S}$	60.00	S3-Pyramid$^{\Xi\triangle}$	75.76	ROUGE-SU* F-Score$^{\Xi\S}$	80.29
	ROUGE-SU* Recall$^{\Xi\S}$	60.00	ROUGE-S* Recall$^{\Xi\S}$	75.76	ROUGE-S* Recall$^{\Xi\S}$	80.24
	ROUGE-S* Recall$^{\Xi\S}$	60.00	ROUGE-SU* Recall$^{\Xi\S}$	75.76	ROUGE-SU* Recall$^{\Xi\S}$	80.23
	ROUGE-W-1.2 F-Score$^{\Xi\S}$	60.00	CHRF$^{\Xi\S}$	74.55	BLEU$^{\Xi\S}$	79.89
CH	BaryScore-SD-0.001$^{\Xi\varepsilon}$	77.78	BaryScore-SD-0.001$^{\Xi\varepsilon}$	92.73	BaryScore-SD-0.01$^{\Xi\varepsilon}$	88.15
	BaryScore-SD-5$^{\Xi\varepsilon}$	68.89	BaryScore-SD-5$^{\Xi\varepsilon}$	78.18	BaryScore-W$^{\Xi\varepsilon}$	87.99
	BaryScore-SD-10$^{\Xi\varepsilon}$	68.89	BaryScore-SD-10$^{\Xi\varepsilon}$	78.18	BERTScore F1$^{\Xi\varepsilon}$	87.91
	BaryScore-SD-1$^{\Xi\varepsilon}$	64.44	BaryScore-SD-1$^{\Xi\varepsilon}$	75.76	DepthScore$^{\Xi\varepsilon}$	87.38
	BaryScore-SD-0.5$^{\Xi\varepsilon}$	60.00	BERTScore F1$^{\Xi\varepsilon}$	74.55	MoverScore$^{\Xi\varepsilon}$	86.95
EM	BaryScore-SD-0.001$^{\Xi\varepsilon}$	77.78	BaryScore-SD-0.001$^{\Xi\varepsilon}$	92.73	BaryScore-SD-0.01$^{\Xi\varepsilon}$	90.01
	BERTScore F1$^{\Xi\varepsilon}$	73.33	BERTScore F1$^{\Xi\varepsilon}$	84.24	BaryScore-W$^{\Xi\varepsilon}$	89.96
	BaryScore-SD-0.01$^{\Xi\varepsilon}$	73.33	BaryScore-SD-0.01$^{\Xi\varepsilon}$	84.24	BERTScore F1$^{\Xi\varepsilon}$	88.67
	MoverScore$^{\Xi\varepsilon}$	73.33	MoverScore$^{\Xi\varepsilon}$	81.82	SUPERT-Golden$^{\Xi\triangle}$	88.10
	BaryScore-W$^{\Xi\varepsilon}$	68.89	BaryScore-W$^{\Xi\varepsilon}$	80.61	ROUGE-WE-3 F-Score$^{\Xi\varepsilon}$	87.93
SU	BaryScore-SD-0.001$^{\Xi\varepsilon}$	77.78	BaryScore-SD-0.001$^{\Xi\varepsilon}$	90.30	BARTScore-SH$^{\Xi\triangle}$	92.65
	BaryScore-SD-5$^{\Xi\varepsilon}$	68.89	BaryScore-SD-5$^{\Xi\varepsilon}$	83.03	BERTScore Recall$^{\Xi\varepsilon}$	91.09
	BaryScore-SD-10$^{\Xi\varepsilon}$	68.89	BaryScore-SD-10$^{\Xi\varepsilon}$	83.03	DepthScor$^{\Xi\varepsilon}$e	90.71
	BaryScore-SD-1$^{\Xi\varepsilon}$	64.44	BaryScore-SD-1$^{\Xi\varepsilon}$	79.39	SUPERT-Golden$^{\Xi\triangle}$	89.83
	BaryScore-SD-0.5$^{\Xi\varepsilon}$	60.00	BaryScore-SD-0.5$^{\Xi\varepsilon}$	76.97	Compression$^{¤\S}$	89.24
EG	BaryScore-SD-0.001$^{\Xi\varepsilon}$	77.78	BaryScore-SD-0.001$^{\Xi\varepsilon}$	92.73	DepthScore$^{\Xi\varepsilon}$	93.44
	BaryScore-SD-5$^{\Xi\varepsilon}$	68.89	BaryScore-SD-5$^{\Xi\varepsilon}$	78.18	BARTScore-SH$^{\Xi\triangle}$	92.44
	BaryScore-SD-10$^{\Xi\varepsilon}$	68.89	BaryScore-SD-10$^{\Xi\varepsilon}$	78.18	SUPERT-Golden$^{\Xi\triangle}$	92.21
	BaryScore-SD-1$^{\Xi\varepsilon}$	64.44	BaryScore-SD-1$^{\Xi\varepsilon}$	75.76	MoverScore$^{\Xi\varepsilon}$	92.07
	BaryScore-SD-0.5$^{\Xi\varepsilon}$	60.00	BERTScore F1$^{\Xi\varepsilon}$	74.55	BERTScore F1$^{\Xi\varepsilon}$	91.74
CX	BaryScore-SD-10$^{\Xi\varepsilon}$	76.41	BaryScore-SD-10$^{\Xi\varepsilon}$	91.19	DepthScore$^{\Xi\varepsilon}$	95.63
	BaryScore-SD-5$^{\Xi\varepsilon}$	76.41	BaryScore-SD-5$^{\Xi\varepsilon}$	91.19	BERTScore Recall$^{\Xi\varepsilon}$	95.49
	BaryScore-SD-1$^{\Xi\varepsilon}$	71.91	BaryScore-SD-1$^{\Xi\varepsilon}$	87.54	Compression$^{¤\S}$	94.31
	CHRF$^{\Xi\S}$	67.42	Novelty-1$^{¤\S}$	87.54	BARTScore-SH$^{\Xi\triangle}$	93.83
	Novelty-1$^{¤\S}$	67.42	BaryScore-SD-0.5$^{\Xi\varepsilon}$	85.11	ROUGE-1 F-Score$^{\Xi\S}$	93.35

Tab. 9: Top 5 metrics per criterion per system-level correlation coefficient.

H Weighted macro F1-scores between automatic metrics and human criteria

Here we report the weighted macro F1-scores between automatic metrics and human criteria obtained through the paired bootstrap resampling test.

Fig. 26: Weighted macro F1-scores of paired bootstrap resampling

I Williams tests between automatic metrics

Here we report the p-values of the Williams tests between automatic metrics for each criterion with story-level and system-level Pearson correlations.

Fig. 27: p-values (%) of the Williams tests between automatic metrics for the RE criterion with story-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

Fig. 28: p-values (%) of the Williams tests between automatic metrics for the CH criterion with story-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

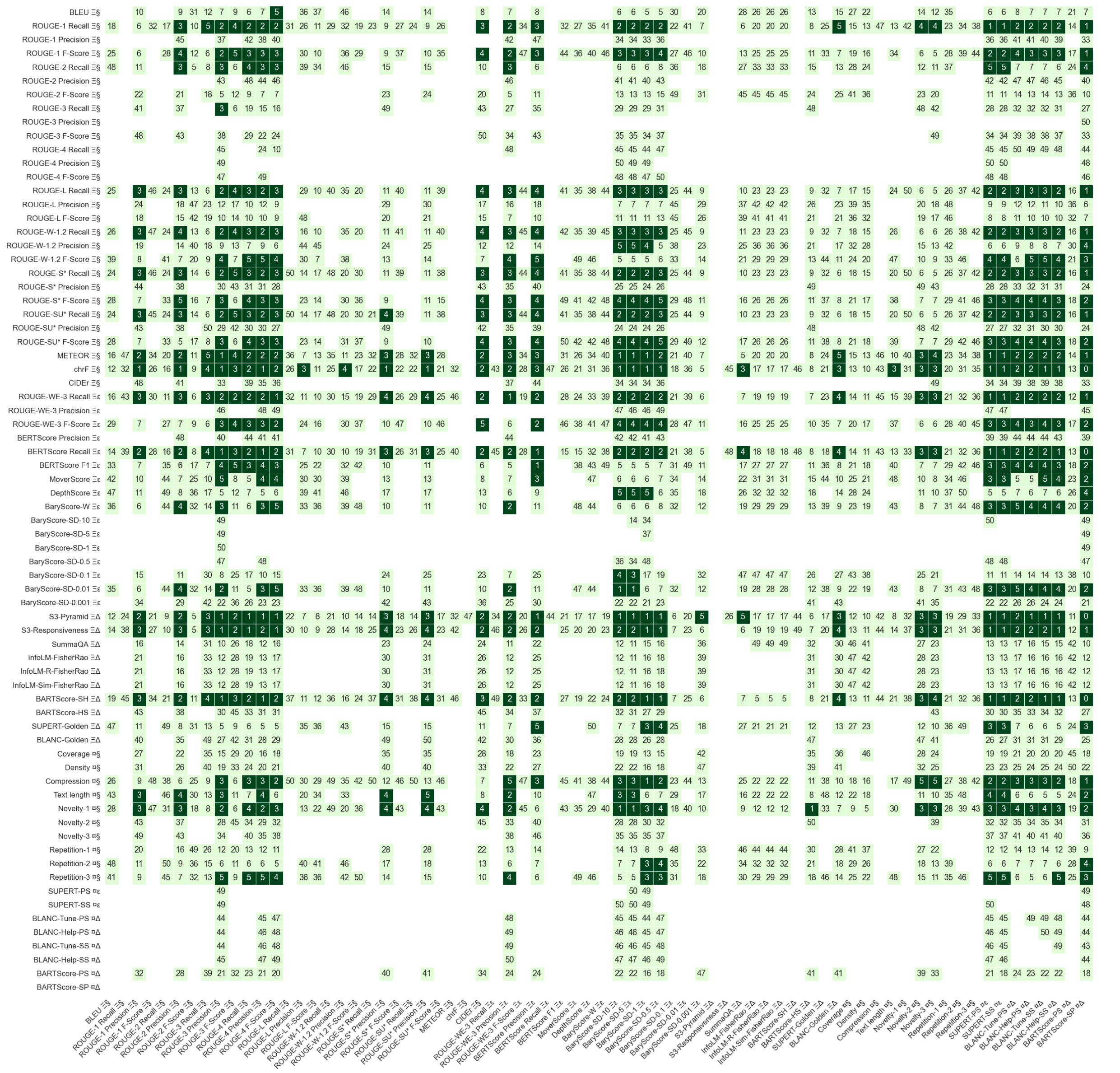

Fig. 29: p-values (%) of the Williams tests between automatic metrics for the EM criterion with story-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

Fig. 30: p-values (%) of the Williams tests between automatic metrics for the SU criterion with story-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

Fig. 31: p-values (%) of the Williams tests between automatic metrics for the EG criterion with story-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

Fig. 32: p-values (%) of the Williams tests between automatic metrics for the CX criterion with story-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

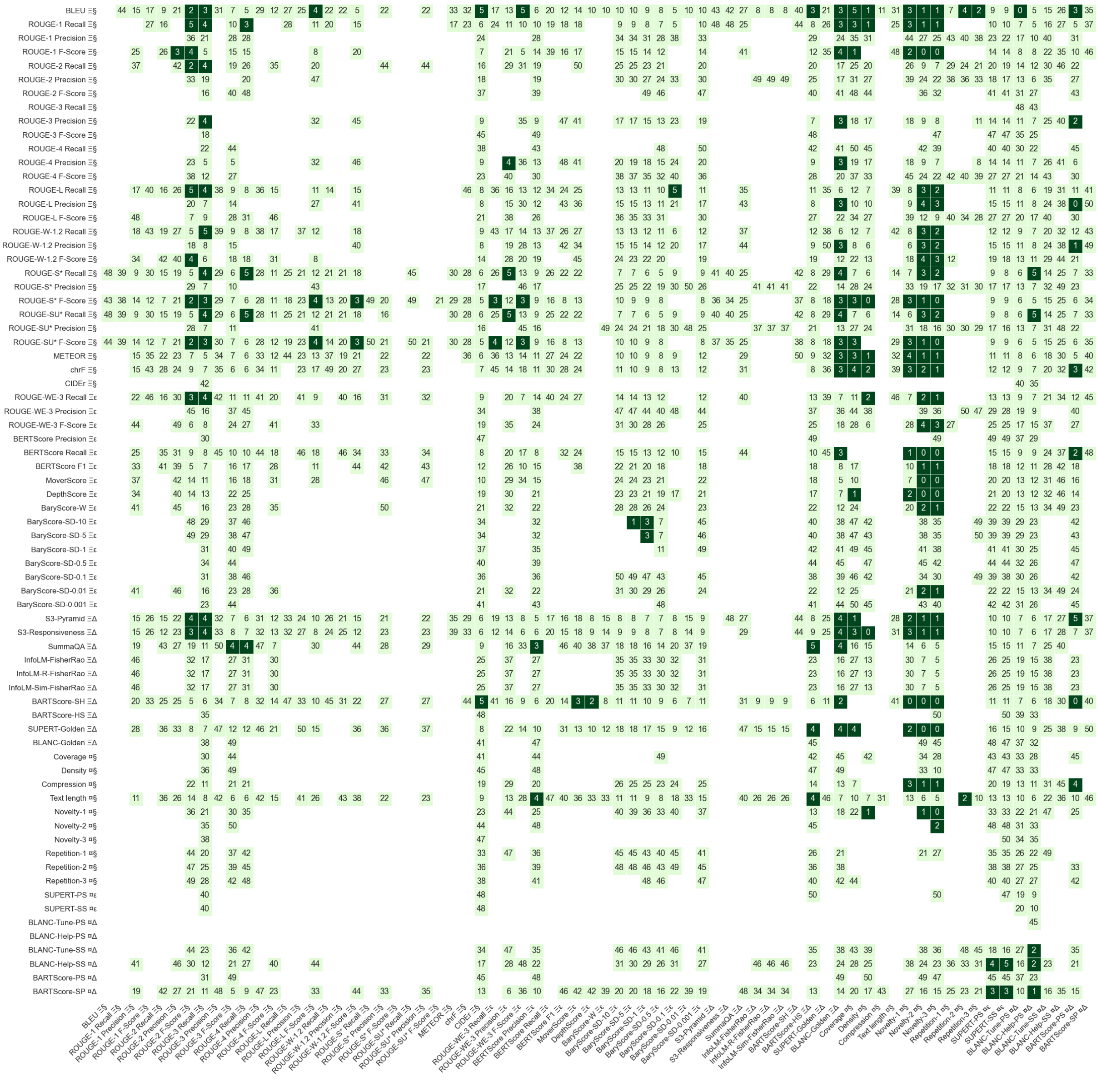

Fig. 33: p-values (%) of the Williams tests between automatic metrics for the RE criterion with system-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

Fig. 34: p-values (%) of the Williams tests between automatic metrics for the CH criterion with system-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

Fig. 35: p-values (%) of the Williams tests between automatic metrics for the EM criterion with system-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

Fig. 36: p-values (%) of the Williams tests between automatic metrics for the SU criterion with system-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

Fig. 37: p-values (%) of the Williams tests between automatic metrics for the EG criterion with system-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).

Fig. 38: p-values (%) of the Williams tests between automatic metrics for the CX criterion with system-level Pearson correlations. Green case means that the row metric has a higher correlation than the column metric, dark green means the increase is statistically significant ($p < 0.05$).