Do Language Models Enjoy Their Own Stories? Prompting Large Language Models for Automatic Story Evaluation

Cyril Chhun[1] **Fabian M. Suchanek**[1] **Chloé Clavel**[2]
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris
[2]ALMAnaCH, INRIA Paris
cyril.chhun@telecom-paris.fr
fabian.suchanek@telecom-paris.fr
chloe.clavel@inria.fr

Abstract

Storytelling is an integral part of human experience and plays a crucial role in social interactions. Thus, Automatic Story Evaluation (ASE) and Generation (ASG) could benefit society in multiple ways, but they are challenging tasks which require high-level human abilities such as creativity, reasoning, and deep understanding. Meanwhile, Large Language Models (LLMs) now achieve state-of-the-art performance on many NLP tasks. In this paper, we study whether LLMs can be used as substitutes for human annotators for ASE. We perform an extensive analysis of the correlations between LLM ratings, other automatic measures, and human annotations, and we explore the influence of prompting on the results and the explainability of LLM behaviour. Most notably, we find that LLMs outperform current automatic measures for system-level evaluation but still struggle at providing satisfactory explanations for their answers.

1 Introduction

The task of Automatic Story Generation (ASG) (Li et al., 2013) consists in the creation of a narrative from a short sentence. Previous research showed that storytelling enables a narrator to communicate honestly with their audience (Rowcliffe, 2004) and to provide listeners with an engaging and instructive experience (Miller and Pennycuff, 2008). Indeed, the process of story creation is a salient testimony of human creativity, requiring both the discovery of interesting ideas and their adept expression through a carefully-built narrative. Strong automatic story generating systems could therefore be useful for a variety of applications, such as gaming (Turner, 2014), education (Lombardo and Damiano, 2012), mental health (George et al., 2014), and marketing (Júnior et al., 2023).

Meanwhile, over the last few years, advances in natural language processing (NLP) have been spearheaded by the development of large language models (LLM) such as GPT-3 (Brown et al., 2020), LaMDA (Thoppilan et al., 2022), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023a). Upon release, these models have been setting new state-of-the-art performance standards for a wide array of NLP tasks, e.g., question answering, summarization, and translation. In particular, for ASG, LLMs are now able to produce convincing stories, so much so that they can be hard to distinguish from human stories (Clark et al., 2021). As their performance improves, they may become valuable assistants to our creative process; already, writing contests have been shown to encourage their use (Edilivre, 2023).

The increased availability of LLMs to the general public underlines the need for reliable story evaluation methods that can be used to improve both the performance of ASG models and our understanding of their strengths and weaknesses. Since the human annotation of stories is costly and time-consuming (Celikyilmaz et al., 2020), Automatic Story Evaluation (ASE) systems could provide an efficient and scalable replacement for human evaluation. However, current automatic measures have been shown to be poorly correlated with human judgment for ASG (Chhun et al., 2022).

In this paper, we investigate whether LLMs themselves can be used as substitutes for human annotators for story evaluation. To that end, we perform several annotation experiments where we ask different LLMs to rate stories according to different criteria and to explain their rating. We show an example in Figure 1 and a schema of our experiments in Figure 2.

Our contributions are the following:

1. **A comparison between LLMs and current ASE measures.** We compute and analyze the correlations between LLM ratings with human annotations on criteria specific to

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1122–1142, 2024. https://doi.org/10.1162/tacl_a_00689
Action Editor: Ehud Reiter. Submission batch: 12/2023; Revision batch: 3/2024; Published 9/2024.
© 2024 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Eval-Prompt
Prompt: A mirror shows your reflection and your future soulmate. You only see your reflection.
Story: Carnival music seemed to dance in the background. A cacophony of color and knick knacks decorated the fairgrounds. [...]
Rate the story on a scale from 1 to 5 on Empathy and explain your answer.

LLM Answer
I would rate the story a 2 on Empathy. While I could relate to the character's feelings of indifference towards the mirror, I did not feel a strong emotional connection to any of the characters. [...]

Figure 1: Example Eval-Prompt and answer from our experiments. ''Prompt'' inside the Eval-Prompt refers to the story-prompt.

story evaluation; we find that, while overall correlations are moderate to weak, system-level correlations are very high, suggesting that LLMs can produce reliable model rankings for ASE.

2. **An analysis of the influence of prompt engineering on LLM performance.** We examine the effects of using different Eval-Prompts on the consistency and distribution of LLM ratings. We find that adding detailed guidelines does not necessarily improve performance and that LLMs are remarkably self-consistent.

3. **Insights on LLM explainability for ASE.** We analyze the explanations provided by LLMs through different methods, including a user study, and find that LLMs' understanding of the ASE task is perfectible. Most notably, they struggle at explaining their ratings with substantiated claims.

4. **An analysis of LLM performance in ASG.** The high system-level correlations of LLMs with human ratings enable us to use them to rate other LLMs for ASG. We find that LLMs perform at least as well as humans for the generation of short stories, and that their performance may be explained by their tendency to produce output that is similar to their pretraining data.

Our methodology can be found in Section 3.1. We release our data and code on GitHub.[1] Our data consists of:

- **ASE experiments**: ∼150k rating and explanation annotations (1,056 stories, 6 criteria, 4 Eval-Prompts, 3 tries, 2 models);
- **User study**: 1,500 human annotations of LLM explanations;
- **ASG experiment**: 384 stories generated by Llama models with corresponding LLM annotations to expand the HANNA dataset of Chhun et al. (2022).

This paper is structured as follows: In Section 2, we review the related work. In Section 3, we lay out our methodology and experimental details. In Section 4, we perform our analysis of the results. In Section 5, we discuss the state of LLMs in ASG and ASE. Finally, in Section 6, we conclude with practical takeaways for researchers, the limitations of our work, and future research directions.

2 Related Work

2.1 Human Evaluation

Evaluating stories is a difficult task (McCabe and Peterson, 1984; Dickman, 2003). In the social sciences literature, multiple criteria have been suggested, often divided into cognitive and emotional factors (Bae et al., 2021). However, the consensus around the criteria to be used in the NLP literature is still weak (Fan et al., 2018; Guan et al., 2020; Rashkin et al., 2020; Goldfarb-Tarrant et al., 2020). Chhun et al. (2022) distill the indicators used in the social sciences literature into 6 criteria (Relevance, Coherence, Empathy, Surprise, Engagement, Complexity), which we will use in our paper as well.

While human evaluation remains the gold standard of evaluation, it is costly and time-consuming. We therefore need to develop automatic measures that can act as substitutes for human judgment, ideally for each of the criteria. Such automatic measures could be used to improve language models, e.g., as a loss function or for chain-of-thought prompting (Wei et al. 2022b)

[1] https://github.com/dig-team/hanna-benchmark-asg.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

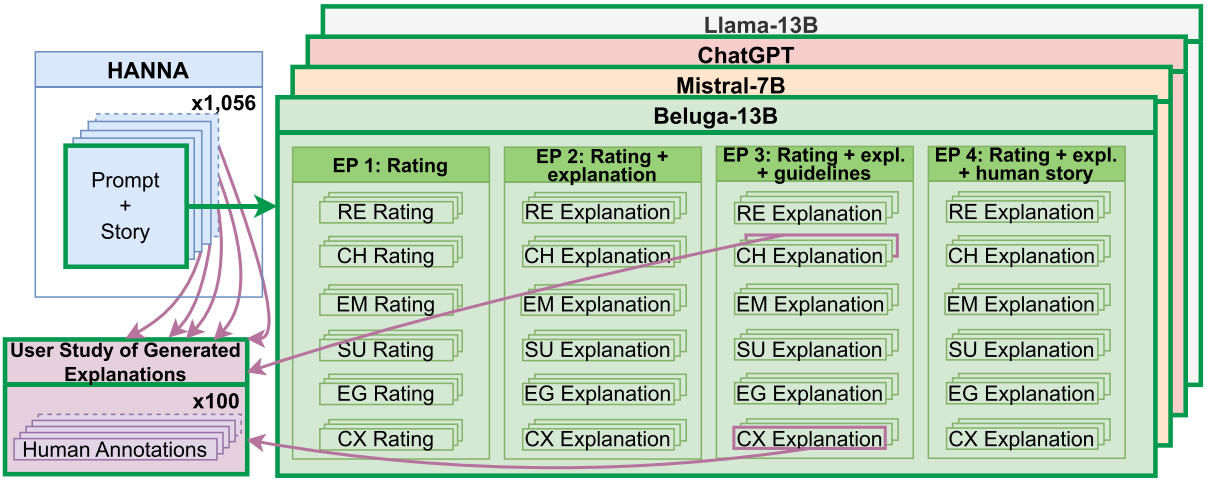

Figure 2: Schema of the performed ASE experiments. RE, CH, etc. are the considered human criteria (Section 3.1). ‘‘EP’’ means ‘‘Eval-Prompt’’, defined in Section 3.1. For the user study (Section 3.3), we randomly sampled 100 explanations from our experiments.

2.2 Automatic Evaluation

Automatic measures (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021)) have been repeatedly shown to correlate moderately to poorly with human judgment, especially when applied to tasks other than the one they were designed for (Zhang et al., 2004; Novikova et al., 2017; Colombo et al., 2023). Deutsch et al. (2022) put forth the particular limitations of reference-free measures. For ASE, Guan et al. (2020) and Chhun et al. (2022) also observe weak correlations between automatic and human ratings, whether they be reference-based or reference-free. This highlights the need for better automatic evaluation methods. To tackle this issue, this paper investigates the use of LLMs to annotate stories with ratings w.r.t. a given criterion.

2.3 Automatic Annotation

LLMs are increasingly being tested for automatic text annotation, e.g., for sentiment analysis (Qureshi et al., 2022), named entity recognition (Enkhsaikhan et al., 2021), or event structure modeling (Vauth et al., 2021). Wang et al. (2021) demonstrate that labeling performed by GPT-3 can achieve the same performance as human labeling and be up to 96% more cost-efficient. Ding et al. (2023) show that GPT-3 performs well for text classification tasks, but struggles with more complex tasks such as named entity recognition. Chakrabarty et al. (2023) design a test for creativity and show that LLM-generated stories pass fewer tests than human stories, and that using LLMs for ASE yields no positive correlations.

We seek to generalize their findings through the use of source-available models and a finer analysis and discussion of LLM performance.

2.4 Prompt Engineering

The importance of designing efficient prompts for large language models such as GPT-3 has been extensively investigated in recent years. Reynolds and McDonell (2021) notably find that zero-shot prompting can perform similarly to few-shot prompting, and even exceed it. They explore the design of metaprompts that prime the language model to better solve a given problem. Zhou et al. (2023b) treat the prompt engineering process as an optimization problem, use search algorithms guided by LLMs to solve it and attain human-level performance. Wei et al. (2022a) and White et al. (2023) review different strategies that have been applied to augment large language model abilities, e.g., least-to-most prompting (Zhou et al., 2023a), ask-me-anything prompting (Arora et al., 2023), and zero-shot chain-of-thought reasoning (Kojima et al., 2022).

We choose to investigate whether LLMs perform better with simple or detailed guidelines, and with zero- or one-shot Eval-Prompts.

3 Meta-Evaluation of LLMs for ASE

3.1 Methodology for ASE

The ASG task commonly involves the generation of a story from a short sentence called a

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

prompt (Alabdulkarim et al., 2021), which we will henceforth call *story-prompt*.

ASE Definition. Given an evaluation measure m (e.g., a scoring algorithm, an LLM,. . .), a story-prompt i, and a story y_i, we define the ASE task as the production of an evaluation score $m(y_i)$.

In this paper, we choose to use LLMs as ASE measures. We will refer to the prompt that is fed to the LLM as the *Eval-Prompt*, to distinguish it from the story-prompt. See Figure 1 for an example of the use of an LLM for story evaluation.

ASE Criteria. We use the criteria introduced by Chhun et al. (2022), who designed HANNA, a benchmark for story evaluation. They compiled a set of six orthogonal criteria from the social sciences literature:

1. **Relevance** (RE, how well the story matches its prompt),
2. **Coherence** (CH, how much the story makes sense),
3. **Empathy** (EM, how well the reader understood the character's emotions),
4. **Surprise** (SU, how surprising the end of the story was),
5. **Engagement** (EG, how much the reader engaged with the story),
6. **Complexity** (CX, how elaborate the story is).

Methodology. Given the importance of good prompt engineering (Zhao et al., 2021), we design four different Eval-Prompts for the generation of ratings. For each of our Eval-Prompts, we provide the model with a story-prompt and a corresponding story. Then:

Eval-Prompt 1 (simple rating): we ask the model to rate the story on a scale from 1 to 5 on one of the six criteria;

Eval-Prompt 2 (rating with explanation): same as Eval-Prompt 1, and we ask the model to explain its answer;

Eval-Prompt 3 (rating with explanation and guidelines): same as Eval-Prompt 2, and we provide the model with the detailed guidelines from the original annotation protocol by Chhun et al. (2022);

Eval-Prompt 4 (rating with explanation and human story): same as Eval-Prompt 2, and we provide the model with the human story associated with the same story-prompt. We explicitly tell the model that the human story is given only for reference purposes.

Different Eval-Prompt examples are shown in Figure 3.

3.2 Meta-Evaluation Measures

Notations. For S systems and N story-prompts, let y_i^j be the story generated by system $j \in \{1, \ldots, S\}$ for story-prompt $i \in \{1, \ldots, N\}$. For a (human or automatic) measure m, we denote by $m(y_i^j)$ the score associated to y_i^j. Let K be a correlation coefficient, e.g., Pearson's r (Pearson, 1895), Spearman's ρ (Spearman, 1961), or Kendall's τ (Kendall, 1938). We note h_k the measure provided by the k-th human annotator.

A naive method to compare ratings from two measures would be to compute how much they differ from each other for each story, e.g., by calculating the average L1 distance between a given evaluation method m and the human ratings, e.g., $\frac{1}{3}\sum_{k=1}^{3}\mathcal{L}_1(m, h_k)$. However, this method suffers from the central tendency bias—the tendency of an individual to rate most items on a survey in the middle of a rating scale—which is often observed in Likert scales (Stevens, 1971) and could be explained by the participants' tendency to base their judgment on a least mean squares estimator rather than a maximum a posteriori estimator (Douven, 2018). We therefore choose more robust measures of meta-evaluation: system-level and overall correlations.

System-level correlation ($K^{\text{sys}}_{m_1,m_2}$). We take the correlation of the vectors containing the mean score of all stories for each system, for m_1 and m_2. This strategy measures how much m_1 and m_2 agree when comparing different systems. Formally:

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

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

The segment-level correlation, often used in conjunction with the system-level one in the meta-evaluation literature (Ma et al., 2019; Bhandari et al., 2020), is not adapted to ASE since stories generated from the same story-prompt are not required to be similar, while, e.g., translations of a sentence should look alike. We therefore use the overall correlation, which we define below.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Eval-Prompt 1

Prompt: You have become death, destroyer of worlds.

Target Story: You look up to see all of them in fear. You just must fix this soon. Slowly, just like your Father always had instructed him, you look down and see all your foes dead and beaten down. You can't resist the urge to touch the wounds. For there is nothing you can do about it. [...]

Rate the story on a scale from 1 to 5 on Surprise (how surprising the end of the story was). Rating:

Eval-Prompt 3

Prompt: You have become death, destroyer of worlds.

Target Story: You look up to see all [...]

Guidelines:
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.

Rate the story on a scale from 1 to 5 on Surprise (how surprising the end of the story was) and explain your answer. Use the provided guidelines. Rating:

Eval-Prompt 4

Prompt: You have become death, destroyer of worlds.

Target Story: You look up to see all [...]

Human Story: I saw the button. It was simple, red, no words on it as I already knew what it did. I mean I built the button, I built what happens [...]

Rate the target story on a scale from 1 to 5 on Surprise (how surprising the end of the story was) and explain your answer. Do not rate the human story; it is here only for reference. Rating:

Figure 3: Example Eval-Prompts for the Surprise criterion. Eval-Prompt 2 is the same as Eval-Prompt 1 with ''explain your answer'' added at the end. ''Prompt'' (bold) refers to the story-prompt.

Overall Correlation (K_{m_1,m_2}). We take the correlation between the full vectors containing the scores of m_1 or m_2 for a given story for every system. Formally:

$$K_{m_1,m_2} \triangleq K(\mathbf{C}_{m_1}, \mathbf{C}_{m_2}), \quad (2)$$

$$\text{where} \quad \mathbf{C}_m \triangleq \left[\left(m(y_i^j) \right)_{(i,j)\in\{1,\ldots,N\}\times\{1,\ldots,S\}} \right].$$

Statistical Testing (Section 4.1). Correlations between two automatic measures on the same annotated dataset are not independent. As advised by Graham and Baldwin (2014), we use the Williams test (Williams, 1959; Moon, 2019) to evaluate the strength of an increase in dependent correlations (Steiger, 1980).

Given three features X_1, X_2, and X_3 of a population of size n, Williams's t test for whether the correlation between X_1 and X_2 equals the correlation between X_1 and X_3 is formulated as follows:

$$t = \frac{(r_{12} - r_{13})\sqrt{(n-1)(1+r_{23})}}{\sqrt{2K\frac{(n-1)}{(n-3)} + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}},$$

where r_{ij} is the correlation between X_i and X_j and

$$K = 1 - {r_{12}}^2 - {r_{13}}^2 - {r_{23}}^2 + 2\, r_{12}\, r_{13}\, r_{23}.$$

Williams's t statistic follows a Student's t-distribution with $n-3$ degrees of freedom. In particular, the Williams test takes the correlations between X_2 and X_3 into account.

Furthermore, since we perform a large quantity of tests, we choose to correct p-values for multiplicity. As advised by Jafari and Ansari-Pour (2019), we control the false discovery rate using the Benjamini-Hochberg (BH) method (Benjamini and Hochberg, 1995) . Given n p-values $p_1, \ldots, p_n$ sorted in increasing order, the BH method consists in computing adjusted p-values $p_k^\star = p_k \frac{m}{k}$ and replacing the p-values from largest to smallest.

Following recent recommendations to move beyond simplistic ''statistical significance'' tests (Amrhein et al., 2019; Wasserstein et al., 2019; McShane et al., 2019), we report all p-values for transparency. We choose to use a gradual notion of

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

evidence for our statistical analysis, as suggested by Muff et al. (2022).

3.3 Human Evaluation of ASE Explanations

We conduct a user study in which we ask human raters to identify potential issues in LLM explanations. Dou et al. (2022) introduced an error annotation schema called SCARECROW that we adapted for ASE. We manually reviewed a random sample of 20 explanations from Beluga-13B on Eval-Prompt 3 and selected the most relevant error types. Then, we randomly sampled another 100 explanations and, for each explanation, we asked 3 human workers to annotate it w.r.t. the following five error categories:

1. **Poor Syntax**: parts of the explanation are grammatically incorrect or wrongly worded;
2. **Incoherence**: parts of the explanation are self-contradictory, logically wrong, or simply do not make sense and do not fit the other categories;
3. **Wrong Guideline**: the explanation does not respect the provided guidelines;
4. **Superfluous Text**: parts of the explanation contain text that repeats itself or generation artefacts;
5. **Unsubstantiated Claims**: the explanation fails to make explicit references to the story to substantiate its reasoning.

We recruited workers on Amazon Mechanical Turk. We estimated that a HIT would take around one minute, so we set the reward at $0.20 per HIT, so about $12 per hour. To ensure that annotators spoke fluent English, we restricted access to the experiment to the UK, the US, Canada, Australia, and New Zealand.

3.4 Experimental Details

Dataset. We use the HANNA dataset (Chhun et al., 2022), which contains 1,056 stories generated from story-prompts from the Writing-Prompts dataset (Fan et al., 2018), with both pretrained language models: **BERTGeneration** (Rothe et al., 2020), **CTRL** (Keskar et al., 2019), **GPT** (Radford et al., 2019), **GPT-2** (Radford et al., 2019), **RoBERTa** (Liu et al., 2019), and **XLNet** (Yang et al., 2019); and ASG-specific models: **Fusion** (Fan et al., 2018), **HINT** (Guan et al., 2021), and **TD-VAE** (Wilmot and Keller, 2021). These stories were annotated with scores from human raters on the six criteria introduced in Section 3.1 and 72 automatic measures. We reproduce the original procedure from Chhun et al. (2022): for reference-based evaluation measures (e.g., BLEU), we use the human story from HANNA as the reference for the generated story. Because of space constraints, we display only the evaluation measures that are the most used in the literature: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), chrF (Popović, 2015), BERTScore (Zhang et al., 2020), SUPERT (Gao et al., 2020), BLANC (Vasilyev et al., 2020), BARTScore (Yuan et al., 2021), BaryScore (Colombo et al., 2021). The results are similar for the other automatic measures.

ASG Models. Since the release of the HANNA dataset, language models have made significant advancements. We therefore felt the need to expand HANNA with more recent models. We selected **Llama-2-7b-chat-hf** (Llama-7B) as a new baseline and 4 high-performing models (at the time of selection) of different sizes on the HuggingFace Open LLM Leaderboard:[2] **Platypus2-70B-instruct** (Platypus2), **Llama-30b-instruct-2048** (Llama-30B), **StableBeluga-13B** (Beluga-13B), and **Mistral-7B-OpenOrca** (Mistral).

ASE Models. We submit each of the four Eval-Prompts 3 times on all 1,056 stories on each of the 6 criteria, and we then extract the ratings automatically from the generated answer via a regular expression. Since story evaluation on multiple prompts and multiple criteria was more computationally demanding, we limited our experiments to the smaller 13B and 7B models. We used the 4 following models: Beluga-13B, Mistral, **Llama-2-13b-chat-hf** (Llama-13B), and **Gpt-3.5-turbo** (ChatGPT). We also ran the ASE experiments with Llama-7B, which failed at the task too often for the results to be exploitable, e.g., by generating nonsensical conversations between itself and the user. We use $(\text{temperature}, \text{top\_p}) = (1, 0.95)$ for Llama models and $(0.7, 1)$ for ChatGPT (default suggested values).

Llama2 (Touvron et al., 2023b) models were trained on a closed ''new mix of data from

[2] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

publicly available sources''. Beluga-13B and Mistral-7B are Llama2 models fine-tuned on Orca-style datasets which contain triplets of ''System message–User query–LLM response'' for a large collection of tasks (Mukherjee et al., 2023). Beluga-13B is fine-tuned on StabilityAI's closed internal dataset, while Mistral-7B is fine-tuned on the open OpenOrca dataset (Lian et al., 2023). ChatGPT (Brown et al., 2020; Ouyang et al., 2022) is a closed-source model trained on a closed internal dataset that includes the CommonCrawl, Books1 and Books2 datasets.

We used the **transformers** library (Wolf et al., 2020) and the OpenAI API for our experiments.

4 Analysis of the Results

Our work aims at answering five important questions for ASE and ASG:

- **ASE1**: How do LLMs compare w.r.t. current evaluation methods, both human and automatic?
- **ASE2**: How does the Eval-Prompt influence the consistency and distribution of LLM ratings?
- **ASE3**: How explainable is the evaluation performed by LLMs?
- **ASG1**: Relying on ASE results, how do LLMs perform at ASG?
- **ASG2**: How does pretraining data help predict ASG performance?

4.1 ASE1: Comparison with Current Evaluation Measures

4.1.1 Automatic Annotation Consistency

First, we want to verify if LLMs provide stable answers. The default decoding strategy for LLMs (both Llama models and ChatGPT) is top-p sampling, which involves random variability in the generation process. We evaluate how consistent LLMs are with themselves through an inter-rater reliability (IRR) estimation. For each task, we interpret the three different LLM ratings as coming from three different annotators and we use the intra-class correlation coefficient (ICC), which is the most relevant one for our case study: Unlike Cohen's and Fleiss's kappas (Cohen, 1960; Fleiss, 1971) or Krippendorff's alpha (Hayes and Krippendorff, 2007), which quantify IRR based on all-or-nothing agreement, the ICC incorporates the magnitude of the disagreement to compute its IRR estimate, with larger-magnitude disagreements resulting in lower ICC than smaller-magnitude disagreements (Hallgren, 2012). We specifically use the ICC for *average random raters* (ICC2k) (Vallat, 2018); with the assumption that the *random* aspect can approximate the random aspect of the generation.

Crit.	Beluga-13B	Mistral-7B	Human
RE	0.88 ±0.01	0.86 ±0.01	0.48 ±0.30
CH	0.93 ±0.01	0.90 ±0.01	0.29 ±0.28
EM	0.88 ±0.01	0.87 ±0.02	0.34 ±0.09
SU	0.80 ±0.02	0.63 ±0.03	0.28 ±0.12
EG	0.91 ±0.01	0.87 ±0.01	0.46 ±0.12
CX	0.85 ±0.01	0.78 ±0.02	0.56 ±0.08

Table 1: Intra-class coefficients type 2k for Eval-Prompt 1 ratings with 95% confidence interval. Higher is better.

ICC2k values for Eval-Prompt 1 for Beluga-13B, Mistral-7B, and human ratings are displayed in Table 1. Comparing LLM consistency and human inter-rater agreement values should be done with caution: Human raters may have subjective appreciations of the Likert scale despite guidelines, while LLM consistency depends mostly on parameters that dictate output variability, e.g., temperature or top-p. That said, we reckon that it is still useful to display human IRR values as a baseline. We observe that LLMs have very high consistency overall for all criteria; the lowest value is Mistral-7B's ICC for Surprise (0.66), which is still fairly high. Confidence intervals are also smaller than for human ratings.

4.1.2 Correlations with Human Annotations

Here, we study the Kendall correlations between LLM and human ratings on corresponding criteria. For the ''Beluga-13B 1'' column in Figure 4, the first value is the correlation between Beluga-13B Relevance ratings and averaged human Relevance ratings for Eval-Prompt 1, then Coherence ratings, etc.

Assuming we want an automatic measure to perform as well as an individual human rater would, we need a baseline for comparison. Therefore, we also compute the average correlations between individual human ratings and average human ratings, which we compiled into the same figures for the sake of readability (the ''Human'' column). Since the individual human rating is

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

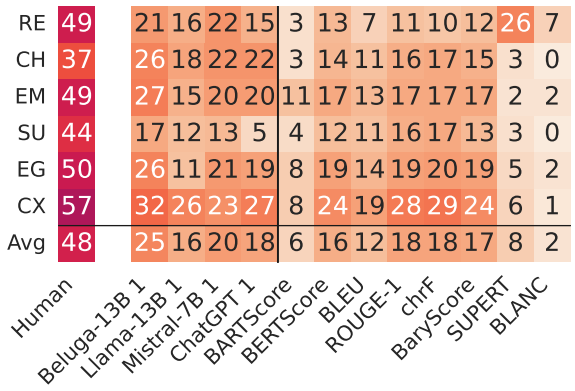

Figure 4: Overall absolute Kendall correlations between evaluation measures and human ratings. Higher is better. The black vertical line separates LLMs (left) and non-LLMs (right). Coefficient values are multiplied by 100 for readability; we will symbolize this with ‘‘(×100)’’ in the next figures.

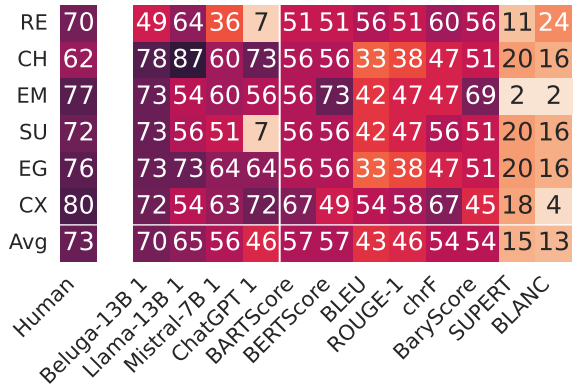

Figure 5: System-level absolute Kendall correlations (×100) between evaluation measures and human ratings. Higher is better. The white vertical line separates LLMs (left) and non-LLMs (right).

included in the average human rating, both measures are not independent, so the column acts as an upper-bound.

Overall Correlations (Figure 4). LLM ratings generally correlate with human ratings similarly to automatic measures, if not better. Overall, Beluga-13B is the best performer, achieving higher correlations (0.25 on average) than both other LLMs and automatic measures (≤0.18). The better results (as compared to Llama-13B (0.16) and Mistral-7B (0.20)) suggest a positive influence of fine-tuning and model size respectively. The inferior performance of ChatGPT (0.18) is difficult to explain since OpenAI does not disclose the details of its architecture, its training process and, most importantly, its training data. Nonetheless, an important takeaway is that current source-available models can effectively compete with closed-source models: this is good news for NLP research, since observations made on closed-source models cannot easily be generalized.

System-level Correlations (Figure 5). First, we observe that human baseline correlations are noticeably higher than non-LLM automatic measures: While human annotators tend to reach a consensus when ranking systems (averaging correlations of 0.73), non-LLM automatic measures are moderately to poorly correlated from human judgment (with values ranging from 0.13 to 0.57).

Meanwhile, Llama models display very high correlations, with Beluga-13B performing almost as well as human raters (0.70 vs 0.73). ChatGPT shows a somewhat erratic performance (correlations range from 0.07 to 0.73), which is overall comparable or inferior to Llama models. Also, LLMs generally outperform other automatic measures (0.70 for Beluga-13B compared to 0.57 for `BARTScore`).

The fact that correlations are sometimes higher than the baseline can be explained by the subjective nature of the task: human annotators may exhibit higher variability in their ratings than the stable LLMs.

Statistical Testing. Figure 6 shows the BH-adjusted p-values of the Williams tests for the increase in correlations with a given criterion between Beluga-13B average Eval-Prompt 1 ratings (row) and other measures (column).

For overall correlations, there is strong statistical evidence that Beluga-13B correlates better with human judgment than many non-LLM automatic measures ($p < 0.01$ for many tests). Evidence is more moderate to weak when comparing Beluga-13B and other LLMs. For instance, between Beluga-13B and ChatGPT, p-values lie between 0.01 and 0.14. While the performance of Beluga-13B still leaves a lot of room for improvement, it performs better than non-LLM automatic measures.

For system-level correlations, statistical evidence for better performance appears weaker: $p > 0.11$ for all tests. However, one should keep in mind that the ratings (averaged over more than 1,000 stories) used to compute system-level correlations hold more information than the individual

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

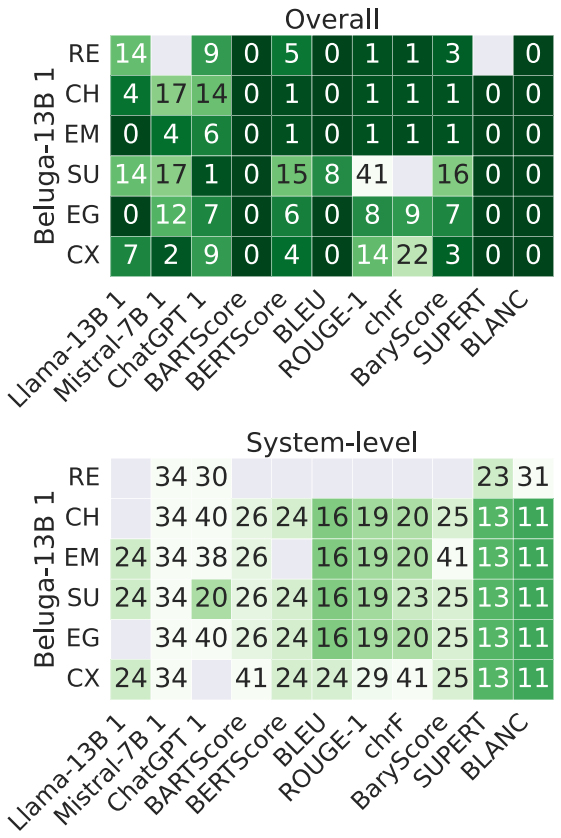

Figure 6: BH-adjusted p-values ($\times 100$) of the Williams tests for overall and system-level Kendall correlations. Lower is better. ''0'' means $p < 0.01$.

ratings of the overall correlations. Therefore, while statistical evidence is weaker, the averaged nature of the correlations and the significant numeric increases in correlations (0.70 for Beluga-13B vs 0.57 for `BARTScore`/`BERTScore`) suggest that Beluga-13B is more reliable at ordering systems compared to non-LLM measures.

4.1.3 Takeaways

First, LLMs show very high self-consistency. Overall correlations remain weak, although LLMs display marginal improvements over non-LLM automatic measures, backed with strong statistical evidence. At the system-level, LLM correlations with human judgment are high, but statistical evidence is weaker. In conclusion, while LLMs still cannot be relied upon to evaluate a single story, they appear more reliable than non-LLM automatic measures for comparing different models and selecting the best one.

4.2 ASE2: Influence of the Eval-Prompt

In this section, we discuss the influence of the Eval-Prompt on the consistency and distribution of the generated LLM ratings.

4.2.1 Influence on Consistency

Here, we analyze the influence of the Eval-Prompt on LLM consistency. ICC2k values for Beluga-13B ratings w.r.t. the different Eval-Prompts are shown in Table 2 (other LLMs display similar behavior). The influence of Eval-Prompts appears limited: providing guidelines (Eval-Prompt 3) tends to decrease self-consistency for all criteria except Complexity with a discernible effect (as shown by the confidence intervals), but ICC values remain very high. LLMs are therefore remarkably consistent in their grading, no matter the Eval-Prompt.

4.2.2 Influence on Ratings

We show the average Likert ratings per LLM per Eval-Prompt on Table 3. Compared to Eval-Prompt 1, Eval-Prompt 2 seems to have limited influence on the ratings for all models, often leading to overlapping confidence intervals. Eval-Prompt 3 causes a statistically discernible decrease in ratings for Beluga-13B and Llama-13B, and a discernible increase for ChatGPT. Eval-Prompt 4 has a similar effect, with the decrease also observable with Mistral-7B. The significantly lower ratings of ChatGPT partly stem from the fact that it was not asked to rate the new Llama-generated stories, which were generally highly-rated.

Overall, it seems that more detailed Eval-Prompts (3 and 4) tend to decrease the ratings for Llama-models while having an opposite effect for ChatGPT. We tried to separate ratings per generative model or per criterion but were unable to identify a more specific pattern: We therefore chose to show only the aggregated results for the sake of clarity.

4.2.3 Influence on Correlations

Here we analyze the influence of Eval-Prompts on correlations between LLM ratings and human ratings.

Overall Correlations (Figure 7). Eval-Prompt 2 overall correlations are very close to Eval-Prompt 1 correlations for all models: simply asking for an explanation has limited influence on correlations. Eval-Prompt 3 tends

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Criterion	Eval-Prompt 1	Eval-Prompt 2	Eval-Prompt 3	Eval-Prompt 4
Relevance	0.88 ±0.01	0.90 ±0.01	0.85 ±0.02	0.92 ±0.01
Coherence	0.93 ±0.01	0.94 ±0.01	0.87 ±0.01	0.93 ±0.01
Empathy	0.88 ±0.01	0.88 ±0.01	0.83 ±0.02	0.91 ±0.01
Surprise	0.80 ±0.02	0.79 ±0.02	0.70 ±0.03	0.85 ±0.01
Engagement	0.91 ±0.01	0.92 ±0.01	0.79 ±0.02	0.93 ±0.01
Complexity	0.85 ±0.01	0.86 ±0.01	0.85 ±0.01	0.89 ±0.01

Table 2: Intra-class coefficients type 2k for Beluga-13B ratings with 95% confidence interval. Higher is better.

LLM	Eval-Prompt 1	Eval-Prompt 2	Eval-Prompt 3	Eval-Prompt 4
Beluga-13B	3.48 ±0.04	3.38 ±0.03	3.06 ±0.03	3.28 ±0.04
Llama-13B	3.48 ±0.03	3.52 ±0.03	3.21 ±0.02	2.82 ±0.03
Mistral-7B	3.47 ±0.03	3.51 ±0.03	3.46 ±0.03	3.28 ±0.03
ChatGPT*	1.52 ±0.03	1.47 ±0.03	1.62 ±0.02	1.60 ±0.03

Table 3: Average Likert ratings per LLM per Eval-Prompt. The asterisk signals the fact that ChatGPT was only asked to rate the original HANNA dataset without Llama-generated stories. Higher is better.

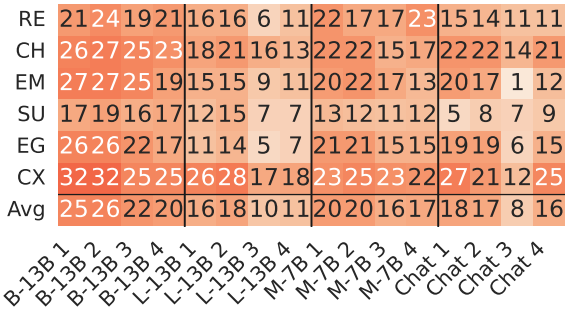

Figure 7: Overall absolute Kendall correlations (×100) between LLMs and human ratings for different Eval-Prompts. Higher is better. B-13B = Beluga-13B, L-13B = Llama-13B, M-7B = Mistral-7B and Chat = ChatGPT.

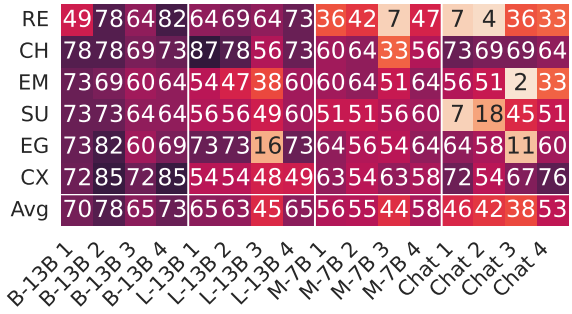

Figure 8: System-level absolute Kendall correlations (×100) between LLMs and human ratings for different Eval-Prompts. Higher is better. B-13B = Beluga-13B, L-13B = Llama-13B, M-7B = Mistral-7B and Chat = ChatGPT.

to decrease correlations for all models: Providing guidelines makes the model less accurate, counter-intuitively. Eval-Prompt 4 (providing a human story for reference) has a similar effect.

System-level Correlations (Figure 8). Eval-Prompt 2 has limited effect on correlations again, except for Beluga-13B for whom it seems to increase correlations. Eval-Prompt 3 decreases correlations, with a marked effect in Llama-13B. Finally, Eval-Prompt 4 seems to cause a small increase in correlations, contrary to its decreasing effect on overall correlations.

4.2.4 Takeaways

First, regardless of Eval-Prompt complexity, LLMs behave consistently when prompted multiple times. Asking for an explanation (Eval-Prompt 2) has negligible effect on ratings, while more complex Eval-Prompts (3 - providing guidelines and 4 - providing a reference human story) have a more discernible influence (positive or negative). As for correlations with human ratings, providing guidelines (Eval-Prompt 3) consistently seems to lower correlations, whereas providing a human story for reference (Eval-Prompt 4) has opposite effects for overall or system-level correlations.

4.3 ASE3: Explainability of Ratings

In this section, we analyze to what extent the explanations provided by LLMs are consistent w.r.t. their ratings, e.g., whether they differ from criterion to criterion, whether they are semantically relevant and, for Eval-Prompt 3, whether they are compliant with the provided guidelines. We

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

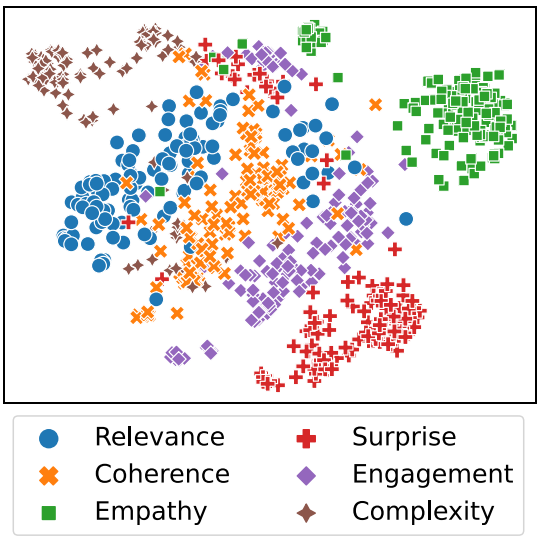

Figure 9: UMAP projection of Beluga-13B explanations.

will focus on Beluga-13B since it had the best correlations with human judgment, as shown in Section 4.1.

4.3.1 Visualization of Explanation Embeddings

First, we want to ascertain whether Beluga-13B provides different explanations for each of the human criteria. We gather the explanations provided by Beluga-13B on human stories for each criterion and use the **SentenceTransformers** library (Reimers and Gurevych, 2019) to compute their corresponding embeddings. We then use a 2D UMAP projection (McInnes et al., 2018) (with parameters n_neighbors = 300 and metric = euclidean) to visualize how the embeddings are distributed. Figure 9 shows the visualization of the UMAP projection: Beluga's explanations are overall well-separated w.r.t. their corresponding criteria.

4.3.2 Keyword Analysis

Since Beluga's explanations seem to vary from one criterion to another, we evaluate whether they make sense from a semantic point of view. We use the YAKE! keyword extractor, which significantly outperforms other state-of-the-art methods (Campos et al., 2020): We show selected 3-gram keywords from the top-30 per criterion in Table 4. The results are consistent with Figure 9: keywords are overall different for each criterion. We can also see here that they are semantically relevant.

Crit.	Keywords
RE	story, prompt, roughly matches, target, weak relationship, connection, weak
CH	story, coherence, make sense, difficult to understand, clear narrative structure
EM	empathy, emotions, understand the characters, depth, emotional connection
SU	story, surprise, ending, predictable, rate, unexpected, twist, completely obvious
EG	story, mildly interesting, engagement, difficult, found, characters, fully engage
CX	story, characters, intricate plot, difficult to understand, straightforward, depth

Table 4: Selected keywords from Beluga-13B explanations w.r.t. a specific criterion.

Error Type	Rate	AC1
Poor Syntax	0.02	0.93 0.97 1.00
Incoherence	0.11	0.73 0.81 0.89
Wrong Guideline	0.13	0.85 0.90 0.96
Superfluous Text	0.20	0.55 0.66 0.78
Unsubstantiated Claims	0.31	0.47 0.60 0.74

Table 5: Error rates of Beluga-13B Eval-Prompt 3 on a sample of 100 explanations. Lower is better.

4.3.3 User Study on LLM Explanations

We display the results of our user study (designed in Section 3.3) in Table 5. We also display the IRR, which we computed using Gwet's agreement coefficient 1 (AC1) (Gwet, 2008; Fergadis and Scheffler, 2022). Gwet's AC1 is known to perform well for IRR estimation on binary classification tasks such as our user study: it was designed to be more stable and less affected by prevalence and marginal probability than Cohen's kappa, and this was confirmed by practical experiments (Wongpakaran et al., 2013).

We can see that Beluga-13B produces near-impeccable syntax, at least according to annotators (2% of ''Poor Syntax''). It also does a good job at producing coherent text (11% of ''Incoherence''), and mostly understands the guidelines (13% of ''Wrong Guideline''). However, it tends to repeat itself somewhat (20% of ''Superfluous Text'') and, most notably, tends not to substantiate its claims with direct references to the story (31% of ''Unsubstantiated Claims''). Overall, annotators tend to agree

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Model	RE	CH	EM	SU	EG	CX	Average
Human	3.37 ±0.12	3.55 ±0.11	3.42 ±0.11	3.11 ±0.13	3.58 ±0.10	3.48 ±0.10	3.42 ±0.06
Platypus2-70B	4.09 ±0.05	4.31 ±0.05	3.92 ±0.06	**3.69** ±0.07	4.19 ±0.05	3.88 ±0.05	4.01 ±0.03
Llama-30B	**4.19** ±0.05	**4.38** ±0.04	**4.04** ±0.06	**3.63** ±0.09	**4.31** ±0.05	**3.98** ±0.05	**4.08** ±0.03
Beluga-13B	4.06 ±0.08	4.10 ±0.06	3.75 ±0.08	3.54 ±0.08	3.90 ±0.08	3.69 ±0.07	3.84 ±0.05
Mistral-7B	4.12 ±0.05	4.25 ±0.05	3.86 ±0.06	3.56 ±0.08	4.11 ±0.05	3.82 ±0.04	3.95 ±0.03
Llama-7B	4.07 ±0.06	4.24 ±0.05	3.90 ±0.06	3.58 ±0.06	4.09 ±0.05	3.79 ±0.05	3.95 ±0.03
GPT-2	2.57 ±0.13	2.36 ±0.11	2.72 ±0.11	2.59 ±0.14	2.67 ±0.12	2.89 ±0.12	2.63 ±0.07
HINT	1.57 ±0.10	1.31 ±0.07	1.59 ±0.10	1.49 ±0.10	1.58 ±0.09	1.43 ±0.08	1.49 ±0.06

Table 6: Average Beluga-13B ratings for Eval-Prompt 1 with 95% confidence interval. Higher is better.

Model	RE	CH	EM	SU	EG	CX	Average
Human	3.48 ±0.11	3.50 ±0.10	3.69 ±0.08	3.24 ±0.11	3.42 ±0.10	3.45 ±0.07	3.46 ±0.05
Platypus2-70B	**4.26** ±0.08	**4.31** ±0.08	**4.05** ±0.07	**3.46** ±0.10	**3.94** ±0.06	3.55 ±0.07	**3.93** ±0.03
Llama-30B	4.15 ±0.10	**4.29** ±0.07	**4.02** ±0.07	**3.46** ±0.09	**3.94** ±0.06	**3.65** ±0.07	**3.92** ±0.03
Beluga-13B	4.07 ±0.09	4.14 ±0.07	3.98 ±0.07	**3.50** ±0.09	3.74 ±0.08	3.59 ±0.07	3.84 ±0.03
Mistral-7B	4.15 ±0.10	4.22 ±0.08	**4.02** ±0.07	**3.51** ±0.11	**3.94** ±0.07	**3.67** ±0.07	**3.92** ±0.04
Llama-7B	4.13 ±0.10	4.14 ±0.09	3.90 ±0.08	**3.48** ±0.09	3.78 ±0.08	3.56 ±0.08	3.83 ±0.05
GPT-2	2.40 ±0.10	2.37 ±0.09	2.74 ±0.10	2.85 ±0.11	2.60 ±0.09	2.88 ±0.09	2.64 ±0.05
HINT	2.12 ±0.11	2.13 ±0.08	2.23 ±0.10	2.28 ±0.11	2.05 ±0.08	2.05 ±0.09	2.15 ±0.06

Table 7: Average Mistral-7B ratings for Eval-Prompt 1 with 95% confidence interval. Higher is better.

with one another, as showed by the high values of Gwet's AC1.

The substantial rate of ''Unsubstantiated Claims'' and the fact that **40% of all Eval-Prompt 3 ratings are not supported by an explanation**—despite the Eval-Prompt explicitly asking for it—beg the question of whether Beluga-13B truly understands the given task. We discuss this question further in Section 5.

Takeaways. LLM explanations seem to be specific to each considered human evaluation criterion; however, a finer analysis with a user study reveals that LLMs often struggle with following guidelines and substantiating their explanations.

4.4 ASG1: LLM Performance in ASG

In this section, we discuss the performance of LLMs at the ASG task compared to human and previous models' performance, as we expanded the HANNA dataset with stories generated from more recent models. Since Beluga-13B and Mistral-7B display very high system-level correlations with human ratings (see Figure 5), we use their ratings as proxy for human ratings. Table 6 and Table 7 show the average Beluga-13B and Mistral-7B ratings for Eval-Prompt 1 per model per criterion for a few HANNA models (GPT-2, HINT) and the Llama models.

We observe that LLMs perform remarkably well, getting higher ratings than older models (GPT-2) and even human stories. Beluga-13B and Mistral-7B both seem to prefer the outputs from larger LLMs (Platypus2-70B, Llama-30B) to their own outputs, suggesting that the LLM grading process cannot be explained simply by a proxy for perplexity. Interestingly, in both tables, Mistral-7B gets slightly higher ratings than Beluga, with some differences being statistically discernible, which could be explained by differences in fine-tuning data.

Takeaways. Larger models (Platypus2-70B, Llama-30B) exhibit the best ASG performance, with LLM ratings at least equal to those of human stories. However, our setting involves short stories of between 500 and 1,000 words; generating longer stories may prove more difficult since maintaining large-scale coherence may become an issue.

4.5 ASG2: Influence of Pretraining Data on ASG Performance

In this section, we verify whether the LLM pretraining data contains the WritingPrompts dataset to check for model contamination, as advised by Magar and Schwartz (2022), and to what extent

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Model	Contamination (%)
Platypus2-70B	0.80
Llama-30B	1.80
Beluga-13B	4.40
Mistral-7B	2.50
Llama-7B	10.10

Table 8: Predicted contamination rates of the WritingPrompts sample.

Model	AUC (%)
Platypus2-70B	92.1
Llama-30B	81.3
Beluga-13B	70.1
Mistral-7B	51.2
Llama-7B	55.1

Table 9: AUC detection score on the BooksMIA dataset.

ASG performance is related with data exploitation, e.g., through reproduction of training examples.

We use the MIN-K% PROB detection method (Shi et al., 2024), which is based on the hypothesis that unseen data will contain more outlier words with low probability than seen data. Furthermore, it does not require additional training. Given a sentence and an LLM's probability distribution of the next token, MIN-K% PROB selects the top-k% of tokens with the highest negative log-likelihood and computes their average log-likelihood. We can then detect if the sentence was included in pretraining data by thresholding this average. We follow Shi et al. (2024) and use $k = 20$ for our two experiments.

Model Contamination. We sample 1,000 stories from the WritingPrompts dataset (Fan et al., 2018), from which the HANNA human stories come. Table 8 shows the predicted contamination rates of the WritingPrompts sample. Since they are very low, this strongly suggests that the WritingPrompts sample was not included in the pretraining data of the evaluated models. We can reasonably surmise that the same applies to the whole WritingPrompts dataset.

Data Reproduction. We use the BooksMIA dataset (Shi et al., 2024), which contains 9,870 samples of books labeled 0 if included in the Books3 dataset (commonly used for pretraining LLMs) or 1 if released in or after January 2023. Since the BooksMIA data is labeled, we compute the area under the ROC curve (AUC) obtained with MIN-K% PROB thresholding. Results are shown on Table 9.

We observe that the AUC detection score is higher for larger models, e.g., it is easier to detect if a book was in the pretraining data of a larger LLM. The definition of the MIN-K% PROB measure also means that larger LLMs tend to produce text that is more similar to their pretraining data, such as fiction books, which could help explain their better ASE ratings.

Takeaways. The better performance of larger LLMs for ASG may be partially explained by their tendency to generate text that is more similar to their pretraining data, e.g., existing novels.

5 Discussion on LLM Performance

Our work is part of the ongoing research on the general ability of LLMs for understanding and thinking.

Mahowald et al. (2024) distinguish formal (the statistical features of language) and functional linguistic competence (the ability to use language in the world) and show that LLMs are very successful on formal linguistic tasks but struggle at functional linguistic tasks. Bubeck et al. (2023) argue that LLMs do display impressive performance at a wide variety of tasks but lack ''slow thinking'' capabilities, referring to the System 1–System 2 dichotomy introduced by Kahneman (2011).

Thus, the high performance of LLMs at ASE should be interpreted with caution: we hypothesize that the ''rating'' part of our story evaluation experiments could be linked to formal linguistic competence and the fast, automatic System 1, while the ''explanation'' part would correspond to functional linguistic competence and the slow, conscious System 2.

This analogy would explain the good correlations of LLM ratings with human ratings: The internal criterion of LLMs for story evaluation may be formal quality (vocabulary, syntax, grammar), regardless of the criterion mentioned in the Eval-Prompt. Indeed, the six criteria from Chhun et al. (2022) are mostly orthogonal but not completely independent: Their correlation with one another may be related to the general ''System 1'' tendency of human raters to favor stories that display better formal qualities. In that sense, LLMs

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

may reflect a human bias towards easy, intuitive thinking. By contrast, the less convincing performance of LLMs at explaining their ratings may highlight their weaker System 2 capabilities as argued by Mahowald et al. (2024) and Bubeck et al. (2023).

6 Conclusions

6.1 Practical Takeaways

1. **Used with prompts based on specific criteria, LLMs are currently the best proxy for human evaluation of story generation (Section 4.1.2).** In particular, LLMs display very high system-level correlations with human judgment.
2. **LLMs are remarkably self-consistent (Section 4.1.1),** exhibiting very high intra-class coefficient values;
3. **LLMs understand the ASE task only partially (Section 4.3.3):** They struggle to explain their answers with substantiated claims.
4. **For ASE, providing detailed guidelines (Eval-Prompt 3) did not lead to improved correlations with human ratings (Section 4.2.3).** Providing a human story for reference (Eval-Prompt 4) yields mixed results.
5. **LLM stories have at least equal ASE ratings to human stories (Section 4.4),** with larger LLMs exhibiting the best performance.
6. **Pretraining data helps explain LLM performance at ASG (Section 4.5):** The higher ratings of larger LLMs may be due to their ability to produce output similar to existing books.

6.2 Limitations and Future Directions

The ASE task is a very subjective one: LLM performance at ASE and ASG must be seen as a reflection of *average* preferences and may therefore include biases, e.g., from their pretraining data.

Furthermore, we performed most of our experiments in a zero-shot setting without further training; it would be interesting to compare our results with future work involving fine-tuning or reinforcement learning with human feedback on data specific to ASE.

Also, we did not conduct our experiments with LLMs that were optimized for long inputs and outputs, such as GPT-4.

Finally, we mainly used source-available LLama models and found that they performed at least as well as ChatGPT, a proprietary model. We encourage the NLP community to favor the use of such models, as the growing presence of closed models hinders research transparency and reproductibility.

Acknowledgments

This work was performed using HPC resources from GENCI-IDRIS (grant 2022-AD011013105R1) and was partially funded by the grants ANR-20-CHIA-0012-01 (‘‘NoRDF’’) and ANR-23-CE23-0033-01 (‘‘SINNet’’).

We would also like to convey our appreciation to TACL action editor Ehud Reiter, as well as to our anonymous reviewers, for their valuable feedback.

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. https://doi.org/10.18653/v1/2021.nuse-1.8

Valentin Amrhein, Sander Greenland, and Blake McShane. 2019. Scientists rise up against statistical significance. *Nature*, 567(7748):305–307. https://doi.org/10.1038/d41586-019-00857-9, PubMed: 30894741

Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Re. 2023. Ask me anything: A simple strategy for prompting language models. In *The Eleventh International Conference on Learning Representations*.

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. https://doi.org/10.1007/978-3-030-92300-6_45

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

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. https://doi.org/10.18653/v1.emnlp-main.751

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*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712v5*.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289. https://doi.org/10.1016/j.ins.2019.09.013

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *ArXiv preprint*, abs/2006.14799v2.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023. Art or artifice? Large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556v1*.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296,

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.565

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. https://doi.org/10.1177/001316446002000104

Pierre Colombo, Maxime Peyrard, Nathan Noiry, Robert West, and Pablo Piantanida. 2023. The glass ceiling of automatic evaluation in natural language generation. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 178–183, Nusa Dua, Bali. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-ijcnlp.16

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021. 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. https://doi.org/10.18653/v12021.emnlp-main.817

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.753

Robert Dickman. 2003. The four elements of every successful story. *Reflections - Society for Organizational Learning*, 4(3):51–58. https://doi.org/10.1162/15241730360580212

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.626

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.501

Igor Douven. 2018. A Bayesian perspective on Likert scales and central tendency. *Psychonomic Bulletin & Review*, 25:1203–1211. https://doi.org/10.3758/s13423-017-1344-2, PubMed: 28752379

Edilivre. 2023. Concours de nouvelles 2023. Accessed: 2023-10-12.

Majigsuren Enkhsaikhan, Wei Liu, Eun-Jung Holden, and Paul Duuring. 2021. Auto-labelling entities in low-resource text: A geological case study. *Knowledge and Information Systems*, 63:695–715. https://doi.org/10.1007/s10115-020-01532-6

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. https://doi.org/10.18653/v1/P18-1082

Aris Fergadis and Benedikt Scheffler. 2022. Chance-corrected agreement coefficients.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378. https://doi.org/10.1037/h0031619

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. https://doi.org/10.18653/v1/2020.acl-main.124

Daniel R. George, Heather L. Stuckey, and Megan M. Whitehead. 2014. How a creative

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

storytelling intervention can improve medical student attitude towards persons with dementia: A mixed methods study. *Dementia*, 13(3):318–329, https://doi.org/10.1177/1471301212468732, PubMed: 24770946

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. https://doi.org/10.18653/v1/2020.emnlp-main.351

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. https://doi.org/10.3115/v1/D14-1020

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. https://doi.org/10.1162/tacl_a_00302

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. 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. https://doi.org/10.18653/v1/2021.acl-long.499

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48. https://doi.org/10.1348/000711006X126600, PubMed: 18482474

Kevin A. Hallgren. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1):23. https://doi.org/10.20982/tqmp.08.1.p023, PubMed: 22833776

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89. https://doi.org/10.1080/19312450709336664

Mohieddin Jafari and Naser Ansari-Pour. 2019. Why, when and how to adjust your P values? *Cell Journal (Yakhteh)*, 20(4):604–607. https://doi.org/10.22074/cellj.2019.5992

João Ricardo de Oliveira Júnior, Ricardo Limongi, Weng Marc Lim, Jacqueline K. Eastman, and Satish Kumar. 2023. A story to sell: The influence of storytelling on consumers' purchasing behavior. *Psychology & Marketing*, 40(2):239–261. https://doi.org/10.1002/mar.21758

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93. https://doi.org/10.2307/2332226

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*, abs/1909.05858v2.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):598–604. https://doi.org/10.1609/aaai.v27i1.8649

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and ''Teknium''. 2023. OpenOrca: An open dataset of GPT augmented FLAN reasoning traces.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text*

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Summarization Branches Out, pages 74–81, Barcelona, Spain. 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*, abs/1907.11692v1.

Vincenzo Lombardo and Rossana Damiano. 2012. Storytelling on mobile devices for cultural heritage. *New Review of Hypermedia and Multimedia*, 18(1-2):11–35. https://doi.org/10.1080/13614568.2012.617846

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. https://doi.org/10.18653/v1/W19-5302

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.18

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540. https://doi.org/10.1016/j.tics.2024.01.011, PubMed: 38508911

Allyssa McCabe and Carole Peterson. 1984. What makes a good story. *Journal of Psycholinguistic Research*, 13(6):457–480. https://doi.org/10.1007/BF01068179

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861. https://doi.org/10.21105/joss.00861

Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. 2019. Abandon statistical significance. *The American Statistician*, 73(sup1): 235–245. https://doi.org/10.1080/00031305.2018.1527253

Sara Miller and Lisa Pennycuff. 2008. The power of story: Using storytelling to improve literacy learning. *Journal of Cross-Disciplinary Perspectives in Education*, 1(1):36–43.

Jihyung Moon. 2019. Significance test of increase in correlation for NLP evaluations in Python.

Stefanie Muff, Erlend B. Nilsen, Robert B. O'Hara, and Chloé R. Nater. 2022. Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution*, 37(3):203–210. https://doi.org/10.1016/j.tree.2021.10.009, PubMed: 34799145

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint arXiv:2306.02707v1*.

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. https://doi.org/10.18653/v1/D17-1238

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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. https://doi.org/10.3115/1073083.1073135

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Karl Pearson. 1895. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347–352):240–242. https://doi.org/10.1098/rspl.1895.0041

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. https://doi.org/10.18653/v1/W15-3049

Muhammad Aasim Qureshi, Muhammad Asif, Mohd Fadzil Hassan, Ghulam Mustafa, Muhammad Khurram Ehsan, Aasim Ali, and Unaza Sajid. 2022. A novel auto-annotation technique for aspect level sentiment analysis. *Computers, Materials and Continua*, 70(3):4987–5004. https://doi.org/10.32604/cmc.2022.020544

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.

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. https://doi.org/10.18653/v1/2020.emnlp-main.349

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7. https://doi.org/10.1145/3411763.3451760

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. https://doi.org/10.1162/tacl_a_00313

Stephen Rowcliffe. 2004. Storytelling in science. *School Science Review*, 86(314):121.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net.

Charles Spearman. 1961. The proof and measurement of association between two things. In J. J. Jenkins and D. G. Paterson, editors, *Studies in Individual Differences: The Search for Intelligence*, pages 45–58. Appleton-Century-Crofts. https://doi.org/10.1037/11491-005

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251. https://doi.org/10.1037/0033-2909.87.2.245

Stanley S. Stevens. 1971. Issues in psychophysical measurement. *Psychological Review*, 78(5):426–450. https://doi.org/10.1037/h0031324

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. LaMDA: Language models for dialog applications.*ArXiv preprint*, abs/2201.08239v3.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971v1*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288v2*.

Scott R. Turner. 2014. *The Creative Process: A Computer Model of Storytelling and Creativity*. Psychology Press.

Raphael Vallat. 2018. Pingouin: Statistics in Python. *Journal of Open Source Software*, 3(31):1026. https://doi.org/10.21105/joss.01026.

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. https://doi.org/10.18653/v1/2020.eval4nlp-1.2

Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. 2021. Automated event annotation in literary texts. In *Proceedings of the Conference on Computational Humanities Research, CHR2021, Amsterdam, The Netherlands, November 17–19, 2021*, volume 2989 of *CEUR Workshop Proceedings*, pages 333–345. CEUR-WS.org.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.354

Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. 2019. Moving to a world beyond ‘‘p<0.05’’. *The American Statistician*, 73(sup1):1–19. https://doi.org/10.1080/00031305.2019.1583913

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models.*Transactions on Machine Learning Research*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *ArXiv preprint*, abs/2302.11382v1.

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

David Wilmot and Frank Keller. 2021. A temporal variational model for story generation. *ArXiv preprint*, abs/2109.06807v1.

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.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025

Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. 2013. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13:16. https://doi.org/10.1186/1471-2288-13-61, PubMed: 23627889

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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual*, pages 27263–27277.

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).

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00689/2470807/tacl_a_00689.pdf by guest on 16 March 2025