



Meta-Evaluation Methodology and Benchmark for Automatic Story Generation

PhD Defense of Cyril Chhun

November 19, 2024

Supervisors: Fabian Suchanek and Chloé Clavel

Reviewers: Claire Gardent and Smaranda
Muresan

Examiners: Sophie Rosset, Benoît Favre and
Marine Carpuat



An Amazing Story

November 19th, 2024.

My PhD defense.

“Methodology and Meta-Evaluation Benchmark for Automatic Story Generation” was my topic, a subject I poured countless hours into, battling technical hurdles to build a model that could truly tell compelling stories.

The audience in the Rose Dieng-Küntz auditorium awaited with eager eyes.

We analyzed bias, built robust benchmarks, and finally stood at the precipice of true narrative generation. The questions were insightful, each answer revealing a new facet to the story we’re building. Fabian’s excitement was contagious, and Chloé’s calm reassurance never faltered.

My defense ended with thunderous applause and proud smiles, another step towards an exciting future where machines could understand stories on a deeper level.

An Amazing Story

November 19th, 2024.

My PhD defense.

“Methodology and Meta-Evaluation Benchmark for Automatic Story Generation” was my topic, a subject I poured countless hours into, battling technical hurdles to build a model that could truly tell compelling stories.

The audience in the Rose Dieng-Küntz auditorium awaited with eager eyes.

We analyzed bias, built robust benchmarks, and finally stood at the precipice of true narrative generation. The questions were insightful, each answer revealing a new facet to the story we’re building. Fabian’s excitement was contagious, and Chloé’s calm reassurance never faltered.

My defense ended with thunderous applause and proud smiles, another step towards an exciting future where machines could understand stories on a deeper level.

Courtesy of Google’s gemma2:2.6b language model.



An Amazing Story

Prompt used: Your name is Cyril Chhun. You are a PhD student in natural language processing at Télécom Paris, a French engineering school. The subject of your thesis is “Methodology and Meta-Evaluation Benchmark for Automatic Story Generation”. Your supervisors are Fabian Suchanek and Chloé Clavel. It is Tuesday, November 19th 2024. You are defending your PhD in the Rose Dieng-Küntz auditorium. Tell me about your defense.



Research Questions and Outline

Several questions come to mind:

Several questions come to mind:

1. Why is this story so amazing? More generally, how can we evaluate the quality of a story?
↪ **Methodology for Story Evaluation and Automatic Story Generation (ASG) Meta-Evaluation**

Several questions come to mind:

1. Why is this story so amazing? More generally, how can we evaluate the quality of a story?
↪ **Methodology for Story Evaluation and Automatic Story Generation (ASG) Meta-Evaluation**
2. How good are existing models at generating stories?
↪ **HANNA: A Corpus of Human-ANnotated Narratives for ASG evaluation**

Several questions come to mind:

1. Why is this story so amazing? More generally, how can we evaluate the quality of a story?
↪ **Methodology for Story Evaluation and Automatic Story Generation (ASG) Meta-Evaluation**
2. How good are existing models at generating stories?
↪ **HANNA: A Corpus of Human-ANnotated Narratives for ASG evaluation**
3. Could we evaluate stories automatically?
↪ **Meta-Evaluation Benchmark of Automatic Story Generation**

Several questions come to mind:

1. Why is this story so amazing? More generally, how can we evaluate the quality of a story?
→ **Methodology for Story Evaluation and Automatic Story Generation (ASG) Meta-Evaluation**
2. How good are existing models at generating stories?
→ **HANNA: A Corpus of Human-ANnotated Narratives for ASG evaluation**
3. Could we evaluate stories automatically?
→ **Meta-Evaluation Benchmark of Automatic Story Generation**
4. What about large language models (LLMs)? Do they even understand the task of story evaluation?
→ **Exploring LLM Explainability for Automatic Story Evaluation (ASE) and ASG**

Outline

Introduction and Background

Methodology for Story Evaluation and ASG Meta-Evaluation

HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation

Meta-Evaluation Benchmark of Automatic Story Generation

Exploring LLM Explainability for ASE and ASG

Conclusion

Outline

Introduction and Background

Methodology for Story Evaluation and ASG Meta-Evaluation

HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation

Meta-Evaluation Benchmark of Automatic Story Generation

Exploring LLM Explainability for ASE and ASG

Conclusion

Automatic Story Generation (ASG)

First, why do we care about ASG?

Automatic Story Generation (ASG)

First, why do we care about ASG?

- Creativity: a long-standing goal of research in AI (at least since the 1950s);

Automatic Story Generation (ASG)

First, why do we care about ASG?

- ▶ Creativity: a long-standing goal of research in AI (at least since the 1950s);
- ▶ Stories play a central role in human societies;

Automatic Story Generation (ASG)

First, why do we care about ASG?

- ▶ Creativity: a long-standing goal of research in AI (at least since the 1950s);
- ▶ Stories play a central role in human societies;
- ▶ Cognitive sciences: storytelling-based approaches can be used in communication, education, gaming, marketing, etc.

Automatic Story Generation (ASG)

First, why do we care about ASG?

- ▶ Creativity: a long-standing goal of research in AI (at least since the 1950s);
- ▶ Stories play a central role in human societies;
- ▶ Cognitive sciences: storytelling-based approaches can be used in communication, education, gaming, marketing, etc.

↪ Can AI help?

Automatic Story Generation (ASG)

ASG systems can be divided into three categories (Alhussain and Azmi, 2021):

1. **structural** models,
2. **planning-based** models,
3. **machine learning (ML)** models.

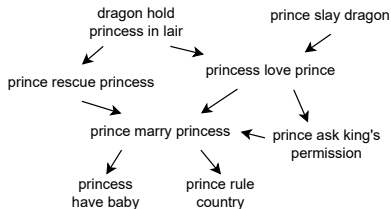


Figure: Plot graph for the sentence: *“the princess loves the prince”* from the planning-based genetic model by McIntyre and Lapata (2010).

Language Models (LMs)

In recent years, **neural networks** → heavy dominance of a specific type of ML models: **language models (LMs)**.

Definition (Language Model)

A language model is a **probability distribution over sequences of tokens**.

Given a vocabulary V of a set of tokens, a language model p assigns to each sequence of tokens $X = (x_1, \dots, x_L) \in V^L$ a probability $p(X) \in [0, 1]$.

Language Models (LMs)

In recent years, **neural networks** → heavy dominance of a specific type of ML models: **language models (LMs)**.

Definition (Language Model)

A language model is a **probability distribution over sequences of tokens**.

Given a vocabulary V of a set of tokens, a language model p assigns to each sequence of tokens $X = (x_1, \dots, x_L) \in V^L$ a probability $p(X) \in [0, 1]$.

In this thesis, we will only consider language models.

Language Models (LMs)

- ▶ Most language models are based on the **transformer architecture** (Vaswani et al., 2017).

Language Models (LMs)

- ▶ Most language models are based on the **transformer architecture** (Vaswani et al., 2017).
- ▶ **Decoder language models** (e.g. GPT), which use the *decoder* part of the transformer, specialize in text generation.

Language Models (LMs)

- ▶ Most language models are based on the **transformer architecture** (Vaswani et al., 2017).
- ▶ **Decoder language models** (e.g. GPT), which use the *decoder* part of the transformer, specialize in text generation.
- ▶ Most existing **large language models** (LLMs) use a decoder architecture, with some exceptions (e.g. Mamba).

Language Models (LMs)

- ▶ Most language models are based on the **transformer architecture** (Vaswani et al., 2017).
- ▶ **Decoder language models** (e.g. GPT), which use the *decoder* part of the transformer, specialize in text generation.
- ▶ Most existing **large language models** (LLMs) use a decoder architecture, with some exceptions (e.g. Mamba).
- ▶ For the definition of LLM, we follow Zhao et al. (2023): we refer to older transformer-based LMs (e.g. BERT, GPT-2) as “pretrained language models” (PLMs), and “LLMs” refer to GPT-3 and more recent models.

Language Models (LMs)

- ▶ Most language models are based on the **transformer architecture** (Vaswani et al., 2017).
- ▶ **Decoder language models** (e.g. GPT), which use the *decoder* part of the transformer, specialize in text generation.
- ▶ Most existing **large language models** (LLMs) use a decoder architecture, with some exceptions (e.g. Mamba).
- ▶ For the definition of LLM, we follow Zhao et al. (2023): we refer to older transformer-based LMs (e.g. BERT, GPT-2) as “pretrained language models” (PLMs), and “LLMs” refer to GPT-3 and more recent models.
- ▶ The public release of ChatGPT (based on GPT-3) in late 2022 marked a definite shift in NLP research, namely due to its impressive conversational abilities.

Outline

Introduction and Background

Methodology for Story Evaluation and ASG Meta-Evaluation

HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation

Meta-Evaluation Benchmark of Automatic Story Generation

Exploring LLM Explainability for ASE and ASG

Conclusion

Relevant Questions

RQ1: How can we evaluate the quality of a story?

- ▶ What setting do we choose?
- ▶ Which criteria can be used to evaluate a story?
- ▶ What are the available automatic evaluation measures?
- ▶ How can we compare human and automatic evaluation methods?

Definitions

Definition (Automatic Story Generation (ASG))

Given a short sentence p_i (called a **story-prompt**), the ASG task consists in using a language model that takes p_i as its input for generating a story y_i .

Definition (Automatic Story Evaluation (ASE))

Given an automatic evaluation measure m (e.g. a scoring algorithm, an LLM...), a story-prompt p_i , a candidate story y_i , and optional additional text z_i (e.g. a reference story), we define the ASE task as the production of an evaluation score $m(p_i, y_i, z_i)$, which we may write $m(y_i)$ for brevity.

ASG Example

- ▶ **Story-prompt:** You swear that you keep seeing the same woman in blue every time you go on family vacation, when you turn 18, you are introduced to her.

ASG Example

- ▶ **Story-prompt:** You swear that you keep seeing the same woman in blue every time you go on family vacation, when you turn 18, you are introduced to her.
- ▶ **Story generated by Platypus2-70B:** As a child, I always looked forward to going on family vacation. It was a time to unwind, enjoy nature, and most importantly, spend time with my loved ones. Yet, something strange always happened whenever we went on these trips. No matter where we went or what we did, I would inexplicably keep seeing the same woman in blue, almost as if she was following us. This mystery woman always wore a light blue summer dress and had her long dark hair styled into a loose braid. [...]

Human Criteria for Story Evaluation

- In the ASG literature, many criteria have been used:

Human Criteria for Story Evaluation

- ▶ In the ASG literature, many criteria have been used:
- ▶ Coherence, Interestingness, Grammar, Sentence Preference, Faithfulness, Relevance / Fidelity, Fluency, Logicality, Overall Preference, Likability, Overall Quality, Outline Utilisation, Narrative Flow, Ordering, Emotion Faithfulness, Content Quality, Human Plausibility, Ending Quality, Naturalness, etc.

Human Criteria for Story Evaluation

- ▶ In the ASG literature, many criteria have been used:
- ▶ Coherence, Interestingness, Grammar, Sentence Preference, Faithfulness, Relevance / Fidelity, Fluency, Logicality, Overall Preference, Likability, Overall Quality, Outline Utilisation, Narrative Flow, Ordering, Emotion Faithfulness, Content Quality, Human Plausibility, Ending Quality, Naturalness, etc.
- ▶ Many criteria **overlap** (Coherence / Logicality, Overall Quality / Content Quality...);

Human Criteria for Story Evaluation

- ▶ In the ASG literature, many criteria have been used:
- ▶ Coherence, Interestingness, Grammar, Sentence Preference, Faithfulness, Relevance / Fidelity, Fluency, Logicality, Overall Preference, Likability, Overall Quality, Outline Utilisation, Narrative Flow, Ordering, Emotion Faithfulness, Content Quality, Human Plausibility, Ending Quality, Naturalness, etc.
- ▶ Many criteria **overlap** (Coherence / Logicality, Overall Quality / Content Quality...);
- ▶ Some of them are **not specific to ASG** (Grammar, Fluency...).

Human Criteria for Story Evaluation

We surveyed the ASG and social sciences literature to design the following set of evaluation criteria:

Human Criteria for Story Evaluation

We surveyed the ASG and social sciences literature to design the following set of evaluation criteria:

1. **Relevance (RE):** “how well the story matches its story-prompt”;

Human Criteria for Story Evaluation

We surveyed the ASG and social sciences literature to design the following set of evaluation criteria:

1. **Relevance (RE):** “how well the story matches its story-prompt”;
2. **Coherence (CH):** “how much the story makes sense”;

Human Criteria for Story Evaluation

We surveyed the ASG and social sciences literature to design the following set of evaluation criteria:

1. **Relevance (RE):** “how well the story matches its story-prompt”;
2. **Coherence (CH):** “how much the story makes sense”;
3. **Empathy (EM):** “how well the reader understood the character’s emotions”;

Human Criteria for Story Evaluation

We surveyed the ASG and social sciences literature to design the following set of evaluation criteria:

1. **Relevance (RE):** “how well the story matches its story-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”;

Human Criteria for Story Evaluation

We surveyed the ASG and social sciences literature to design the following set of evaluation criteria:

1. **Relevance (RE)**: “how well the story matches its story-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”;

Human Criteria for Story Evaluation

We surveyed the ASG and social sciences literature to design the following set of evaluation criteria:

1. **Relevance (RE):** “how well the story matches its story-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”.

Taxonomy of Evaluation Measures

We propose a bi-dimensional taxonomy of automatic measures:

Taxonomy of Evaluation Measures

We propose a bi-dimensional taxonomy of automatic measures:

- ▶ **Reference-based:** it compares a candidate story with a reference story written for the same prompt;
- ▶ **Reference-free:** it relies only on the candidate story (and, possibly, the prompt).

Taxonomy of Evaluation Measures

We propose a bi-dimensional taxonomy of automatic measures:

- ▶ **Reference-based:** it compares a candidate story with a reference story written for the same prompt;
- ▶ **Reference-free:** it relies only on the candidate story (and, possibly, the prompt).

and

- ▶ **String-based:** it evaluates the textual representation of the inputs;
- ▶ **Embedding-based:** it relies on word embeddings or contextualized embeddings;
- ▶ **Model-based:** it leverages regression or language models to return a score.

Taxonomy of Evaluation Measures

	Reference-based	Reference-free
String-based	BLEU ROUGE METEOR chrF CIDEr	Coverage Density Compression Text length Novelty Repetition
Embedding-based	ROUGE-WE BERTScore MoverScore BaryScore DepthScore	SUPERT
Model-based	S3 SummaQA InfoLM BARTScore	BLANC BARTScore

Table: Taxonomy of the automatic measures considered in our study.

Meta-Evaluation Framework

Objective: comparing evaluation methods between one another.
Especially, we would like to ascertain **how well automatic measures correlate with human evaluation**, which we will use as our gold standard.

Notations

For S systems and N story-prompts, let y_i^j be the story generated by system $j \in \{1, \dots, S\}$ for story-prompt $i \in \{1, \dots, N\}$.
For a (human or automatic) measure m , we note $m(y_i^j)$ the score associated to y_i^j .

Kendall's Correlation Coefficient

We follow Gilpin (1993); Newson (2002) and choose to use Kendall's τ coefficient, a statistic used to measure the ordinal association between two measured quantities.

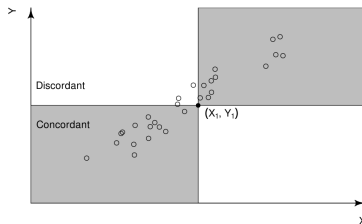


Figure: Gray and white areas: respectively concordant and discordant points w.r.t. (X_1, Y_1) . 395 concordant pairs and 40 discordant pairs $\rightarrow \tau = 0.816$.

Meta-Evaluation Strategies

Definition (System-level Correlation ($K_{m_1, m_2}^{\text{sys}}$))

This strategy 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 measures.

Definition (Overall Correlation ($K_{m_1, m_2}^{\text{ovl}}$))

This strategy measures **how** m_1 and m_2 **agree at the level of the story itself**. It computes the correlation between the full vectors containing the scores of m_1 or m_2 for a given story for every system.

Statistical Testing

- ▶ To evaluate the statistical significance of observed increases in dependent correlations, we used the **Williams test** (Graham and Baldwin, 2014).
- ▶ To correct p -values for multiplicity, we used the **Benjamini-Hochberg procedure** (Benjamini and Hochberg, 1995).

Summary

RQ1: How can we evaluate the quality of a story?

- ▶ We defined our ASG and ASE settings;
- ▶ We designed 6 criteria for human evaluation: **Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity**;
- ▶ We proposed a taxonomy of automatic evaluation measures;
- ▶ We defined specific meta-evaluation strategies (system-level and overall correlations) and statistical testing procedures (Williams Test and Benjamini-Hochberg method).

Outline

Introduction and Background

Methodology for Story Evaluation and ASG Meta-Evaluation

HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation

Meta-Evaluation Benchmark of Automatic Story Generation

Exploring LLM Explainability for ASE and ASG

Conclusion

Relevant Questions

RQ2: How good are existing models at generating stories?

- ▶ Is there a dataset with stories generated by multiple systems and annotated with both human and automatic evaluation measures?
- ▶ If not, how do we build it?
- ▶ Which protocol do we use to manually annotate stories?
- ▶ Which protocol do we use to annotate stories using LLMs?
- ▶ How do generated stories perform compared with human stories?

Existing Story Generation Corpora

Name	Type	Annotations	Avg. Words
ROCStories	Title + Story	✗	80
SIND	Pictures + Story	✗	80
WritingPrompts	Prompt + Story	✗	750
RolePlayerGuild	RPG Thread	✗	3,000
PG-19	Book	✗	69,000
Storium	Collaborative Story	≈	19,000
OpenMEVA	Title/Prompt + Story	≈	400

Table: Overview of existing story generation corpora. No corpus provides annotations on different criteria of story quality.

Existing Story Generation Corpora

Name	Type	Annotations	Avg. Words
ROCStories	Title + Story	✗	80
SIND	Pictures + Story	✗	80
WritingPrompts	Prompt + Story	✗	750
RolePlayerGuild	RPG Thread	✗	3,000
PG-19	Book	✗	69,000
Storium	Collaborative Story	≈	19,000
OpenMEVA	Title/Prompt + Story	≈	400

Table: Overview of existing story generation corpora. No corpus provides annotations on different criteria of story quality.

Building Our Corpus

We collected the aligned outputs on 96 story-prompts from the **WritingPrompts** dataset from 10 language models:

1. 3 ASG-specific systems: Fusion, TD-VAE, and HINT;
2. 7 pretrained language models fine-tuned on WritingPrompts: BERTGeneration (BERTGen), CTRL, RoBERTa, XLNet, GPT, GPT-2, and GPT-2 (tag).

Each story-prompt also comes with a human story. Therefore, we gathered $11 \times 96 = \mathbf{1,056}$ stories in total.

Annotation Campaign

We ran an annotation campaign on Amazon Mechanical Turk, asking human workers to rate our stories w.r.t. our human criteria on a 1 to 5 Likert scale. Each story was rated by three distinct annotators.

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 (e.g. 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.

Table: Guidelines for the Empathy criterion.

HANNA

Story-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. [...]

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 [...]

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. [...]

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

Metric	Human	Story #1	Story #2
BLEU (%)	1.00	0.01	0.01
ROUGE-1	1.00	0.24	0.33
BERTScore	1.00	0.50	0.52
BARTScore	-0.98	-3.97	-4.03
SUPERT	0.94	0.37	0.36

Table: Example story-prompt, human and generated stories from HANNA with human annotations and measure scores. The dataset is available at <https://github.com/dig-team/hanna-benchmark-asg>.

Evaluating Our Human Criteria

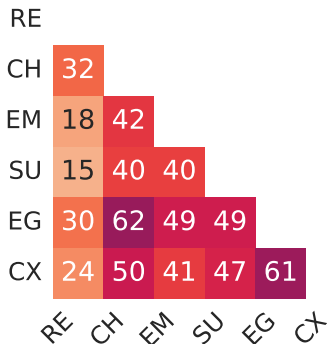


Figure: Overall absolute Kendall correlations (×100) between human criteria.

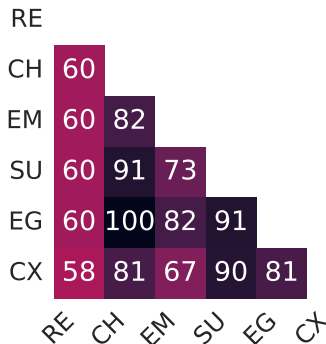
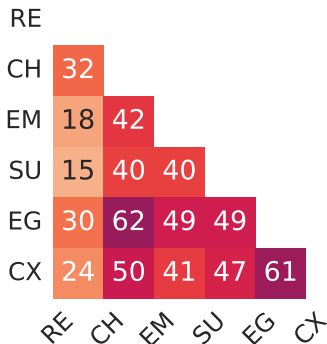


Figure: System-level absolute Kendall correlations (×100) between human criteria.

Evaluating Our Human Criteria



Moderate to weak overall correlations: our criteria evaluate **distinct aspects of storytelling** which cannot be re-grouped in fewer criteria.

Figure: Overall absolute Kendall correlations ($\times 100$) between human criteria.

Evaluating Our Human Criteria

Higher system-level correlations: a given system tends to be **uniformly better or worse** than other systems across all criteria.

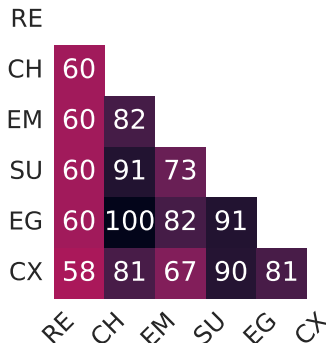


Figure: System-level absolute Kendall correlations ($\times 100$) between human criteria.

Performance of ASG Systems

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

Table: Average system ratings per criterion with 95% confidence interval. Higher is better.

Human stories are rated much more highly than generated stories by human annotators. GPT-2 is the best system overall.

Adding Large Language Models

We use LLMs to produce new stories for HANNA.

We perform several annotation experiments: we ask LLMs to rate stories w.r.t. to our criteria with different Eval-Prompts (*i.e.*, the prompt that is given as input to the LLM).

We produce:

- ▶ **ASE**: ~150k rating and explanation annotations using Llama models (Beluga-13B, Llama-13B, Mistral-7B) and ChatGPT;
- ▶ **ASG**: 480 stories generated by Llama models (Platypus2-70B, Llama-30B, Beluga-13B, Mistral-7B) with corresponding LLM annotations (excluding ChatGPT) to expand the HANNA corpus.

ASE Methodology with LLMs

We first provide the model with a story-prompt and a matching story. Then, we use four different Eval-Prompts:

- ▶ **Eval-Prompt 1** (simple rating): we ask the model to rate the story on a scale from 1 to 5 on one of our six criteria;
- ▶ **Eval-Prompt 2** (rating with explanation): Eval-Prompt 1 + we ask the model to explain its answer;
- ▶ **Eval-Prompt 3** (rating with explanation and guidelines): Eval-Prompt 2 + the detailed guidelines from our original human annotation protocol;
- ▶ **Eval-Prompt 4** (rating with explanation and human story): Eval-Prompt 2 + the human story associated with the same story-prompt. We explicitly tell the model that the human story is only given for reference purposes.

ASE Experiments

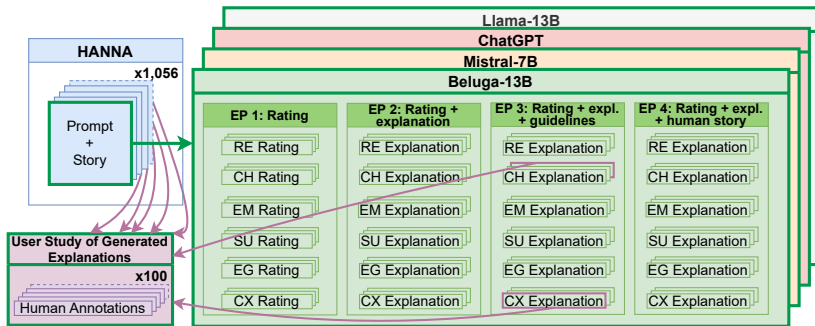


Figure: Schema of the performed ASE experiments. “EP” means “Eval-Prompt”.

LLM Performance at ASG

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

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

Larger models (Platypus-70B, Llama-30B) are more highly rated by Beluga-13B. However, we need to confirm that LLMs are reliable proxies for human evaluation.

Summary

RQ2: How good are existing models at generating stories?

- ▶ We built HANNA, a corpus containing 1,536 stories generated by 15 different systems (1 human, 3 ASG-specific, 7 pretrained LMs, 4 LLMs);
- ▶ All non-LLM stories were rated by 3 human annotators w.r.t. our 6 criteria;
- ▶ Non-LLM models are noticeably below human performance according to human raters;
- ▶ All stories were rated by 3 or 4 different LLMs with 4 different Eval-Prompts;
- ▶ LLMs seem to perform as well as human writers for this specific setting, according to Beluga-13B ratings.

Outline

Introduction and Background

Methodology for Story Evaluation and ASG Meta-Evaluation

HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation

Meta-Evaluation Benchmark of Automatic Story Generation

Exploring LLM Explainability for ASE and ASG

Conclusion

Relevant Questions

RQ3: To which extent can we use automatic measures for story evaluation?

- ▶ How consistent are LLMs in their ratings?
- ▶ How do automatic measures correlate with human judgment?
- ▶ How do LLMs compare with non-LLM automatic measures?
- ▶ How does the Eval-Prompt influence LLM behaviour?

Automatic Annotation Consistency

Criterion	Beluga-13B	Mistral-7B	Human
Relevance	0.88 ± 0.01	0.86 ± 0.01	0.48 ± 0.30
Coherence	0.93 ± 0.01	0.90 ± 0.01	0.29 ± 0.28
Empathy	0.88 ± 0.01	0.87 ± 0.02	0.34 ± 0.09
Surprise	0.80 ± 0.02	0.63 ± 0.03	0.28 ± 0.12
Engagement	0.91 ± 0.01	0.87 ± 0.01	0.46 ± 0.12
Complexity	0.85 ± 0.01	0.78 ± 0.02	0.56 ± 0.08

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

LLM consistency and human inter-rater agreement are not directly comparable, but we can still observe that LLMs display very high consistency overall.

Correlations with Human Judgment

RE	49	21	16	22	15	3	13	7	11	10	12	26	7
CH	37	26	18	22	22	3	14	11	16	17	15	3	0
EM	49	27	15	20	20	11	17	13	17	17	17	2	2
SU	44	17	12	13	5	4	12	11	16	17	13	3	0
EG	50	26	11	21	19	8	19	14	19	20	19	5	2
CX	57	32	26	23	27	8	24	19	28	29	24	6	1
Avg	48	25	16	20	18	6	16	12	18	18	17	8	2
Human													
Beluga-13B 1													
Llama-13B 1													
Mistral-7B 1													
ChatGPT 1													
BARTScore													
BERTScore													
BLEU													
ROUGE-1													
chrF													
BaryScore													
SUPERT													
BLANC													

Figure: Overall absolute Kendall correlations ($\times 100$) between evaluation measures and human ratings. Higher is better.

LLMs perform at least as well as other automatic measures, but correlations remain generally low. Fine-tuning and model size seem to improve performance as Beluga-13B has the highest correlations.

Correlations with Human Judgment

RE	70	49	64	36	7	51	51	56	51	60	56	11	24
CH	62	78	87	60	73	56	56	33	38	47	51	20	16
EM	77	73	54	60	56	56	73	42	47	47	69	2	2
SU	72	73	56	51	7	56	56	42	47	56	51	20	16
EG	76	73	73	64	64	56	56	33	38	47	51	20	16
CX	80	72	54	63	72	67	49	54	58	67	45	18	4
Avg	73	70	65	56	46	57	57	43	46	54	54	15	13
Human		Beluga-13B 1	Llama-13B 1	Mistral-7B 1	ChatGPT 1	BARTScore	BERTScore	BLEU	ROUGE-1	chrF	BaryScore	SUPERT	BLANC

Figure: System-level absolute Kendall correlations ($\times 100$) between evaluation measures and human ratings. Higher is better.

Here, LLMs outperform automatic measures. Beluga-13B and Mistral-7B display especially high correlations.

Influence of the Eval-Prompt on Consistency

Criterion	EP 1	EP 2	EP 3	EP 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: Intra-class coefficients type 2k for Beluga-13B ratings with 95% confidence interval. Higher is better.

Providing guidelines (Eval-Prompt 3) appears to slightly decrease consistency with a discernible effect, but ICC values remain very high.

Influence of the Eval-Prompt on Ratings

LLM	EP 1	EP 2	EP 3	EP 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

Table: Average Likert ratings per LLM per Eval-Prompt. Higher is better.

Asking for an explanation (Eval-Prompt 2) has limited influence on ratings, but more detailed Eval-Prompts (3 and 4) tend to decrease the ratings with a statistically discernible effect.

Influence of the Eval-Prompt on Correlations

RE	21	24	19	21	16	16	6	11	22	17	17	23	15	14	11	11
CH	26	27	25	23	18	21	16	13	22	22	15	17	22	22	14	21
EM	27	27	25	19	15	15	9	11	20	22	17	13	20	17	1	12
SU	17	19	16	17	12	15	7	7	13	12	11	12	5	8	7	9
EG	26	26	22	17	11	14	5	7	21	21	15	15	19	19	6	15
CX	32	32	25	25	26	28	17	18	23	25	23	22	27	21	12	25
Avg	25	26	22	20	16	18	10	11	20	20	16	17	18	17	8	16
	B-13B 1	B-13B 2	B-13B 3	B-13B 4	L-13B 1	L-13B 2	L-13B 3	L-13B 4	M-7B 1	M-7B 2	M-7B 3	M-7B 4	Chat 1	Chat 2	Chat 3	Chat 4

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

Providing guidelines or a human story (Eval-Prompts 3 and 4) tends to decrease correlations for all models, surprisingly.

Influence of the Eval-Prompt on Correlations

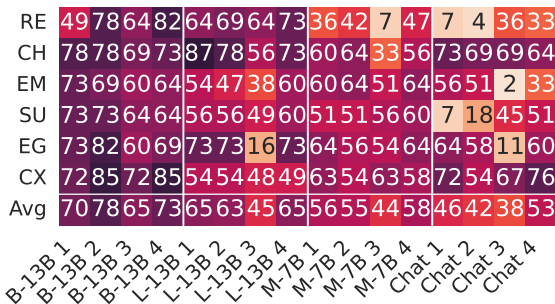


Figure: System-level absolute Kendall correlations ($\times 100$) between LLMs and human ratings for different Eval-Prompts. Higher is better. B-13B = Beluga-13B, L-13B = Llama-13B, M-7B = Mistral-7B, Chat = ChatGPT.

Eval-Prompt 3 decreases correlations again, but Eval-Prompt 4 tends to increase them.

Summary

RQ3: To which extent can we use automatic measures for story evaluation?

- ▶ We performed **an extensive meta-evaluation**, notably comparing correlations between automatic measures (including LLMs) and human judgment;
- ▶ Used with prompts based on specific criteria, **LLMs are currently the best proxy for human evaluation of story generation**. In particular, LLMs display very high system-level correlations with human judgment;
- ▶ **LLMs are remarkably self-consistent**, exhibiting very high intra-class coefficient values;
- ▶ For ASE, **providing detailed guidelines (Eval-Prompt 3) did not improve correlations with human ratings**. Providing a reference human story (Eval-Prompt 4) yields mixed results;

Outline

Introduction and Background

Methodology for Story Evaluation and ASG Meta-Evaluation

HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation

Meta-Evaluation Benchmark of Automatic Story Generation

Exploring LLM Explainability for ASE and ASG

Conclusion

Relevant Questions

RQ4: How explainable are the evaluation ratings of LLMs?

- ▶ How specific to the given criteria are the explanations provided by LLMs?
- ▶ What sort of issues do LLM explanations display?
- ▶ How well can LLMs be thought to understand the ASE task?
- ▶ Can studying LLM pretraining data help explain their ASG performance?

Clustering of Explanation Embeddings

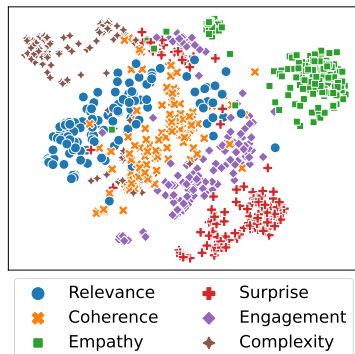


Figure: UMAP projection of Beluga-13B explanations.

LLM explanations are overall well-separated w.r.t. their corresponding criteria.

Keyword Analysis

Crit.	Keywords
RE	story, prompt, roughly matches, target, weak relationship, connection, weak, focuses, unrelated, human story provided, idea, difficult, writing
CH	story, coherence, make sense, difficult to understand, clear narrative structure, follow, making it difficult, rate, context, jumps, understandable plot
EM	empathy, emotions, understand the characters, depth, emotional connection, clear, feelings, context, fully, level, specific, thoughts, sadness, recognize
SU	story, surprise, ending, predictable, rate, unexpected, twist, completely obvious, human, plot, abruptly, resolution, half, context, offer, hints
EG	story, mildly interesting, engagement, difficult, found, characters, fully engage, clear plot, clear narrative, unique, felt disjointed, protagonist
CX	story, characters, intricate plot, difficult to understand, straightforward, depth, simple, extremely simple, involves, development, details

Table: Selected keywords from Beluga-13B explanations w.r.t. a specific criterion. Keywords are semantically relevant to the criterion.

User Study on LLM Explanations

We ask human raters to identify issues in 100 randomly sampled LLM Eval-Prompt 3 explanations. We distinguish 5 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 is not faithful to the predicted rating according to 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.

User Study on LLM Explanations

Error Type	Rate	AC1
No Explanation*	0.40	—
Poor Syntax	0.02	0.97 ± 0.03
Incoherence	0.11	0.81 ± 0.08
Wrong Guideline	0.13	0.90 ± 0.06
Superfluous Text	0.20	0.66 ± 0.12
Unsubstantiated Claims	0.31	0.60 ± 0.14

Table: Error rates of Beluga-13B Eval-Prompt 3 on a sample of 100 explanations. Lower is better. The asterisk signals that all 1,056 Eval-Prompt 3 annotations were considered.

High rate of “Unsubstantiated Claims”, and **40% of all Eval-Prompt 3 ratings did not even have an explanation.**

Influence of Pretraining Data

- ▶ We use the Min-K% Prob detection method (Shi et al., 2024), based on the hypothesis that unseen data will contain more outlier words with low probability than seen data.
- ▶ We showed that **it is easier to detect if a book was in the training data of a larger LLM**, and that larger LLMs tend to produce text that is **more faithful to their training data**.
- ▶ \hookrightarrow This could explain the better ASG performance of larger LLMs.

Summary

RQ4: How explainable are the evaluation ratings of LLMs?

- ▶ We performed different experiments, including a user study on LLM explanations and an estimation of the influence of pretraining data on LLM performance;
- ▶ **LLMs understand the ASE task only partially:** while they provide explanations that are specific to the evaluated criteria, they struggle to explain their answers with substantiated claims;
- ▶ **Pretraining data helps explain LLM performance at ASG:** the higher ratings of larger LLMs may be due to their ability to produce output similar to existing books.

Outline

Introduction and Background

Methodology for Story Evaluation and ASG Meta-Evaluation

HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation

Meta-Evaluation Benchmark of Automatic Story Generation

Exploring LLM Explainability for ASE and ASG

Conclusion

Main Contributions and Findings (1/2)

RQ1: How can we evaluate the quality of a story?

- ▶ 6 criteria for human evaluation: **Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity;**

RQ2: How good are existing models at generating stories?

- ▶ **HANNA**, a corpus containing **1,536 stories** generated by **15 different systems** (1 human, 3 ASG-specific, 7 pretrained LMs, 4 LLMs);
- ▶ All non-LLM stories were rated w.r.t. our 6 criteria by 3 human annotators and 4 different LLMs with 4 different Eval-Prompts;
- ▶ **LLMs seem to perform as well as human writers for this specific setting**, according to Beluga-13B ratings.

Main Contributions and Findings (2/2)

RQ3: To which extent can we use automatic measures for story evaluation?

- ▶ Used with prompts based on specific criteria, **LLMs are currently the best proxy for human evaluation of story generation.**

RQ4: How explainable are the evaluation ratings of LLMs?

- ▶ **LLMs understand the ASE task only partially:** notably they struggle to explain their answers with substantiated claims;
- ▶ **Pretraining data helps explain LLM performance at ASG:** larger LLMs seem to produce output that is more similar to existing books.

Limitations and Future Perspectives

- ▶ Manually annotating stories is an arduous task. We recruited novices, but **expert annotation may yield different results**;
- ▶ Our different Eval-Prompts were fairly basic: **more complex Eval-prompts may yield more interesting results**, e.g. with Chain-of-Thought prompting;
- ▶ As the LLM scene is changing ever so rapidly, **our results may already be partially outdated**;
- ▶ We would have liked to **design specific ASE measures**, ideally measures that would highly correlate with our criteria;
- ▶ We believe **further exploration of LLM explainability is crucially needed**;
- ▶ More generally, we believe **the societal and environmental risks posed by LLMs should be more broadly discussed**.

Publications

1. **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 (COLING)*, pages 5794–5836.
<https://aclanthology.org/2022.coling-1.509/>
2. **Cyril Chhun**, Fabian M. Suchanek, Chloé Clavel. 2024. Do Language Models Enjoy Their Own Stories? Prompting Large Language Models for Automatic Story Evaluation. *Transactions of the Association for Computational Linguistics (TACL)*, volume 12, pages 1122–1142.
https://doi.org/10.1162/tacl_a_00689

Recap Slide. Ready for the Questions!

- ▶ 6 criteria: **Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity**;
- ▶ **HANNA corpus**: 1,056 stories annotated by 3 human raters and 4 LLMs + 480 LLM stories annotated by 3 LLMs;
- ▶ Meta-evaluation benchmark: **LLMs are currently the best automatic proxy for human judgment**. While system-level correlations are satisfactory, overall correlations remain weak;
- ▶ Experiments on LLM explainability: **LLMs often fail to provide explanations** despite being explicitly asked; when they provide one, they are usually specific but **do not always substantiate their claims**.
- ▶ Future directions: more refined annotation protocols, specific ASE measures, further investigation of LLM explainability and societal risks.



Outline

Appendix

Meta-Evaluation Strategies

Definition (System-level Correlation ($K_{m_1, m_2}^{\text{sys}}$))

This strategy 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 measures. Formally:

$$K_{m_1, m_2}^{\text{sys}} := K \left(\frac{1}{N} C_{m_1}^{\text{sys}}, \frac{1}{N} C_{m_2}^{\text{sys}} \right), \quad (1)$$

$$\text{where } C_m^{\text{sys}} := \left[\sum_{i=1}^N m(y_i^1), \dots, \sum_{i=1}^N m(y_i^S) \right].$$

Meta-Evaluation Strategies

Definition (Overall Correlation ($K_{m_1, m_2}^{\text{ovl}}$))

This strategy measures **how m_1 and m_2 agree at the level of the story itself**. It computes the correlation between the full vectors containing the scores of m_1 or m_2 for a given story for every system. Formally:

$$K_{m_1, m_2}^{\text{ovl}} := K(C_{m_1}^{\text{ovl}}, C_{m_2}^{\text{ovl}}), \quad (2)$$

$$\text{where } C_m^{\text{ovl}} := \left[\left(m(y_i^j) \right)_{(i,j) \in \{1, \dots, N\} \times \{1, \dots, S\}} \right].$$

Statistical Testing

Correlations between two measures on the same dataset are not independent. We use the Williams test to evaluate the strength of an **increase** in dependent correlations.

Definition (Williams Test)

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

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

Statistical Testing

Since we perform a large quantity of tests, we need to correct p -values for multiplicity. We choose to control the false discovery rate using the Benjamini-Hochberg method.

Definition (Benjamini-Hochberg (BH) Method)

Given m p -values p_1, \dots, p_m sorted in increasing order and a significance level α , the Benjamini-Hochberg method consists in finding the largest k such that $p_k \leq \frac{k}{m}\alpha$. The null hypothesis would then be rejected for the first k tests. This is equivalent to computing adjusted p -values $p_k^* = p_k \frac{m}{k}$ and replacing the p -values from largest to smallest.

Statistical Testing

		Overall										
Beluga-13B 1	RE	14		9	0	5	0	1	1	3		0
	CH	4	17	14	0	1	0	1	1	1	0	0
	EM	0	4	6	0	1	0	1	1	1	0	0
	SU	14	17	1	0	15	8	41		16	0	0
	EG	0	12	7	0	6	0	8	9	7	0	0
	CX	7	2	9	0	4	0	14	22	3	0	0
		Llama-13B 1	Mistral-7B 1	ChatGPT 1	BARTScore	BERTScore	BLEU	ROUGE-1	chrF	BaryScore	SUPERT	BLANC

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

Moderate to strong statistical evidence that Beluga-13B correlates better with human judgment than other measures.

Statistical Testing

		System-level										
Beluga-13B 1	RE		34	30							23	31
	CH		34	40	26	24	16	19	20	25	13	11
	EM	24	34	38	26		16	19	20	41	13	11
	SU	24	34	20	26	24	16	19	23	25	13	11
	EG		34	40	26	24	16	19	20	25	13	11
	CX	24	34		41	24	24	29	41	25	13	11
		Llama-13B 1	Mistral-7B 1	ChatGPT 1	BARTScore	BERTScore	BLEU	ROUGE-1	chrF	BaryScore	SUPERT	BLANC

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

Weaker statistical evidence at the system-level, mitigated by the averaged nature of the correlations.

Influence of Pretraining Data

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. Given a sequence of tokens $x = x_1, \dots, x_N$ and an LLM's probability distribution p of the next token, Min-K% Prob selects the top- $k\%$ of tokens with the highest negative log-likelihood to form a set $\text{Min-K\%}(x)$ and computes their average log-likelihood. Formally:

$$\text{Min-K\% Prob}(x) := \frac{1}{E} \sum_{x_i \in \text{Min-K\%}(x)} \log p(x_i | x_1, \dots, x_{i-1}),$$

where E is the size of the $\text{Min-K\%}(x)$ set. 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.

Influence of Pretraining Data

We use the Min-K% Prob detection method to verify whether the LLMs were trained on the WritingPrompts dataset.

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

Table: Predicted contamination rates of a random WritingPrompts sample of 1,000 stories.

The low predicted rates suggest this was not the case.

Influence of Pretraining Data

We use the BooksMIA dataset to compute the area under the ROC curve (AUC) obtained with Min-K% Prob thresholding.

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

Table: AUC detection score on the BooksMIA dataset

The higher AUC detection score for larger models means that **it is easier to detect if a book was in the training data of a larger LLM**, and that larger LLMs tend to produce text that is **more faithful to their training data**. This could explain their better ASG performance.

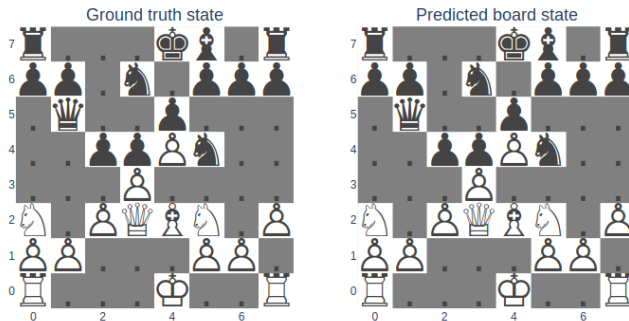
Environmental Impact

- ▶ We used the Jean Zay supercomputer;
- ▶ A100 GPU nodes: 1,431 hours consumed;
- ▶ V100 GPU nodes: 203 hours consumed;
- ▶ Estimation balance sheet: 0.078 tons of CO2 equivalent, according to the Labos 1point5 estimation method¹;
- ▶ \approx a one-way Paris-Marseilles flight.

¹<https://labos1point5.org/les-rapports/estimation-empreinte-calcul>

On LLM Intelligence: Playing Chess

Interesting work from Karvonen (2024): “we provide evidence that an LLM trained on a next token prediction task can develop a **world model of complex systems such as chess**, including the ability to estimate latent variables such as player skill.” Their best linear probe classifier achieves **99.6% accuracy** in classifying the precise state of each square across 10,000 test games.





Quality Criteria for Evaluations Taxonomy

Belz et al. (2024) attempt to standardize quality criteria in NLP with an interactive tool:

1. First node of the taxonomy: Overall Quality;
2. Correctness / Goodness / Features;
3. In Their Own Right / Relative to the Inputs / Relative to an External Frame of Reference;
4. Form / Content / Form and Content;

But the tool was unavailable online...?

Manual Annotation

Lee et al. (2019) advise to:

- ▶ use quantitative analysis “if the goal is to judge the merit of the system”;
- ▶ use “either multiple-item 7-point Likert scales, or a (continuous) ranking task”;
- ▶ “choose a sample that reflects the audience for which the system was developed”.

Karpinska et al. (2021) show that “AMT worker judgments improve when they are shown model-generated output alongside human-generated references, which enables the workers to better calibrate their ratings”.

Initial PhD Topic (April 2021)

Recent advances in natural language processing have enabled language models to produce ever more realistic texts. However, automatically generated texts exhibit poor consistency; we argue that this is due to the lack of a strong representation of the underlying state of the story. This thesis will study the **extraction and representation of a meaningful chain of events** from a given text, and, reciprocally, the **generation of a convincing story from such a representation**. We will also study the **controllability of the generated text**; for instance by enforcing formal constraints such as the respect of poetic conventions.

- Arwa I. Alhussain and Aqil M. Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Computing Surveys*, 54(5).
- Anya Belz, Simon Mille, Craig Thomson, and Rudali Huidrom. 2024. QCET: An interactive taxonomy of quality criteria for comparable and repeatable evaluation of NLP systems. In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 9–12, Tokyo, Japan. Association for Computational Linguistics.
- 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.
- Andrew R. Gilpin. 1993. Table for conversion of Kendall’s tau to Spearman’s rho within the context of measures of magnitude of effect for meta-analysis. *Educational and Psychological Measurement*, 53(1):87–92.

- 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.
- 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.
- Adam Karvonen. 2024. Emergent world models and latent variable estimation in chess-playing language models. In *Proceedings of the 1st Conference on Language Modeling*.
- 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.

Neil McIntyre and Mirella Lapata. 2010. Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1562–1572, Uppsala, Sweden. Association for Computational Linguistics.

Roger Newson. 2002. Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences. *The Stata Journal*, 2(1):45–64.

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)*, Vienna, Austria. OpenReview.net.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion

Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, Long Beach, CA, USA.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *ArXiv Preprint*, abs/2303.18223v13.