

Gauging Large Language Models for Generative Story Evaluation

IT 469

Group 5

[Total: 8 points]

Team members

1. Deem Alshaye 443200583
2. Norah Alwohaibi 443200753
3. Ghaina Alhassnan, 443200495
4. Raghad Alharkan, 443200477
5. Shaug Alyahya, 443200888

Table of Contents

1. Introduction:	3
2. Experiment setup:	3
2.1 Dataset:	4
2.2 Methodology	6
3. Evaluation and results	8
4. Discussion	9
5. Conclusion	12
6. References:	13

Table of Tables

Table 1: Distribution of Human-Written and AI-Generated Stories in the HANNA Dataset	5
Table 2: corpus/topic or genera distribution	5
Table 3: Dataset Sample	6
Table 4: Model Details	7
Table 5: Evaluation Results	8

Table Of Figures

Figure 1 GPT-2	10
Figure 2 Mistral-7B	10
Figure 3 GPT-2	11
Figure 4 Mistral-7B	11
Figure 5 GPT-2	11
Figure 6 Mistral-7B	12

1. Introduction:

Large Language Models (LLMs) have recently demonstrated remarkable capabilities in various generative tasks, including story generation [1]. However, despite their success in producing coherent narratives, evaluating the quality of generated stories remains a complex challenge. Traditional evaluation methods often depend on human judgment, which can be subjective, time-consuming, and difficult to scale across large datasets.

This project focuses on exploring the potential of LLMs to self-evaluate the emotional quality, particularly empathy, in the stories they generate. Specifically, the project examines how different models (GPT-2 and Mistral-7B-Instruct-v0.2) perform in generating empathetic fictional stories and how their outputs align with an empathy prototype.

To ensure a consistent and fair evaluation, we utilize the HANNA Stories Annotations dataset [2], which contains 1,056 stories annotated by human raters across six narrative criteria: Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity. The evaluation process involves comparing the generated stories against an empathetic reference using sentence-BERT (SBERT) embeddings [3]. This project aims to provide an objective, automated alternative to traditional human-based story evaluation. The findings contribute to advancing research in automatic narrative assessment and enhancing the emotional intelligence of future LLM-based storytelling systems.

2. Experiment setup:

To evaluate the effectiveness of large language models in generating emotionally empathetic narratives, we designed a structured experimental setup that ensures fairness, reproducibility, and rigorous evaluation. The experiment consists of two main stages: first, generating fictional short stories using two different models (GPT-2 and Mistral-7B-Instruct-v0.2), and second, quantitatively evaluating the empathetic quality of these stories through semantic similarity measures. The HANNA Stories Annotations dataset is utilized as the standardized input source, enabling consistent prompt conditions across models. This section outlines the techniques, methods, and dataset characteristics that form the basis of the experimental framework.

2.1 Dataset

The HANNA dataset consists of 1,056 machine-generated stories created from 96 prompts selected from the WritingPrompts dataset, a popular corpus for narrative generation tasks.

Each story was evaluated by three human annotators on six qualitative criteria for a total of contains 19,008 annotations:

- **Relevance** (RE)
- **Coherence** (CH)
- **Empathy** (EM)
- **Surprise** (SU)
- **Engagement** (EG)
- **Complexity** (CX)

Motivation

The HANNA dataset was created to support the evaluation of stories generated by large language models using both human annotations and model-based assessments. It addresses a key gap in the field of automatic story generation (ASG) evaluation. HANNA enables systematic assessment of narrative quality across six core criteria (Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity) as stated in the dataset's description. Additionally, each story is annotated not only by human raters but also by multiple language models and evaluated with 72 automatic metrics, providing a rich foundation for comparing human and machine-based story evaluations.

Funding Information

This work was performed using HPC resources from GENCI-IDRIS (Grant 2022-AD011013105R1) and was partially funded by the grant ANR-20-CHIA-0012-01 ("NoRDF")

What preprocessing/cleaning was done?

According to the dataset README, the stories in HANNA were grouped by generation model - *"Story ID is the ID of the story (from 0 to 1,055). Stories are grouped by model (0 to 95 are the Human stories, 96 to 191 are the BertGeneration stories, etc.)"*-. Annotations were collected through Amazon Mechanical Turk (mTurk), and the dataset includes data such as Worker ID, Assignment ID, time spent on task (in seconds), and named characters (the name entered by the worker for the first mentioned character in the story). However, the README does not mention any additional text preprocessing steps such as tokenization, part-of-speech tagging, or linguistic feature extraction. Therefore, it can be inferred that minimal preprocessing was applied beyond organizing and labeling the data for evaluation.

Ethical considerations

The dataset includes human annotations collected via Amazon Mechanical Turk(mTurk). While the project is released under the MIT License -which allowed reuse and redistribution- it does not specify whether the crowdworkers were explicitly informed about the public release or research use of their data. The README and license do not mention any mechanism for revoking consent or limiting use, suggesting that participation fell under the general terms of Amazon Mechanical Turk rather than project-specific informed consent procedures.

Will the dataset be updated? How often, by whom?

According to the dataset README, the HANNA repository has received an update on 2024/05/13 for TACL Paper since its initial commit that was on 2022/08/24. The dataset appears to be actively maintained by the authors Cyril Chhun, Fabian Suchanek and Chloé Clavel. However there is no mention of a regular update schedule or future release plans.

Dataset Details:

The dataset consists of two types of stories:

Table 1: Distribution of Human-Written and AI-Generated Stories in the HANNA Dataset

Type of Story	Number of Stories	Average Length (tokens)
Human-Written	96	~480
AI-Generated	960 (96 per model)	~230

- **Corpus/Topic Distribution**

Table 2: corpus/topic or genera distribution

Topic (model)	tokens	Unique terms	Avg tokens per story
BertGeneration	59142	3015	205.354167
CTRL	60051	2884	208.510417
Fusion	34755	1162	120.677083
GPT	55344	3052	192.166667
GPT-2	93321	4028	324.031250
GPT-2 (tag)	87069	3593	302.322917
HINT	25356	786	88.041667
Human	139386	6740	483.979167
RoBERTa	56280	3123	195.416667
TD-VAE	90846	5268	315.437500
XLNet	87786	3434	304.812500

Note: In this analysis (as suggested in the [Corpus_topic Colab]), the token count is calculated using whitespace splitting only. No advanced tokenization or preprocessing (such as punctuation removal or stemming) has been applied during the calculation process.

• Dataset Sample

Table 3: Dataset Sample

Story ID	Prompt	Human	Story	Model	R E	C H	E M	S U	E G	C X	Worker ID	Assignment ID	Work time in seconds	Name
0	When you die..	3,000 years have...	3,000 years have I ..	Human	4	4	3	2	4	4	A2VE5I V9OD2 SK1	3X87C8JF VHIT235K Q4UTS826 4I6SQJ	579	None
0	When you die the afterlife is an..	3,000 years have I been...	3,000 years have I been..	Human	5	5	1	3	4	1	A1IZ4N X41GK U4X	3DR23U6 WEGL5K0 SU6D4J8W 9EM9LTE7	82	none

2.2 Methodology

In this study, we aim to evaluate and compare the ability of two large language models to generate fictional stories with high levels of emotional empathy. The methodology implemented is a text generation and evaluation system, where models generate short stories based on prompts, and their outputs are evaluated against an empathy prototype using semantic similarity measures.

The approach consists of two main components:

1. **Text Generation:** Using pre-trained large language models (GPT-2 and Mistral-7B) to generate short fictional stories from given prompts.
2. **Semantic Similarity Evaluation:** Using Sentence-BERT (SBERT) embeddings and cosine similarity to compare each generated story to a predefined empathetic reference prototype. A higher similarity score indicates higher perceived empathy.

The system does not involve text classification or retrieval but focuses on generating narratives and quantitatively assessing their empathetic content.

Baseline Model

The baseline model for this evaluation is GPT-2. GPT-2 was chosen as it represents a well-known autoregressive language model that was widely used prior to the emergence of instruction-tuned models.

Proposed Model

The proposed model is Mistral-7B-Instruct-v0.2, an instruction-tuned language model with significantly larger scale and improved alignment capabilities compared to GPT-2.

Model Details (following Mitchell et al., 2019 guidelines)

Table 4: Model Details

Category	GPT-2	Mistral-7B-Instruct-v0.2
Developer	OpenAI	Mistral AI
Model Date	February 2019	October 2023
Model Version	GPT-2 (final release)	Mistral-7B v0.2
Model Type	Transformer-based Causal Language Model	Transformer-based Causal Language Model (Instruction-tuned)
Training Algorithms	Autoregressive language modeling	Autoregressive language modeling with instruction fine-tuning
Number of Parameters	1.5 billion	7 billion
Training Dataset	WebText (diverse internet text corpus)	Public internet data + instruction-following datasets
Fairness Constraints	None explicitly applied	No fairness constraints during generation; designed for general instruction-following
Key Features	Next-token prediction, generic generation	Instruction-following, enhanced context adherence, richer emotional and structural generation
Paper / Resource	Radford et al., 2019 — <i>Language Models are Unsupervised Multitask Learners</i>	Mistral AI Model Card — <i>Mistral-7B and Mistral-7B-Instruct</i> (2023)
Citation	Radford, A., et al. (2019). <i>Language Models are Unsupervised Multitask Learners</i> .	Mistral AI (2023). <i>Introducing Mistral 7B and Mistral 7B Instruct</i> .
License	Open License	Apache License 2.0

By generating stories for the same set of prompts and evaluating the generated outputs against an empathetic prototype using SBERT embeddings and cosine similarity, we aim

to quantify and compare the models' abilities to produce emotionally rich and empathetic content.

The baseline comparison (GPT-2) serves as a control for evaluating the advancements introduced by instruction-tuning and larger model architectures (Mistral-7B).

3. Evaluation and results

Cosine similarity was employed as an automatic evaluation metric to measure the semantic closeness between generated stories and empathy prototype embedding. Higher similarity indicated greater alignment with the desired empathetic quality. Based on these similarity scores, a win rate was calculated by comparing each story pair to determine which model (GPT-2 or Mistral-7B) produced a story closer to the empathy prototype, allowing a direct comparison of model performance across many examples. Finally, a paired t-test was conducted on the similarity scores to statistically validate whether the observed differences between the models were significant. These metrics were chosen because in real-world applications, especially in areas like empathetic storytelling and user interaction, the emotional quality and depth of model outputs have a much greater impact than traditional accuracy measures. Evaluating based on similarity to human-like empathy and validating differences statistically ensures that the chosen model aligns better with real-world needs.

Table 5: Evaluation Results

Model	Win Count	Win Rate (%)	Mean Cosine Similarity	Statistical Significance
GPT-2	3	6.0%	Lower than Mistral-7B	Not better
Mistral-7B	43	86.0%	Higher than GPT-2	Significant ($p < 0.001$)
Tie	4	8.0%	—	—

The table above summarizes the evaluation of each model's performance in generating empathetic stories. Mistral-7B outperforms GPT-2 with a significantly higher win rate of 86.0% compared to GPT-2's 6.0%, while ties accounted for 8.0% of the cases. The mean cosine similarity between Mistral-7B's outputs and the empathetic references was notably higher than that of GPT-2. Furthermore, a paired t-test confirmed that the difference between the two models was statistically significant ($T = 8.3648$, $p \approx 0.0000$), providing strong evidence that Mistral-7B produces more empathetic responses.

4. Discussion

The results obtained in this experiment indicate a clear and statistically significant advantage of the Mistral-7B model over GPT-2 in generating emotionally empathetic fictional narratives. Out of 50 evaluated prompts, Mistral-7B won 43 comparisons (86%), while GPT-2 won only 3 (6%), and 4 outputs resulted in a tie. These outcomes were supported by a paired T-test, which yielded a t-statistic of 8.3648 and a p-value of 0.0000, providing strong evidence that the observed difference is not due to chance.

This performance disparity can be primarily attributed to the differences in model scale and training strategy. Mistral-7B is a 7-billion parameter model that has been instruction-tuned, enabling it to follow prompts more effectively and to generate coherent, contextually appropriate responses. GPT-2, by contrast, is smaller in scale (1.5 billion parameters) and lacks instruction tuning, often resulting in outputs that are disorganized and emotionally inconsistent.

Case 1: Emotional Coherence and Reflection

In the first example, prompted by *“A girl meets a boy that changes her life forever. Her first kill,”* the two models produced markedly different narratives. The GPT-2 output (Figure 1) lacked emotional structure and coherence, featuring abrupt transitions, graphic content, and unclear emotional development. In contrast, Mistral-7B’s response (Figure 2) was more reflective and emotionally resonant, portraying the protagonist’s vulnerability and internal transformation with a consistent narrative tone. The emotional clarity and pacing of Mistral-7B’s story better aligned with the qualities of empathetic storytelling.

Case 2: Identity and Social Isolation

The second comparison, based on the prompt *“You live in a world where people’s shadows show who they truly are... You are the only person in the world with no shadow,”* further highlights the performance gap. GPT-2’s story (Figure 3) exhibited structural disorganization and incoherent dialogue, ultimately failing to convey any deep emotional insight. Mistral-7B’s story (Figure 4), however, explored themes of identity, difference, and belonging through a consistent and poetic narrative. The model demonstrated a sophisticated handling of abstract emotional concepts, resulting in a more human-like and empathetic output.

Case 3: Empathy in Cultural Conflict

In the third case, the prompt *“A warrior from a planetary culture with no concept of surrender sits in a POW camp”* was used to evaluate the models’ ability to generate empathy in a speculative fiction setting. The GPT-2 response (Figure 5) was fragmented and disjointed, lacking a clear plot or emotional perspective. Conversely, Mistral-7B’s story (Figure 6) offered a well-structured and emotionally introspective narrative that humanized the warrior figure and conveyed themes of honor, captivity, and internal conflict with nuance and empathy.

These comparisons underscore the importance of instruction tuning and large-scale training in enhancing a model’s ability to produce emotionally rich and coherent narratives. Mistral-7B

consistently generated content that reflected human-like emotional intelligence, particularly in terms of narrative structure, empathy, and context relevance.

The evaluation process, which used SBERT cosine similarity against an empathy prototype, was effective for automatic empathy measurement. However, future work could benefit from incorporating human evaluation alongside automated metrics, particularly to capture subtleties in emotional interpretation that may be missed by embedding-based methods. Additionally, expanding the scope of emotional evaluation to include other affective dimensions—such as compassion, guilt, or forgiveness—would provide a more comprehensive understanding of a model’s empathetic capabilities.

```
--- Prompt ---
A girl meets a boy that changes her life forever. Her first kill.

--- GPT-2 Story ---
She crosses a red light before the iron pommel of his bayonet hits her temple. It sends a cold chill down her spine as she shakes from the pain. Her tears stream down her face as she slowly moves to stand upright. She has practiced. A naked boy with a canvas covered face and a gun rests on his right hand. A neat and comfortable man in a business suit covers his face. One second she's seeing him looking past her. The next, she sees him quietly walking towards her. He was silent as she moves towards him. She always feared that this man would speak to her or call her name. It isn't uncommon for a man to return to her, taking them with him. She speaks her name quietly in this man's mind and with a puzzled look on his face. He quietly whispered her name and as she moved towards him she managed to open her mouth to speak. He turned and gave her a solid kiss. "Maria," he said softly. He turned and walked over to her with the pistol in hand. He reached out and kissed her neck. "We are meant for each other" he whispered as he walked over to her. "I can feel it in your bones. It is so close now. Don't you think that maybe I will stay here for a while?" He pulled away from her and gave her a final kiss before walking away. That was three years ago. The red lights got worse and worse, the corpses started coming closer. They are coming from every direction at once, the screams rising from the living and eating away at her spine. She is a ghost. She feels her heart begin to race. All she wants to do is sit here with this man. Sit back in peace. But that's not true. The lives of all those who he was planning to kill won't come. They all started coming from one place, and now every morning and every night he steps out of his chair and walks straight into her soul. He pulls out his shotgun and shoots himself in the head

--- Mistral-7B Story ---
Write a fictional short story (~400 words) based on this prompt:

A girl meets a boy that changes her life forever. Her first kill.
```

Figure 1 GPT-2

```
Story:

Emma sat alone on the cold, muddy bank of the pond, staring at the calm water. The sun was setting, and the last light of the day bounced off the still surface, creating ripples that spread out in all directions. The silence was broken only by the sound of crickets chirping in the grass behind her.

She had come to this place every day since her mother's passing, hoping to find some solace in the tranquility of nature. But today, as she watched the sun dip below the horizon, something unexpected happened. A figure appeared at the edge of the clearing, a boy no older than her, with a look of awe on his face as he took in the beauty of the scene before him.

Emma looked away, feeling a familiar sting of vulnerability in her chest. She had grown used to being alone, to keeping her emotions hidden away, like a dark secret. But this

Winner: B
```

Figure 2 Mistral-7B

--- Prompt ---

You live in a world where people's shadows show who they truly are at their core. Some shadows look like monsters, some look like animals. You are the only person in the world with no shadow.

--- GPT-2 Story ---

"Come on," I told her, "And watch. You can almost see what my shadow looks like. You know that you should see what your shadow is like. I'm just the same color as the opposite color. It would seem that my shadow is much, much darker." "I love you," she said. "I miss you so much." She leaned in. "You can't say that, Peter. I'm really not a girl. I'm the opposite of you. I'm not a shadow. I'm just a person." I kissed her on the cheek. Her light brown hair fell back into her shoulders, and from the sight of her, I knew I was a woman. "Go home, home." And that's how I became a shadow. And all the people around me, having turned my shadow to form a monster, have turned themselves into a boy and a girl. All because of a twisted accident. The dead will be cremated as a means of memorializing those they loved. Write a story about it. I watched the casket, like all the other dead bodies. The casket was caved in, which I did, but I was still unsure about how to bury it, so I changed the soggy black suit my car had. I raised the casket to my chest. The casket fell, revealing nothing more than the empty void. "We used to celebrate all the dead being cremated for us." I remember the way her hair fell. The way the casket creaked with every step. But now, the casket was completely hollow. Not the lifelessness of an empty casket, but a hollow dead pile of souls. You are in a hotel room when you notice a girl with long blonde hair slowly starts moving around in the corner of the room. She seems quite ordinary, but then you notice how she is carrying a small knife in her hands. You are in danger and decide to get rid of the intruder. There is a screen in the middle of the room with some text in bold. It reads: "Press release" I pushed the button in front of me and opened the door. Behind the door was a

--- Mistral-7B Story ---

Write a fictional short story (~400 words) based on this prompt:

You live in a world where people's shadows show who they truly are at their core. Some shadows look like monsters, some look like animals. You are the only person in the world with no shadow.

...

Figure 3 GPT-2

Story:

In the world of shadows, I was an anomaly. A person with no shadow of my own. I was the only one, the lone human in a land of beings who wore their souls on their sleeves, or rather, on their backs. They were a curious lot, these shadow-dwellers. Each one was a reflection of their core; monstrous or gentle, depending on what lay at their heart.

I, however, was a mystery. I wandered the land without a trace of myself behind me. It was a lonely existence, but I had grown accustomed to it. I had learned to hide my true nature, to blend in with the others. But there was one creature, one particular shadow, who had always intrigued me.

It was a creature of the night, a massive beast with razor-sharp fangs and talons. Its eyes glowed an eerie green, and its ro

Winner: B

Figure 4 Mistral-7B

--- Prompt ---

A warrior from a planetary culture with literally no concept of surrender or quarter sits detained in a POW camp.

--- GPT-2 Story ---

"Shit, this isn't going according to plan. I need to execute my section then! Fuck" "Don't ever open your eyes, there is no escape from this situation. Look, I will negotiate if you will but we must sacrifice any part of the skirmish." "Fuck! Please, just listen to me now." The DPS inspector from Kyuu shook his head and felt the atmosphere envelope the location where he had previously been, "J-Hey there! Come!" A few seconds later, the interrogator brought out a syringe. "... Lieutenant Salipsa. Prisoner 42290975. Suffering from Perio^UH-69a" "What?" "What the hell is that?" "Send him to hang in the hut, he's gone to die of thirst" "Anything?" "I'm the saviour of Earth." DPS inspected or created an unprecedented situation, which should make him our savior. "This is a humiliation... But don't worry, I have a plan" He made up his mind and floated to the front of the battalion, "I need a piece of our squad for last night's skirmish" General Ross realised he mustn't continue with his goons with his war plan, Your character has the ability to time travel, but only when they're in danger. You've managed to convince them that saving the day was worth it and they must do it

--- Mistral-7B Story ---

Write a fictional short story (~400 words) based on this prompt:

A warrior from a planetary culture with literally no concept of surrender or quarter sits detained in a POW camp.

Figure 5 GPT-2

Story:

In the vast expanse of the cosmos, on a planet known as Kaldor, a unique culture thrived. They were the Kaldorians, a proud and noble race that prided themselves on their unyielding spirit and unwavering resolve. In their society, the concept of surrender was foreign, and the very notion of quarter was unheard of.

Among the Kaldorians was a mighty warrior named Tharok. He was a towering figure, his muscles rippling beneath his gleaming armor. His heart was as hard as the iron he wore, and his spirit was unbreakable. He had led countless battles against their ancient enemies, the Xoraxians, and had never once wavered in his resolve.

But even the mightiest of warriors could not withstand the tide of fate. One fateful day, as Tharok led his troops against the Xoraxian horde,

Winner: B

Figure 6 Mistral-7B

5. Conclusion

We investigated the ability of large language models to generate emotionally empathetic fictional stories and assessed their performance using an automated semantic similarity approach. Through a structured evaluation framework based on the HANNA benchmark dataset and SBERT-based embedding comparisons, we demonstrated that instruction-tuned models like Mistral-7B-Instruct-v0.2 significantly outperform earlier models like GPT-2 in producing empathetic narratives.

Mistral-7B achieved a notably higher win rate and cosine similarity score compared to GPT-2, and statistical testing confirmed the significance of the performance difference. These results highlight the impact of model scale and instruction fine-tuning on enhancing emotional depth and narrative coherence.

Future work can build on these findings by incorporating human evaluations alongside automatic methods, exploring broader emotional dimensions beyond empathy, and applying similar evaluation frameworks to multilingual or domain-specific storytelling tasks.

6. References:

- [1] Ibm. (2025, April 16). Large Language Models. a. <https://www.ibm.com/think/topics/large-language-models>
- [2] Dig-Team. (n.d.). GitHub - dig-team/hanna-benchmark-asg: HANNA, a large annotated dataset of Human-ANnotated Narratives for ASG evaluation. GitHub. <https://github.com/dig-team/hanna-benchmark-asg>
- [3] SentenceTransformers Documentation — Sentence Transformers documentation. (n.d.). <https://www.sbert.net/>