Overview of Long Story Generation Challenge (LSGC) at INLG 2024

Aleksandr Migal[♠♣], Daria Seredina[♠], Ludmila Telnina[♠], Nikita Nazarov[♠], Anastasia Kolmogorova[♠], Nikolay Mikhaylovskiy[◊♠]

[♠]National Research University Higher School of Economics, Saint-Petersburg, Russia, 190068

[◊]Higher IT School, Tomsk State University, Tomsk, Russia, 634050

[♠]NTR Labs, Moscow, Russia, 129594

amigal@ntr.ai, nickm@ntrlab.com

Abstract

This report describes the setup and results of the shared task of human-like long story generation, the LSG Challenge, which asks to generate a consistent, human-like long story (a Harry Potter fanfic in English for a general audience) given a prompt of about 1,000 tokens. We evaluated the submissions using both automated metrics and human evaluation protocols. The automated metrics, including the GAPELMAPER score, assessed the structuredness of the generated texts, while human annotators rated stories on dimensions such as relevance, consistency, fluency, and coherence. Additionally, annotators evaluated the models' understanding of abstract concepts, causality, the logical order of events, and the avoidance of repeated plot elements. The results highlight the current strengths and limitations of state-of-the-art models in long-form story generation, with key challenges emerging in maintaining coherence over extended narratives and handling complex story dynamics. Our analysis provides insights into future directions for improving long story generation systems.

1 Introduction

This report presents an analysis of the results of the Long Story Generation Challenge (LSGC), where participants showcased their systems for creating extended stories. With this shared task, we aimed to advance the generation of long-form literary texts. Our evaluation was based on two main criteria: statistical metrics and a human evaluation protocol. The LSGC was originally proposed by Mikhaylovskiy (2023); this report follows the cited work closely.

Over 110 years ago, mathematician Andrei Markov demonstrated how to study effectively the text using mathematical methods (Markov, 1913). In his work, he examined the relationship between vowels and consonants in the early chapters of Eugene Onegin. He later gave his name to processes known as Markov chains. Markov chains formed the basis of early text generation algorithms that generated basically nonsense based on the probabilistic distribution of words in a text.

Today, text generation has advanced tremendously. Autoregressive probabilistic large language models (LLMs) have become a cornerstone for solving every task in computational linguistics through few-shot learning (Brown et al., 2020) or prompt engineering (Sanh et al., 2021). Many users now interact with advanced commercial models such as GPT, Claude, or Google Bard in chat setting regularly. However, these models still have many deficiencies. Despite the targeted effort, they can generate false information, propagate social stereotypes, and produce toxic language (Taori et al., 2023).

Specifically, current autoregressive language models fail to catch long-range dependencies in the text consistency. While the autoregressive window for commercial models reaches tens or even hundreds of thousands of tokens at the time of writing, which is a lot, it, however, does not allow them to generate long coherent texts. While relevance, consistency, fluency and coherence are relatively easily achieved by the latest

Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, pages 47–53
September 23 –27, 2024.

autoregressive generative models on short texts (under 10K tokens), all the current models fail when one tries to generate a long story in a single pass. Modeling long stories requires many additional abilities compared to short texts (Guan et al., 2022), including (1) commonsense reasoning regarding characters' reaction and intention, and knowledge about physical objects (e.g., "river") and abstract concepts (e.g., "irony"); (2) modeling discourse-level features such as inter-sentence relations (e.g., causality) and global discourse structures (e.g., the order of events); and (3) the generation coherence and controllability, which require both maintaining a coherent plot and adhering to controllable attributes (e.g., topics).

Several authors have shown theoretically and empirically (Lin and Tegmark, 2017, Alvarez-Lacalle et al., 2006, Mikhaylovskiy and Churilov, 2023) that the power law autocorrelations decay is closely connected to the hierarchical structure of texts. Indeed, the hierarchical structure of, for example, Leo Tolstoy's War and Peace consists of at least 7 levels: the whole novel, books, parts, chapters, paragraphs, words, and letters. There are strong reasons to think that this structure reflects an important aspect of human thinking: people do not generate texts autoregressively. Writing a long text requires some thinking ahead, and going back to edit previous parts for consistency. This going back and forth can be reflected by navigating a tree-like structure. The autoregressive nature of the current state-of-the-art models does not reflect this; for example, S4 model (Gu et al., 2021) exhibits clear exponential autocorrelations decay (Mikhaylovskiy and Churilov, 2023).

2 Task Description

The LSG Challenge task required participants to provide a system that could output a coherent, human-like long story (a Harry Potter fanfiction for a general audience of at least 40,000 words) given a prompt of about 1,000 tokens. The organizers provided a set of story starters for developers. Systems were evaluated based on text generated from these starters, written by volunteers and imitating the stylistic features of Harry Potter fan fiction. The starters were designed from scratch specifically for this task.

It is important to note that no copyrighted texts were used in the creation of our dataset. The evaluation protocol below also does not require the usage of any of the original Harry Potter texts. It is based on the assumption that the assessors have a general knowledge of the Harry Potter universe, and this is enough to rate the texts using the provided questionnaries.

We employ both automatic and human evaluation to evaluate the quality of the texts. In particular, we used GAPELMAPER (Mikhaylovskiy, 2023) as an unreferenced automatic, statistical metric of the text structuredness. We adopt multiple human evaluation metrics to better measure model performance. Similarly to Kryscinski et al. (2019), we ask annotators to rate the texts across four dimensions:

1. Relevance (of topics in the text to the expected ones),
2. Consistency (alignment between the parts of the text),
3. Fluency (quality of individual sentences)
4. Coherence (quality of sequence of sentences).

Extending Guan and Huang (2020) we ask annotators to rate repeating similar texts. Finally, we asked the annotators to evaluate the creative dimensions of the resulting texts:

5. Doubt of the characters of the text or the narrator in their own rightfulness
6. Expression of the strong positions of the text (beginning/end of the text, beginning/ end of the chapter)
7. General idea of the text
8. Usage of idioms
9. Creativity of the text
10. Emotionality of the text

3 Dataset Description

Story starters were created by undergraduate students majoring in Linguistics as a part of their coursework with a proper credit. For testing and development purposes, we presented participants with five distinct story starters.

4 Shared Task Timeline

The LSGC was planned throughout the recent academic year. The key dates of the shared task were:

- SEP, 2023: The shared task is announced at the INLG 2023 conference.
- DEC, 2023: The task website is up; participants can register to the task.
- JULY 15, 2024: The submission is closed; organizers conduct manual evaluation.
- AUG, 2024: The LSG Challenge shared task is fully completed. Organizers submit participant reports and challenge reports to INLG 2024.

5 Baseline

We developed a baseline, published at https://lsgc.vercel.app/baseline, that generates a fan fiction text complying to the shared task requirements to make sure that the shared task is feasible. In light of the shared task's objective to create a lengthy, coherent fanfiction, we incorporated a hierarchical prompting system into the baseline to ensure the narrative's "completeness".

The baseline implements a process that begins with a "story starter". By establishing a clear narrative structure, we create a framework for generating additional content, with the aim of remaining faithful to the original story in the fanfiction we produce. After setting up the narrative framework, we then focus on fleshing out the details of each section, creating chapter outlines that outline the events to be included (see Figure 1). The number of chapters produced will depend on the capabilities of the generative model to generate believable text. This includes:

- Introduction: Establishing the protagonist's world and introducing key themes.
- Development: Presenting obstacles, conflicts, and character growth.
- Climax: A turning point where the protagonist faces a critical challenge or revelation.

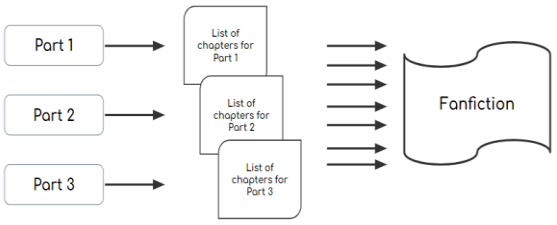

Figure 1: Chapter development

- Resolution: Tying up loose ends, providing closure and a sense of accomplishment.
- Conclusion: Offering a satisfying denouement, wrapping up the narrative and leaving a lasting impression on the reader.

6 Participants

Two teams participated in the challenge. Each team submitted one story generated using their systems. All texts were anonymized prior human evaluation to ensure objective evaluation. Each text was assessed using the GAPELMAPER metric and the human evaluation described below.

Team 1 (Decision Stump, Boriskin, Galimzianova, 2024) – The approach does not include any fine-tuning and utilizes Llama 3 with 70b parameters with special prompting scheme for the text generation. Team 1 developed the baseline in the direction of generating of the book components. The full pipeline consists of 2 parts – summary generation and generation of chapters in a loop with the transmission of context about previous events in the book via the system prompt.

The team presented a text consisting of 14 chapters, each chapter spanning 10 pages. This design aims to have only 14 potential points of discontinuity (at the junctions between chapters) where plot inconsistencies might arise, such as repeated scenes. For instance, at the end of Chapter 1, the main character Theo encounters the heroine Pansy, and at the beginning of Chapter 2, the model again describes their meeting. However, even such minor flaws blend reasonably harmoniously into the overall context. Throughout the fanfic, the narrative thread is maintained, making it

Team	Power law MAPE	Exp law MAPE	GAPELMAPER
Team 1 (Decision Stump)	0.52	0.57	0.91
Team 2 (Neurowling)	0.17	0.40	0.44
Baseline	0.17	0.31	0.57

Table 1: GAPELMAPER metrics of solutions

challenging to distinguish the text from that of a real author.

Team 2 (Neurowling, Seredina, 2024) – Approach is based on fine-tuning the Mistral-7B-Instruct-v0.2-GPTQ model with Supervised Learning (SL). The final text of a fanfiction was generated with the fine-tuned model and the prompts following the baseline.

The team also delivered commendable results: their text comprises numerous short chapters with rapidly unfolding action, unlike the first team's story. This format makes for easier reading, but due to the brevity of each chapter (approximately one page), inconsistencies and contextual discrepancies can be noticed at the chapter boundaries. For example, a character's gender might be female in one chapter and male in another. Nevertheless, the provided structure, which involved preliminary generation of all chapters according to a unified concept (outline), ensured reasonable consistency of the narrative.

7 Evaluation

7.1 GAPELMAPER Metric

GAPELMAPER (GloVe Autocorrelations Power/Exponential Law Mean Absolute Percentage Error Ratio) is a metric designed to assess text coherence based on the autocorrelation of embeddings. It helps determine whether the text is intrinsically structured, based on the decay patterns of the autocorrelations. The results of evaluating the submitted texts with GAPELMAPER are listed in the Table 1.

Mikhaylovskiy and Churilov (2023) state that "GAPELMAPER less than 1 means that the autocorrelations decay according to a power law and the text is structured in a way. GAPELMAPER more than 1 means that the autocorrelations decay according to an exponential law and the text is unstructured". From this viewpoint, the text produced by the system by Decision Stump is on a verge of being structured, while Neurowling's text exhibits a clear long-distance structure to a level that exceeds the baseline.

7.2 Human Assessment

To assess the results of our shared task from a human perspective, we asked a group of undergraduate students majoring in Linguistics to read several fanfics about "Harry Potter", including texts written by humans and those generated by language models participating in our shared task. The average age of the evaluators is 21 years old; all of them are confident English speakers (B2 to C1 level as assessed via prior coursework). The native language of all evaluators is Russian. Some respondents had only read "Harry Potter" in Russian and have never read any "Harry Potter" books in English and were therefore surprised by the absence of explanations and hints about the characters' backstories, with many terms, such as "Sorting Hat", being unfamiliar to them. This lack of context sometimes led to difficulties in understanding the narrative and its underlying themes.

Each evaluator evaluated three texts, randomly selected between participant submissions, baseline and three human-written fan fictions. The number of persons who evaluated the work of the Decision-Stump team was 10, while only 5 persons evaluated the text of the Neurowling team. The respondents were asked to answer a series of questions about the texts they read (the results can be seen in the table) and provide any additional comments they might have.

The evaluators analyzed the texts for literary quality, originality, style, cohesion and coherence of the generated texts and overall perception. Each evaluator assessed the text according to the documented criteria on a scale from 1 to 5, where 1 is the worst rating possible and 5 is the highest. The tables 2 and 3 show the average scores of the calculated based on all expert assessments of the data. We present the results of "Harry Potter and the Slytherin Selection" (DrizzleWizzle, 2012) evaluation as "Fan Fiction" line for comparison.

	Relevance		Consistency	The order of events	Repeating similar texts	Fluency	Coherence
Team	**Correlation between the fanfic title and its content**	**Compatibility of chapter and sub-chapter titles with the overall style of the text**	**The strength of the stylistic connection between all the elements of the text**	**The pace of the plot**	**Word repetitions**	**Text composition**	**Text syntax**
Decision Stump	1.75	**3.6**	2	2.3	2.8	2.3	3
Neuro-wling	**3.25**	3.2	2.6	1.8	2.6	2.8	3.2
Baseline	2.25	3.5	3	2.8	3.6	3.2	3.8
Fan Fiction	2.3	2.7	**3.9**	**3**	**3.9**	**4.3**	**3.9**

Table 2: Human evaluation results – text quality

Text 1 ("Decision Stump")

The majority of respondents noted the presence of narrative inconsistencies in the texts, stating that "instances of redundancy occur not only on a lexical level but also on a semantic level: the same event can be described multiple times using slightly different words or with different (not very original) details, which may be indicative of a lack of cohesive narrative structure". Additionally, respondents pointed out the lack of character dialogue in the texts, which made the stories seem less engaging: "The story is driven not by the characters and their actions, but by the narrative itself, resulting in a sense of detachment from the characters' experiences". Respondents who had read the original books in English or were fans of the series noted stylistic discrepancies: "There are moments that stand out to a reader immersed in the lore, indicating that the text was clearly not written by an expert (for example, the way Hagrid speaks, which deviates from his characteristic mannerisms and speech patterns in the original books)".

Nevertheless, many respondents noted that the text has some strong aspects, such as a well-structured beginning and conclusion, and a moderate use of complex syntactic structures (embedded clauses, subordinate clauses of various types, participial phrases, impersonal or indefinite-personal sentences, and ellipses). The text employs conventional stylistic devices, but the language itself is not sufficiently creative. Respondents were unable to discern the main and overarching idea or theme in the text, although occasional glimpses of an idea did emerge in certain sections. Furthermore, the text also contains elements that appear to be logically integrated into the narrative, but ultimately prove to be inconsequential to the overall plot. These elements seem to be introduced with a specific purpose in mind, but fail to contribute meaningfully to the story's development or resolution, leaving the reader wondering about their significance. On the other hand, respondents praised the harmonious combination of chapter and subchapter titles with the overall style of the text.

Text 2 ("Neurowling")

The informants highly praised the semantic correspondence between the fanfiction title and the subsequent text, as well as the combination of chapter and subchapter titles with the overall style and content of the chapters and subchapters. They noted the presence of hints at a common idea or theme, although it was challenging to pinpoint a single, unified concept. However, the informants were less impressed with the pacing of the plot, which they found to be either too fast or too slow at times, with the rhythm sometimes changing in a way that didn't align with the unfolding narrative.

On a lexical and grammatical level, the text exhibited repetition, which came across as a limited vocabulary. Nevertheless, the text featured a

Team	Doubt of the characters of the text or the narrator in their own rightness	Expression of the strong positions of the text (beginning/end of the text, beginning/ end of the chapter)	General idea of the text	Usage of idioms	Creativity of the text	Emotionality of the text
Decision Stump	2.1	2.9	2.1	2.4	2	2.6
Neuro-wling	2.4	3.2	3.2	3	3	3
Baseline	3.1	3.2	3.2	3.1	3.5	3.6
Fan Fiction	3.5	3.4	4	4.2	3.3	4.4

Table 3: Human evaluation results – creative aspects

sufficient number of complex constructions, including parenthetical phrases, subordinate clauses, and participial phrases. The text also employed conventional metaphors, comparisons, and familiar clichéd oxymorons, but nothing beyond that.

When asked if they could summarize the main plot of the text, some informants responded positively, which suggests that there are indeed signs of a cohesive narrative. The majority of informants also praised the strong opening and conclusion of the text. Regarding the emotional resonance of the text, this aspect of literary writing still leaves room for improvement, as the emotions expressed in the text change, but in a somewhat abrupt and peculiar manner.

Furthermore, the chapters often repeated each other's plot, which led one informant to comment, "This makes me think that it wasn't written by a human. If it weren't for this, I would say that the text was written by a teenager who is very fond of the Harry Potter universe."

8 Conclusions

Both teams have demonstrated their capacity to generate long-form narratives with structured coherence, as evidenced by their GAPELMAPER scores. However, based on the combined quantitative and qualitative evaluations, Team 2 ("Neurowling") emerges as the stronger contender. Both teams not very significantly departed from the baselines in terms of the system architecture. The results of both teams also only sparsely improved on the baseline.

The GAPELMAPER score of 0.44 for Team 2 indicates a significantly more cohesive narrative structure compared to Team 1's score of 0.91. Although both texts exhibited certain narrative inconsistencies, Team 2's shorter chapter format and faster pacing made the text more accessible to readers, even if this format occasionally led to contextual discrepancies. Moreover, the manual evaluation highlighted that Team 2's text maintained a better alignment between chapter titles and content, as well as a clearer thematic structure.

While Team 1 ("Decision Stump") produced a more extensive narrative, the manual assessment revealed that this length led to redundancy and a lack of emotional engagement, as well as difficulties for readers unfamiliar with the "Harry Potter" universe. In contrast, Team 2's text, despite its flaws, was more favorably received in terms of readability and structure.

The evaluators easily detect the generated texts. The generated texts are still behind even non-professionally writing humans in terms of text quality and creativity.

Nevertheless, we can say that our expectations for this challenge were reasonably justified. The results of this study show the difference between using fine-tuning and prompt engineering approaches in text generation and demonstrate the advantages and disadvantages of each. In future, we would like to continue this research with a larger data set, and see more diverse text generation approaches from participants. This would allow us to get closer to understanding the linguistic nature

of the generated text and, possibly, the nature of the text itself.

References

Enric Alvarez-Lacalle, Beate Dorow, Jean-Pierre Eckmann, and Elisha Moses. 2006. *Hierarchical structures induce long-range dynamical correlations in written texts*. PNAS, 103(21):7956–7961.

Aleksandr Boriskin, Daria Galimzianova, 2024. *The LSG ChallengeWorkshop at INLG 2024: Prompting Techniques for Crafting Extended Narratives with LLMs*. In Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, Tokyo, Japan. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. *Language models are few-shot learners.* In Advances in Neural Information Processing Systems, volumes 2020-Decem, pages 1877–1901.

DrizzleWizzle. 2012. *Harry Potter and the Slytherin Selection*. Retrieved from: https://www.fanfiction.net/s/8666085/1/Harry-Potter-and-the-Slytherin-Selection

Albert Gu, Karan Goel, and Christopher Ré. 2021. *Efficiently Modeling Long Sequences with Structured State Spaces.* International Conference on Learning Representations. 2021:1–32.

Jian Guan and Minlie Huang. 2020. *UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation.* In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9157–9166, Online. Association for Computational Linguistics.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2022. *LOT: A Story-Centric Benchmark for Evaluating Chinese Long Text Understanding and Generation.* Transactions of the Association for Computational Linguistics, 10:434–451.

Henry W. Lin and Max Tegmark. 2017. *Critical behavior in physics and probabilistic formal languages.* Entropy, 19(7):1–25.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. *Neural text summarization: A critical evaluation.* 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Andrei Markov, 1913. *An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains.* Science in Context. 2006. Vol. 19, no. 4. pages 591–600. DOI 10.1017/S0269889706001074.

Nikolay Mikhaylovskiy. 2023. *Long Story Generation Challenge*. In Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges, pages 10–16, Prague, Czechia. Association for Computational Linguistics.

Nikolay Mikhaylovskiy and Ilya Churilov. 2023. *Autocorrelations Decay in Texts and Applicability Limits of Language Models*. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023"

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, et al. 2021. *Multitask Prompted Training Enables Zero-Shot Task Generalization.* ICLR 2022.

Daria Seredina, 2024. *A Report on LSG 2024: LLM Fine-Tuning for Fictional Stories Generation.* In Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, Tokyo, Japan. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. *Stanford Alpaca: An Instruction-following LLaMA Model.* https:github.com/tatsu-lab/stanford_alpaca, 2023.