Long Story Generation Challenge

Nikolay Mikhaylovskiy
Higher IT School, Tomsk State University, Tomsk, Russia, 634050
NTR Labs, Moscow, Russia, 129594
nickm@ntr.ai

Abstract

We propose a shared task of human-like long story generation, *LSG Challenge,* that asks models to output a consistent human-like long story (a Harry Potter generic audience fanfic in English), given a prompt of about 1K tokens. We suggest a novel statistical metric of the text structuredness, GloVe Autocorrelations Power/ Exponential Law Mean Absolute Percentage Error Ratio (GAPELMAPER) and the use of previously-known UNION metric and a human evaluation protocol. We hope that *LSG Challenge* can open new avenues for researchers to investigate sampling approaches, prompting strategies, autoregressive and non-autoregressive text generation architectures and break the barrier to generate consistent long (40K+ word) texts.

1 Task Overview

The human-like long story generation (*LSG*) task asks models to output a consistent human-like long story (a Harry Potter generic audience fanfic in English), given a prompt of about 1K tokens. The text will be evaluated by automated metrics described in Section 3.1, and a human evaluation protocol described in Section 3.2.

2 Motivation

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 (Sahn et al., 2021). Many users now interact with such models as ChatGPT, 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).

The LLM deficiency we particularly want to attack is their inability to produce a human-grade long text. Current autoregressive language models fail to catch long-range dependencies in the text consistently. Large language models such as GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023), ALPACA (Taori et al., 2023) push the boundary of "short text" rather far, but do not remove the problem. Commercial instruction-following language models such as ChatGPT, GPT-4, Claude and Google Bard are targeted at the use in a dialogue (and probably that is not for nothing). They generate a limited number of tokens per user input, and only generate further text after additional prompting. While the autoregressive window for commercial models at the time of writing reaches 32K tokens for OpenAI and even 100K tokens for Anthropic, which is a lot, it does not allow them to generate long coherent texts.

While relevance, consistency, fluency and coherence are easily achieved by the latest 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).

Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges, pages 10–16
September 11-15, 2023 ©2023 Association for Computational Linguistics

Mikhaylovskiy and Churilov (2023) have recently studied autocorrelations in long texts using pretrained word vectors. That allowed to study a wide range of autocorrelation distances in human-written and model-generated texts and show that the autocorrelations in human-written literary texts decay according to power laws on distances from 10 to 10K words independently from the language. On the other hand, the behavior of autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts. Large language models often exhibit Markovian (Markov, 1913) behavior with exponential autocorrelations decay.

Several authors have shown theoretically and empirically (Lin and Tegmark, 2017, Alvarez-Lacalle et al., 2006) 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 Pease 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).

We hope that this challenge can gain interest from the NLG community and advance sampling approaches, prompting strategies, autoregressive and non-autoregressive text generation architectures and other subfields of text generation.

3 Task Description

Formally, the task of LSG Challenge asks participants to provide a system that can output a consistent human-like long story (a Harry Potter generic audience fanfic at least 40K words long), given a prompt of about 1K tokens. A set of at least three dev prompts will be provided by organizers. The systems will be evaluated on a withheld test prompt. The prompts similar to the beginnings of human-written fan fiction will be developed from scratch specifically for the task.

	Power law MAPE	Exp law MAPE	GAPEL-MAPER
The Adventures of Tom Sawyer	0.21	0.55	0.38
The Republic	0.13	0.38	0.34
Don Quixote	0.20	0.44	0.45
War and Peace	0.09	0.42	0.21
Critique of Pure Reason	0.14	0.25	0.56
The Iliad	0.19	0.54	0.35
Moby-Dick or, The Whale	0.15	0.47	0.32
S4 generated text	0.062	0.050	1.24

Table 1: MAPE of power and exp law approximations of texts in English, and resulting GAPELMAPER

It is important to note that no copyright-eligible texts will be used in the shared task. The evaluation protocol below does not require using the original Harry Potter texts, and subjective evaluation relies on the fact that judges have read Harry Potter books/seen the films, but no factual knowledge of Harry Potter books is also required for the evaluation criteria below.

Given the open-ended and cutting-edge nature of the generation task and ongoing discussion on the best corpora and approaches to training LLMs, we feel that constraining the training set can be harmful to the task performance and participants are open to train their models on any dataset, as long as it is described in the system report.

We employ both automatic and human evaluation, described below to evaluate the quality of the texts.

3.1 GloVe Autocorrelations Power/ Exponential Law Mean Absolute Percentage Error Ratio (GAPEL-MAPER) Metric

Suppose we have a sequence of N vectors $V_i \in R^d, i \in [1, N]$. Autocorrelation function $C(\tau)$ is the average similarity between the vectors as a function of the lag $\tau = i - j$ between them. The simplest metric of vector similarity is the cosine distance $d(V_i, V_j) = \cos\angle(V_i, V_j) = \frac{V_i \cdot V_j}{\|V_i\|\|V_j\|}$, where $\cdot$ is a dot product between two vectors and $\|\ \|$ is an Euclidean norm of a vector. Thus,

$$C(\tau) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} \frac{V_i \cdot V_{i+\tau}}{\|V_i\| \|V_{i+\tau}\|}. \quad (5)$$

$C(\tau)$ ranges from -1 for perfectly anticorrelated sequence (for $\tau = 1$ and $d = 1$ that would be $1, -1, 1, -1$ etc.) to 1 for a perfectly correlated one (for $\tau = 1$ and $d = 1$ that would be $1, 1, 1, 1$ etc.).

A distributional semantic assigns a vector to each word or context in a text. Thus, a text is transformed into a sequence of vectors, and we can calculate an autocorrelation function for the text. Two distributional semantics approaches have been proposed for word-level autocorrelation computations: Alvarez-Lacalle et al. (2006) proposed a bag-of-words (BOW) model, and Mikhaylovskiy and Churilov (2023) have suggested the use of pretrained GloVe (Pennington et al., 2014) vectors. Unlike BOW, which only allows measuring long distance correlations, the latter approach allows to measure autocorrelations at any word distance starting with 1. Thus, we suggest using GloVe for autocorrelation measurement.

Mikhaylovskiy and Churilov (2023) have found that autocorrelations in long human-written texts decay according to a power law at ranges from 10 to 10K words. We suggest measuring the structuredness of a generated text by comparing how well the autocorrelations decay is approximated by power law and exponential law. To do so, one can compute autocorrelations in this range, approximate these points by a straight line in log-log and log-linear coordinates using the least squares regression and evaluate the goodness of fit of these regressions by MAPE (Mean Absolute Percentage Error). The ratio of these two errors constitute a metric we call GloVe Autocorrelations Power/Exponential Law Mean Absolute Percentage Error Ratio (GAPELMAPER):

$$\text{GAPELMAPER} = \frac{MAPE_{power}}{MAPE_{exp}}$$

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 a exponential law and the text is unstructured. As a matter of example, we take Table 3 from Mikhaylovskiy and Churilov (2023) and compute GAPELMAPER in Table 1.

The metric proposed above does not require any gold standard, it is a statistical metric of the text itself. Thus, in terms of Guan and Huang (2020) it is an unreferenced metric.

3.2 UNION Metric

UNION is an unreferenced metric for evaluating open-ended story generation, proposed by Guan and Huang (2020). Built on top of BERT, UNION is trained to distinguish human-written stories from negative samples. The negative samples are programmatically constructed using Repetition, Substitution, Reordering and Negation Alteration.

3.3 Human Evaluation Approach

A single number is not enough to evaluate the quality of a long story. 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), and
4. coherence (quality of sequence of sentences).

Additionally, extending (Guan et al., 2022), we ask annotators to rate

5. knowledge about physical objects (LLM generated failure example: “I was on shore in a boat; but I was not in the water. I was not in the water. I was in the water.”)
6. knowledge about abstract concepts (LLM generated failure example: “The twenty-eighth one is a twenty-eighth one. The twenty-nineteenth one is a twenty-eighth one. The twenty-ninth one is a twenty-ninth one. The twenty-tenth one is a twenty-tenth one.”)
7. causality (LLM generated failure example: “The first part was pretty easy. The second one, on the other hand, took a lot of practice. I had a lot of difficulty with the first one.”)
8. the order of events (LLM generated failure example: “This is the way all voyages of travel are done in all ages of the earth; they come to it and lay it down in the same fashion: — They get a wind, sail about awhile, and

gather what stores are sufficient for a week, or for one night's stay.")

Finally, extending Guan and Huang (2020) we ask annotators to rate

9. repeated plots (repeating similar texts)

A detailed evaluation manual will be developed as a part of the competition preparation and provided to judges, including a checklist conforming to suggestions of Howcroft et al., (2020).

Each text will be rated by 3 distinct judges with the final score obtained by averaging the individual scores. We plan to hire linguistics/philology students with English knowledge level at least C1 as the judges in at least two low-cost countries. Where possible, the judge assignment will be included into coursework. Small non-government/donation funding will be made available to cover judging expenses where the above approach is not possible.

3.4 Protocol

We propose the following schedule:

- **Phase 1** (from Sep, 2023): The shared task is announced at the INLG 2023 conference, and the data are available on the shared task website; participants can register to the task.
- **Phase 2** (from Dec, 2023): The leaderboard is open; participants can submit their systems to the organizers and the online leaderboard keeps updating the best performance using automatic evaluation metrics.
- **Phase 3** (from Mar, 2024): The submission is closed; organizers conduct manual evaluation.
- **Phase 4** (Jul, 2024): The LSG Challenge shared task is fully completed. Organizers submit participant reports and challenge reports to INLG 2024 and present at the conference.

For fairness and reproducibility, participants should specify what and how external resources are used in their system reports. In Phase 3, after the submission deadline, the organizers will start to evaluate summaries generated by final submitted models with the help from linguistic experts.

Please note that the above schedule can be modified accordingly when the schedule of INLG 2024 is released. The leaderboard and the detailed schedule will be announced on the shared task website.

4 Related work

Shaham et al. (2022) introduced SCROLLS, a suite of tasks that require reasoning over long texts. It includes earlier introduced works of Huang et al. (2021), Chen et al. (2022), Zhong et al. (2021), Dasigi et al. (2021), Kočiský et al. (2018), Pang et al. (2022), and Koreeda and Manning (2021). While all are related to long texts, none of these datasets and tasks asks to generate a long text.

Gehrmann et al. (2021) introduced GEM, a living benchmark for natural language Generation (NLG), its Evaluation, and Metrics. GEM provides an environment in which models can easily be applied to a wide set of tasks and in which evaluation strategies can be tested and consists of 11 datasets/tasks. Tay at al. (2020) proposed Long Range Arena, a suite of tasks consisting of sequences ranging from 1K to 16K tokens, encompassing a wide range of data types and modalities such as text, natural, synthetic images, and mathematical expressions requiring similarity, structural, and visual-spatial reasoning. None of these tasks asks to generate a long text as well.

Very recently Köksal et al. (2023) introduced the LongForm dataset, which is created by leveraging English corpus examples with augmented instructions. No evaluation protocol or competition is suggested in the cited paper.

On the unreferenced metrics front, Guan and Huang (2020) proposed UNION metric described in Section 3.2. Huang et al. (2020) proposed a metric dubbed GRADE, which stands for Graph-enhanced Representations for Automatic Dialogue Evaluation. Gao, Zhao, and Eger (2020) suggested SUPERT, which rates the quality of a summary by measuring its semantic similarity with a pseudo reference summary. Vasilyev, Dharnidharka, and Bohannon (2020) suggested BLANC that measures the performance boost gained by a pre-trained language model with access to a document summary while carrying out its language understanding task on the document's text.

The most similar effort to ours was most likely made by Guan et al. (2022), who proposed a story-centric benchmark named LOT for evaluating Chinese long text modeling. The benchmark aggregates two understanding tasks and two generation tasks. The authors constructed new datasets for these tasks based on human-written Chinese stories. Unlike our proposal, LOT benchmark is limited to texts hundreds of words long, and Chinese language.

5 Conclusion

We propose the LSG Challenge to address the task of long text generation, with the hope that it can open new avenues for researchers to investigate sampling approaches, prompting strategies, autoregressive and non-autoregressive text generation architectures and break the barrier to generate consistent long (40K+ token) texts, and the frontier of text generation can be pushed further.

Acknowledgments

Thanks to Simon Mille and anonymous referees for the useful comments that helped to improve this paper.

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.

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.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A Dataset for Abstractive Screenplay Summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4599–4610, Online. Association for Computational Linguistics.

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

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, et al.. 2021. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. In Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), pages 96–120, Online. Association for Computational Linguistics.

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 2020, Online, November 16-20, 2020, pages 9157–9166. 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.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In Proceedings of the 13th International Conference on Natural Language Generation, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9230–9240, Online. Association for Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient Attentions for Long Document Summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, pages 1419–1436, Online. Association for Computational Linguistics.

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

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. Transactions of the Association for Computational Linguistics, 6:317–328.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abdullatif Köksal, Timo Schick, Anna Korhonen and Hinrich Schütze. LongForm: Optimizing Instruction Tuning for Long Text Generation with Corpus Extraction. ArXiv abs/2304.08460 (2023)

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 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"

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question Answering with Long Input Texts, Yes!. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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*.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized CompaRison Over Long Language Sequences. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 12007–12021, Abu Dhabi, United Arab Emirates. 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.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long Range Arena: A benchmark for efficient transformers. In International Conference on Learning Representations.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions. https://arxiv.org/abs/2212.10560

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5905–5921, Online. Association for Computational Linguistics.