# Long Story Generation (LSG) Challenge

Nikolay Mikhaylovskiy

Higher IT School, Tomsk State University, Tomsk, Russia, 634050

NTR Labs, Moscow, Russia, 129594

nickm@ntr.ai

# We live in boomers’ Sci-Fi

- To mogę zrobić - rzekł Trurl. - Ale czemu chcecie aż trzech maszyn?
- Życzylibyśmy sobie - rzekł, z lekka tocząc się to w jedną, to w drugą stronę Synchrofazy - aby pierwsza opowiadała historie zawiłe, lecz pogodne, druga chytre I dowcipne, trzecia zaś otchłanne i poruszające.

Stanisław Lem
**Bajka o trzech maszynach opowiadających króla Genialona**
**1967**

"I can do that," Trurl said. - But why do you want three machines?
"We would like," Synchrofazy said, turning slightly from the one side to the other, “for one to tell intricate but comforting stories, the other to tell stories cunning and witty, and the third one – stories profound and moving.

Stanisław Lem
***Tale of the Three Storytelling Machines of King Genius***
**1967**

**You are here** →

- Machine can write a story
- Machine can write a long coherent story
- Machine can write a long interesting story
- Machine can write an intricate but comforting story
- Machine can write a cunning and witty story
- Machine can write a profound and moving story

# The LSG Challenge Task

- 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
- Prompts will be developed from scratch

- No constraints on training dataset

# Evaluation: Manual Complements Automatic

**Automatic**

- GAPELMAPER Metric of text structuredness
- UNION Metric for evaluating open-ended story generation

**Manual**

- relevance (of topics in the text to the expected ones),
- consistency (alignment between the parts of the text),
- fluency (quality of individual sentences), and
- coherence (quality of sequence of sentences)
- knowledge about physical objects
- knowledge about abstract concepts
- causality
- the order of events
- repeated plots (repeating similar texts)

# Manual Evaluation

- A detailed evaluation manual will be developed
  - 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
  - the judge assignment will be included into coursework where possible

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.

# Protocol

- **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.

# Questions?

- Nick Mikhaylovskiy
- nickm@ntr.ai