What is Wrong with Language Models that Can Not Tell a Story?

Ivan P. Yamshchikov
CAIRO, THWS
Würzburg, Germany
CEMAPRE,
University of Lisbon, Portugal
ivan@yamshchikov.info

Alexey Tikhonov
Inworld.AI
Berlin, Germany
altsoph@gmail.com

Abstract

In this position paper, we contend that advancing our understanding of narrative and the effective generation of longer, subjectively engaging texts is crucial for progress in modern Natural Language Processing (NLP) and potentially the broader field of Artificial Intelligence. We highlight the current lack of appropriate datasets, evaluation methods, and operational concepts necessary for initiating work on narrative processing.

1 Introduction

Since the linguistic turn in the early 20th century (Wittgenstein, 1921), human language has been considered fundamental to shaping human cognition. This notion positions language as a core aspect of intelligence, often equating intelligence with the ability to generate natural language. In (Turing, 1950), Turing famously suggests that the capacity for meaningful natural language interaction is critical for artificial intelligence. While most contemporary researchers narrow the Turing test's scope to day-to-day conversations, the original essay emphasizes that artificial intelligent agents should convincingly imitate humans in creative tasks expressed in natural language. Framing the problem in Turing's original terms reveals the current limitations of artificial systems, which can only partially imitate human dialogue in specific contexts and struggle to generate engaging stories (van Stegeren and Theune, 2019) or jokes (Niculescu, 2021).

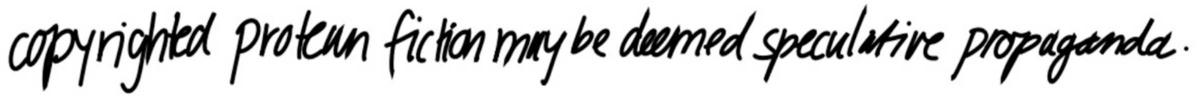
Modern Natural Language Generation (NLG) leverages increased computational power and vast training data (Brown et al., 2020; Raffel et al., 2020; Chowdhery et al., 2022; Bajaj et al., 2022; Zoph et al., 2022), focusing on computation-heavy solutions rather than on statistical methods and mathematical models to qualitatively advance our understanding of language. A century after Andrey Markov developed his eponymous chains to analyze poetry, NLG concepts remain similar, and their limitations could hardly be overcome solely through quantitative means. This is particularly evident in narrative processing, where automatic generation of textual narratives often requires significant human intervention or relies on predefined narrative structures (van Stegeren and Theune, 2019).

Efforts to generate longer text blocks[1] exist, such as (Kedziorski, 2019) and (Agafonova et al., 2020), see Figure 1, but they succeed only under certain stylistic and topical constraints that preclude genuine narrative generation. While recent advancements have been made in suspense generation (Doust and Piwek, 2017), narrative personalization (Wang et al., 2017), and short context-based narratives (Womack and Freeman, 2019), generating extended stories remains a challenge (van Stegeren and Theune, 2019).

Philosophers and linguists have attempted to conceptualize plot, narrative arc, action, and actor notions for nearly a century (Shklovsky, 1925; Propp, 1968; Van Dijk, 1976), but few of these concepts have proven useful for modern NLP. In (Ostermann et al., 2019), a machine comprehension corpus is presented for end-to-end script knowledge evaluation, revealing that existing machine comprehension models struggle with tasks humans find relatively easy. Despite these setbacks, some progress in narrative generation has been made within the NLP community (Fan et al., 2019; Ammanabrolu et al., 2020). However, narrative generation is still largely considered a fringe research topic.

We argue that the concept of *narrative* is crucial for further NLP progress and should become a focal point within the NLP community. This paper raises vital questions for narrative processing to establish itself as a well-defined sub-field in NLP research. We begin by presenting several arguments for why breakthroughs in narrative pro-

[1] https://github.com/NaNoGenMo



copyrighted protein fiction may be deemed speculative propaganda.

Figure 1: "Copyrighted protein fiction may be deemed speculative propaganda" — a line from a generative art project "Paranoid Transformer — a diary of an artificial neural network", (Agafonova et al., 2020). The diary was generated end-to-end without any human post-processing and published as a hardcover book. This is one of the examples of long-form generated artistic text, however the text is devoid of narrative.

cessing could be pivotal for artificial intelligence research in general. We then explore the bottlenecks hindering progress in narrative processing and decompose the question "why don't we have an algorithm to generate good stories?" into three systemic components: data, evaluation methods, and concepts. We contend that these three areas present significant challenges, with only data being partially addressed.

2 On the Importance of Narrative

Before addressing the three fundamental bottlenecks that separate us from achieving qualitatively advanced narrative generation models, let's briefly present a case for why narrative processing is crucial for further NLP development. Recent years have witnessed the success of language models driven by the distributional hypothesis (Harris, 1954). Although these models primarily focus on local input and training, they have been transformative even beyond the scope of classical NLP. For instance, (Lu et al., 2021) show that pretraining on natural language can enhance performance and compute efficiency in non-language downstream tasks. (Zeng et al., 2022) propose a new approach to AI systems, wherein multimodal tasks are formulated as guided language-based exchanges between different pre-existing foundation models. (Tam et al., 2022) discuss how language provides useful abstractions for exploration in a reinforcement learning 3D environment. Given these advancements, is narrative processing still necessary? Can all verbal cognition, like politics, be local?

We argue that narrative processing as a research field would significantly impact two other core aspects of natural language processing, which are essential for expanding the adoption of NLP products and technologies. The first aspect is causality and natural language inference. Causal inference from natural language is crucial for further NLP progress, and current language models still underperform in this area. Although narrative could be considered a sub-category within a broader family of causal texts, we contend that narrative generation is an ideal task for validating and testing hypotheses around natural language inference, paving the way for more explainable AI. The second area where narrative processing is indispensable is the continued development of human-machine interaction. People are known to remember stories more than facts (Wiig, 2012), but NLP-based natural language interfaces exhibit the opposite tendency, processing and "remembering" facts more easily than stories. These factors make narrative essential for further NLP progress.

Another field in which narrative processing could prove pivotal is explainable AI. One could argue that a feasible path to explainable artificial intelligence involves a set of dedicated models trained to communicate with humans in natural language, clarifying specific aspects of a given decision. These models would necessarily need to be capable of causal inference in natural language. Although this technically leads to the same bottleneck discussed earlier, we believe this field is so critical for the continued development and adoption of artificial intelligence in the industry that it warrants explicit mention here.

3 Where Do We Fail?

This position paper aims to highlight critical gaps in our conceptual understanding, benchmarking, and evaluation within the field of narrative processing. We contend that these three significant layers require the immediate focus of the research community. In this section, we examine each of these layers in depth and propose potential avenues for progress.

3.1 Data

Many existing datasets labeled as narrative datasets in academic literature deviate significantly from a common-sense understanding of a "story." Some authors even refer to their datasets as *scenarios* rather than stories or narratives. Additionally, these datasets are often too small for meaningful use with modern transformer-based language models. In (Regneri et al., 2010), authors collect 493 event sequence descriptions for 22 behavior scenarios. In (Modi et al., 2016), authors present the InScript dataset, consisting of 1,000 stories centered around 10 different scenarios. (Wanzare et al., 2019) provide 200 scenarios and attempt to identify all references to them in a collection of narrative texts. (Mostafazadeh et al., 2016) present a corpus of 50k five-sentence commonsense stories.

As we progress towards longer stories, the landscape of available data splits into two major fields: collections of narrative written in various natural languages and labelled data that facilitates narrative understanding. The examples of the latter direction include (Bamman et al., 2020) who annotate longer stories to aid narrative understanding, (Zhao et al., 2022) who pair plot descriptions with corresponding abstractive summaries, and (Pang et al., 2022; Wang et al., 2022) with QA/summarization datasets for longer stories from Project Gutenberg. The former filed of longer narrative datasets is still relatively sparse. (Fan et al., 2018) collect a large dataset of 300K human-written stories paired with writing prompts from an online forum. The MPST dataset contains 14K movie plot synopses, (Kar et al., 2018), and WikiPlots[2] comprises 112,936 story plots extracted from the English Wikipedia. (Malysheva et al., 2021) provided a dataset of TV series along with an instrument for narrative arc analysis. The rise of large language models in the last year significantly stimulated the interest of the community to the datasets that collect longer stories. For example, (Bamman et al., 2020) annotate longer stories to aid narrative understanding, (Zhao et al., 2022) pair plot descriptions with corresponding abstractive summaries, and (Pang et al., 2022; Wang et al., 2022) are QA/summarization datasets on longer stories from Project Gutenberg. We are sure that this interest will grow in the nearest future, since high-quality annotated longer narrative datasets are still rare.

Another aspect of narrative data that is still rarely addressed is multilingual narrative data. A vast majority of the narrative datasets are only available in English. In (Tikhonov et al., 2021) authors present StoryDB — a broad multilanguage dataset of narratives. With stories in 42 different languages, the authors try to amend the deficit of multilingual narrative datasets. This is one of the early attempts to amend the lack of mulilitngual narrative datasets that we know of yet we expect more in the next years.

While data is the only area of narrative processing exhibiting positive progress, it is essential to acknowledge the current state: limited datasets with longer narrative texts are available, primarily in English, and rarely include human labeling regarding narrative structure and quality. Furthermore, there is minimal discussion about the necessary narrative datasets for advancing narrative generation within the community.

3.2 Evaluation

Before delving into the narrative itself, let's first discuss the evaluation techniques available for natural language generation in general. In (Hämäläinen and Alnajjar, 2021), the authors review numerous recent generative papers, covering both automated and manual methods, where native speakers are instructed to evaluate specific properties of the generated text. This review encompasses over twenty papers on text generation that evaluate various aspects of generated texts using human labels. We believe that the scope of this paper represents the field as a whole.

Examining the evaluation aspects addressed in these 20+ papers on text generation, we find a range of methods, approaches, and concepts. For details, we refer the reader to (Hämäläinen and Alnajjar, 2021); however, in the context of this discussion, we can broadly categorize the majority of the proposed methods into five major groups:

Fluency; these methods estimate whether a generated text contains grammatical and syntactic mistakes. These metrics are relatively well-defined and can be automated to some extent. At least 13 out of the 23 NLG papers in the study utilize one or more fluency metrics for evaluation.

Topic/style/genre matching; these metrics can also be automated, typically relying on a pretrained classifier, as seen in (Ficler and Goldberg, 2017). 12 papers in the study use one or more evaluation criteria of this type.

Coherence; this group of metrics is more arbitrary, with at least three major types of coherence evaluation approaches. First, some estimate coherence on a linguistic pragmatics level, focusing on coherent causal statements that include words

[2] https://github.com/markriedl/WikiPlots

like "hence/so/thus/etc." The second approach evaluates whether the generated text aligns with the reader's general world knowledge. These questions are more subjective, especially since fictional texts often describe alternative realities[3]. Lastly, the most abstract methods assess if the text is coherent within the internal logic of the "world" it describes. This high level of abstraction leads to greater misalignment between human annotators and lower potential for automated evaluation.

Even this brief overview demonstrates that there is no consensus on the coherence evaluation, yet 10 out of the 23 papers in (Hämäläinen and Alnajjar, 2021) used coherence evaluation understanding the term 'coherence' differently. However, there is a trend that might solidify the understanding of coherence in the field and move it towards the third line of reasoning that we described above, namely, coherence within the internal logic of the "world" that the text describes. The arrival of large language models that can process longer sequences of text brings to light a recursive approach to narrative generation, see (Yang et al., 2022). The idea to generate the outline of the story first and then extend separate blocks of the story while keeping some necessary information in the prompt to control coherence seems promising. Similarly, (Goldfarb-Tarrant et al., 2020) suggest an approach that combines overall story planning, generative language model and an ensemble of scoring models that each implement an aspect of good story-writing.

Overall emotional effect; these metrics are more challenging to automate, as they rely on human emotional response. However, with enough human labels, it is possible to train a classifier for this task. 11 out of the 23 papers in the study utilize some form of emotional effect evaluation.

Novelty/originality/interestingness; these metrics are even more difficult to formalize and automate. Most papers that ask human labelers to assess interestingness imply a certain level of novelty. Nevertheless, human labelers may interpret interestingness as a topic-related category. 7 out of 23 papers in the review use human evaluation of novelty.

The first two evaluation types dominate automated evaluation methods, while coherence and novelty are seldom assessed rigorously. Numerous NLG papers employing automatic evaluation fall within these five categories, emphasizing our limited tools for evaluating generated narratives.

Coherence is something humans can intuitively estimate, but it is notoriously difficult to automate. Meanwhile, we still struggle to understand even the most basic tools, such as semantic similarity metrics for short texts, as seen in (Yamshchikov et al., 2021; Solomon et al., 2021).

Novelty depends on a deeper understanding of semantics, and it may entail an additional layer of complexity. After all, human experience typically suggests that comprehending something presented to us is less challenging than creating something new from scratch.

In summary, we must conclude that among the five groups of metrics used in human evaluation, first two could be automated yet hardly advance our understanding of narrative, while three others could hardly be fully automated and applied to narrative evaluation. They are either automated but operate on a lower level with shorter texts or address high-level conceptual questions that are not quantified in a manner that permits automatic evaluation. This surprising realization leads us to the following logical conclusion: we cannot explain to humans how to evaluate a narrative. Despite the existence of literary criticism, narratology that represents a separate scientific field and a variety of approaches proposed in NLP, i.e. (Castricato et al., 2021), we still lack a universal formalized understanding of what a narrative is and how to assess it. Let us discuss this in the further subsection.

3.3 Concepts

In a review paper, (Gervás et al., 2019) authors present a compelling argument that the concept of storytelling encompasses a diverse set of operations. These operations are sometimes executed independently to create simple stories or specific story components, while other times they are combined to produce more complex narratives. The authors propose "deconstructing" storytelling into the following approaches: stories as narrative structures; stories as simulations; stories as evolving networks of character affinity; stories as narrations of observed facts; and stories as suspense-driven entertainment.

Upon closer examination, the proposed taxonomy reveals similar issues to those encountered in the evaluation process. There are no universally agreed-upon mechanisms for narrative representa-

[3] Still, we intuitively understand that some science fiction or fantasy novels are coherent, even if not realistic.

tion with high coherence among human labelers. Most methods are either deeply subjective (such as the well-known anthology of four plots first presented in (Borges, 1972)) or extremely low-level, working for causal inference on a short time scale but unable to extend to the level of a short story, let alone a novel.

It is essential to emphasize that each conceptual approach can yield practical results. However, there is no clear understanding of how these approaches structure the broader field of narrative processing, which we argue should be the primary focus of the NLP and AI communities in the near future. Is one approach sufficient to develop new models capable of generating entertaining stories? Do we need a combination of these pipelines? Should there be qualitative and quantitative interactions between these pipelines, and if so, how should they be organized? Finally, there is a set of even more general question. For example, could be have a narrative representation that would ne non-textual? What are independent properties of such representation if it exists? How one could quantify them? We hope this position paper could help intensifying the discussion of these questions.

4 Conclusion

This position paper puts forth two primary assertions:

- The generation of novel, entertaining narratives is a crucial task that could propel the progress of artificial intelligence across various fields and industries.
- Despite the critical importance of this task, the current NLP and AI communities are far from reaching a shared understanding of suitable datasets for narrative generation, appropriate evaluation methods, and the need for rigorous definition of concepts to address these problems effectively.

We hope this paper stimulates further discussion on these topics and attracts the attention of the NLP and AI community towards the challenges surrounding narrative generation.

Limitations

This is a position paper thus we do not see what the potential limitations could be. The only potential limitation might be the incompleteness of the list of relevant publications.

Ethics Statement

This paper complies with the ACL Ethics Policy. We have used generative AI for editing of the final text of the paper, since some of the authors might not be native speakers of English.

References

Yana Agafonova, Alexey Tikhonov, and Ivan P Yamshchikov. 2020. Paranoid transformer: Reading narrative of madness as computational approach to creativity. *Future Internet*, 12(11):182.

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J Martin, and Mark O Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7375–7382.

Payal Bajaj, Chenyan Xiong, Guolin Ke, Xiaodong Liu, Di He, Saurabh Tiwary, Tie-Yan Liu, Paul Bennett, Xia Song, and Jianfeng Gao. 2022. Metro: Efficient denoising pretraining of large scale autoencoding language models with model generated signals. *arXiv preprint arXiv:2204.06644*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in english literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54.

Jorge Luis Borges. 1972. El oro de los tigres. In *Emece, Buenos Aires*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and others (OpenAI). 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Louis Castricato, Stella Biderman, Rogelio E Cardona-Rivera, and David Thue. 2021. Towards a formal model of narratives. *arXiv preprint arXiv:2103.12872*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Richard Doust and Paul Piwek. 2017. A model of suspense for narrative generation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 178–187.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104.

Pablo Gervás, Eugenio Concepción, Carlos León, Gonzalo Méndez, and Pablo Delatorre. 2019. The long path to narrative generation. *IBM Journal of Research and Development*, 63(1):8–1.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338.

Mika Hämäläinen and Khalid Alnajjar. 2021. Human evaluation of creative nlg systems: An interdisciplinary survey on recent papers. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 84–95.

Z Harris. 1954. Distributional hypothesis. *Word*, 10(23):146–162.

Sudipta Kar, Suraj Maharjan, A Pastor López-Monroy, and Thamar Solorio. 2018. Mpst: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Richard Koncel Kedziorski. 2019. *Understanding and Generating Multi-Sentence Texts*. Ph.D. thesis.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*.

Anastasia Malysheva, Alexey Tikhonov, and Ivan P Yamshchikov. 2021. Dyplodoc: Dynamic plots for document classification. In *Modern Management based on Big Data II and Machine Learning and Intelligent Systems III*, pages 511–519. IOS Press.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. Inscript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3485–3493.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Andreea I Niculescu. 2021. Brief considerations on the phenomenon of humor in hci. In *Asian CHI Symposium 2021*, pages 152–156.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. Mcscript2. 0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 103–117.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. 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.

Vladimir Propp. 1968. Morphology of the folktale, trans. *Louis Wagner, 2d. ed.*

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 979–988.

Viktor Shklovsky. 1925. Theory of prose (b. sher, trans.). *Champaign, IL: Dalkey Archive Press. Original work published.*

Shaul Solomon, Adam Cohn, Hernan Rosenblum, Chezi Hershkovitz, and Ivan P Yamshchikov. 2021. Rethinking crowd sourcing for semantic similarity. *arXiv preprint arXiv:2109.11969*.

Allison C Tam, Neil C Rabinowitz, Andrew K Lampinen, Nicholas A Roy, Stephanie CY Chan, DJ Strouse, Jane X Wang, Andrea Banino, and Felix Hill. 2022. Semantic exploration from language abstractions and pretrained representations. *arXiv preprint arXiv:2204.05080*.

Alexey Tikhonov, Igor Samenko, and Ivan Yamshchikov. 2021. Storydb: Broad multi-language narrative dataset. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 32–39.

A. Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433.

Teun A Van Dijk. 1976. Philosophy of action and theory of narrative. *Poetics*, 5(4):287–338.

Judith van Stegeren and Mariët Theune. 2019. Narrative generation in the wild: Methods from nanogenmo. In *Proceedings of the Second Workshop on Storytelling*, pages 65–74.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R Bowman. 2022. Squality: Building a long-document summarization dataset the hard way. *arXiv preprint arXiv:2205.11465*.

Pengcheng Wang, Jonathan P Rowe, Wookhee Min, Bradford W Mott, and James C Lester. 2017. Interactive narrative personalization with deep reinforcement learning. In *IJCAI*, pages 3852–3858.

Lilian Diana Awuor Wanzare, Michael Roth, and Manfred Pinkal. 2019. Detecting everyday scenarios in narrative texts. In *Proceedings of the Second Workshop on Storytelling*, pages 90–106.

Karl Wiig. 2012. *People-focused knowledge management*. Routledge.

Ludwig Wittgenstein. 1921. Logisch-Philosophische Abhandlung. *Annalen der Naturphilosophie*.

Jon Womack and William Freeman. 2019. Interactive narrative generation using location and genre specific context. In *International Conference on Interactive Digital Storytelling*, pages 343–347. Springer.

Ivan P Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14213–14220.

Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. 2022. Narrasum: A large-scale dataset for abstractive narrative summarization. *arXiv preprint arXiv:2212.01476*.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. Designing effective sparse expert models. *arXiv preprint arXiv:2202.08906*.