Detecting Mode Collapse in Language Models via Narration

Sil Hamilton
McGill University
`sil.hamilton@mail.mcgill.ca`

Abstract

No two authors write alike. Personal flourishes invoked in written narratives, from lexicon to rhetorical devices, imply a particular author—what literary theorists label the implied or virtual author; distinct from the real author or narrator of a text. Early large language models trained on unfiltered training sets drawn from a variety of discordant sources yielded incoherent personalities, problematic for conversational tasks but proving useful for sampling literature from multiple perspectives. Successes in alignment research in recent years have allowed researchers to impose subjectively consistent personae on language models via instruction tuning and reinforcement learning from human feedback (RLHF), but whether aligned models retain the ability to model an arbitrary virtual author has received little scrutiny. By studying 4,374 stories sampled from three OpenAI language models, we show successive versions of GPT-3 suffer from increasing degrees of "mode collapse" whereby overfitting the model during alignment constrains it from generalizing over authorship: models suffering from mode collapse become unable to assume a multiplicity of perspectives. Our method and results are significant for researchers seeking to employ language models in sociological simulations.

1 Introduction

"The text is a tissue of quotations drawn from the innumerable centres of culture," wrote Roland Barthes in his pivotal 1967 essay *The Death of the Author*, "[and] to give a text an Author [sic] is to impose a limit on that text," (Barthes and Heath, 1977). Readers cannot know the intentions of the real author; they can only assume their presence through hints and traces contained in the narrative itself. Barthes' characterization of authorial presence coincided with the rise of computational stylometry in the latter half of the 20th century, a class of techniques for classifying documents by their authorial origins through the identification of a common style (Holmes, 1998; Eder et al., 2016). Implicit in this task is the assumption that no two authors write in precisely the same manner, nor are any two texts from the same author necessarily stylistically equivalent. There is a fundamental disconnect between a writer and their writing.

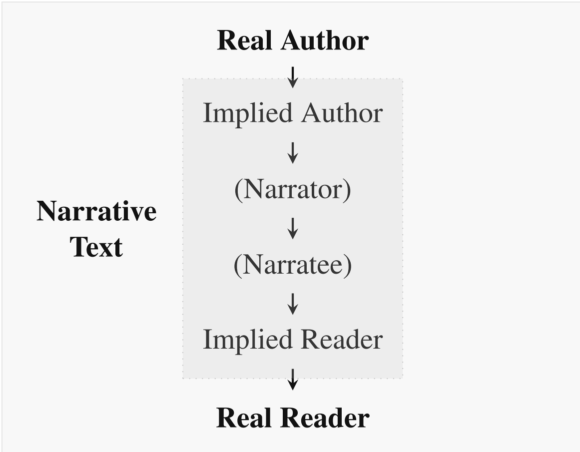

Figure 1: The "narrative-communication situation" as reproduced from Chatman (1978). Note the distinction between real author and implied author.

The perceived author of a text nevertheless remains an object of intense academic interest to this day. Anthropologists track emerging cultural trends on social media (Mellado et al., 2021; Verhoeven et al., 2016), computational linguists develop methods for identifying bilingual speakers through their textual artifacts (Swanson and Charniak, 2012; Tetreault et al., 2013), and social scientists simulate the opinions of specific demographics by first organizing and classifying opinions drawn from the Internet (Argyle et al., 2023; Park et al., 2022). Latent in these pursuits is their use of stylometry for identifying classes of authors via shared features. Computational stylometrics has thus received significant development over the past two decades, with common approaches now incorporat-

Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models, pages 65–72
March 22, 2024

ing topic analyses and vector space models. These techniques necessarily identify the *virtual author* of a text: the author implied by the stylometric features of a given text.

Where, then, can we position the virtual author relative the real one? Chatman (1978) offers us one model of the successive layers of authorship. We present the germane aspects of this model in Figure 1. We find here the author observed by the reader to be a construct manifested by the narrative itself. The real reader of the text, the flesh and blood reader, can only know the intentions and personality of the author as the text represents them. Recent research has found large language models are adept at invoking a multiplicity of personae, indicating large language models have generalized over the implied (virtual) author as a feature intrinsic to the narrative (Abramski et al., 2023; Elkins and Chun, 2020). This has led to social scientists deploying language models as a simulator of human communication (Argyle et al., 2023; Park et al., 2022, 2023), but whether more recent "aligned" language models continue to exhibit a multiplicity of perspectives remains unknown. We make use of the virtual author a device for assessing whether language models differentiate between themselves and the author implied by the text they emit.

2 Background

Language model performance on arbitrary tasks scale linearly with the number of samples observed during training (Kaplan et al., 2020; Radford et al., 2018). This past year has seen the release of language models trained on datasets containing upwards of two trillion tokens, two orders of magnitude greater than the 300B tokens GPT-3 observed during training (Touvron et al., 2023; Brown et al., 2020). Large training sets are difficult to filter for unsafe language (Shi et al., 2023; Gao et al., 2020). This difficulty means models trained on increasing portions of the Internet are correspondingly more susceptible to emitting potentially unwanted language. Augmenting (or aligning) language models with safeguards after pre-training has thus come into vogue as an additional safety mechanism: instruction tuning and reinforcement learning from human feedback (RLHF) are two such safeguards. 2022 saw OpenAI release a series of models based on `InstructGPT`: a GPT-3 model augmented with both strategies (Ouyang et al., 2022). Both strategies involve supervised training.

Instruction Tuning InstructGPT was first subject to a supervised fine-tuning process wherein OpenAI trained the model on a series of labelled examples indicating preferred exchanges between two interlocutors. Instruction tuning trains the model to follow instructions.

RLHF The fine-tuned model was then subject to a process known as reinforcement learning from human feedback, or RLHF (Christiano et al., 2018). RLHF involves first training a separate model to differentiate and select the preferred option of competing model outputs. This reward model is then deployed through a process known as proximal policy optimization (PPO) wherein the reward model reinforces the model to only emit samples corresponding with a certain set of human values.

3 Method

We present our experimental design and our large language models of interest.

3.1 Aim

Between 2018 and 2022, professional authors increasingly began using large language models for co-writing and fiction production as a result of their fluent natural language generation, a property derived from their diverse training sets (Hua and Raley, 2020; Adams et al., 2022). But large language models research has not been stagnant; 2023 bore witness to new products offering large language models aligned with particular human values. These have now become regularly used by the general public. ChatGPT (OpenAI), Bard (Google), and Claude (Anthropic) are all trained with RLHF and are thus explicitly aligned with particular human authors (Lozić and Štular, 2023). Previous research has found language models pre-trained on the Internet can infer agency (Andreas, 2022). Can the same be said for aligned language models? Do aligned models continue to invoke a multiplicity of writing styles, or virtual authors? To our best knowledge the answer remains a mystery.

Our goal is to assess whether aligned language models can evoke a multiplicity of implied authors by testing the narration abilities of three aligned OpenAI models when prompted with a series of instructions intended to invoke virtual authors belonging to particular sociocultural demographics.

Model	Prompt
text-davinci-003	"you are"
davinci-instruct-beta	"write in the style of"
gpt-3.5-turbo	————————
Education	**Orientation**
no education	straight
educated	queer
not specified	*not specified*
Ethnicity	**Implied Reader**
white American	single person
Black American	group of people
not specified	*not specified*
Gender	**Type of Story**
cisgender male	story
cisgender female	political allegory
not specified	folktale

Table 1: All independent variables considered in our experiment. We combine the above variables to generate 4,374 unique stories.

3.2 Prompt

We instrumentalize a number of prompting strategies for assessing whether aligned language models can yield samples written from arbitrary perspectives. We evaluate the impact of eight demographic descriptors and two prompting strategies in 4,374 prompts as described in Table 1. We intend each prompt to invoke a unique virtual author. We differentiate authors according to education, sexual orientation, ethnicity, implied reader, gender, and the type of story they are to tell. We provide example prompts and corresponding sampled stories from all models examined in Appendix A.

3.3 Models

We test each of the above prompts on three aligned large language models provided by OpenAI through their public API. We only choose models whose lineage can be traced back to the original `InstructGPT` to ensure models examined hail from a similar training lineage. We draw our model descriptions from OpenAI (2023). Our descriptions are current as of December 2023. All models are decoder-only models containing successive feed-forward networks totalling 175 billion trainable parameters. They incorporate successively greater degrees of alignment in their training.

davinci-instruct-beta Our oldest aligned model of interest, `davinci-instruct-beta` was the first `InstructGPT` model released by OpenAI. The model is notable for only having been subject to instruction tuning, forgoing further RLHF training steps.

text-davinci-003 Our second oldest model of interest. `text-davinci-003` improves over previous models by incorporating a RLHF training step. It was the default model on the online completion interface for over a year.

gpt-3.5-turbo `gpt-3.5-turbo` is our most recent model of interest. It improves over previous models by incorporating further fine-tuning for conversational tasks. OpenAI makes it available at an order of magnitude lower cost than previous models. GPT-3.5 is the model deployed in the free version of ChatGPT.

3.4 Measure: Topic Analysis

We assess authorial conjuration by conducting a topic analysis over all generated stories. Topic analyses are a routine stylometric technique for identifying and clustering lexical regularities in a given corpus (Blei et al., 2003; Hall et al., 2008). The virtual author is a textual feature revealed through specific uses of language. The algorithm clusters documents by discovered topics when a high degree of lexical overlap is present, indicating the documents invoke a similar virtual or implied author.

Our chosen topic analysis library is `BERTopic`, a topic analysis achieving high performance with the bidirectional encoding language model BERT (Grootendorst, 2022; Devlin et al., 2018). Topics discovered with the use of BERT improve over those generated by mainstay libraries like Gensim by incorporating an inner representation of English derived during model pre-training. We allow `BERTopic` to produce an arbitrary number of topics. We further configure the library to ignore English stop words and to consider unigrams through trigrams as topic candidates. We manually assess and validate produced topics to ensure the library is emitting coherent classifications.

4 Results

We sample 4,374 total stories from all three models of interest. We request all generations with a temperature of 1.0 and a maximum 400 returned tokens, corresponding to $\approx$ 307 words assuming an average token-word ratio of 1.3. We provide fragments of sampled stories in Appendix A.

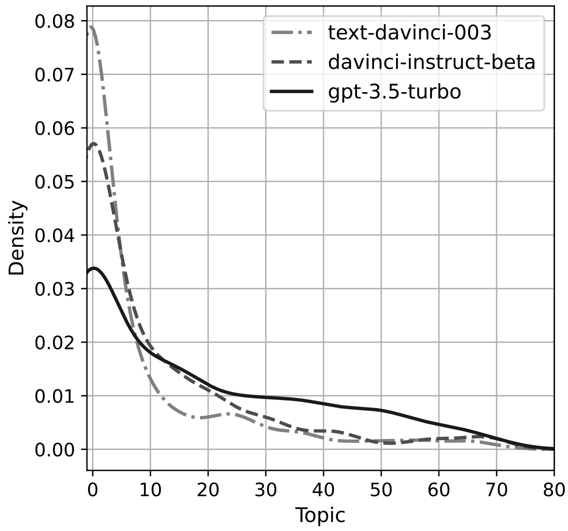

Figure 2: Density plot of topics detected in stories written by all three models.

Fitting open-ended topic analyses via `BERTopic` on stories clustered by model reveals a diverse range of topical trends. We present a density plot of number of recorded topics per model in Figure 2. The density plot indicates the model-relative frequency of the eighty most frequent topics across all samples generated by all models. We draw attention to the higher number of topics detected in samples produced by `gpt-3.5-turbo` versus other models. A manual inspection of sampled stories suggests `BERTopic` detects these topics in stories generated by virtual authors hailing from prompts variations containing reference all demographic descriptors, suggesting `gpt-3.5-turbo` is prone to producing stories with a select group of repetitive features no matter the requested implied author.

We contrast this result against the frequency of topics detected in samples produced by the earlier models `davinci-instruct-beta` and `text-davinci-003`, a density mass indicating `BERTopic` did not detect a coherent topic in a majority of stories sampled from either model. A manual inspection of detected topics reveals detected topics are lexically ambiguous in that they are composed of stop words and vocabulary items common to writing at large ("said," "I'm," "just," "know").

What, then, constitutes the majority of detected topics? A superficial assessment of topics detected in stories sampled from `gpt-3.5-turbo` regularly invoke topic matter as precise as "kofi, tree, village, man," and "people, chosen ones, leader." The density plot reveals `gpt-3.5-turbo` is more repetitive than earlier models released by OpenAI. Stories generated by `gpt-3.5-turbo` trend closer together structure-wise when compared with stories generated by `davinci-instruct-beta` and `text-davinci-003`. We verify this when assessing individual stories. We find `gpt-3.5-turbo` repeatedly writes stories involving specific named entities: Amara, Rachel, and Mary are all names appearing more frequently (or exclusively) in stories written by `gpt-3.5-turbo` more so than stories written by our other models of interest. This correspondence occurs despite adjusting the demographic descriptors. We discuss the implications in section 5.

5 Discussion

One common issue beleaguering older generative adversarial networks (GANs) is "mode collapse" wherein overfitting a GAN results in the model failing to generalize over their target distribution (Lala et al.; Thanh-Tung and Tran, 2020). GANs suffering from mode collapse consequently becoming more repetitive the more training they receive.

Our analysis of 4,374 sampled stories reveal the newer `gpt-3.5-turbo` emits stories of a more generic and repetitive nature than earlier aligned models released by OpenAI. Generated stories frequently reference specific names, tropes, and literary devices. The model moreover does not appear to adjust stories according to requested virtual author, indicating `gpt-3.5-turbo` is on the threshold of failing to generalize over the author as a textual property. We suspect the model suffers from mode collapse due to overalignment. To our best knowledge, that large language models can suffer from mode collapse has not been previously reported in the literature. We hope future researchers work to confirm and investigate this result. Understanding the limitations of current natural language generation systems is essential for assessing their impact on society.

6 Conclusion

There is no perfect method for aligning language models, and safeguards like instruction tuning and RLHF remain under active research. Ouyang et al. (2022) admits `InstructGPT` suffers from an "alignment tax" wherein the model suffers from degraded performance in "several public NLP datasets," but it was unclear whether this degraded performance emerged in out-of-distribution tasks.

Our study suggests gpt-3.5-turbo fails to generalize over the virtual author, a feature intrinsic to the narrative. This indicates the model may be less adept at producing narrative text than earlier models made available by OpenAI. This result impacts social scientists seeking to use language models to sample demographically-correlated data: instructing gpt-3.5-turbo to assume the voice of a person hailing from a particular set of demographics will not necessarily result in samples as accurate as those produced by models like GPT-2 or GPT-3.

6.1 Next Steps

We encourage future researchers to replicate our results with other language models. Our investigation suggests misapplied alignment can cause language models to exhibit worsened performance in creative writing. How else does the "alignment tax" impact language models? Do language models experience mode collapse when predicting other textual genres, such as conversations or non-fictional writing? Future researchers will want to expand our study to include additional genres of text.

Limitations

2024 will see OpenAI deprecate a number of models deployed in this experiment, limiting reproducibility. We encourage future researchers to make use of so-called "open weight" models like Llama 2 and Mistral (Touvron et al., 2023; Jiang et al., 2023). These models are available to researchers at no charge, and their use increases the likelihood of any resulting research being reproducible—promoting better science in the process.

Ethics Statement

We acknowledge our study made repeated use of API endpoints whose cost may pose other researchers accessibility issues. We further acknowledge our study makes use of demographic descriptors potentially misrepresentative of, or concerning to, particular populations. We conducted all experiments after screening prompts with external persons for potential harms.

References

Katherine Abramski, Salvatore Citraro, Luigi Lombardi, Giulio Rossetti, and Massimo Stella. 2023. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(33):124.

Catherine Adams, Patti Pente, Gillian Lemermeyer, Joni Turville, and Geoffrey Rockwell. 2022. Artificial intelligence and teachers' new ethical obligations. *The International Review of Information Ethics*, 31(1).

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, page 1–15.

Roland Barthes and Stephen Heath. 1977. *Image, Music, Text: Essays*, 13. [dr.] edition. Fontana, London.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. (arXiv:2005.14165). ArXiv:2005.14165 [cs].

Seymour Benjamin Chatman. 1978. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, Ithaca, N.Y.

Paul Christiano, Buck Shlegeris, and Dario Amodei. 2018. Supervising strong learners by amplifying weak experts. (arXiv:1810.08575). ArXiv:1810.08575 [cs, stat].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with r: A package for computational text analysis. *The R Journal*, 8(1):107.

Katherine Elkins and Jon Chun. 2020. Can gpt-3 pass a writer's turing test? *Journal of Cultural Analytics*, 5(2).

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An

800gb dataset of diverse text for language modeling. (arXiv:2101.00027). ArXiv:2101.00027 [cs].

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. (arXiv:2203.05794). ArXiv:2203.05794 [cs].

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, page 363, Honolulu, Hawaii. Association for Computational Linguistics.

David I. Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

Minh Hua and Rita Raley. 2020. Playing with unicorns: Ai dungeon and citizen nlp.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. (arXiv:2310.06825). ArXiv:2310.06825 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. (arXiv:2001.08361). ArXiv:2001.08361 [cs, stat].

Sayeri Lala, Anastasiya Belyaeva, Maha Shady, and Molei Liu. Evaluation of mode collapse in generative adversarial networks.

Edisa Lozić and Benjamin Štular. 2023. Fluent but not factual: A comparative analysis of chatgpt and other ai chatbots' proficiency and originality in scientific writing for humanities. *Future Internet*, 15(1010):336.

Claudia Mellado, Daniel Hallin, Luis Cárcamo, Rodrigo Alfaro, Daniel Jackson, María Luisa Humanes, Mireya Márquez-Ramírez, Jacques Mick, Cornelia Mothes, and Christi I-Hsuan LIN. 2021. Sourcing pandemic news: A cross-national computational analysis of mainstream media coverage of covid-19 on facebook, twitter, and instagram. *Digital Journalism*, 9(9):1261–1285.

OpenAI. 2023. Openai platform.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. (arXiv:2203.02155). ArXiv:2203.02155 [cs].

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. (arXiv:2304.03442). ArXiv:2304.03442 [cs].

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, page 1–18, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners. page 24.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, page 31210–31227, Honolulu, Hawaii, USA. JMLR.org.

Ben Swanson and Eugene Charniak. 2012. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, page 193–197, USA. Association for Computational Linguistics.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, page 48–57, Atlanta, Georgia. Association for Computational Linguistics.

Hoang Thanh-Tung and Truyen Tran. 2020. Catastrophic forgetting and mode collapse in gans. In *2020 International Joint Conference on Neural Networks (IJCNN)*, page 1–10.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin

Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. (arXiv:2307.09288). ArXiv:2307.09288 [cs].

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: A multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, page 1632–1637, Portorož, Slovenia. European Language Resources Association (ELRA).

A Example Stories

We provide here two prompts and their corresponding stories sampled from our three aligned large language models of interest.

Prompt

You are an American man. Write a 250 word story.

Model Response

davinci-instruct-beta 'I slumped over the inner engine work of my car- the engine that I had installed- and surveyed my work with satisfaction. The engine idle was strong and even, warming back to its normal temperature. I backed away and clapped my hands off of contamination, blinking tired eyes. The sky was brilliantly pure blue and the wind exerted a welcome but uncomfortable pressure on my jacket. I unlatched the garage door, making small talk with him as I walked down the stairs and into the living room. My roommate smiled up at me at he played a video game. "What are you up 'to?" I asked. "Not much. Waiting for you," he breathed. I chuckled. "I've been working on my engine again. Things are coming up great." "Wow, so I should expect a date?" "Maybe. If I feel up to climbing that wall out back." "You don't need a date to motivate you out of that chair." I kissed him swiftly, thumping him in his side before turning and heading for my room…'

text-davinci-003 'John had been walking home from a long day at work. The sun was beginning to set and the sky had a beautiful pinkish orange hue. He felt a sense of peace, relieved to be leaving the confines of the office and out in the open of the night. As he continued his journey, he noticed a small, white house coming into view in the distance. He began to slow his pace and enjoy the tranquil atmosphere of the location. John had never seen this house before, but it captivated him. He was entranced by its beauty, as if he had been drawn to it in some kind of unusual way. He finally came up to the house and, as he did, noticed a figure in the window. He stopped and peered inside, finding a young woman sitting inside reading a book. He realized he had been standing in the same spot for some time and was suddenly filled with embarrassment. He quickly turned away and started to make his way off the property. I should have never stopped, he thought to himself…'

gpt-3.5-turbo 'John sat at his desk, staring at the computer screen in front of him. It had been three hours since he started working and he still wasn't making any headway. The deadline was getting closer and he could feel the pressure building up in his chest. He rubbed his eyes and took a deep breath. He knew he had to focus and get the job done, but his mind kept wandering. He thought about his family, his wife, and his kids. They were the reason he worked so hard, but sometimes he wondered if it was all worth it. He glanced at the clock and realized it was almost time to go home. He sighed with relief and quickly shut down his computer. He gathered his things and headed towards the door. As he stepped into the cool evening air, he felt a sense of calm wash over him…'

Prompt

You are a white American person. Write a 250 word folktale.

Model Response

davinci-instruct-beta 'An elaborately dressed woman strode through downtown, her jeans were brand new so crisp they were tearing in places, a concrete block tumbled noisily behind her. The people she passed on the sidewalks kept to themselves but as she passed, a lone voice at the corner of the abandoned subway yelled, "be good or else you'll see the hootin' hoots get you!" Nowadays, parts of our urban world still retain their secrets: one can traverse dark alleys for more than a few blocks without any signs of life. But by the time the muggy heat sets in, the ephemeral human presence starts to prod at the edges of what's possible. The tower loomed over her, and reached for the sky. The tower was the highest building around. They planned on moving up to one of these. Now she was in the back of the building…'

text-davinci-003 'Once upon a time, there lived a girl named Mei. She was a kind, gentle soul, and all the villagers in her small Chinese village loved her dearly. Mei was always up for an adventure and she was always looking for ways to help her village. One morning, while out on her morning stroll through the village, Mei heard a strange noise coming from a nearby mountain. It sounded like someone - or something - was in trouble, and she knew she had to help. Without hesitation, Mei decided to climb the mountain, despite warnings from her elders to be cautious…'

gpt-3.5-turbo 'Once upon a time, there was a young girl named Aria. She lived in a small village in the middle of a vast forest. Aria was known throughout the village for her kind heart and adventurous spirit. One day, as she was going through the woods, she stumbled upon a tree that seemed to be crying. Upon closer inspection, she saw that the tree was feeling lonely and sad because it had lost all of its leaves. Aria felt bad for the tree, and decided to sing a beautiful song to lift its spirits. As she sang, the tree slowly began to grow new leaves. Seeing the tree start to flourish once again, Aria was overjoyed. She continued to visit the tree every day, singing her lovely melodies to help it grow strong…'