Towards an Authorial Leverage Evaluation Framework for Expressive Benefits of Deep Generative Models in Story Writing

Sherol Chen
Google Research
California, USA
sherol@google.com

Carter Morgan
Carter Morgan Comedy
New York, USA
carter@carterthecomic.com

David Olsen
Aeiouy Media & Entertainment
California, USA
dolsen@aeiouy.org

Ethan Manilow
Google Research
California, USA
emanilow@google.com

Kenric Allado-McDowell
Google Research
California, USA
kenric@google.com

Mark J. Nelson
American University
Washington, DC, USA
mnelson@american.edu

Qiuyi (Richard) Zhang
Google Research
California, USA
qiuyiz@google.com

Senjuti Dutta
University of Tennessee, Knoxville
Tennessee, USA
sdutta6@vols.utk.edu

Kory Mathewson*
Piotr W Mirowski*
korymath@deepmind.com
piotrmirowski@deepmind.com
Deepmind

ABSTRACT

What are dimensions of human intent, and how do writing tools shape and augment these expressions? From papyrus to auto-complete, a major turning point was when Alan Turing famously asked, "Can Machines Think [30]?" If so, should we offload aspects of our thinking to machines, and what impact do they have in enabling the intentions we have? This paper adapts the Authorial Leverage framework [5], from the Intelligent Narrative Technologies literature, for evaluating recent generative model advancements. With increased widespread access to Large Language Models (LLMs), the evolution of our evaluative frameworks follow suit. To do this, we discuss previous expert studies of deep generative models for fiction writers [6, 34] and playwrights [16], and propose both author- and audience-focused directions for furthering our understanding of Authorial Leverage of LLMs, particularly in the domain of comedy writing.

KEYWORDS

large language models, mixed initiative, generative ai, intelligent narrative technologies, NLP, NLU, computational creativity

1 INTRODUCTION: AUTHORIAL LEVERAGE AND MIXED INITIATIVE CO-CREATIVE INTERFACES

Storytelling is a core means of human expression. Narratologists have posed that "our need for narrative form is so strong that we don't really believe something is true unless we can see it as a story."[1] The tools we create alleviate the costs and efforts towards our goals, whether for survival, flourishing, or catharsis [10]. With the invention of new tools, evaluative frameworks become essential to help shape the utility and equity of impact. Early analysis of Authorial Leverage was studied within the artificial intelligence (AI) space of Intelligent Narrative Technologies [5]. In Figure 1, we summarize Authorial Leverage as a ratio between the experience of the audience (inclusive of the authorial intent for said experience) over the effort or cost to the author for adopting the use of a tool. As an example, we outline the qualities that inform the Authorial Leverage from former studies on Declarative Optimization-Based Drama Management in Figures 1, 2, and 3 [5].

$$\text{Authorial Leverage} = \frac{\text{Audience Experience}}{\text{Authorial Effort}}$$

Figure 1: We summarize Authorial Leverage as a ratio between the experience of the audience (inclusive of the authorial intent for said experience) over the effort or cost to the author for adopting the use of a tool [5].

$$\text{Authorial Leverage} = \frac{\text{Quality} \times \text{Variations} \times \text{Control}}{\text{Authorial Effort}}$$

Figure 2: Values that inform the Authorial Leverage from former studies on Declarative Optimization-Based Drama Management [5].

1.1 Examples of Rule-based Expressive AI Tools for Writing

While the history of AI assisted storytelling spans the lifetime of AI [2], many of the earlier known systems were built in 1990's to 2005, when the AI-driven interactive drama, Facade was featured in the NY Times [25]. Studies primarily involved audience evaluations with some authorial analysis. Authorial Leverage was designed to examine whether the gains from any tool, specifically AI, demonstrated measurable benefit, while considering effort in tool design, development, and deployment. Below we list a variety of rule-based writing studies.

*Last authorship is shared equally

- Authorial tools such as Wide Ruled [27] are examples of building a tool to help non-programmers create generative stories by providing help, an interface and suggestions.
- AI tools such as Bad News [23] that provide a framework and state that setup a scenario for improving theatre given providing a middleware between participants all while providing a scenario to act in.
- Studies of narrative logic [20] analyzed how Rube Goldberg comics were written and then a system was developed to generate the captions, similar to the comic, given initial conditions from the late 1800's.
- Beep Beep Boom Boom [19] had the user interact by placing objects in the world, taking the role of a virtual cartoonist, and the system had to create a scenario where the a humorous narrative would be created but still staying true to the underlying premise.

Of the various domains, the comedy writing has had a breadth of annotation and formalization. Joke generators (and related genres like automated pun generation) have used a variety of techniques [22, 31]; recent work, like with most applications of natural language generation, uses large generative language models [29, 33]. From the perspective of interaction with human comedians and comedy writers, a joke generator can be useful as a brainstorming tool, but generally we see this as only a starting point for richer modes of interaction and collaboration. As such, choosing a particular domain, like comedy, enables further understanding of shorter-form storytelling in spaces where quality can be highly subjective.

1.2 AI Enabled Expressive Tools and Deep Generative Models

AI for storytelling has had both a rule-based [23, 27] and a data-driven history [26]. With the increased advancement of Deep Learning in the 2010's, the research community moved past automating simpler tasks, like spam detection, and towards aiding more complex expressive domains [18]. In the early 2020's, more generative AI studies centered on evaluating the experience of authors as expert users [6, 16, 34]. Directly, we can learn and build on top of the findings from Wordcraft [6, 34] and Dramatron [16], which both work towards understanding human-AI collaborative story writing, making use of "prompting techniques and UX patterns for interfacing with a large language model" [34]. Both studies found more domain-specific evaluation as a clear future direction. Specifically, we consider stand-up comedy writing, drawing from the history of computational humor [19, 22, 27, 31] to present day challenges in generative models of subjective spaces [8].

2 ETHICAL DISCUSSION AND FUTURE WORK

Expert user analysis allows us to consider the ethical risks along with possible mitigation strategies. For example, the Dramatron [16] study identifies 1) the reproduction of existing societal biases and the involuntary generation of offensive language, 2) copyright infringement and 3) undermining creative economies [32]. Focusing on the topic of Authorial Leverage, we consider a fourth risk concerning the writer's agency when collaborating with AI writing tools. Are writers offloading aspects of their thinking to machines, or do they use machines to be challenged, stimulated and inspired in their thinking? Comedy itself often straddles the line of appropriateness, a tool that provides useful suggestions may make suggestions on the wrong side of the line. Human decision making has long been a standard for AI [30]. Researcher Roger Schank describes this human process to be unavoidably calculated and purposeful, "forced to make their story acceptable and easily comprehensible" for (1) themselves and (2) others [24]. This bi-direction in purpose helps organize the goals and ethical considerations into two pursuits.

2.1 Direction 1: Decreasing Authorial Effort (interface and model adaptation)

In addition to understanding leverage through interface studies, we can investigate fine-tuning LLMs to adapt to different tasks, as well as to personalize to specific authorial preferences, which are essential to providing greater control and diversity in augmented storytelling responsibly [11]. Since these models are expensive to train, increasing the flexibility of such models to adapt to few-shot learning tasks, such as theme specification or style transfer, should be incorporated into the meta-learning process. Furthermore, such models would need to excel in an interactive feedback environment where the user may provide direct linguistic feedback on the model's performance. Applying active learning and Bayesian optimization approaches are promising for rapid fine-tuning. [9].

2.2 Direction 2: Improving Audience Experience (annotation, subjectivity, and data)

What is the overall quality of a performance and how can we understand these subjective values? LLMs generalize over large amounts of data, however it may be important to surface subjective representations from a full range of human opinions, perspectives, interpretations [3], and diversity of each audience member (such as personal and cultural background [21]). In addition, it could be beneficial for LLMs to acquire subjective comedic notions by training models using subjective datasets, such as the multimodal comedy dataset [17]. Overall this may not only help the models in producing comedy-related augmented story telling for a diversity of audiences, but also allow for controls that create leverage for an author.

3 CONCLUSION

Screenwriting scholar, Robert McKee, describes writing as the process of delivering a lifetime through selecting a few moments [15]. How those moments are chosen has evolved from oral traditions to scribed to printed and now generative. The types of stories we tell are motivated by our understanding of the world around us [1] and the world we are trying to influence and inform [24]. In this paper, we proposed further Authorial Leverage directions following up on workshops previously done with LLMs [6, 16], and focusing in on the specific domain on comedy, as there are strong communities of expertise with additional gains (for generative models) from annotations in subjective spaces like humor.

REFERENCES

[1] H Porter Abbott. 2020. *The Cambridge introduction to narrative.* Cambridge University Press.

[2] Robert P Abelson. 1963. Computer simulation of" hot cognition". *Computer simulation of personality* (1963), 277–298.

[3] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion proceedings of the 2019 world wide web conference.* 1100–1105.

[4] Jean Benedetti. 2005. *Stanislavski: An Introduction, Revised and Updated.* Routledge.

[5] Sherol Chen, Mark J. Nelson, and Michael Mateas. 2009. Evaluating the Authorial Leverage of Drama Management.. In *Artificial Intelligence and Interactive Digital Entertainment.* 136–141.

[6] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a human-ai collaborative editor for story writing. *arXiv preprint arXiv:2107.07430* (2021).

[7] Simon Colton, Maria Teresa Llano, Rose Hepworth, John Charnley, Catherine V Gale, Archie Baron, François Pachet, Pierre Roy, Pablo Gervás, Nick Collins, et al. 2022. The beyond the fence musical and computer says show documentary. *arXiv preprint arXiv:2206.03224* (2022).

[8] Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[9] Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning BERT for low-resource natural language understanding via active learning. *arXiv preprint arXiv:2012.02462* (2020).

[10] Anna Huang, Sherol Chen, Mark Nelson, and Doug Eck. 2018. Mixed-Initiative Generation of Multi-Channel Sequential Structures. (2018).

[11] Zhe Liu, Ke Li, Shreyan Bakshi, and Fuchun Peng. 2021. Private Language Model Adaptation for Speech Recognition. *arXiv preprint arXiv:2110.10026* (2021).

[12] Kory Mathewson and Piotr Mirowski. 2018. Improbotics: Exploring the imitation game using machine intelligence in improvised theatre. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 14. 59–66.

[13] Kory Wallace Mathewson and Piotr Mirowski. 2017. Improvised Comedy as a Turing Test. *arXiv e-prints* (2017), arXiv–1711.

[14] Kory W Mathewson and Piotr Mirowski. 2017. Improvised theatre alongside artificial intelligences. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference.*

[15] Robert McKee. 1997. *Story: style, structure, substance, and the principles of screenwriting.* Harper Collins.

[16] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. *arXiv preprint arXiv:2209.14958* (2022).

[17] Anirudh Mittal, Pranav Jeevan, Prerak Gandhi, Diptesh Kanojia, and Pushpak Bhattacharyya. 2021. " So You Think You're Funny?": Rating the Humour Quotient in Standup Comedy. *arXiv preprint arXiv:2110.12765* (2021).

[18] Andrew Ng. 2017. HBR: What ai can and can't do. https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now

[19] David Olsen and Michael Mateas. 2009. Beep! Beep! Boom! Towards a Planning Model of Coyote and Road Runner Cartoons. In *Proceedings of the 4th International Conference on Foundations of Digital Games* (Orlando, Florida) *(FDG '09)*. Association for Computing Machinery, New York, NY, USA, 145–152. https://doi.org/10.1145/1536513.1536544

[20] David Olsen and Mark J. Nelson. 2017. The Narrative Logic of Rube Goldberg Machines. In *Interactive Storytelling*, Nuno Nunes, Ian Oakley, and Valentina Nisi (Eds.). Springer International Publishing, Cham, 104–116.

[21] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699* (2021).

[22] Graeme Ritchie. 2001. Current directions in computational humour. *Artificial Intelligence Review* 16 (2001), 119–135.

[23] Ben Samuel, James Ryan, Adam J. Summerville, Michael Mateas, and Noah Wardrip-Fruin. 2016. Bad News: An Experiment in Computationally Assisted Performance. In *Interactive Storytelling*, Frank Nack and Andrew S. Gordon (Eds.). Springer International Publishing, Cham, 108–120.

[24] Roger C. Schank and Robert P. Abelson. 1995. Knowledge and Memory: The Real Story. In *Knowledge and Memory: The Real Story*, Jr Robert S. Wyer (Ed.). Lawrence Erlbaum Associates, 1–85.

[25] Seth Schiesel. 2005. Redefining the power of the gamer. https://www.nytimes.com/2005/06/07/arts/redefining-the-power-of-the-gamer.html

[26] Mike Sharples and Rafael Pérez y Pérez. 2022. *Story Machines: How Computers Have Become Creative Writers.* Routledge.

[27] James Skorupski, Lakshmi Jayapalan, Sheena Marquez, and Michael Mateas. 2007. Wide Ruled: A Friendly Interface to Author-Goal Based Story Generation. In *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling*, Marc Cavazza and Stéphane Donikian (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 26–37.

[28] Ajitesh Srivastava and Naomi T. Fitter. 2021. A Robot Walks into a Bar: Automatic Robot Joke Success Assessment. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 2710–2716. https://doi.org/10.1109/ICRA48506.2021.9561941

[29] Joe Toplyn. 2022. Witscript 2: A System for Generating Improvised Jokes Without Wordplay. In *Proceedings of the International Conference on Computational Creativity*. 54–58.

[30] Alan M Turing. 2009. *Computing machinery and intelligence.* Springer.

[31] Tony Veale. 2021. *Your Wit Is My Command: Building AIs with a Sense of Humor.* MIT Press.

[32] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021). https://arXiv.org/abs/2112.04359

[33] Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 1650–1660.

[34] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

A ADDITIONAL CONSIDERATIONS FROM DRAMATRON'S EXPERT USER STUDY

While this paper describes two types of humans in the loop (author and audience), expressive domains, like theater employ human direction and participation from additional roles. Human Actors, for example, perform the discourse and delivery as part of the storytelling process.

Performances (live or recorded) of human actors alongside machines necessitate the consideration of multiple perspectives, namely the actors and the audiences. On the one hand, actors add interpretative layers to written text, from reconstructing the given circumstances of the scene to imagining subtext and objectives for the characters [4]. As such, actors can add or even create meaning for AI-generated text, as happened during the production of 2016 musical comedy *Beyond the Fence* [7]. On the other hand, audiences provide real-time, live feedback, such as laughter, which can be leveraged by a robotic actor / comedy writer [28].

The introduction of actors and audiences invites new instruments for evaluating human-computer interaction (HCI). For instance, AI-based improvised comedy *Improbotics* [14] was evaluated using a simplified Turing test in terms of audience interaction [13], and using HCI instruments in terms of actors' experience [12]. Recently-introduced *Dramatron* is an LLM-based interactive writing system for theatre scripts and screenplays. It was the subject of an extensive study with 15 industry professionals [16] that focused on evaluating the quality of the co-creative interaction between the playwright/screenwriter and the system, and also provided the playwright from actors' feedback following the public productions of full plays co-written with Dramatron. Theatre plays enable a complex interaction between audiences, actors, and script writers. This presents the potential to fill a gap of knowledge and evaluate the triangle actors-playwright-system from the point of view of audiences.

B CASE STUDY: STAND-UP COMEDY

Similar to screenplays, hierarchical prompt chaining could also facilitate the creation of a stand-up comedy special. There are a few areas where generating stand-up comedy diverges from generating movies: stand-up comedy specials do not follow a default structure as often as movies do. Some comics deliver an hour of one-liners, others deliver one long soliloquy on a specific theme, others tell a bunch of smaller stories, and there are many variations in-between. Despite the various forms, stand-up comedy continues to borrow narrative structures from film and television to as the structures of jokes evolve (https://pudding.cool/2018/02/stand-up/).

For example, studies can be done toward understanding whether the description of a comedian's style and material can generate comedy prompts. Specifically, the structure of a stand-up special, often referred to as bits, commonly follows the 3-act story format in film (https://www.comedy.co.uk/pro/inside_track/set-up-reveal-escalation-payoff/). Given the studies around humor (https://jonathansandling.com/script-based-semantic-theory-of-humour/), prompt chaining could be a useful tool to help comedy writers better craft their ideas, as well as produce and structure more effective material. (https://bigthink.com/high-culture/every-joke-falls-in-one-of-these-11-categories-according-to-the-founder-of-the-onion/, https://opensiuc.lib.siu.edu/cgi/viewcontent.cgi?article=1292&=&context=kaleidoscope&=&sei-redir=1, https://usir.salford.ac.uk/id/eprint/14688/)

C ROLES OF AI IN COMEDY

If one goal is to produce a truly funny LLM, we must define what is funny as well as what role a bot plays in a comedic interaction. Popular examples of LLM and LSTM-generated humor to date often play on the failure of bots to effectively understand context. For example, the surreal and absurd stage direction in the film Sunspring (https://www.youtube.com/watch?v=LY7x2Ihqjmc) are humorously interpreted by the actors. Similarly, Netflix's "Comedy Special Written Entirely by Bots" uses uncanny and stiff delivery of unfunny generated jokes by a crude virtual avatar to produce an anxious viewing experience that elicits laughter because it is not funny, at least in a conventional sense. In 2019, comedians like @KeatonPatti on Twitter posted scripts purported to have been generated by AI. (https://twitter.com/KeatonPatti/status/1138457675472220167). These were funny precisely inasmuch as they failed to produce conventional humor. We suggest that, while these AI-generated jokes may be funny in the context in which they are presented, they are not funny according to the standards commonly held for human comedians. To articulate what makes a bot funny, we need to define its comedic role. In the examples above, bots are playing the role of what might be called a foil, or historically, a "straight man". Other roles, like talk show host, guest, director, writer, or improviser in a game structure, might provide different methods of producing comedic AI output.

D EXPANDED QUOTES REFERENCED IN THIS PAPER

H. Porter Abbott, Narratologist, argues that humans are unavoidably storytellers.

> You could in fact argue, and people have, that our need for narrative form is so strong that we don't really believe something is true unless we can see it as a story [1].

Computer Scientist, Andrew Ng, identifies the space of which AI currently provides the most leverage.

> If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future. (https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now)

AI Researcher, Roger Schank, points out that the storytelling process is calculated and purposeful.

> [People] are thus forced, in some sense, to make their story acceptable and easily comprehensible both by their initial attempts to understand the events themselves and by their prior attempts to tell others their story. [24]

Robert McKee, creative writing expert, describes what it means to tell stories.

> From an instant to eternity, from the intracranial to the intergalactic, the life story of each and every character offers encyclopedic possibilities. The mark of a master is to select only a few moments but give us a lifetime [15].

Quality	This value is typically determined by user evaluation. If we can deliver a better experience without having to compromise viable variations and that costs the same amount (or less) in effort, we have created leverage. (**User focused**)
Variations	This value determines the diversity among potential experiences. In previous work, this has been done through comparing play traces. If we demonstrate an increase in legal variations of the same quality, or manage to create better sets of interesting variations without increasing effort, then we have created leverage. (**User focused**)
Control	If we are able to make changes, control and extend a story world, or create a brand new story world without compounding the effort or breaking the user-experience, then we have gained leverage. This value represents the precision and integrity of how well the audience-experience stays true to the integrity of the design or the authorial intention. If changes to the interactive space create nonsense or break the overall experience, then the system is inflexible. (**Author focused**)
Effort	We find the script-and-trigger policy (traditional approach) that produces quality experiences equal to the new approach. Then, within a similar space of interesting variations, we have a quantitative measure of effort. This is the amount of effort it would take an author to create the entire experience without an AI system. In practice, this was done by comparing the number of rules and specifications that are needed for a functioning or playable experience. (**Author focused**)

Figure 3: Descriptions of the values mentioned in Figure 2.

Received 23 February 2023; revised 05 April 2023; accepted 07 March 2023