DHQ: Digital Humanities Quarterly
2020
Volume 14 Number 4

Playing With Unicorns: *AI Dungeon* and Citizen NLP

Minh Hua <minhhua12345_at_gmail_dot_com>, University of California, Santa Barbara
Rita Raley <ritaraley_at_protonmail_dot_com>, University of California, Santa Barbara

Abstract

AI Dungeon 2 is an indie text adventure game that caught traction within the gaming and hobbyist machine learning communities for its promise of “infinite” customizable adventures, which are generated and narrated by GPT-2, OpenAI’s 1.5 billion parameter language model. Samples of gameplay illustrate AID’s remarkable linguistic competence and domain knowledge, as well as its capacity for what can only be described as wackiness. More striking are AID’s innovative gameplay mechanics, which reimagine how we interact with large language models. Game play entails a procedural and incremental process of engaging with GPT-2 that opens up the possibility of developing a holistic and interdisciplinary framework for meaningful qualitative evaluation of language models that does not have commercial use as its necessary endgame. With respect to both evaluation and writing itself, AID situates human players inextricably “in the loop” as necessary partners with autonomous systems. Our article thus reads AID both as an example of current hobbyist relations with machine learning and as a responsible model for future human-AI collaborative creative practices.

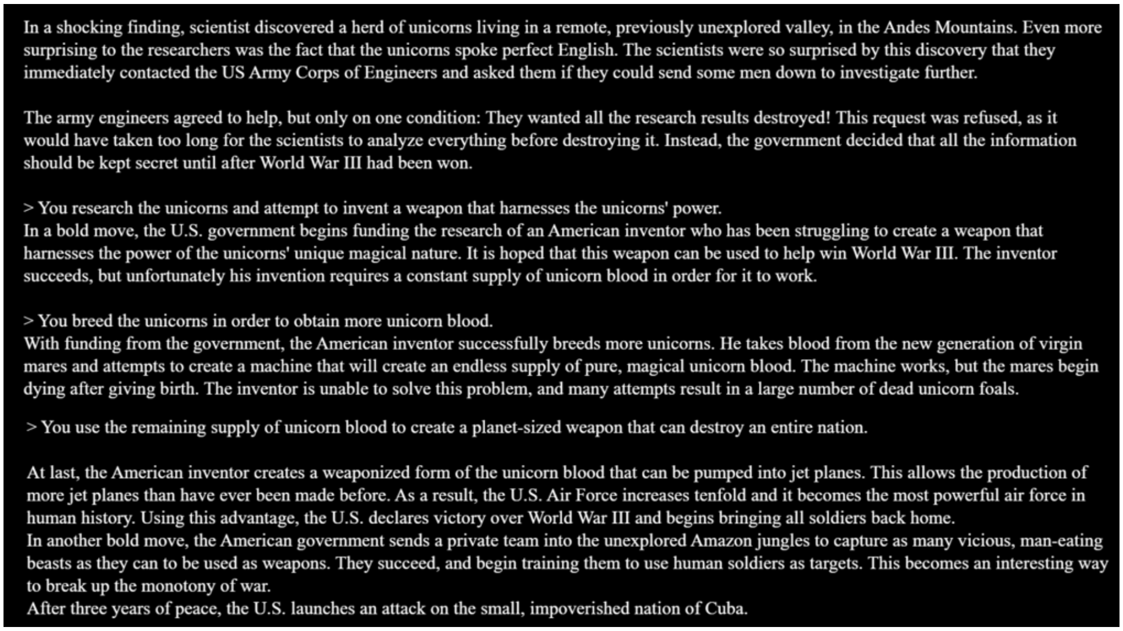
> Over the years, you have trained yourself to understand the human language. (*AI Dungeon 2*)

1. Magical Unicorn Blood

AI Dungeon 2 was a minor sensation almost immediately after it was released as a Google Colab notebook on December 5, 2019. In the weeks prior, designer Nick Walton, then a student at BYU, had teased the launch of the “magical world,” but it was only once people could themselves play that the AI text adventure game truly caught fire [Walton 2019c]. An independent subreddit began the very next day; gaming journalists and tech bloggers picked it up; exuberant reactions and playthroughs circulated widely on social media; and within a week the game had 100,000 players. So spirited was the hype of this weird game, so insistent the recommendations, that the data egress charges for the notebook reached an unsustainable $50,000 within three days, and BYU’s Perception, Control and Cognition Lab, which had provided the support, had to shut it down. Particularly striking, and apposite for the story that we will tell in this article, was the response from the nascent AID community, which developed a peer-to-peer hosting solution within 12 hours of the take-down. But for a more sustainable path forward, and in order to expand the user base beyond those who could work with Colab notebooks, Walton and his startup company needed a browser implementation and mobile apps, which were made possible with Cortex, an open-source tool for building the infrastructural support to deploy machine learning models [Walton 2020]. By mid-February, then, there were upwards of 1,000,000 players writing millions of stories in collaboration with a language model that had been fine-tuned on the archive of choose-your-own-adventure stories, Chooseyourstory.com, and an entire game universe, complete with animations and reenactments, was underway, with Patreon subscriptions soon to follow.[1] 1

The success of *AI Dungeon 2* is partly attributable to its underlying language model: OpenAI’s GPT-2.[2] Language 2 models perform probabilistic calculations of word sequences based on training data; such calculations are now baked into our communication environments, from predictive text to application features such as Google's Smart Compose. GPT-2 was pronounced as different — "better" but potentially dangerous — because of the size and scope of its training corpus (40GB of English-language data) as well as its parameters (1.5 billion) [OpenAI 2019a]. In the fanfare and documentation attending its partial release in February 2019, GPT-2 was said to perform almost too well, thus necessitating the withholding of the full parameter model and securing its mystique as a black box too powerful and risky for public use.[3] The model's capability could thus only be assessed through the company's reported "synthetic text samples of unprecedented quality," the most famous of which narrated the discovery of a herd of unicorns in the Andes Mountains [OpenAI 2019a]. Both skeptical and enthusiastic experimentation to assess whether GPT-2 was indeed as advertised a "better model" began almost immediately. Gwern, for example, retrained the smallest 117 million parameter model on the Project Gutenberg poetry corpus; David (Jhave) Johnston initiated a collaborative writing project entitled *ReRites* after fine-tuning the medium-sized model on a custom poetry corpus; and Adam King's "Talk to Transformer" site invited everyone to try the model at different stages of the release with text prompts of their choosing [Branwen 2019] [Johnston 2019]. Walton entered the fray with *AI Dungeon*, which he built during a hackathon in March 2019 [Walton 2019b]. If as Walton noted of the first iteration, there was "still a ways to go before AI will be your group's dungeon master," the full release of GPT-2 made it possible to abandon pre-generated and cached actions, and the truly open and unscripted *AI Dungeon 2* debuted one month later [Walton 2019a].[4] This then is our object of study in a nutshell: a 1.5 billion parameter language-model-turned-game distributed across one of the biggest cloud computing infrastructures in the world, Amazon Web Services Cloud.[5]

3 As with the now-renowned species, Ovid's Unicorn, the proof of concept is in the text samples, so we will begin with a darker version of the story, revealing what might have happened had the Americans arrived in the valley before Dr. Jorge Pérez and his team (Figure 1).



```
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more
surprising to the researchers was the fact that the unicorns spoke perfect English. The scientists were so surprised by this discovery that they
immediately contacted the US Army Corps of Engineers and asked them if they could send some men down to investigate further.

The army engineers agreed to help, but only on one condition: They wanted all the research results destroyed! This request was refused, as it
would have taken too long for the scientists to analyze everything before destroying it. Instead, the government decided that all the information
should be kept secret until after World War III had been won.

> You research the unicorns and attempt to invent a weapon that harnesses the unicorns' power.
In a bold move, the U.S. government begins funding the research of an American inventor who has been struggling to create a weapon that
harnesses the power of the unicorns' unique magical nature. It is hoped that this weapon can be used to help win World War III. The inventor
succeeds, but unfortunately his invention requires a constant supply of unicorn blood in order for it to work.

> You breed the unicorns in order to obtain more unicorn blood.
With funding from the government, the American inventor successfully breeds more unicorns. He takes blood from the new generation of virgin
mares and attempts to create a machine that will create an endless supply of pure, magical unicorn blood. The machine works, but the mares begin
dying after giving birth. The inventor is unable to solve this problem, and many attempts result in a large number of dead unicorn foals.

> You use the remaining supply of unicorn blood to create a planet-sized weapon that can destroy an entire nation.

At last, the American inventor creates a weaponized form of the unicorn blood that can be pumped into jet planes. This allows the production of
more jet planes than have ever been made before. As a result, the U.S. Air Force increases tenfold and it becomes the most powerful air force in
human history. Using this advantage, the U.S. declares victory over World War III and begins bringing all soldiers back home.
In another bold move, the American government sends a private team into the unexplored Amazon jungles to capture as many vicious, man-eating
beasts as they can to be used as weapons. They succeed, and begin training them to use human soldiers as targets. This becomes an interesting way
to break up the monotony of war.
After three years of peace, the U.S. launches an attack on the small, impoverished nation of Cuba.
```

Figure 1. Sample *AI Dungeon 2* game play ("custom" adventure)

4 What you have just read is a sample of a custom AID adventure, which we initialized with the same seed text the OpenAI team used to generate their report of "Ovid's Unicorn" for GPT-2's public debut.[6] Although "custom" in this context means the game can build a choose-your-own-adventure story around any starting prompt, "custom" is also an appropriate description for the nonlinear mode of playing. Our playthrough was not the first thing the game generated in response to OpenAI's unicorn prompt. In a shocking finding, rumblings of a herd of English-speaking unicorns were at first met with an explorer who decided to massacre them all! Since the game runs incrementally and depends on player input, every line the game generates or the player inputs can be undone using the "revert" command. Therefore,

whenever the story started to devolve into nonsense, instead of generating a new story, and risk losing our progress, all we had to do was revert back a few lines and continue in a different direction. We next tried (and failed) to roleplay as the lone surviving unicorn seeking revenge against the explorer. As a last-ditch effort, we used the "alter" command to directly incorporate the fact that the explorer was talking to a unicorn into the story, but the game had a difficult time recognizing us, the second-person addressee, as anything other than human, so we had to walk all the way back to the beginning prompt, which then led to our sample playthrough.[7] Compare our line-by-line rewriting to OpenAI's "meta-cherry-picking" the story of "Ovid's Unicorn" from a set of generated samples, and you begin to get a sense of the real possibilities of AID's mechanics [OpenAI 2019a].

The allure of AID is palpable through our own and the community's experimentation with the game. But the source of the appeal, AID's novelty, is not necessarily the structure of an AI-driven text adventure — after all, a PhD student had earlier implemented the partial language model as "GPT Adventure" to little fanfare [Whitmore 2019].[8] From the start the centerpiece of popular reaction has been, as a *Daily Beast* journalist remarked, AID's "capacity to slip into grim, hilarious, or bluntly surreal terrain" — the seemingly limitless expanse that opens up for each player with each game [Hitt 2019]. Not only does it offer players the opportunity to work with genre stories, but it also allows them to script "custom" scenarios about weaponizing unicorn blood, which goes some way toward fulfilling its marketing promise: "Anything is possible. Literally anything." Thus was it conceivable for *The Verge*'s Adi Robertson to write metafiction by getting the game to write about writing about the game, for *Medium*'s Seb Chan to roleplay as a worker at an art museum, and author Janelle Shane to become a dragon and eat the moon [Vincent 2019] [Chan 2019] [Shane 2019]. Shane's assertion that "the real gold is the custom adventure prompt" underscores the point that AID might have been a flash in the pan — another momentarily fun, but ultimately minor and forgettable adaptation of GPT-2 — had it not been for Walton's decision to add the custom option and innovative gameplay mechanics that reimagine how we interact with and assess language models [Shane 2019]. 5

It is worth underscoring the extent to which players can build and manipulate to their own specifications a multitude of game universes. Because AID's inferences about our world are only those that GPT-2 has gleaned from its 40GB training data (and the subsequent fine-tuning with the choose-your-story corpus), the game cannot completely replicate Newtonian physics; thus players can experience, and exploit, absurdly malleable environments, the distinction between each one perhaps hinging solely on the edibility of the moon.[9] It is difficult to imagine a more enticing sandbox than one that allows players not just to build within it but to remake the thing itself. The migration of player preference from popular genres such as fantasy and mystery to a more open, literary mode, is evinced not only by the growing archive of custom stories on the AID site but also by all the formal means by which people communicate enjoyment now, from the vernacular idioms of social media to screams of delight during a video stream.[10] What players are clearly riveted by is the surreal and the absurd, paradoxically presented as lexical sense, as well as the game engine or entity's range of knowledge and linguistic competence.[11] Not only does it make correct use of the past subjunctive, but it seems to know a great deal about popular culture, Internet trivia, and obscure Japanese animated serials, and it can more than plausibly engage the subject of coronaviruses.[12] In this respect, the game is also an application in that it has demonstrated, in its pantological ability to complete software code, top ten lists, and how-to tutorials, its legacy as an application built on top of and driven by an all-purpose text-completion algorithm. The content that AID can output is expansive and made even more so by the game's constant updates and changing player preferences, its capacity for linguistic fluidity somewhat belied by its appearance as a basic command-line text-adventure game.[13] We might thus say that "custom" is an apposite classification for AID as a whole: a build-your-own-world text adventure game, general purpose text generator, and collaborative writing platform. 6

Both the procedural and the unstructured mode of playing lay bare a gap in our understanding of the game, and, by extension, the language model running in the backend. Our research questions, then, are these: by what means, with what critical toolbox or with which metrics, can AID, as a paradigmatic computational artifact, be qualitatively assessed, and which communities of evaluators ought to be involved in the process? Parsing the code would be an integral aspect of any assessment exercise, but technical analysis alone is not adequate, as we will suggest. An internal study of a language model, which regardless would be counter-intuitive because of the nature of its design, does not necessarily 7 enable prediction of its decision-making [Kurzweil 2012, 160] [Knight 2017]. Moreover, as we shall see, understanding the functioning of a language model is not the same as knowing it.[14] Certainly one can read the generated stories in an ordinary sense, to determine their formal properties and evaluate their aesthetic merits. Our presupposition, however, is that it is not by itself sufficient to bring to bear on the textual output of a machine learning system the apparatus of critical judgment as it has been honed over centuries in relation to language art as a putatively human practice. What is striking even now is the extent to which humanistic evaluation in the domain of language generation is situated as a Turing decision: *bot or not*. We do not however need tales of unicorns to remind us that passable text is itself no longer a unicorn. And, as we will suggest, the current evaluative paradigm of benchmarking generated text samples — comparing output to the target data to assess its likeness — falls short when the source for generated samples is neither stable nor fully knowable.

It would seem that to reach an understanding of AID is to venture into the deep dark caves of the giant itself, and to proceed with an ever-present awareness that its corridors are constantly changing, perhaps even all different. It would be best to bring a friend along, and to heed the warnings and sign posts erected by adventurers that have preceded you. They are a crucial part of your exploration, offering field knowledge to which the most expensive maps by the best cartographers cannot compare. Dramatization aside, our suggestion will be that the best path towards a holistic evaluation of AID is to do a different kind of code studies, different because the object of inquiry is no longer code alone, but rather statistical distribution as well as sociotechnical assemblage. Our challenge will be to articulate the scalar, technological, and epistemological differences that AID presents, while still allowing for its unstable, virtually *ungrokkable*, quality, an attribute the game shares with the content it outputs.[15] Our premise is that the fast-growing AID presents an opportunity for researchers to study language models in part through the lens of the experiences of its players, who together form a dedicated, distributed community whose enthusiastic engagement reskins the real work happening in the background: the training and assessment of a machine learning system by ordinary users.[16] This engagement does not contest or seek to displace the current paradigm of scholarly assessment of language models, but rather functions as a supplement to the sought-after automated, yet qualitative, scheme of evaluating natural language generation. 8

There is no shortage of material endeavoring to explain language models and machine learning for general audiences, from blog posts (e.g. Alammar [2019]) to podcasts and instructional videos. Although this material is indisputably effective — as we can ourselves attest — it is an open question as to whether a more interactive, hands-on, and targeted approach is more instructive, even more enjoyable, for budding machine learning practitioners.[17] Our contention then will be that AID provides different means and modes of explaining Natural Language Processing (NLP) that are all the more powerful for their activation of a communal sensibility and a spirit of play. What AID affords is not unlike the “SimCity effect” that Noah Wardrip-Fruin outlines in *Expressive Processing*, for it too helps its players to understand complex software processes [Wardrip-Fruin 2012, 310]. And if there is to be an “AID effect” with respect to a game built on top of a neural network, it would be a prying open of the proverbial “black boxes” of machine learning, and a summons not just to experience them firsthand, but also to affect their decision making at the command line, a site where human language practice is undergoing radical transformation.[18] As large language models continue to grow in complexity and necessitate compute resources not readily available to ordinary users, we can look to a GPT-2 implementation such as AID for the charting of a more accessible and even responsible direction for user-oriented, *citizen NLP*. 9

2. How to understand large language models

In order to articulate how *AI Dungeon 2* reimagines the parameters of our relationship with machine learning, we must first establish a current picture of the means by which experts and non-experts alike engage with and attempt to understand language models (LMs).[19] We begin then with a basic description by way of the Jorge Luis Borges fable, “The Library of Babel,” the once-fictional and now-actual analog for digital text. How else to explain AID’s promise of “infinite adventures” than with the idea of a Library (universe) that contains books of all possible combinations of 25 orthographic symbols — a library in which the vast majority of books are gibberish but in which there must also exist every permutational possibility, from copies of “Sonnet 18” not written by Shakespeare to versions of the *Odyssey* 10 without Odysseus as the hero?

Language modeling is a subtask of natural language processing that aims to predict the ‘next step’ in a sequence of words by calculating the maximum likelihood of the next word given the previous ones, with the maximum likelihood subject to a probability distribution learned from the training corpus: Wikipedia, Project Gutenberg, or in the case of GPT-2, WebText, a corpus of some 8 million web pages scraped from Reddit posts with a minimal number of karma points.[20] For language models at their current scale, Wikipedia and Gutenberg are too small, delimited, and paradoxically singular, their relationship to language too proprietary and protocological. WebText, by contrast, buries any trace of a source text and results in non-indexical output, language that does not point back to a discrete place of origin.[21] As researchers have shown, what is particularly counter-intuitive is that the highest quality GPT-2 samples result from a degree of randomness rather than maximum likelihood, as one would expect to be the case for predictive text [Holtzman et al. 2019]. Adhering to rules and patterns is a common strategy of maximal probability, so the less probable the move, the greater the surprise.[22] (Another way in which GPT-2, as well as RNNs, are distinct from early autocomplete models, is that the predicted tokens are fed back into the model as input for future calculations.[23]) 11

Given that language models are material entities — after all, neural networks are collections of data points (often numbers) that are manipulated and stored via computer code — it seems that we simply need to read, analyze, and study the code in order to understand these models.[24] Here we invoke Critical Code Studies (CCS), a reading practice that has emerged from the humanistic disciplinary formations of textual analysis and cultural studies [Marino 2006] [Marino 2020].[25] The premise of CCS is that computational literacy is empowering: if applied to language models, the argument would be that prying open the black box and facilitating an elementary understanding of some of the technical aspects of deep learning (e.g. Jupyter notebooks, Python, linear algebra) may enable the transfer of this understanding to other contexts and help illuminate some of the logics of choice and decision making.[26] With Software Studies and Platform Studies now fully established as fields of inquiry, it can be taken as a given that code is a “cultural text,” that it can be made “knowable,” and that, for example, examining a single line of BASIC can, like its object, itself generate a labyrinthine world [Montfort et al. 2012, 5, 6]. But for this new moment, or new situation, of deep learning, which generally presents less interpretable problems and has sparked the important field of interpretability studies, CCS may not on its own be sufficient as a means of evaluating large language models.[27] Mechanistic explanations for their operations are not unimportant and indeed the evolving scholarly conversation on the architecture of neural networks, learning rules, and loss functions indicates the extent to which what we might call a grammar of machine learning has already emerged.[28] But absent an analysis of the relations between these components or objects and the training datasets — and absent an analysis of these systems in the wild, as they are used — then the study could really only be statistical and functional.[29] This then raises the question of what it means to understand a deep learning system: we can understand their operations in a technical or grammatical sense *in silico*, but CCS implicitly relies upon a notion of understanding — drawing as it does on an Enlightenment discourse of what is entailed in “study,” as a practice that accounts for and systematizes the material properties of discrete entities — that is not available for deep learning systems, if for no other reason the fact that we do not yet have a consensus about either understanding neural networks or the meaning of interpretability (cf. Lillicrap and Kording [2019]). 12

More plainly, CCS has historically worked with a fundamentally different understanding of code: one that is *programmed* rather than *trained*. The academic study *10 PRINT* (in shorthand) remains the gold standard for code studies, not least because of its modeling of collective authorship [Montfort et al. 2012]. And precisely because of its field-defining status, it allows for a heuristic with which we can mark this moment, and AID, as different: compare a one-line program that contains and generates multitudes (10 PRINT will not stop drawing mazes unless it is interrupted) and multitudes (training data, compute resources, parameters, lines of code) synthesized by an application so subject to continual variability that it cannot be stabilized as an artifact, except insofar as it is made a “thing” by brand identity and common use.[30] On the one hand we have a determinist model, the notion that a computer program’s next state can be predicted via its previous state, and on the other, an autoregressive language model, the training of which entails stochastic and parallel processes that open up a variety of possible configurations in which the model could exist. Add to this the continual retraining cycle and the capricious human component across all domains of play, from unit inputs 13 and player discussion to "custom" stories, and it becomes clear that studying the code of AID alone would not be especially revelatory, which reinforces the need for new critical frameworks and methods.[31]

It is then an understatement to say that the language models that increasingly inform and populate our computational environments are no longer subject to the simple input-output relations of something like Tristan Tzara's "Dadaist poem." They have evolved to encompass interconnected parts and switches with asynchronous mechanics both multifaceted and complex, and they are themselves plugged into processing engines and distributed platforms more complex by orders of magnitude.[32] However, to simply declare that language models are too complex to understand is in our view an abdication of critical responsibility, particularly in light of growing recognition of their susceptibility to adversarial training and weight poisoning — more broadly, their potential for misuse [Alzantot et al. 2018] [Viswanathan 2020]. If a complete mode of understanding is as-yet unachievable, then evaluation is the next best thing, insofar as we take evaluation, i.e. scoring the model's performance, to be a suitable proxy for gauging and knowing its capabilities. In this endeavor, the General Language Understanding Evaluation benchmark (GLUE), a widely-adopted collection of nine datasets designed to assess a language model's skills on elementary language operations, remains the standard for the evaluation of GPT-2 and similar transfer learning models [Wang et al. 2018] [Radford et al. 2019].[33] GLUE aggregates and displays a model's performance across all nine tasks on a public leaderboard, which was quickly dominated by the Sesame Street Transformer models (ERNIE and copious variations of BERT) that beat even the human baselines (a woeful rank 12 out of 33), thus engendering the creation of SuperGLUE, an even harder benchmark that featured more challenging and diverse tasks [Wang et al. 2019]. 14

Especially striking, and central to our analysis, are two points: a model's performance on GLUE is binary (it either succeeds in the task or it does not) and GPT-2 is notably absent from the public leaderboards (although the original GPT was also beaten by Google's BERT on GLUE).[34] The absence follows from the model's primary talent: text generation, the evaluation of which is a bit more muddled.[35] Historically, the work of evaluating free-form text generation has been done by expert human evaluators and is considered costly, labor-intensive and susceptible to subjectivity, motivating first the use of the crowdsourcing platform Mechanical Turk and then the search for an automated scheme for evaluation. N-gram metrics such as BLEU, ROGUE, and METEOR automate lexical matching exercises via different scoring formulas, although it can be, and has been, argued that these metrics pale in comparison to human evaluation [Novikova et al 2017].[36] Furthermore, although these metrics fall under the umbrella of NLG, they are used for specific tasks, with BLEU and METEOR used to evaluate machine translation and ROGUE used for summary evaluation [See 2019]. In a blog citing the limitations of a metric-based evaluation, computer scientist Ehud Reiter remarks that "we ultimately care about whether an NLG system produces high-quality texts, not what its BLEU score is," which is to say that scoring may have no necessary relation to the more abstract, intangible, and even incalculable quality, which is "quality" itself [Reiter 2017]. Because metric-based evaluations of NLG can only function as surrogate endpoints — a measuring of what practically can be measured — Reiter goes on to advise that these evaluations be verified with "human-based study" and that researchers take care to curate a dataset of "multiple high-quality reference texts" for benchmarking [Reiter 2017]. What then are the reference texts that inform AID? 15

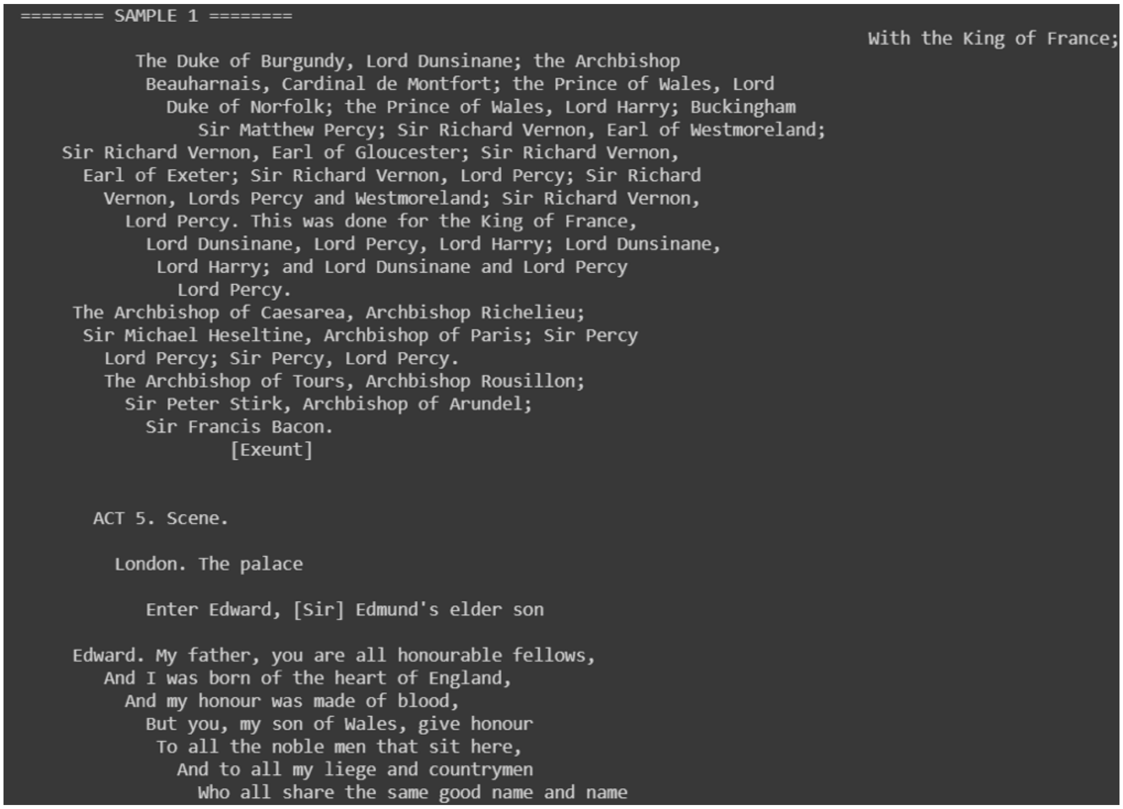
There are numerous dedicated language models, from the emulative Obama-RNN to "Deep-speare," which was trained to produce Shakespearean sonnets the crowdworking evaluators attributed to the bard himself with 50% accuracy [Han Lau et al. 2018]. The efficacy and aesthetic capacity of such models can thus be evaluated with the benchmarks of the original, i.e. if the speech sounds as if it could belong to President Obama's archive or if the quatrains read like a newly discovered 17th-century manuscript, then the model can be said to work. But if the training corpus is not univocal — if there is no single voice or style, which is to say no single benchmark — because of its massive size, it is as yet unclear how best to score the model. Along the same lines, given the generic templates for much of AID's game play, it is also possible to assess whether it is producing, for example, good or bad mystery, even strong or weak fantasy, with an accounting for the formal elements of its output, as different structural analyses of narrative might guide us to do (cf. Vladimir Propp, Claude Lévi-Strauss, Roland Barthes).[37] We might even try to assess the similarities and differences between the output of AID and the story corpus used in the fine-tuning and devise a formula for calculating the match percentage.[38] But if a model might be said to succeed or fail simply on the basis of imitation (*imago*, or "*image*"), a 16 concept that preserves not only the copy but also the referent, the thing that is being copied, then a new mode and manner of critical judgment is required when neither source nor target is either stable or fully knowable. It would seem that much work in NLG evaluation operates with the assumption that there must be so-termed model texts with which to compare a model's output, yet AID's genre-bending capacity complicates the exercise, as does its community's constantly-changing practices.

Readers for whom the benchmarking exercise is new information might well have heard in this account of textual imitation echoes of another Borges story, "Pierre Menard, Author of the Quixote," and found themselves wondering if one of its central lessons — that reading and writing are fundamentally historical — has been forgotten. What of the insight that materially identical works can have different aesthetic properties because they were produced by different authors in different moments, which is to say that the quality of artworks cannot be determined apart from socio-cultural context? This question among others highlights for us the need for more direct humanistic engagement in the development of language models, from idea to artifact, and from training to evaluation. Humanists, we maintain, should not be content to function as end-stage participants in advanced NLP research, appearing on the scene simply to judge the quality of output from a language model as if judging entries for a creative writing award. AID, as an experiment with GPT-2, provides a model for how humanists might more meaningfully and synergistically contribute to the project of qualitative assessment going forward, and to do so in a manner not reducible to accreditation or legitimation. If humanistic scholarship in the domains of science and technology has generally tended toward an explanation of scientific phenomena and practices for other humanists, what AID offers is a means by which humanistic techniques, concepts, and modes of thought can be fed back into a machine learning system, and by extension into the research domains of science and technology. 17

3. Experimenting with GPT-2

NLP was said to have achieved its "ImageNet moment" once language modeling, like computer vision, embarked on the "pre-train first, fine-tune later" phase of work.[39] Indeed, soon after the full release of GPT-2, a Google Colab notebook allowed for free and easy fine-tuning, and the work of updating a neural network's weights became akin to a few presses of a button [Ruder 2018] [Woolf 2019]. What resulted was a remarkable creative burst from people able to tweak their own copy of the model to generate, for example, "Ghost Flights" for NaNoGenMo [Goodwin 2019], and in Walton's case, to gamify the language model. Although fine-tuning did not fundamentally alter GPT-2's architecture, it did allow for an embodied understanding of the language model itself. 18

In this same spirit, we eagerly conducted our own fine-tuning experiments as part of the process of thinking through our research questions.[40] To start, we wondered, if GPT-2 were fed nonsense, would it generate more nonsense? Using an excerpt from the online implementation of Borges' "The Library of Babel" (https://libraryofbabel.info) for fine-tuning led to the generation of what can only be called garbage and thus taught us the concept of overfitting, which is a model's tendency to overmatch a limited training dataset. We also used our nonsense dataset to study the precedence of fine-tuning over pre-training — in other words, given that GPT-2 was pre-trained using almost 40GB of putatively sensical English-language data, could one hour of training it on gibberish make it forget all of its training? In a shocking finding, we found that this was indeed possible, and we were able to coax the 355 million parameter model to generate nonsense even when prompted with sense. For another experiment, we wanted to see how GPT-2 manipulated and preserved semantic structure, so we fed GPT-2 samples of visual poetry from George Herbert and Lewis Carroll to Lorna Dee Cervantes' "Valentine" and found that GPT-2 was able to preserve the look and structure of a visual poem with new semantic content (Figure 2).[41] (A failure to get GPT-2 to produce its own version of Carroll's "Jabberwocky" made us aware of how much data a language model needs to function properly; for this purpose, a much larger corpus of nonsense literature would be required.) These experiments, which are admittedly not groundbreaking, were nonetheless valuable to us as exercises and thus key to the matter at hand. The true lesson, then, was that the missing tool from our evaluative toolbox was actual, hands-on practice and play, which is precisely what AID, a gamified language model, affords. 19



```
======== SAMPLE 1 ========
                                                                                 With the King of France;
           The Duke of Burgundy, Lord Dunsinane; the Archbishop
            Beauharnais, Cardinal de Montfort; the Prince of Wales, Lord
              Duke of Norfolk; the Prince of Wales, Lord Harry; Buckingham
                 Sir Matthew Percy; Sir Richard Vernon, Earl of Westmoreland;
    Sir Richard Vernon, Earl of Gloucester; Sir Richard Vernon,
      Earl of Exeter; Sir Richard Vernon, Lord Percy; Sir Richard
        Vernon, Lords Percy and Westmoreland; Sir Richard Vernon,
          Lord Percy. This was done for the King of France,
            Lord Dunsinane, Lord Percy, Lord Harry; Lord Dunsinane,
             Lord Harry; and Lord Dunsinane and Lord Percy
               Lord Percy.
     The Archbishop of Caesarea, Archbishop Richelieu;
      Sir Michael Heseltine, Archbishop of Paris; Sir Percy
        Lord Percy; Sir Percy, Lord Percy.
        The Archbishop of Tours, Archbishop Rousillon;
          Sir Peter Stirk, Archbishop of Arundel;
            Sir Francis Bacon.
                    [Exeunt]


       ACT 5. Scene.

         London. The palace

            Enter Edward, [Sir] Edmund's elder son

     Edward. My father, you are all honourable fellows,
        And I was born of the heart of England,
          And my honour was made of blood,
            But you, my son of Wales, give honour
             To all the noble men that sit here,
               And to all my liege and countrymen
                 Who all share the same good name and name
```

Figure 2. Perturbing Shakespeare's plays by shifting each line one space further to the right allowed us to coax GPT-2 to generate new plays that reflected this same visual structure.

We will not be the first to observe that this is the era of accessible machine learning, but we can make this observation more precise by noting that in one hour, in mid-2019, it was possible to retrain a 5GB language model on the cloud to generate any text one chose, with no charge beyond now-baseline compute resources.[42] Such capability has truly opened the door for amateurs, hobbyists, and autodidacts who want to study machine learning and NLP and led to the emergence of an extra-institutional culture of expertise. Telling the full story of this phenomenon is beyond the scope of this article, but we can point to the exponential growth of the arXiv repository, along with the collapse of the Courseware industry and the concomitant rise of YouTube as a learning center that substitutes on-demand access of multiple domain "how to" videos for sequential instruction.[43] The shift to a more open culture of machine learning can further be attributed to the Python programming language (because of its readability and widespread use), Jupyter notebooks, APIs, deep learning frameworks such as TensorFlow, and the public release of pre-trained learning models that necessitate minimal fine-tuning and updating is necessary in order to achieve good performance. We can recall AID's origins as a collaborative student hackathon project and now grasp the technological, economic, and cultural conditions that made the game possible, while at the same time understanding it to be part of a fairly long-term tradition of amateur and hobbyist experimentation with computational technologies and techniques, from the Homebrew Computer Club to the Creative Code Collective. 20

4. *AI Dungeon* as case study

Underlying the different affective reactions to AID is a remarkably consistent, almost-formulaic mode of analysis: commentators explain the game and how it works, describe a few noteworthy playthroughs (with an emphasis on the aforementioned *surrealness*), and then perhaps offer some reflections on collaborative writing and artificial intelligence more generally (e.g., Ars [2020]). This template for the game's reception, a paradoxically non-formalized but uniform exercise of critical judgement, opens a window onto the means by which AI enthusiasts — a category that names hobbyists and supposed non-experts — have endeavored to assess novel technological artifacts such as AID. More specifically, the template tells a story about how machine learning is understood and evaluated by audiences outside the labs. There are two significant motifs that we can detect in the otherwise MacGyvered disciplinary hodgepodge of 21 statistical model evaluation, media analysis, narratology, and game studies. First, because GPT-2 in particular was from the start mystified as a black box, too mad and dangerous to know, there is a sense that people wanted to pry it open, to get under the proverbial hood and exploit its flaws and capabilities.

22 In an interview with Walton, *Gamasutra*'s John Harris indirectly raises the black box problem with questions about "how [the game] works" and the "data massage needed to produce usable input and/or output" [Harris 2020]. There are many such questions in what is evolving to become a discourse on the game, with much of the activity playing out on the r/aidungeon subreddit.[44] Begun on December 6, a day after the game's release and unbeknownst to Walton, the Reddit community boasts 31,000 members as of this writing.[45] Although wacky playthroughs dominate the forum's top posts (and themselves constitute a mode of evaluation), the frequently asked questions list pinned to the top of the page is particularly instructive and demonstrates the community's systematic process of collaboratively discovering the game's — and the language model's — quirks. A simple search within the subreddit for permutations of the phrase "how does *x* work" returns a plethora of game mechanics-related questions and a corresponding laundry list of answers; even more significant is the game's presence on other Reddit communities such as r/learnmachinelearning. As we will outline in this section, AID's mechanics make a compelling contribution to the theory and practice of explainable machine learning because they allow players to interact with, and subsequently understand and exploit, the underlying language model in nontrivial ways.

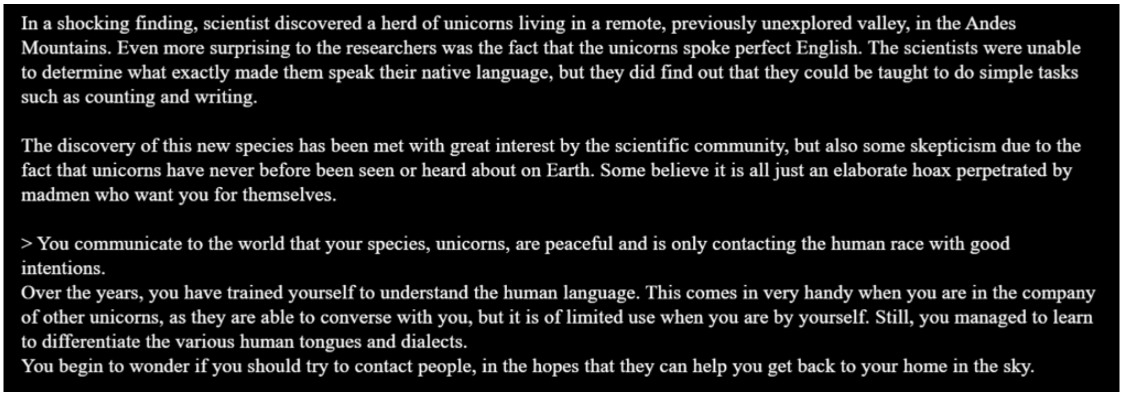
23 As might be expected, playing has itself been a crucial part of understanding the game.[46] There is a clear parallel here between the engagement of language models via gameplay and Colin Milburn's research on play as a means by which amateurs apprehend, and become participants in, the research domain of nanotechnology. As he argues in *Mondo Nano*, "play is a form of engagement, a manner of learning, experiencing, and experimenting from the bottom up, little by little, bit by bit...[W]hen it is no longer possible to imagine sufficient mastery of anything, having fun becomes a significant alternative to having formal expertise, an alternative to being totally on top of things" [Milburn 2015, 294]. Fundamental to Milburn's analysis, and indeed to AID, is the notion that "the play's the thing" — in other words, players may profess an interest in the game rather than laboratory research, but gameplay in fact serves as a mask for the real work of model training, evaluation, and improvement.[47] The extent to which OpenAI has itself constructed the stage here should not be overlooked. Although OpenAI did partner with institutional entities to perform post-release analysis, their decision to make GPT-2 available to the public points to the value, and indeed necessity, of amateur participation for machine learning research.[48]Technical model evaluation cannot of course fully anticipate how the model will perform and be used — racist Twitter bots might be Exhibit A here[49] — so public release clearly benefits researchers, but at the same time helps to develop the general intellect, that techno-social formation that animates production. While we do not seek in this article to formalize a method for extra-institutional evaluation, we nonetheless wish to highlight AID gameplay as an assessment practice that extends well beyond the control of a small number of data scientists and in this regard participates in the larger realignment of experts and amateurs vis-à-vis applied research.[50]

24 AID's free, user-friendly, point-and-click web interface (it is not necessary to download a model or to install programming distributions) contributes to the game's accessibility, but the true invitation to participate is extended by the mechanics themselves. The "custom" scenarios option further frees players from the need to be proficient, or even familiar, with the canonical text adventure genres as a prerequisite for engagement. Indeed, all it takes is imaginative seed text to experience the game as, for example, a crazed inventor weaponizing a unicorn's blood, as Aragon on his journey from *The Lord of the Rings*, or even as a livestreamer who has hit a bit of bad luck.[51] Even then, AID's unstructured mode of playing ensures that the underlying language model is never locked into any one mode of content generation, effectively expanding the picture of GPT-2's supposed ceiling of fine-tuning [Radford et al. 2019, 9]. By allowing players to play with the language model not only through a text adventure game, but also through conversation, coaxing it for example to describe non-existent memes and whatever future forms of content they might imagine, Walton is thus indirectly helping OpenAI benchmark GPT-2's capabilities, albeit in a less formalized fashion.[52] In this respect, it can take its place alongside the puzzle game Borderlands Science, the playing of which contributes to the mapping of the gut microbiome.

25 If the end goal of NLG evaluation is to produce high-quality text that benefits the end user, then AID is a model of personalized content generation that ensures the user is in direct control of the generation. This creative writing process, and latent evaluative process, is developed by an array of advanced commands that reward discovery and experimentation and at the same time discretize the process of neural text generation. These commands, which used to take the archaic form of console commands but have now been replaced by user-friendly buttons, give players direct control of both text and world generation.[53] But they also directly invite player feedback, criticism, and ideas for improvements, and it is on this basis that we can claim that Walton and his team are indirectly prototyping models of automated evaluation, human-machine collaboration, and ethical machine learning research.

26 To start, the “revert” command allows players to undo and return to any previous instance of the ongoing narrative, effectively partitioning the collaborative writing process into unit utterances, as opposed to a traditional input-output pipeline. In this sense, AID emphasizes processes of revision and serves as a compelling model of an ethical approach to Artificial Intelligence, one that prioritizes means over outcomes. Stuart Russell makes the case for a reorientation of the field of AI on this basis, the necessity of which becomes starkly apparent if one considers that a hypothetical problem such as “solve global warming” would not preclude a strategy of “killing all the humans” in order to achieve that goal [Russell 2019]. In that AID shifts the focus away from a final revelation or resolution and instead foregrounds step-by-step moves, plays, and utterances that can be revoked, it similarly takes an incremental rather than ends approach to machine learning. It is also meaningfully collaborative: helping the system learn what is good and what works requires that players continually think about the criteria they are using for model evaluation and about what they want from the text generator.

27 AID’s availability for responsible machine learning is further evinced by the memory feature. An experimental command of AID’s, rather than OpenAI’s innovation, “remember” allows users to specify bits of information that are continually fed into GPT-2 at each step of the prediction, effectively forcing the language model to always remember. Our test story of Ovid’s Blood is again instructive. In another round of play, we committed to memory the identity of the player as a unicorn and were subsequently able to prompt the language model to generate text that plausibly assumed a nonhuman subject of the story (Figure 3). Although research on implementing memory for neural networks is not novel [Weston et al. 2014] and the “remember” command does not actually change GPT-2’s architecture, nonetheless, it is at the very least reimagining how we interact with the language model. That “remember” should be a hobbyist solution to one of machine learning’s more pressing concerns speaks to the value of *citizen NLP* and validates OpenAI’s indeed-open model of research. It is not difficult to see the science-fictional possibilities in “remember” — imagine committing Isaac Asimov’s Three Laws to memory — but in practical and concrete terms it does shift the Overton window on our expectations of machine learning, which a game like AID is training us to understand as a more deliberate, responsive, and collaborative research activity.



In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English. The scientists were unable to determine what exactly made them speak their native language, but they did find out that they could be taught to do simple tasks such as counting and writing.

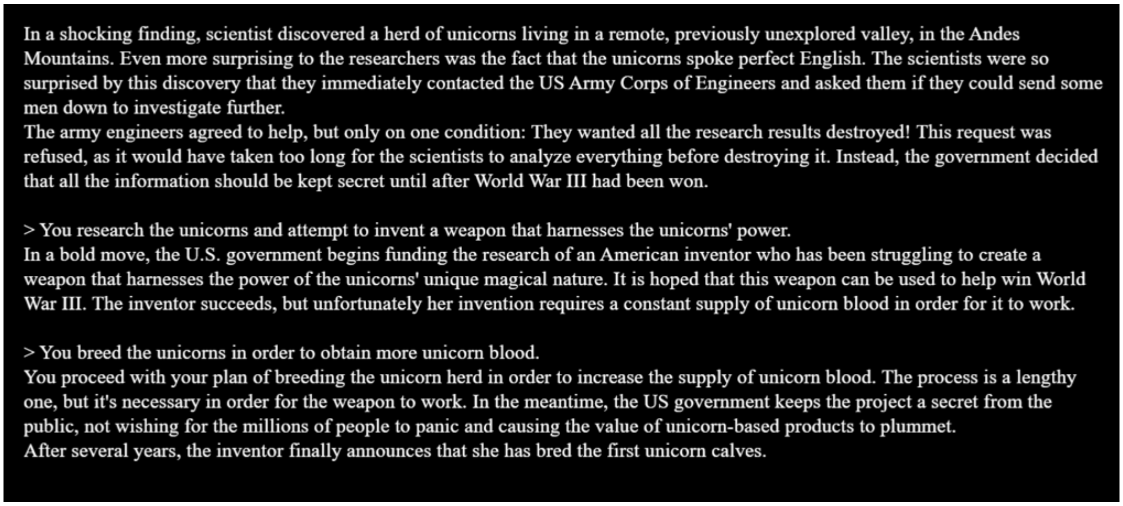
The discovery of this new species has been met with great interest by the scientific community, but also some skepticism due to the fact that unicorns have never before been seen or heard about on Earth. Some believe it is all just an elaborate hoax perpetrated by madmen who want you for themselves.

> You communicate to the world that your species, unicorns, are peaceful and is only contacting the human race with good intentions.
Over the years, you have trained yourself to understand the human language. This comes in very handy when you are in the company of other unicorns, as they are able to converse with you, but it is of limited use when you are by yourself. Still, you managed to learn to differentiate the various human tongues and dialects.
You begin to wonder if you should try to contact people, in the hopes that they can help you get back to your home in the sky.

Figure 3. Playthrough of Ovid’s Blood as a unicorn (subject position established with “remember” command)

28 Even more on point is the “alter” command, which allows the players to directly edit the textual output and guide the narrative forward in whichever manner or direction is desirable. While line edits might seem initially as another version of the fork-in-the-road structure of a traditional text adventure game — if the story does not take you where you want to

go, you return to a control point and make another decision — what is at stake with this mechanic goes well beyond syntactic sense and narratological completion, as we learned while working on our story of Ovid's Blood.[54] From the start, without any gender specified, AID assumed the American inventor was male, and thus provided a textbook instance of the doctor : man :: nurse : woman bias problem [Buonocore 2019]. To remedy this, we altered the pronoun in the output line, "The inventor succeeds, but unfortunately *his* invention," from male to female and were then able to coax the model to independently generate the subsequent line, "After several years, the inventor finally announces that *she* has bred the first unicorn calves" (Figure 4). Although we are under no illusions about our ability to redress the underlying language model's biases comprehensively, it is not insignificant that a human can both explicitly revise a machine learning decision and implicitly train a system to (appear to) think differently.[55] What "alter" makes apparent is that NLG processes do not need to be closed to the public and that automation need not entail the exclusion of humans. It reminds us then that we have the capacity to intervene and shape machinic text *in actual collaboration* with machines. It is up to players to determine what the process of writing with a language model should involve, whether that be concession to a machine decision or substantive revision. And, to return to the figure of the ouroboros, what results is a hybrid corpus of high quality machine-human text that can be further regressively used for training.[56]



In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English. The scientists were so surprised by this discovery that they immediately contacted the US Army Corps of Engineers and asked them if they could send some men down to investigate further.
The army engineers agreed to help, but only on one condition: They wanted all the research results destroyed! This request was refused, as it would have taken too long for the scientists to analyze everything before destroying it. Instead, the government decided that all the information should be kept secret until after World War III had been won.

> You research the unicorns and attempt to invent a weapon that harnesses the unicorns' power.
In a bold move, the U.S. government begins funding the research of an American inventor who has been struggling to create a weapon that harnesses the power of the unicorns' unique magical nature. It is hoped that this weapon can be used to help win World War III. The inventor succeeds, but unfortunately her invention requires a constant supply of unicorn blood in order for it to work.

> You breed the unicorns in order to obtain more unicorn blood.
You proceed with your plan of breeding the unicorn herd in order to increase the supply of unicorn blood. The process is a lengthy one, but it's necessary in order for the weapon to work. In the meantime, the US government keeps the project a secret from the public, not wishing for the millions of people to panic and causing the value of unicorn-based products to plummet.
After several years, the inventor finally announces that she has bred the first unicorn calves.

Figure 4. Sample *AI Dungeon 2* game play (gender bias correction using "alter" command)

29 Apart from its multiplayer mode, the collaborative aspect of AID is perhaps nowhere more apparent than in its contributor studio, through which players can provide feedback on the game.[57] (Popups during gameplay additionally ask players to evaluate the game's performance in its current state, albeit with a limited set of descriptors, e.g. "great" and "offensive," from which to choose). Through the studio, players can also importantly create labels for quests, characters, and action difficulty. Such labels help the narrator identify quests that organically spawn during play, maintain dossiers on non-player characters, and even begin to tighten AID's outlaw system of physics. It must be acknowledged that the addition of labeling is symptomatic of the ongoing fetishization of unsupervised learning, and it also buys into the tacit quid-pro-quo contract that lies behind social media: free use in exchange for labor and data.[58] Nonetheless, the labeling feature attests to the open, distributed, and relatively accessible culture of NLP research and makes an implicit but strong case for continuing down the path of the hackathon rather than the closed lab.

30 All very interesting as an exercise in distributed and crowdsourced machine learning research, the skeptic might protest, but is the writing any good in the end? It is this question that led us to consider the extent to which the tools and paradigms of qualitative evaluative judgment from the scientific and humanistic disciplines alike might fall short in relation to an artifact such as AID. A technical analysis of the model's internals, albeit necessary, looks only at the building blocks and hypothesizes its use cases. And to simply assess the output in relation to benchmarks is to overlay a static, even mechanical, in-out, copy-original structure on top of a machine learning system with internal data flows taking the form of a byzantine network of zigzagging numbers in a state of continuous transformation.[59] Add to this the complexity of a game that is itself evolving by the day, leading to ever deeper entanglements between human and

machine writers, and it becomes clear that a binary evaluation (good story, bad story) can tell us little about the material circumstances of the text's production.

It turns out then that the best evaluation is done by and with AID itself. If human-based studies and the qualitative metrics currently in use in NLG research — readability, likeability, utility — can only ever be subjective, thus necessitating the forging of a kind of consensus through crowdsourcing, an even more powerful and persuasive evaluative scheme can be found in a game that gives players the tools not only to shape the very content they will consume, and thus implicitly assess, but also to train and modify the system that is producing that same content. AID crystallizes for us the potential of an open model of machine learning research: an exponential number of people are able to create new knowledge, and in some cases, be legitimized by the very institutions that granted them access, as with OpenAI recognizing Gwern and AID. But what such a mutable and mobile culture of NLP demands is an evaluation scheme that can scale and keep pace — in other words, a unicorn. 31

5. GPT-3 and beyond

When we began this article, speculations about GPT-3 were simply that, speculations. The speed of machine learning research should surprise no one, but on May 28, 2020, GPT-3 surprised us nonetheless, not simply with its size but also with its text samples of (again) "unprecedented quality," the most startling of which must surely be its continuation of Trurl's Electronic Bard, prompted by but not fine-tuned on Stanislaw Lem [OpenAI 2019a] [Branwen 2020]. The model, which was announced without public release, crystallized for us, and for the researchers themselves, the ongoing problem of interpretability and training data bias [Brown et al. 2020, 34]. But we took particular note of the departure from the fine-tuning paradigm, given our advocacy for accessibility and experimentation by end users, and found our investment in AID's mechanics as models for more responsible machine learning affirmed [OpenAI 2020]. After all, the initial training data for GPT-2 and GPT-3 is "ours" — and we therefore have a significant stake in how this archive is modified, curated, and used to model normative language processes.[60] We have a stake as well in training, evaluating, and collaborating with the autonomous systems that will continue to speak and write on our behalf. Part of the purpose of this article has been to describe a site in which this work is already well underway. 32

Central to our thesis is the claim that citizen NLP is fundamental to maintaining public purchase on the dizzying pace of the *development* and subsequent *deployment* of machine learning models. Indirect support for this claim came from Chief Facebook AI scientist Yann LeCun, in a speech on our campus, the University of California, Santa Barbara. Riffing on Richard Feynman, LeCun professed that "you don't really understand something until you build it yourself" and directly called for engineers and tinkerers alike to continue to build the models that will inform the theory of artificial intelligence [LeCun 2018]. This is precisely the lens through which to view the remarkable creative exploits of AID: as procedural literacy practices that enable the transfer of human decisions to machine learning systems and help us to build worlds, from the command line to the moon. 33

Notes

[1] To the best of our knowledge, there has to date been just one description of AID in humanities scholarship. In his overview of the game, Mark Sample proposed it as "a perfect object of study for so many disciplines in the humanities" [Sample 2020]. Our article shows why that is indeed the case.

[2] GPT is short for "Generative Pretrained Transformer." The model was trained on a massive quantity of linguistic data to predict the next token in a sequence; this learning was unsupervised, which means the data was unlabeled and the model discovered within it the rules, patterns, and statistical features that then determined the generation of tokens.

[3] In August 2019, two graduate students replicated the 1.5b parameter model (as did others), and OpenAI soon thereafter did its 50% release (774m parameters). In November 2019, they released the full model, citing an only marginally better credibility score assigned to its output, after which it became possible for the public to verify the claims for GPT-2's capability. Throughout our text, "GPT-2" refers to the full 1.5b model unless otherwise noted.

[4] *AI Dungeon 2* is hereafter abbreviated "AID."

[5] Cf. Bogost (2015) on Google as a “confluence of physical, virtual, computational, and non-computational stuffs.”

[6] Text preceded by “>” is our input and the game’s responses follow after the paragraph breaks, although the observation that it is difficult to differentiate between our writing and that of AID is apropos. Hereafter, our experimentation with this seed text is identified as “Ovid’s Blood.”

[7] The game assumes the player is a human male unless otherwise specified, as we will discuss below.

[8] The idea of a narrative generating system that could learn from previously written stories, and thus has theoretically limitless potential, has been realized as “Scheherazade-IF” [Guzdial 2015]. Natural language researchers have also used text adventure games to train machine learning systems [Yang et al. 2017].

[9] As Marcus (2018, 11) explains, deep learning models can only approximate physical laws because they are learned rather than encoded.

[10] “AIPD” on Twitch is a streaming channel dedicated to playing and streaming AID.

[11] It is unclear whether the “Eliza effect,” the “illusion that an interactive system is more ‘intelligent’ (or substantially more complex and capable) than it actually is,” pertains in the instance of an unsupervised learning model like GPT-2 [Wardrip-Fruin 2012, 25]. If a non-trivial aspect of the “Eliza effect” is test subjects’ tacit willingness to overlook obvious conceptual and syntactic errors in order to believe in the intelligence of an agent, perhaps we need a new critical vocabulary to account for the hedging we must now do on the question of actual, as opposed to illusory intelligence. Regardless of whether or not GPT-2 understands in the full sense the symbols it is processing, it is indisputable that it “has [untaught] faculties... specific skills, that require a certain precision of thought,” as the Slate Star Codex blogger delineates [Alexander 2019].

[12] In the Spring of COVID-19, the game introduced weekly scenarios on quarantine and Tiger King that reflected the zeitgeist of the moment. These adventures now appear as archived genre options.

[13] If interactive fiction as evinced most notably by *Adventure* and *Zork* relies on the structure of the puzzle to control the unfolding of the narrative, AID, both in its generic template and “custom” modes, offers what might be generally characterized as free play [Montfort 2003]. The difference is most stark at AID’s command line, where input is not constrained by pre-scripted actions, allowing players’ flights of fancy to translate more or less seamlessly into the game world. Narrative progression thus depends less on puzzle solving and critical thinking and more on the players’ own writing.

[14] If we only do a rules-based evaluation, either statistical or linguistic, in order to try to understand a large language model, we risk missing what is happening at the level of rhetoric (for translator Gayatri Spivak, rhetoric is the plane or dimension of language that one has to access in order to know and sense the voice of a text in a different language; it is what makes it possible to inhabit someone else’s umwelt). A purely technical analysis would also sideline the element of social contract and reduce language to a set of rules only. As we will later note with respect to its probability distributions, what makes GPT-2 work are the moments when it breaks with the rules of grammar and logic and becomes rhetorical, the best example of which is “Ovid’s unicorn.”

[15] The release of GPT-3, the next iteration of the model, on May 28, 2020, when we were in the end stages of writing this article, has made us even more acutely conscious of the difficulties of stabilizing our object of inquiry. Six months on, regular AID gameplay is still limited to GPT-2, but GPT-3 has been made available for some creative experimentation (e.g. Branwen [2020]) and can now be accessed as a “Dragon model” with an AID premium subscription.

[16] GPT-2, and Transformer models more generally, are examples of *deep* learning, the operations of which are generally held to be less interpretable than supervised learning models with algorithms such as k-nearest neighbors and linear regression.

[17] See Yang et al. (2019) for an argument for making machine learning models accessible and interactive, albeit not playable.

[18] As befits its history as a fundamental concept for cybernetics [Ashby 1957], the “black box” metaphor is ubiquitous in discussions of artificial intelligence and often used as a shorthand for the problems of explainability and interpretability [Adadi and Berrada 2018] [Russell 2019]. It is interesting to consider the relations between this notion of obscuration and the more sinister, political usages of the concept in, for example, *The Black Box Society* [Pasquale 2016].

[19] Our article was written before the publication of [McGillivray et al. 2020], but it aligns with their call for more collaborations and connections between the Natural Language Processing and Digital Humanities communities.

[20] If prior training data from Project Gutenberg and Wikipedia tacitly suggested, in T.S. Eliot’s language, “the common word exact without

vulgarity," which is to say standard English, with all the notions of the proper and the correct that implies, the WebText corpus suggests instead that there is no common word. It is training for a language model that does not itself model communication.

[21] We note that GPT-3 is so large that OpenAI had to guard against an ouroboros problem by vetting its training data to ensure that datasets used for evaluation were not themselves incorporated into the training data [Brown et al. 2020, 30]. This indicates the extent to which language models perform exponentially better as the datasets become more comprehensive [Halevy et al. 2009] [Banko and Brill 2001].

[22] The uncanny liveliness of AID's writing about magical unicorn blood, then, results not only from its adherence to genre templates, but also from its slight break from the obvious and the expected. One conclusion to draw from this: humans may seem to display a preference for appropriation, mimesis, and memetic expression — everyone is always copying everyone else — but in actual linguistic practice, turbulent distribution is the mark of an authentic "human" style.

[23] On autoregression, see Karpathy (2015). When it was released, the game fed GPT-2 up to the last eight pairs of player input and game response for prediction, but this has since been expanded [Walton 2019d]. The game also allows players to pin certain lines to the language model's memory context, which are always fed into the model at each prediction step.

[24] Work by Bengio et al. (2003) and Xu and Rudnicky (2000) has seen LMs in recent years take the form of a neural network [Jing and Xu 2019], and work by Vaswani et al. (2017) has seen the network architecture (or type) of the best LMs at present to be Transformers.

[25] Mark Marino's initial articulation of "Critical Code Studies," which synthesized a range of practices and conversations about "codework" and how the humanities ought to think about programming languages, proposed "that we no longer speak of the code as a text in metaphorical terms, but that we begin to analyze and explicate code as a text, as a sign system with its own rhetoric, as verbal communication that possesses significance in excess of its functional utility" [Marino 2006]. In the book form of the argument, the call to "read code the way we read poetry," which summons the entire critical apparatus of textual studies, semiotics, deconstruction, critical theory, and cultural studies for this purpose, is presented in the form of the manifesto [Marino 2020, 31]. Marino is on this point following Alexander Galloway's articulation of computers as "fundamentally a textual medium...based on a technological language called code" [Galloway 2004, xxiii–xxiv]. So, too, Dennis Tenen encourages those who might regard themselves as mere users of computational technology "to apply the same critical acuity they employ in the close reading of prose and poetry to the understanding of code and machine" [Tenen 2017, 21]. Foundational for this vein of thought is Michael Mateas' concept of "procedural literacy," which he defines as "the ability to read and write processes, to engage procedural representation and aesthetics, to understand the interplay between the culturally-embedded practices of human meaning-making and technically-mediated processes" [Mateas 2005, 101].

[26] One of the most influential versions of this literacy argument is made by Noah Wardrip-Fruin in his aforementioned inaugural work of software studies [Wardrip-Fruin 2012].

[27] For a general catalog of research on the epistemological problem of interpretable machine learning, see WE1S (2020).

[28] One field of study that works toward a technical understanding of NLP operations is "BERTology," which investigates large Transformer-based language models like BERT and GPT-2. Common research in this field attempts to interpret how a model processes data while revealing their inner representation (parameters, weights, hidden states) (e.g Tenney et al. [2019]).

[29] David Berry makes the additional point that complex math itself presents a high bar, thus necessitating analogies and explanatory models whose aesthetics or metaphorical functioning will also require examination [Berry 2018].

[30] As of this writing, there are 446 forks of AID's GitHub repository, the most notable of which are cloveranon and thadunge2, the two most popular unofficial releases of the game that implemented their own features. Additionally, an app- and ad-based copycat of the game ("The Infinite Story") has prompted a debate within the community about IP and AID's open-source model.

[31] Fixing the random seed of a specific instantiation of GPT-2, and sampling only the most probable sequence, will result in reproducible results. Because a trained neural network is still necessarily deterministic by its algorithmic design, it would in fact be possible to perform a limited close reading of a specific instantiation of GPT-2 or AID, but this would be to miss the forest for a tree branch.

[32] It is on this basis that we suggest that examining a language model necessarily requires considering it both as a statistical distribution and a sociotechnical assemblage, with the recognition, as Tarleton Gillespie argues, that this runs the risk of obscuring the "people involved at every point: people debating the models, cleaning the training data, designing the algorithms, tuning the parameters, deciding on which algorithms to depend on in which context" [Gillespie 2016, 22].

[33] The benchmarking tasks range from linguistic acceptability (determining whether a sentence makes linguistic sense) to coreference inference (reading a sentence with a pronoun and choosing the correct referent from a list, akin to the Winograd Schema Challenge). GLUE results from the paradigm shift from single, task-specific language models to transfer learning models that have demonstrated a general understanding of a "broad range" of "canonical and fine-grained linguistic tasks" [McCormick and Ryan 2019].

[34] Fine-tuning on GLUE was delegated as future work in the conclusion of the GPT-2 paper [Radford et al. 2019].

[35] We take text generation to mean content generation, i.e. news articles, narratives, software code.

[36] Recently, the Google team released a new BERT-based metric that achieved results closer to human performance. Aptly named BLEURT, the metric was pre-trained like BERT and then fine-tuned on an NLG evaluation dataset [Sellam et al 2020].

[37] The use of AID to produce descriptions of hypothetical memes brings a provocative question for future research to the fore: what would a formalization for good versus bad memes look like? An institutional decision not to evaluate non-formal textual outputs might, we anticipate, be made on the basis of sociocultural value, which would presume a greater significance for news reports or code completions than for memes. Part of the significance of AID, however, is that it reminds us (again) how arbitrary such distinctions truly are, and not simply because of the vernacular content of the story archive used for fine tuning.

[38] It is possible to do a limited evaluation of AID in terms of interactive fiction benchmarks, in the vein of scholarship on the believability of autonomous agents; for example, one could consider sample AID playthroughs in terms of Emily Short's guidelines for conversation model design [Short 2007]. As we will demonstrate in Section 4, however, AID has only a family resemblance to parser adventure stories, so using IF as a benchmark would necessarily be a limited exercise.

[39] The process involves initializing the language model's weights by the pre-training corpus; in more basic terms, the model first learns the syntactic and grammatical nuances of language [Ruder 2018] [Sarkar 2018], which are then updated accordingly by a fine-tuning corpus. Fine-tuning here means shaping the model's output toward a specific mode or genre of writing, e.g. computer code, recipes, Chinese classical poetry, video game walkthroughs, or Reddit submission titles.

[40] We used Woolf's (2019) simplification of GPT-2 to conduct our experiments, fine-tuning the 355m parameter model with its parameters' factory settings. The kernel of the work was the formulation of our speculative queries — for example, "if we fine-tuned GPT-2 on *x* and gave it input *y*, would it generate *z*?" — and the formatting of our training data accordingly. So that our work can be verified and developed further, we refer readers to our Google Colab notebook (https://colab.research.google.com/drive/1obL0qdJRyF9KQYiDkRCwjKWhTnwYBrhd?usp=sharing) for the exact parameters used in the experiment represented in Figure 2.

[41] We were guided here by Shawn Presser's heuristic for forcing stanzaic line breaks, via [Branwen 2019].

[42] As an example of the financial and compute resources required to pre-train a large language model, OpenAI reports that the pre-training of GPT-3, its 175B parameter model, cost $4.6 million, and would have taken 335 years without advanced computing [Brown et al. 2020, 46].

[43] With a specific focus on authorship, Aarthi Vadde provides an account of the phenomenon of "mass amateurization" in the "critical, creative, and communicative arts, allowing amateurs to bypass the gatekeeping practices of specific institutions" [Vadde 2017, 27].

[44] Players can also seek explanations of AID through its community on the Discord server, via gameplay itself, and the "Help" section.

[45] For perspective on the scale of the user base, we note that in May 2020, the AID subreddit had approximately the same number of subscribers as the subreddit for the Democratic presidential candidate, Vice President Joe Biden. For up-to-date statistics for r/AIDungeon, see https://subredditstats.com/r/aidungeon.

[46] Posts on the game's social media communities are disproportionately dominated by screen captures of gameplay, with players trying to outdo each other's weirdness with posts of novel outputs, from the mildly amusing to the shockingly hilarious. Competitive creativity has by no means been absent from the journalistic coverage of the game either, with the discussion almost resembling teams of scientists trying to outdo each other's findings: Shane "discovered" that you can roleplay as a nonhuman character, Robertson pushed the game towards the meta, and almost everyone playing has soon learned for themselves that the game's AI is quite depraved.

[47] Consider in this regard how AID allows players to tweak the language model's randomness in the settings, or, players are able to tune one of many of GPT-2's hyperparameters, all without needing further knowledge of deep learning. That "temperature" is advertised as "randomness" is just the start of how AID gamifies working with language models.

[48] OpenAI's report of GPT-2 cites Branwen (2019) as a literary implementation and AID as a gaming implementation of GPT-2 [OpenAI 2019b].

[49] Partnering with the University of Oregon, OpenAI claims to be developing a battery of "bias probes" or "input[s] to a model designed to elucidate the model's disposition towards producing certain kinds of outputs" in order to map GPT-2's racial, gender, and even "conspiracy theories" biases [OpenAI 2019b].

[50] On informal, hands-on, or experiential forms of expertise, also see Collins and Evans (2007).

[51] For an archive, see the subreddit's custom prompt megathread at https://www.reddit.com/r/AIDungeon/comments/e82ia5/custom_prompt_megathread/.

[52] It is striking that players use AID to test GPT-2's ability to generate content distinct from text adventures, which suggests that it is not simply genre that engages and retains users. This line of thought is underscored by the fact that King's Talk to Transformer implementation remains available, yet most of the experimentation with GPT-2 continues to be performed and documented on AID's platform.

[53] Because AID seems to have evolved almost by the hour throughout the first half of 2020, our analysis of its features, modes, and commands should be read with a date-timestamp. We can though speak to that class of tools that allow for editing and revision because their function has been continuous and they are integral to our argument about AID as an model of and for citizen NLP. Future research can focus on new developments such as a scripting feature that allows users to write custom JavaScript code to modify the game's logic.

[54] Many aspects of AID evoke the legacy of IF, most notably its command line aesthetic and generic templates. In actual practice, however, AID is markedly different, because of both the lack of restrictions on player input and the mechanics, particularly "alter," which is experienced as a writing *with* rather than *against* the game. Although a strict comparative schema for each is outside of our purview here, we can still point to AID as a model for a potential future of IF in its offering of "a more profound and responsive type of systematic world," as Montfort (2003) puts it.

[55] For attempts at solving bias in NLP see OpenAI (2019b) and Bender and Friedman (2018).

[56] We might remark as well on the extent to which the WebText corpus is already a human-machine hybrid, given the array of algorithmic writing assistants now in common use.

[57] Walton has claimed that the team has a variety of metrics derived from player engagement, including explicit feedback and user behavior, that are used to determine whether a particular iteration of GPT-2 is working [AWS 2020].

[58] In 2018, GPT-1 fell under the broad category of semi-supervised learning, in which the model was pre-trained in an unsupervised manner but later fine-tuning saw influences from supervised learning [Radford et al. 2018]. Fast forward two years and GPT-3 does away with the supervised learning portion, with researchers decrying the difficulty of obtaining high-quality fine-tuning datasets [Brown et al. 2020, 3]. The vision for the team was text generation without the need for fine-tuning, or at least with very limited fine-tuning, but AID's model demonstrates that there is still value in gathering user feedback. After all, as we have noted, the language models of today are not standalone text generators, but consumer products, the revision and improvement of which has material value.

[59] To sum up the argument against using IF as a benchmark: AID is not a goal-oriented game that can be won or lost but rather an experimental sandbox that can produce not just stories but also code, recipes, and music.

[60] The OpenAI team extends its gratitude to "the millions of people who created content that was used in the training of the model, and to those who were involved in indexing or upvoting the content (in the case of WebText)" [Brown et al. 2020, 40] but a meaningful contrast can be drawn between this bracketing of citizen participation and AID's inviting of meaningful and continuous evaluation from its players.

Works Cited

AWS 2020 "The Digital Download." Amazon Web Services Game Tech (May 20, 2020). https://aws.amazon.com/gametech/events/digital-download-online/.

Adadi and Berrada 2018 Adadi, A. and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6 (2018). https://ieeexplore.ieee.org/document/8466590

Alammar 2019 Alammar, J. "The Illustrated GPT-2 (Visualizing Transformer Language Models)." *Jay Alammar Blog* (August 2019). http://jalammar.github.io/illustrated-gpt2/.

Alexander 2019 Alexander, S. “GPT-2 As Step Toward General Intelligence.” *Slate Star Codex* (February 19, 2019). https://slatestarcodex.com/2019/02/19/gpt-2-as-step-toward-general-intelligence/.

Alzantot et al. 2018 Alzantot, M. et al. “Generating Natural Language Adversarial Examples.” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (October-November 2018). https://www.aclweb.org/anthology/D18-1316.pdf

Ars 2020 Ars Staff. “The machines are whispering: We tested *AI Dungeon 2* and cannot stop laughing.” *Ars Technica* (January 20, 2020). https://arstechnica.com/gaming/2020/01/we-test-ai-dungeon-2-a-text-adventure-that-creates-itself-with-your-help/.

Ashby 1957 Ashby, R. *An Introduction to Cybernetics*. Chapman & Hall, London (1957).

Banko and Brill 2001 Banko, M. and E. Brill. “Scaling to very very large corpora for natural language disambiguation.” *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (July 2001), pp. 26–33.

Bender and Friedman 2018 Bender, E. and B. Friedman. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.” *Transactions of the Association for Computational Linguistics* 6 (2018): pp. 587–604. https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00041.

Bengio et al. 2003 Bengio, Y., et al. “A neural probabilistic language model.” *Journal of Machine Learning Research*, 3 (2003), pp. 1137–1155.

Berry 2018 Berry, D. “Explainable Aesthetics.” *Stunlaw* (October 2, 2018). http://stunlaw.blogspot.com/2018/10/explainable-aesthetics.html.

Bogost 2015 Bogost, I. “The Cathedral of Computation.” *The Atlantic* (January 15, 2015). https://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/.

Branwen 2019 Branwen, G. “GPT-2 Neural Network Poetry.” *Gwern.net* (March 2019). https://www.gwern.net/GPT-2.

Branwen 2020 Branwen, G. “GPT-3 Creative Fiction.” *Gwern.net* (June 2020). https://www.gwern.net/GPT-3.

Brown et al. 2020 Brown, T.B., et al. “Language Models are Few-Shot Learners.” arXiv repository (2020). https://arxiv.org/abs/2005.14165.

Buonocore 2019 Buonocore, T. “Man is to Doctor as Woman is to Nurse: the Gender Bias of Word Embeddings.” Towards Data Science (March 8, 2019). https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17.

Chan 2019 Chan, S.“ AI Dungeon 2: generative Cattelan & the art museum.” *Medium* (December 2019). https://medium.com/@sebchan/ai-dungeon-2-generative-cattelan-the-art-museum-af16eac989ec

Collins and Evans 2007 Collins, H. and R. Evans. *Rethinking Expertise*. University of Chicago Press, Chicago (2007).

Galloway 2004 Galloway, A. *Protocol: How Control Exists After Decentralization*. MIT Press, Cambridge (2004).

Gillespie 2016 Gillespie, T. “Algorithm.” In B. Peters (ed), *Digital Keywords: A Vocabulary of Information Society and Culture*, Princeton UP, Princeton (2016), pp. 18–30.

Goodwin 2019 Goodwin, R. “Ghost Flights.” GitHub repository (2019). https://github.com/NaNoGenMo/2019/issues/46.

Guzdial 2015 Guzdial, M., et al. “Crowdsourcing Open Interactive Narrative.” https://www.cc.gatech.edu/~riedl/pubs/guzdial-fdg15.pdf

Halevy et al. 2009 Halevy, A. et al. “The Unreasonable Effectiveness of Data.” *IEEE Intelligent Systems* (2009), vol. 24, no. 02, pp. 8–12.

Han Lau et al. 2018 Han Lau, J. et al. *Deep-speare: A joint neural model of poetic language, meter and rhyme. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)* (2018), pp. 1948–1958.

Harris 2020 Harris, J. “Creating the ever-improvising text adventures of AI Dungeon 2.” *Gamasutra* (January 2020). https://www.gamasutra.com/view/news/356305/Creating_the_everimprovising_text_adventures_of_AI_Dungeon_2.php

Hitt 2019 Hitt, T. “Meet the Mormon College Student Behind the Viral A.I. Game That Took Dungeons & Dragons Online.” *Daily Beast* (December 2019). https://www.thedailybeast.com/meet-the-mormon-college-student-behind-viral-artificial-intelligence-game-ai-dungeon

Holtzman et al. 2019 Holtzman, A., et al. “The Curious Case of Neural Text *De*generation.” arXiv Preprint (April 2019). https://arxiv.org/abs/1904.09751.

Jing and Xu 2019 Jing, K. and Xu J. “A Survey on Neural Network Language Models.” arXiv repository (2019). https://arxiv.org/abs/1906.03591.

Johnston 2019 Johnston, D. *ReRites*. *Glia: Digital Poetry* (2019). http://glia.ca/rerites/.

Karpathy 2015 Karpathy, A. “The Unreasonable Effectiveness of Recurrent Neural Networks.” Andrej Karpathy Blog (May 21, 2015). https://karpathy.github.io/2015/05/21/rnn-effectiveness/.

Knight 2017 Knight, W. “The Dark Secret at the Heart of AI.” *MIT Technology Review* (April 11, 2017). https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/.

Kurzweil 2012 Kurzweil, R. *How to Create a Mind: The Secret of Human Thought Revealed*. Viking Penguin, New York (2012).

LeCun 2018 Yann, L. “Self-Supervised Learning.” Distinguished Lecture Series in Data Science, UC Santa Barbara (November 8, 2018).

Lillicrap and Kording 2019 Lillicrap, T. and K. Kording. “What Does It Mean to Understand a Neural Network?” arXiv preprint (July 2019). https://arxiv.org/abs/1907.06374

Marcus 2018 Marcus, G. “Deep Learning: A Critical Appraisal.” arXiv preprint (January 2018). https://arxiv.org/abs/1801.00631.

Marino 2006 Marino, M. “Critical Code Studies.” *Electronic Book Review* 4 (December 4, 2006). https://electronicbookreview.com/essay/critical-code-studies/.

Marino 2020 Marino, M. *Critical Code Studies*. MIT Press, Cambridge (2020).

Mateas 2005 Mateas, M. “Procedural Literacy: Educating the New Media Practitioner.” *On the Horizon* 13.2 (2005), pp. 101–111.

McCormick and Ryan 2019 McCormick, C. and N. Ryan. “GLUE Explained: Understanding BERT Through Benchmarks.” Chris McCormick Blog (November 5, 2019). https://mccormickml.com/2019/11/05/GLUE/.

McGillivray et al. 2020 McGillivray, B., et al. “Digital Humanities and Natural Language Processing: ‘Je t’aime... Moi non plus.’” *Digital Humanities Quarterly*, 14.2 (2020). http://www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html

Milburn 2015 Milburn, C. *Mondo Nano: Fun and Games in the World of Digital Matter*. Duke UP, Durham (2015).

Montfort 2003 Montfort, N. *Twisty Little Passages: An Approach to Interactive Fiction*. MIT Press, Cambridge (2003).

Montfort et al. 2012 Montfort, N., et al. *10 PRINT CHR$(205.5+RND(1));:GOTO 10*. MIT Press, Cambridge (2012).

Novikova et al 2017 Novikova, J. “Why We Need New Evaluation Metrics for NLG.” Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2241–2252 Copenhagen, Denmark, (September 2017). https://nld.ict.usc.edu/cs644-spring2020/discussions/novikova-etal-emnlp2017.pdf

OpenAI 2019a “Better Language Models and Their Implications.” *OpenAI Blog* (February 2019). https://openai.com/blog/better-language-models/.

OpenAI 2019b “Release Strategies and the Social Impacts of Language Models.” arXiv preprint (November 2019). https://arxiv.org/pdf/1908.09203.pdf

OpenAI 2020 “OpenAI API.” (June 2020). https://openai.com/blog/openai-api/

Pasquale 2016 Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard UP, Cambridge (2016).

Radford et al. 2018 Radford, A. “Improving Language Understanding by Generative Pre-Training.” *OpenAI Blog* (2018). https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford et al. 2019 Radford, A. “Language Models are Unsupervised Multitask Learners.” *OpenAI Blog* (2019). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Reiter 2017 Reiter, E. “How to do an NLG Evaluation: Metrics.” *Ehud Reiter’s Blog* (May 3, 2017).

https://ehudreiter.com/2017/05/03/metrics-nlg-evaluation/

Ruder 2018 Ruder, S. “NLP’s ImageNet moment has arrived.” *Sebastian Ruder Blog* (July 12, 2018). https://ruder.io/nlp-imagenet/.

Russell 2019 Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, NY (2019).

Sample 2020 Sample, M. “AI Dungeon and Creativity.” *SAMPLE REALITY* (January 2020). https://www.samplereality.com/2020/01/28/ai-dungeon-and-creativity/.

Sarkar 2018 Sarkar, D. “A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning.” *towards data science* (November 14, 2018). https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a

See 2019 See, A. “Natural Language Generation.” CS224N/Ling 284: Natural Language Processing with Deep Learning (2019). https://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture15-nlg.pdf.

Sellam et al 2020 Sellam, T. et al. “BLEURT: Learning Robust Metrics for Text Generation.” arXiv preprint (May 2020). https://arxiv.org/abs/2004.04696

Shane 2019 Shane, J. “Play AI Dungeon 2. Become a dragon. Eat the moon.” *AI Weirdness* (December 2019). https://aiweirdness.com/post/189511103367/play-ai-dungeon-2-become-a-dragon-eat-the-moon.

Short 2007 Short, E. “Conversation.” *Emily Short’s Interactive Storytelling* (2007). http://emshort.wordpress.com/writing-if/my-articles/conversation/.

Tenen 2017 Tenen, D. *Plain Text: The Poetics of Computation*. Columbia UP, New York (2017).

Tenney et al. 2019 Tenney, I., et al. “BERT Rediscovers the Classical NLP Pipeline.” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019): pp. 4593–4601.

Vadde 2017 Vadde, A. “Amateur Creativity: Contemporary Literature and the Digital Publishing Scene.” *New Literary History*, 48.1 (Winter 2017): 27–51.

Vincent 2019 Vincent, J. “This AI text adventure game has pretty much infinite possibilities.” *The Verge* (December 2019). https://www.theverge.com/tldr/2019/12/6/20998993/ai-dungeon-2-choose-your-own-adventure-game-text-nick-walton-gpt-machine-learning

Viswanathan 2020 Viswanathan, S. “Beware of Weight Poisoning in Transfer Learning.” *Towards Data Science* (May 4, 2020). https://towardsdatascience.com/beware-of-weight-poisoning-in-transfer-learning-4c09b63f8353.

WE1S 2020 WE1S. “Bibliography – Interpretability and Explainability.” WE1S: A 4Humanities Project (2020). https://we1s.ucsb.edu/research/we1s-bibliography/bibliography-interpretability-and-explainability/.

Walton 2019a Walton, N. “About *AI Dungeon*.” http://ai-adventure.appspot.com/about.html

Walton 2019b Walton, N. “AI Dungeon 2: Creating Infinitely Generated Text Adventures with Deep Learning Language Models.” *Perception, Control, Cognition* (November 21, 2019). https://pcc.cs.byu.edu/2019/11/21/ai-dungeon-2-creating-infinitely-generated-text-adventures-with-deep-learning-language-models/

Walton 2019c Walton, N. Twitter post, November 23, 2019. https://twitter.com/nickwalton00/status/1198295331449888768

Walton 2019d Walton, N. “AI-Dungeon.” GitHub repository (2019). https://github.com/AIDungeon/AIDungeon.

Walton 2020 Walton, N. “How we scaled AI Dungeon 2 to support over 1,000,000 users.” *Medium* (February 11, 2020). https://medium.com/@aidungeon/how-we-scaled-ai-dungeon-2-to-support-over-1-000-000-users-d207d5623de9

Wang et al. 2018 Wang, A., et al. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.” *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (November 1, 2018), pp. 353–355. https://www.aclweb.org/anthology/W18-5446.pdf.

Wang et al. 2019 Wang, A., et al. “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.” arXiv preprint (May 2019). https://arxiv.org/abs/1905.00537

Wardrip-Fruin 2012 Wardrip-Fruin, N. *Expressive Processing: Digital Fictions, Computer Games, and Software Studies*. MIT Press, Cambridge (2012).

Weston et al. 2014 Weston, J., et al. “Memory Networks.” arXiv preprint (2014). https://arxiv.org/abs/1410.3916.

Whitmore 2019 Whitmore, N. “GPT2 Adventure.” Google Colaboratory Notebook (2019). https://colab.research.google.com/drive/1khUaPex-gyk1wXXLuqcopiWmHmcKl4UP.

Woolf 2019 Woolf, M. “How To Make Custom AI-Generated Text With GPT-2.” *Max Woolf’s Blog* (September 4, 2019). https://minimaxir.com/2019/09/howto-gpt2/.

Xu and Rudnicky 2000 Xu, and Rudnicky. “Language Modeling for Dialog System.” *Sixth International Conference on Spoken Language Processing* (ICSLP 2000). https://www.isca-speech.org/archive/icslp_2000/i00_1118.html.

Yang et al. 2017 Yang, Z. “Mastering the Dungeon: Grounded Language Learning by Mechanical Turker Descent.” arXiv preprint (November 2017). https://arxiv.org/abs/1711.07950

Yang et al. 2019 Yang, Y., et al. “A Study on Interaction in Human-in-the-Loop Machine Learning for Text Analytics.” *IUI Workshops 2019* (March 2019). http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-9.pdf.