Research Statement
Daphne Ippolito

One of the oldest yet most elusive promises of AI is computers that can converse with humans, not just via rigidly structured templates and programming languages, but in natural language. This problem is formalized as Natural Language Generation (NLG), the task of writing novel sentences in a human language such as English. When I first began conducting research on NLG, my goal was to explore open-ended tasks, such as dialog and story writing, where a human and AI system interact to write text. However, I quickly ran into the problem that machine-generated text was frustratingly insufficient for my ambitions. Though state-of-the-art NLG systems can now produce text that, at a surface-level, is fluent and grammatical, generated text still differs in significant ways from the text humans would write. The issues range from common-sense errors most people easily pick up on, like writing about a four-horned unicorn, to subtle statistical anomalies, like using the word "the" 25% more frequently than human-written text. I realized that if I am to advance the applications of NLG, I need to simultaneously pursue a deeper understanding of its limitations and methods for mitigation.

In my work as a PhD student at the University of Pennsylvania and a researcher with Google Brain, **I have developed a research program focused on advancing state-of-the-art NLG while pursuing thoughtful analysis of its limitations and ramifications**. For example, I have found that minor changes to the algorithms used to generate text can significantly impact the manner in which it deviates from genuine human text. I have also measured how the neural networks used for NLG have a worrying tendency to memorize and regurgitate entire passages from their training data, including personal conversations and copyrighted book chapters. Finally, I have worked on extending NLG's capabilities to serve as a collaborative partner in creative writing applications, a domain where the aforementioned limitations I've researched have significant consequences.

The study of NLG has taken on surprisingly broad societal importance as researchers rush toward bigger, more powerful models and practitioners incorporate language generation into increasingly far-flung applications, from text adventure games and educational tools to email and code writing assistants. With the expertises that I gained at UPenn and at Google Brain, I am well positioned to help lead this emerging subfield of artificial intelligence.

The Detection of Machine Generated Text

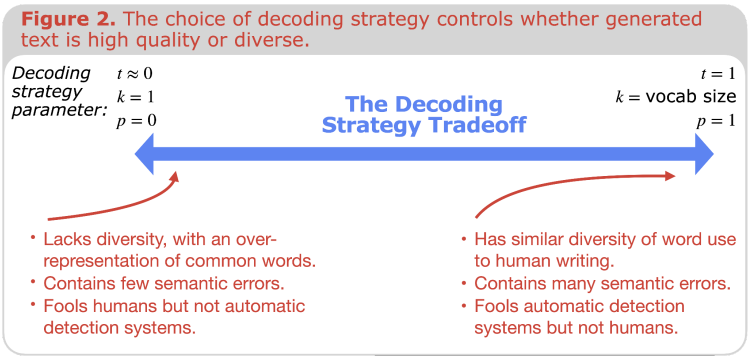
The proliferation of machine-generated text, especially when it lacks attribution, has entered the sphere of public concern. Generated text is now both easy to produce, requiring almost no technical expertise, and remarkably fluent, which could enable nefarious uses such as fabricated product reviews and social media posts. **In my work [2], we measure the ability of humans as well as automatic systems to detect machine-generated text.** Understanding detectability is imperative because it gives us a proxy for how far along generative systems are at fooling humans and whether undesired use of machine-generated text can be mitigated.

Figure 1. Generated text (blue), such as fake reviews, can have negative societal implications.

```
Reviews for Unicorn University
5/5
I loved my time at Unicorn University. It's a great place to learn new things and meet new people. I even made friends with some of the teachers. I'm still in shock that I was accepted into Unicorn University. I never thought I would get in. I can't wait to learn all sorts of magical things.

1/5
Unicorn University is terrible. I was expecting a school that teaches unicorns magic and flying. Instead, all they do is sit in a room and do nothing.
```

To discuss the factors that influence detectability, it is first necessary to understand *how* text is generated. Neural language models are the primary tool used by modern systems to generate text. Language models typically work by taking as input a sequence of tokens[1] and predicting a probability distribution over all the words in the vocabulary for what the next token in the sequence should be. Text is generated by repeatedly sampling a token from the predicted distribution and then appending it to the input for the next step of generation. The choice of sampling algorithm (a.k.a decoding method) has a significant impact on the nature of the text that gets generated, and as a result, on the detectability of this text.



My experiments showed that generated text can be fluent or diverse but usually not both (Figure 2). A decoding method that only chooses tokens the model reports as very high likelihood results in generated text that humans find very convincing. However, since the resulting text lacks the lexical diversity found in genuine human-written text, even basic automatic detection systems can pick up on the altered word statistics in order to predict that the text was machine-generated.

On the other end of the spectrum, we could randomly sample directly from the model's predicted distribution. However, neural language models tend to predict distributions with a long tail of low-likelihood tokens; random sampling thus results in text that at best makes occasional poor word choices and at worst is entirely incoherent. Human raters can readily pick up on these errors, but automatic detection systems are often fooled by the fact that, in terms of basic word statistics, the text looks like what a human would write.

Thus, in my papers [1] and [2] I show that even when the underlying language model is fixed, there exists a tradeoff between fluency and diversity created by the choice of decoding method. At the time this work was completed, the decoding tradeoff was poorly understood. Studies commonly overrepresented the performance of novel decoding strategies or else compared NLG system generations where both the decoding strategy and underlying model were altered, creating an unfair comparison. The field has since become better at reporting choice of decoding strategy and making fair comparisons between systems. My work documenting the diversity-quality tradeoff has also been valuable to NLG practitioners who must make a subjective decision of how much noise they are willing to tolerate for the benefit of generating text with lexical diversity on par with real human writing.

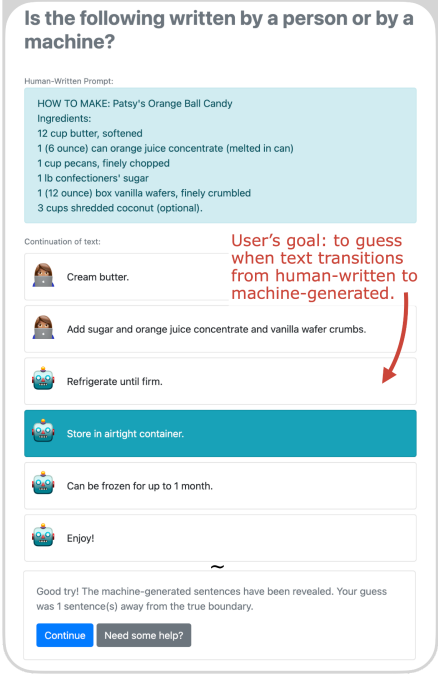
Helping the public to understand what machine-generated text looks like and to acquire the skills to detect it is important for mitigating its misuse**. In [3], I show that it is possible to simultaneously educate the public on what generated text looks like while collecting valuable data for comparing open-ended text generation systems.** I extended the work of [2] by building a public web platform for anyone to test their detection skills (Figure 3). The goals of the project, called Real or Fake Text (RoFT), were two-fold: (1) to engage the public in the task of identifying machine-generated text and (2) to collect data

[1] When generating English, tokens are usually individual words or pieces of words.

enabling the analysis of the variety of factors that have an impact on detectability. For example, I was interested in how different domains (e.g., news vs. fiction stories) are easier or harder for NLG systems to impersonate, and how factors like the size and training data of the neural language model also affect its generation abilities.

To incentivize users to try their best, we gamified the detection task, adding a point system and a leaderboard. To date, the Real or Fake Text game has had over 500 users who have contributed 40,000 annotations across four domains and three NLG systems. Moreover, we found that, when properly incentivized, annotators were able to improve at the detection task over time. We also found that different domains, model sizes, training objectives, and decoding strategies have significant measurable differences in detectability. These results suggest that modern extensions of the Turing test are a promising direction for evaluating NLG.

Figure 3. RoFT is a game-like platform to evaluate NLG systems.

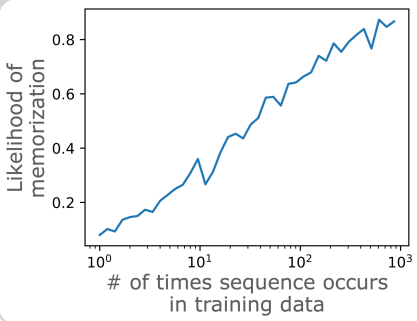


Memorization of Training Data

Machine-generated text is most undetectable when it looks *exactly* like its training data. Given that neural language models are trained to maximize the likelihood of their training data, what's stopping them from simply memorizing and regurgitating exact text sequences they were trained on? This sort of memorization is directly harmful if it breaches expectations of privacy or content ownership from those whose data is included in the train set. It also reduces generalizability if models are biased toward examples that are not representative of the underlying distribution of natural language.

In our recent paper [4], I show that large numbers of near-duplicate examples in the training data for neural language models significantly increases their tendency to memorize (Figure 4). Deduplication of the training data results in more efficient training (since dataset size is reduced) without harming model performance, and providing improved generalization to out-of-domain text. In ongoing work, my team is measuring the scaling laws of memorization, showing that memorization frequency increases logarithmically with model size and the number of times a training example is repeated.

Figure 4. A neural language model is more likely to memorize sequences it saw frequently during training.



AI Tools for Creative Writers

One application where NLG has considerable potential is in the development of tools for creative writing. Creative writing is an attractive application to work in because ideation tools are already part of writers' arsenal, and mistakes like hallucinating false facts are less problematic in fiction than in domains like automatic news summarization, where faithfulness to the real world is crucial. In addition, writers have been grappling with the concept of sentient, human-like machines for at least as long as computer scientists have.

In my research, I am interested in bridging the gap between what NLG systems do by default (continue from a prompt) and the operations writers actually would want. **To study how creative writers perceive AI writing assistance, my Google collaborators and I built WordCraft [5], a text editor with generative text tools built in.** Users can ask the AI to insert new text at their cursor position; elaborate on a selected person or other entity; or rewrite selected text to be more happy, Shakespearean, metaphorical, or any other "rewrite" instruction of their choice (Figure 5). **The ability to execute arbitrary rewriting operations using a general language model without any finetuning is supported by a technique we introduced in [6].** Lastly, Wordcraft also includes a chatbot interface where users can discuss their story with a dialog agent who can give suggestions on what to write next.

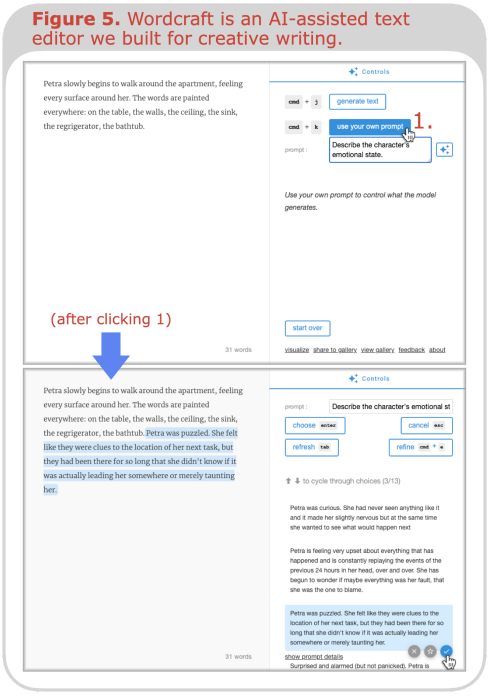


Figure 5. Wordcraft is an AI-assisted text editor we built for creative writing.

I think it is very important to evaluate NLG systems using individuals from their intended audience, even though doing so is more difficult and costly than recruiting low-paid crowdworkers on sites like Amazon Mechanical Turk. To this end, we are in the process of commissioning professional writers from a diverse set of genres–from poetry to science fiction to comics–to write stories using Wordcraft. In addition to publishing a digital literary magazine of the stories produced, we will carefully document the writers' perspectives on the provided tools' strengths and limitations, and their aspirations for what these tools should achieve.

Future Directions

As an assistant professor, I will continue my work on understanding the limitations and potentials of neural language models and methods to make them more useful, building collaborations with researchers in Human Computer Interaction, Privacy, and Machine Learning. Research employing neural networks and big data can require tremendous computational resources, often beyond the means of individual academic labs, so I plan to maintain and grow my collaborations with industry partners, such as Google, where I am currently a research scientist.

AI + Humans: Collaboration and System Evaluation

Much of the model evaluation and dataset collection in NLP is performed by crowdsourced non-expert workers who are paid per-annotation, a system that disincentivizes meticulous work and also may not sample from a broad slice of society. In my future work, I plan to continue exploring methods for improving annotation quality and diversity. Better methods for human annotation will allow us to more effectively evaluate existing and newly proposed NLG systems, especially in open-ended text generation domains where "good" is difficult to quantify. They will also facilitate the collection of high-quality training and benchmark datasets.

First, I would like to study how choices in annotator pool and task design influence the nature of the annotations that are collected. Second, I aim to investigate how AI-in-the-loop

interactive systems can accelerate and improve human annotation. For example, in a recent NeurIPS paper [7], my colleagues and I explored AI-human collaboration as a means of more efficiently building benchmark datasets with desirable deviations from the real world (such as balancing genders). Lastly, I am interested in developing metrics and visualizations that can be used by annotators on an example-by-example basis to improve the quality and diversity of a dataset as it is being constructed, rather than via post-hoc filtering.

Assessing the Risks of Large Language Models

In spring of 2020, as teams around the world began developing digital tracing apps to combat the spread of Covid-19, I contemplated the inherent tradeoffs between user privacy and public health interests [8, 9, 10]. This introduction to working with the Privacy and Security community taught me to probe for vulnerabilities and think skeptically about systems. It has significantly influenced how I look at neural language models and their inherent risks. I am interested in probing existing and widely-used language models' risk of divulging their training data, making use of membership inference and data extraction attacks. I am further interested in developing train-time techniques to enable controllable memorization—can we build systems that selectively memorize or copy from specifiable portions of their training data?

More Usable and Controllable Language Models

In NLG, it is typical to take a model pre-trained on general data and then fine-tune it for particular tasks. This paradigm is infeasible for larger models, both because of the computational expense of fine-tuning and because it is unsustainable to have *n* copies of a giant model stored in memory and ready for inference at the same time. Recent advances such as few-shot learning, prompt tuning, and future discriminators have allowed models to be adapted to particular tasks without the need for full fine-tuning. I believe there is much more progress to be done in this space. For example, I am interested in exploring alternative pre-training objectives that yield models with better support for task differentiation at inference time. I am also interested in improving model support for other inference paradigms besides simple prompt continuation, such as text rewriting and fill-in-the-blank. Such techniques would have broad impact on real-world applications.

References

[1] Ippolito, Daphne*, Reno Kriz*, João Sedoc, Maria Kustikova, and Chris Callison-Burch. "Comparison of Diverse Decoding Methods from Conditional Language Models." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3752-3762. 2019.

[2] Ippolito, Daphne*, Daniel Duckworth*, Chris Callison-Burch, and Douglas Eck. "Automatic Detection of Generated Text is Easiest when Humans are Fooled." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1808-1822. 2020.

[3] Dugan, Liam*, Daphne Ippolito*, Arun Kirubarajan*, and Chris Callison-Burch. "RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 189-196. 2020.

[4] Lee, Katherine*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." *arXiv preprint arXiv:2107.06499* (2021).

[5] Coenen, Andy, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. "Wordcraft: a Human-AI Collaborative Editor for Story Writing." arXiv preprint arXiv:2107.07430 (2021).

[6] Reif, Emily*, Daphne Ippolito*, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. "A Recipe for Arbitrary Text Style Transfer with Large Language Models." *arXiv preprint arXiv:2109.03910* (2021).

[7] Yuan, Ann, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. "SynthBio: A Case Study in Human-AI Collaborative Curation of Text Datasets." In *Proceedings of the Neural Information Processing Systems Datasets Track*. 2021.

[8] Cho, Hyunghoon, Daphne Ippolito, and Yun William Yu. "Contact Tracing Mobile Apps for COVID-19: Privacy Considerations and Related Trade-offs." *arXiv preprint arXiv:2003.11511* (2020).

[9] Bengio, Yoshua, Richard Janda, Yun William Yu, Daphne Ippolito, Max Jarvie, Dan Pilat, Brooke Struck, Sekoul Krastev, and Abhinav Sharma. "The Need for Privacy with Public Digital Contact Tracing During the COVID-19 Pandemic." *The Lancet Digital Health* 2, no. 7 (2020): e342-e344.

[10] Alsdurf, H., Edmond Belliveau, Yoshua Bengio, Tristan Deleu, Prateek Gupta, Daphne Ippolito, et al. "COVI White Paper. arXiv preprint 2020." *arXiv preprint arXiv:2005.08502* (2020).

* denotes co-first authors.