# AI-Assisted Collaborative Story Writing with Human-in-the-Loop

## Plan, Write, and Revise: an Interactive System for Open-Domain Story Generation

Story composition is a challenging problem for machines and even for humans. We present a neural narrative generation system that interacts with humans to generate stories. Our system has different levels of human interaction, which enables us to understand at what stage of story-writing human collaboration is most productive, both to improving story quality and human engagement in the writing process. We compare different varieties of interaction in story-writing, story-planning, and diversity controls under time constraints, and show that increased types of human collaboration at both planning and writing stages results in a 10-50% improvement in story quality as compared to less interactive baselines. We also show an accompanying increase in user engagement and satisfaction with stories as compared to our own less interactive systems and to previous turn-taking approaches to interaction. Finally, we find that humans tasked with collaboratively improving a particular characteristic of a story are in fact able to do so, which has implications for future uses of human-in-the-loop systems.

## Wordcraft: a Human-AI Collaborative Editor for Story Writing

As neural language models grow in effectiveness, they are increasingly being applied in real-world settings. However these applications tend to be limited in the modes of interaction they support. In this extended abstract, we propose Wordcraft, an AI-assisted editor for story writing in which a writer and a dialog system collaborate to write a story. Our novel interface uses few-shot learning and the natural affordances of conversation to support a variety of interactions. Our editor provides a sandbox for writers to probe the boundaries of transformer-based language models and paves the way for future human-in-the-loop training pipelines and novel evaluation methods.

## CharacterChat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot

We present CharacterChat, a concept and chatbot to support writers in creating fictional characters. Concretely, writers progressively turn the bot into their imagined character through conversation. We iteratively developed CharacterChat in a user-centred approach, starting with a survey on character creation with writers (N=30), followed by two qualitative user studies (N=7 and N=8). Our prototype combines two modes: (1) Guided prompts help writers define character attributes (e.g. User: "Your name is Jane."), including suggestions for attributes (e.g. Bot: "What is my main motivation?") and values, realised as a rule-based system with a concept network. (2) Open conversation with the chatbot helps writers explore their character and get inspiration, realised with a language model that takes into account the defined character attributes. Our user studies reveal benefits particularly for early stages of character creation, and challenges due to limited conversational capabilities. We conclude with lessons learned and ideas for future work.

## CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities

Large language models (LMs) offer unprecedented language generation capabilities and exciting opportunities for interaction design. However, their highly context-dependent capabilities are difficult to grasp and are often subjectively interpreted. In this paper, we argue that by curating and analyzing large interaction datasets, the HCI community can foster more incisive examinations of LMs' generative capabilities. Exemplifying this approach, we present CoAuthor, a dataset designed for revealing GPT-3's capabilities in assisting creative and argumentative writing. CoAuthor captures rich interactions between 63 writers and four instances of GPT-3 across 1445 writing sessions. We demonstrate that CoAuthor can address questions about GPT-3's language, ideation, and collaboration capabilities, and reveal its contribution as a writing "collaborator"under various definitions of good collaboration. Finally, we discuss how this work may facilitate a more principled discussion around LMs' promises and pitfalls in relation to interaction design. The dataset and an interface for replaying the writing sessions are publicly available at <https://coauthor.stanford.edu>.

## Wordcraft: Story Writing With Large Language Models

The latest generation of large neural language models such as GPT-3 have achieved new levels of performance on benchmarks for language understanding and generation. These models have even demonstrated an ability to perform arbitrary tasks without explicit training. In this work, we sought to learn how people might use such models in the process of creative writing. We built Wordcraft, a text editor in which users collaborate with a generative language model to write a story. We evaluated Wordcraft with a user study in which participants wrote short stories with and without the tool. Our results show that large language models enable novel co-writing experiences. For example, the language model is able to engage in open-ended conversation about the story, respond to writers' custom requests expressed in natural language (such as "rewrite this text to be more Dickensian"), and generate suggestions that serve to unblock writers in the creative process. Based on these results, we discuss design implications for future human-AI co-writing systems.

## TaleBrush: Sketching Stories with Generative Pretrained Language Models

While advanced text generation algorithms (e.g., GPT-3) have enabled writers to co-create stories with an AI, guiding the narrative remains a challenge. Existing systems often leverage simple turn-taking between the writer and the AI in story development. However, writers remain unsupported in intuitively understanding the AI's actions or steering the iterative generation. We introduce TaleBrush, a generative story ideation tool that uses line sketching interactions with a GPT-based language model for control and sensemaking of a protagonist's fortune in co-created stories. Our empirical evaluation found our pipeline reliably controls story generation while maintaining the novelty of generated sentences. In a user study with 14 participants with diverse writing experiences, we found participants successfully leveraged sketching to iteratively explore and write stories according to their intentions about the character's fortune while taking inspiration from generated stories. We conclude with a reflection on how sketching interactions can facilitate the iterative human-AI co-creation process.

## Topic Description

In recent years, neural language models have grown greatly in effectiveness. They are already being applied to real-world tasks. Large language models have shown unprecedented capabilities in natural language processing tasks. However, due to their highly context-dependent nature, large language models are difficult to grasp and make use of. State-of-the-art large language models, such as GPT-3, have achieved new levels of performance on benchmarks for language understanding and generation. These models have shown that they can perform very well on many downstream tasks without any explicit training. Story writing is one such task which continues to be a challenging problem for machines and even for humans. Existing works include neural narrative generation systems that interact with humans in different ways to generate stories. While large language models, with their improved text generation capabilities, have enabled writers to co-create stories with an AI, guiding the narrative remains a challenge. This topic explores different methodologies to facilitate iterative human-AI co-writing process in an interactive and collaborative manner.

# Automated Story Generation with Pre-Set Control Mechanisms

## Topic Description

In recent years, neural language models have grown greatly in effectiveness. They are already being applied to real-world tasks. Large language models have shown unprecedented capabilities in natural language processing tasks. However, due to their highly context-dependent nature, large language models are difficult to grasp and make use of. State-of-the-art large language models, such as GPT-3, have achieved new levels of performance on benchmarks for language understanding and generation. These models have shown that they can perform very well on many downstream tasks without any explicit training. Story generation is one such task which continues to be a challenging problem for machines. Existing works include neural narrative generation systems that automatically generate stories given some input prompts, keywords, or outlines. While large language models, with their improved text generation capabilities, can generate believable stories; guiding the narrative remains a challenge. This topic explores different methodologies to facilitate automatic generation of a story narrative in a controllable manner.

# Anonymization of Speaker Voice for Privacy Protection

## X-vector anonymization using autoencoders and adversarial training for preserving speech privacy

The rapid increase in web services and mobile apps, which collect personal data from users, has also increased the risk that their privacy may be severely compromised. In particular, the increasing variety of spoken language interfaces and voice assistants empowered by the vertiginous breakthroughs in deep learning have prompted important concerns in the European Union in terms of preserving the privacy of speech data. For instance, an attacker can record speech from users and impersonate them to obtain access to systems that require voice identification. By extracting speaker, linguistic (e.g., dialect), and paralinguistic features (e.g., age) from a speech signal, the speaker profiles can also be hacked from users through existing technology. To mitigate these weaknesses, in this study, we present a speech anonymization method based on autoencoders and adversarial training. Given an utterance, we first extract an x-vector, which is a powerful utterance-level embedding used in state-of-the-art speaker recognition. This original x-vector is transformed by an autoencoder producing a new x-vector, where speaker, gender, and accent information are suppressed through adversarial training. The anonymized speech is finally generated through a neural speech synthesizer driven by the anonymized x-vector, fundamental frequency, and phoneme information extracted from the input speech. For the evaluation, we followed the VoicePrivacy Challenge framework, where anonymization or privacy is measured using automatic speaker verification and the preservation of the intelligibility is evaluated through automatic speech recognition. Our experimental results show that the proposed method achieves higher privacy than the VoicePrivacy baseline (i.e., a higher speaker verification error) while preserving a similar intelligibility for the spoken content (i.e., a similar word error rate).

## Adjustable deterministic pseudonymization of speech

While public speech resources become increasingly available, there is a growing interest to preserve the privacy of the speakers, through methods that anonymize the speaker information from speech while preserving the spoken linguistic content. In this paper, a method for pseudonymization (reversible anonymization) of speech is presented, that allows to obfuscate the speaker identity in untranscribed running speech. The approach manipulates the spectro-temporal structure of the speech to simulate a different length and structure of the vocal tract by modifying the formant locations, as well as by altering the pitch and speaking rate. The method is deterministic and partially reversible, and the changes are adjustable on a continuous scale. The method has been evaluated in terms of (i) ABX listening experiments, and (ii) automatic speaker verification and speech recognition. ABX experimental results indicate that the speaker identifiability among forced choice pairs reduced from over 90% to less than 70% through pseudonymization, and that de-pseudonymization was partially effective. An evaluation on the VoicePrivacy 2020 challenge data showed that the proposed approach performs better than the signal processing based baseline method that uses McAdams coefficient and performs slightly worse than the neural source filtering based baseline method. Further analysis showed that the proposed approach: (i) is comparable to the neural source filtering baseline based method in terms of phone posterior feature based objective intelligibility measure, (ii) preserves formant tracks better than the McAdams based method, and (iii) preserves paralinguistic aspects such as dysarthria in several speakers.

## Self-Supervised Speech Representations Preserve Speech Characteristics while Anonymizing Voices

Collecting speech data is an important step in training speech recognition systems and other speech-based machine learning models. However, the issue of privacy protection is an increasing concern that must be addressed. The current study investigates the use of voice conversion as a method for anonymizing voices. In particular, we train several voice conversion models using self-supervised speech representations including Wav2Vec2.0, Hubert and UniSpeech. Converted voices retain a low word error rate within 1% of the original voice. Equal error rate increases from 1.52% to 46.24% on the LibriSpeech test set and from 3.75% to 45.84% on speakers from the VCTK corpus which signifies degraded performance on speaker verification. Lastly, we conduct experiments on dysarthric speech data to show that speech features relevant to articulation, prosody, phonation and phonology can be extracted from anonymized voices for discriminating between healthy and pathological speech.

## Privacy and Utility of X-Vector Based Speaker Anonymization

We study the scenario where individuals (speakers) contribute to the publication of an anonymized speech corpus. Data users leverage this public corpus for downstream tasks, e.g., training an automatic speech recognition (ASR) system, while attackers may attempt to de-anonymize it using auxiliary knowledge. Motivated by this scenario, speaker anonymization aims to conceal speaker identity while preserving the quality and usefulness of speech data. In this article, we study x-vector based speaker anonymization, the leading approach in the VoicePrivacy Challenge, which converts the speaker's voice into that of a random pseudo-speaker. We show that the strength of anonymization varies significantly depending on how the pseudo-speaker is chosen. We explore four design choices for this step: the distance metric between speakers, the region of speaker space where the pseudo-speaker is picked, its gender, and whether to assign it to one or all utterances of the original speaker. We assess the quality of anonymization from the perspective of the three actors involved in our threat model, namely the speaker, the user and the attacker. To measure privacy and utility, we use respectively the linkability score achieved by the attackers and the decoding word error rate achieved by an ASR model trained on the anonymized data. Experiments on LibriSpeech show that the best combination of design choices yields state-of-the-art performance in terms of both privacy and utility. Experiments on Mozilla Common Voice further show that it guarantees the same anonymization level against re-identification attacks among 50 speakers as original speech among 20,000 speakers.

## Evaluation of Speaker Anonymization on Emotional Speech

Speech data carries a range of personal information, such as the speaker's identity and emotional state. These attributes can be used for malicious purposes. With the development of virtual assistants, a new generation of privacy threats has emerged. Current studies have addressed the topic of preserving speech privacy. One of them, the VoicePrivacy initiative aims to promote the development of privacy preservation tools for speech technology. The task selected for the VoicePrivacy 2020 Challenge (VPC) is about speaker anonymization. The goal is to hide the source speaker's identity while preserving the linguistic information. The baseline of the VPC makes use of a voice conversion. This paper studies the impact of the speaker anonymization baseline system of the VPC on emotional information present in speech utterances. Evaluation is performed following the VPC rules regarding the attackers' knowledge about the anonymization system. Our results show that the VPC baseline system does not suppress speakers' emotions against informed attackers. When comparing anonymized speech to original speech, the emotion recognition performance is degraded by 15% relative to IEMOCAP data, similar to the degradation observed for automatic speech recognition used to evaluate the preservation of the linguistic information.

## Topic Description

Collection of speech data and creation of datasets is a vital step in training speech recognition systems and other speech-based machine learning models. However, the issue of privacy protection is an increasing concern that must be addressed. Speech data carries a range of personal information, such as the speaker's identity and emotional state. These attributes can be used for malicious purposes. The rapid increase in web services and mobile apps, which collect personal data from users, has also increased the risk that their privacy may be severely compromised. In particular, the increasing variety of spoken language interfaces and voice assistants empowered by the recent breakthroughs in deep learning have prompted important concerns in the European Union in terms of preserving the privacy of speech data. For instance, an attacker can record speech from users and impersonate them to obtain access to systems that require voice identification. By extracting speaker, linguistic (e.g., dialect), and paralinguistic features (e.g., age) from a speech signal, the speaker profiles can also be hacked from users through existing technology. Current studies have addressed the topic of preserving speech privacy. One of them, the VoicePrivacy initiative, aims to promote the development of privacy preservation tools for speech technology. The task for the VoicePrivacy Challenge (VPC) is about speaker anonymization. The goal is to hide the source speaker's identity in untranscribed running speech while preserving the linguistic information. This is the problem that this topic tries to solve.

# Anonymization of Personally Identifiable Information (PII) in Image by Replacing with Generated Fakes

## Realistic Full-Body Anonymization with Surface-Guided GANs

Recent work on image anonymization has shown that generative adversarial networks (GANs) can generate near-photorealistic faces to anonymize individuals. However, scaling these networks to the entire human body has remained a challenging and yet unsolved task. We propose a new anonymization method that generates close-to-photorealistic humans for in-the-wild images. A key part of our design is to guide adversarial nets by dense pixel-to-surface correspondences between an image and a canonical 3D surface. We introduce Variational Surface-Adaptive Modulation (V-SAM) that embeds surface information throughout the generator. Combining this with our novel discriminator surface supervision loss, the generator can synthesize high quality humans with diverse appearance in complex and varying scenes. We show that surface guidance significantly improves image quality and diversity of samples, yielding a highly practical generator. Finally, we demonstrate that surface-guided anonymization preserves the usability of data for future computer vision development.

## IdentityDP: Differential private identification protection for face images

Because of the explosive growth of face photos as well as their widespread dissemination and easy accessibility in social media, the security and privacy of personal identity information has become an unprecedented challenge. Meanwhile, the convenience brought by advanced identity-agnostic computer vision technologies is attractive. Therefore, it is important to use face images while taking careful consideration in protecting people's identities. Given a face image, face de-identification, also known as face anonymization, refers to generating another image with similar appearance and the same background, while the real identity is hidden. Although extensive efforts have been made, existing face de-identification techniques are either insufficient in photo-reality or incapable of well-balancing privacy and utility. In this paper, we focus on tackling these challenges to improve face de-identification. We propose IdentityDP, a face anonymization framework that combines a data-driven deep neural network with a differential privacy (DP) mechanism. This framework encompasses three stages: facial representations disentanglement, ∊-IdentityDP perturbation and image reconstruction. Our model can effectively obfuscate the identity-related information of faces, preserve significant visual similarity, and generate high-quality images that can be used for identity-agnostic computer vision tasks, such as detection, tracking, etc. Different from the previous methods, we can adjust the balance of privacy and utility through the privacy budget according to practical demands and provide a diversity of results without pre-annotations. Extensive experiments demonstrate the effectiveness and generalization ability of our proposed anonymization framework.

## Learnable Privacy-Preserving Anonymization for Pedestrian Images

This paper studies a novel privacy-preserving anonymization problem for pedestrian images, which preserves personal identity information (PII) for authorized models and prevents PII from being recognized by third parties. Conventional anonymization methods unavoidably cause semantic information loss, leading to limited data utility. Besides, existing learned anonymization techniques, while retaining various identity-irrelevant utilities, will change the pedestrian identity, and thus are unsuitable for training robust re-identification models. To explore the privacy-utility trade-off for pedestrian images, we propose a joint learning reversible anonymization framework, which can reversibly generate full-body anonymous images with little performance drop on person re-identification tasks. The core idea is that we adopt desensitized images generated by conventional methods as the initial privacy-preserving supervision and jointly train an anonymization encoder with a recovery decoder and an identity-invariant model. We further propose a progressive training strategy to improve the performance, which iteratively upgrades the initial anonymization supervision. Experiments further demonstrate the effectiveness of our anonymized pedestrian images for privacy protection, which boosts the re-identification performance while preserving privacy. Code is available at \url{https://github.com/whuzjw/privacy-reid}.

## Pseudonymization of Text and Image Data to Provide Confidentiality

It is important to maintain secrecy of one’s valuable data when it is loaded in the Internet. Unintended parties can view the data, resulting in breach of confidentiality. Another serious concern is the possibility of misuse of data to the advantage of an unintended party. To overcome these problems, this paper presents an algorithm developed and implemented to make the data that is uploaded to the servers, pseudo anonymous (pseudonymous). Typically, user-sensitive information exists as either texts or images. Hence, such data is used as input to the developed algorithm. From the textual data, the algorithm can detect keywords called tags, such as names, places, financials, and replace the original tag with newly created false tags. Detection of the tags is done by the existing natural language processing (NLP) algorithms. For image input, the algorithm detects face and replaces it with a falsified one. This is done by using generative adversarial networks (GANs). Further, the original or true image is hidden in the falsified image using SteganoGAN. This image will be sent over the network. At the receiver, both anonymized text and image data are de-anonymized as per the developed algorithm.

## Safe Fakes: Evaluating Face Anonymizers for Face Detectors

Since the introduction of the GDPR and CCPA privacy legislation, both public and private facial image datasets are increasingly scrutinized. Several datasets have been taken offline completely and some have been anonymized. However, it is unclear how anonymization impacts face detection performance. To our knowledge, this paper presents the first empirical study on the effect of image anonymization on supervised training of face detectors. We compare conventional face anonymiz-ers with three state-of-the-art Generative Adversarial Network-based (GAN) methods, by training an off-the-shelf face detector on anonymized data. Our experiments investigate the suitability of anonymization methods for maintaining face detector performance, the effect of detectors overtraining on anonymization artefacts, dataset size for training an anonymizer, and the effect of training time of anonymization GANs. A final experiment investigates the correlation between common GAN evaluation metrics and the performance of a trained face detector. Although all tested anonymization methods lower the performance of trained face detectors, faces anonymized using GANs cause far smaller performance degradation than conventional methods. As the most important finding, the best-performing GAN, DeepPrivacy, removes identifiable faces for a face detector trained on anonymized data, resulting in a modest decrease from 91.0 to 88.3 mAP. In the last few years, there have been rapid improvements in realism of GAN-generated faces. We expect that further progression in GAN research will allow the use of Deep Fake technology for privacy-preserving Safe Fakes, without any performance degradation for training face detectors.

## Identity-Preserving Face Anonymization via Adaptively Facial Attributes Obfuscation

With the popularity of using computer vision technology in monitoring system, there is an increasing societal concern on intruding people's privacy as the captured images/videos may contain identity-related information e.g. people's face. Existing methods on protecting such privacy focus on removing the identity-related information from faces. However, this would weaken the utility of current monitoring system. In this paper, we develop a face anonymization framework that could obfuscate visual appearance while preserving the identity discriminability. The framework is composed of two parts: an identity-aware region discovery module and an identity-aware face confusion module. The former adaptively locates the identity-independent attributes on human faces, and the latter generates the privacy-preserving faces using original faces and discovered facial attributes. To optimize the face generator, we employ a multi-task based loss function, which consists of discriminator loss, identify preserving loss, and reconstruction loss functions. Our model can achieve a balance between recognition utility and appearance anonymizing by modifying different numbers of facial attributes according to pratical demands, and provide a variety of results. Extensive experiments conducted on two public benchmarks Celeb-A and VGG-Face2 demonstrate the effectiveness of our model under distinct face recognition scenarios.

## Effective De-identification Generative Adversarial Network for Face Anonymization

The growing application of face images and modern AI technology has raised another important concern in privacy protection. In many real scenarios like scientific research, social sharing and commercial application, lots of images are released without privacy processing to protect people's identity. In this paper, we develop a novel effective de-identification generative adversarial network (DeIdGAN) for face anonymization by seamlessly replacing a given face image with a different synthesized yet realistic one. Our approach consists of two steps. First, we anonymize the input face to obfuscate its original identity. Then, we use our designed de-identification generator to synthesize an anonymized face. During the training process, we leverage a pair of identity-adversarial discriminators to explicitly constrain identity protection by pushing the synthesized face away from the predefined sensitive faces to resist re-identification and identity invasion. Finally, we validate the effectiveness of our approach on public datasets. Compared with existing methods, our approach can not only achieve better identity protection rates but also preserve superior image quality and data reusability, which suggests the state-of-the-art performance.

## Privacy preservation for image data: A GAN-based method

The importance of protecting personal information, like, a person's address or health history, is well known and commonly discussed. However, images also contain sensitive information that can compromise a person's privacy or be used for nefarious purposes. To date, most methods for preserving privacy with images have relied on obfuscation techniques, such as pixelation, blurring, or masking parts of the image. However, new face-recognition technologies driven by deep learning are showing cracks in the old techniques. Moreover, faceless recognition is presenting a whole new set of challenges for image privacy. The core of these issues it is how to ensure privacy while still being able to see and use the image. Our solution is a model based on a generative adversarial network that protects identity information while preserving face features of the original image as much as possible. The premise is to generate a fake image of a face that shares all the same attributes as the original image, for example, a brown-eyed child smiling. With this strategy, the image remains useful, but no person or algorithm could determine the identity of the pictured individual. The framework consists of three parts: a detection module, an image creation module, and an image transformation module. The detection module extracts the attribute labels. The image creation module generates images of faces, and the image transformation module transforms the fake features to match the attributes in the original image. A comprehensive set of experiments shows the effectiveness of the proposed framework.

## Unnoticeable synthetic face replacement for image privacy protection

A rapidly growing amount of personal sensitive information is being released to the public due to unprotected sharing of face images and videos on social networks. Although some pioneering face de-identification techniques, such as face blurring, have been proposed, there is still a long way towards providing full protection of one's facial privacy. In this paper, we propose a novel end-to-end privacy protection approach to seamlessly replace a face in an image with a synthesized face that looks as natural as normal photos yet pertaining very different look from the original face. The synthesized face images will prevent potential attackers from de-identifying the users. Specifically, our approach relies on generative adversarial network and considers both the foreground and background constrains with respect to the input face image to achieve the following two goals: Firstly, to make synthesized images perceptually unaltered, we design a new generative model to effectively fuse a synthesized face with the original background. Secondly, to ensure a synthesized face to be much different from the original face, we define multiple losses to distinguish the synthesized face from the original face. The experimental results on public datasets have validated the effectiveness of our approach compared with the-state-of-the-art.

## Topic Description

The importance of protecting personal information, like, a person's address or health history, is well known and commonly discussed. However, images also contain sensitive information that can compromise a person's privacy or be used for nefarious purposes. To date, most methods for preserving privacy with images have relied on obfuscation techniques, such as pixelation, blurring, or masking parts of the image. However, new face-recognition technologies driven by deep learning are showing cracks in the old techniques. Moreover, faceless recognition is presenting a whole new set of challenges for image privacy. Since the introduction of the GDPR and CCPA privacy legislation, both public and private facial image datasets are increasingly scrutinized. Several datasets have been taken offline completely and some have been anonymized. Conventional anonymization methods unavoidably cause semantic information loss, leading to limited data utility. The core of these issues is how to ensure privacy while still being able to see and use the image for downstream tasks. This the problem that this topic tries to solve.