**Topic Submission Form**

This form should be submitted by the mentioned deadline.

Name:  Paresh Pradhan

Student Number:   PN1038004

Course:  36269 MS - Artificial Intelligence & Machine Learning\_\_\_\_

**Fill your topic/s below**

Project Title/Area 1:   AI-Assisted Collaborative Story Writing with Human-in-the-Loop \_

Dataset:

* WritingPrompts
  + <https://github.com/facebookresearch/fairseq/blob/main/examples/stories/README.md>
* ROCStories
  + <https://cs.rochester.edu/nlp/rocstories/>
* CoAuthor
  + <https://coauthor.stanford.edu/>
* HANNA
  + <https://github.com/dig-team/hanna-benchmark-asg>
* TVRecap
  + <https://github.com/mingdachen/TVRecap>
* MoviePlotEvents
  + <https://paperswithcode.com/dataset/movieplotevents>
  + <http://www.cs.cmu.edu/~ark/personas/>
* Scifi TV Shows
  + <https://huggingface.co/datasets/lara-martin/Scifi_TV_Shows>
* WikiPlots
  + <https://github.com/markriedl/WikiPlots>
* OpenMEVA
  + <https://github.com/thu-coai/OpenMEVA#i-datasets>
* FanFiction Archive
  + <https://archive.org/details/fanfictiondotnet_repack>
* Deep Dungeons and Dragons (DDD)
  + <https://groups.inf.ed.ac.uk/cup/ddd/>
* Lit Bank
  + <https://github.com/dbamman/litbank>
* STORIUM
  + <https://storium.cs.umass.edu/>
  + <https://github.com/dojoteef/storium-gpt2>
* ESTER
  + <https://github.com/PlusLabNLP/ESTER>
  + <https://eventqa.github.io/>
* Cornell Movie Dialog
  + <https://convokit.cornell.edu/>
  + <https://huggingface.co/datasets/cornell_movie_dialog>
* ScriptWriter
  + <https://github.com/DaoD/ScriptWriter>
* NarrativeQA
  + <https://github.com/deepmind/narrativeqa>
  + <https://huggingface.co/datasets/narrativeqa>
* MCTest
  + <https://mattr1.github.io/mctest/data.html>
  + <https://huggingface.co/datasets/sagnikrayc/mctest>
* InSentive
  + <https://github.com/roemmele/InSentive>
* BookCorpus
  + <https://huggingface.co/datasets/bookcorpus>

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 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.

Project Title/Area 2:   Automated Story Generation with Guided Control Mechanisms \_

Dataset:

* WritingPrompts
  + <https://github.com/facebookresearch/fairseq/blob/main/examples/stories/README.md>
* ROCStories
  + <https://cs.rochester.edu/nlp/rocstories/>
* HANNA
  + <https://github.com/dig-team/hanna-benchmark-asg>
* TVRecap
  + <https://github.com/mingdachen/TVRecap>
* MoviePlotEvents
  + <https://paperswithcode.com/dataset/movieplotevents>
  + <http://www.cs.cmu.edu/~ark/personas/>
* Scifi TV Shows
  + <https://huggingface.co/datasets/lara-martin/Scifi_TV_Shows>
* WikiPlots
  + <https://github.com/markriedl/WikiPlots>
* OpenMEVA
  + <https://github.com/thu-coai/OpenMEVA#i-datasets>
* FanFiction Archive
  + <https://archive.org/details/fanfictiondotnet_repack>
* Deep Dungeons and Dragons (DDD)
  + <https://groups.inf.ed.ac.uk/cup/ddd/>
* Lit Bank
  + <https://github.com/dbamman/litbank>
* STORIUM
  + <https://storium.cs.umass.edu/>
  + <https://github.com/dojoteef/storium-gpt2>
* ESTER
  + <https://github.com/PlusLabNLP/ESTER>
  + <https://eventqa.github.io/>
* Cornell Movie Dialog
  + <https://convokit.cornell.edu/>
  + <https://huggingface.co/datasets/cornell_movie_dialog>
* ScriptWriter
  + <https://github.com/DaoD/ScriptWriter>
* NarrativeQA
  + <https://github.com/deepmind/narrativeqa>
  + <https://huggingface.co/datasets/narrativeqa>
* MCTest
  + <https://mattr1.github.io/mctest/data.html>
  + <https://huggingface.co/datasets/sagnikrayc/mctest>
* InSentive
  + <https://github.com/roemmele/InSentive>
* BookCorpus
  + <https://huggingface.co/datasets/bookcorpus>

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 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.

Project Title/Area 3:  Anonymization of Personally Identifiable Information (PII) in Image by Replacing with Generated Fakes

Dataset:

* ImageNet - <https://huggingface.co/datasets/imagenet-1k>
* CelebA - <https://huggingface.co/datasets/huggan/CelebA-faces>
* CelebA-HQ - <https://huggingface.co/datasets/huggan/CelebA-HQ>
* COCO-CSE - <https://github.com/facebookresearch/detectron2/blob/main/projects/DensePose/doc/DENSEPOSE_DATASETS.md#continuous-surface-embeddings-annotations-1>

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 future tasks. This the problem that this topic tries to solve.

**Fill in this section if a member of staff has agreed to be your supervisor:**

Member of Staff:                                                                                   \_\_\_\_

If you have found a supervisor then you and the member of staff who agreed to supervise your project should sign below.

\_Paresh Pradhan\_                                                                        \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student Signature                                                                         Supervisor Signature

\_20-Sep-2022\_                                                                            \_\_\_\_\_\_\_\_\_\_\_\_

Date                                                                                               Date