Figure 1. We aim to determine whether a piece of text was generated by a particular LLM p, such as GPT-3. To classify a candidate passage x, DetectGPT first generates minor **perturbations** of thepassage $\tilde{x}_i$ using a generic pre-trained model such as T5. Then DetectGPT **compares** the log probability under p of the original sample x with each perturbed sample $\tilde{x}_i$. If the average log ratio is high, the sample is likely from the source model.

2022), the articulate nature of such generated text may stillmake LLMs attractive for replacing human labor in some contexts, notably student essay writing and journalism. At least one major news source has released AI-written content with limited human review, leading to substantial factual er-rors in some articles (Christian, 2023). Such applications ofLLMs are problematic for a variety of reasons, making fairstudent assessment difficult, impairing student learning, and proliferating convincing-but-inaccurate news articles. Un- fortunately, humans perform only slightly better than chance when classifying machine-generated vs human-written text(Gehrmann et al., 2019), leading researchers to consider automated detection methods that may identify signals dif-ficult for humans to recognize. Such methods might give teachers and news-readers more confidence in the human origin of the text that they consume.

As in prior work (Jawahar et al., 2020), we study the machine-generated text detection problem as a binary clas-sification problem. Specifically, we aim to classify whether

Figure 1. We aim to determine whether a piece of text was generated by a particular LLM p, such as GPT-3. To classify a candidate passage x, DetectGPT first generates minor **perturbations** of thepassage $\tilde{x}_i$ using a generic pre-trained model such as T5. Then DetectGPT **compares** the log probability under p of the original sample x with each perturbed sample $\tilde{x}_i$. If the average log ratio is high, the sample is likely from the source model.

2022), the articulate nature of such generated text may stillmake LLMs attractive for replacing human labor in some contexts, notably student essay writing and journalism. At least one major news source has released AI-written content with limited human review, leading to substantial factual er-rors in some articles (Christian, 2023). Such applications ofLLMs are problematic for a variety of reasons, making fairstudent assessment difficult, impairing student learning, and proliferating convincing-but-inaccurate news articles. Un- fortunately, humans perform only slightly better than chance when classifying machine-generated vs human-written text(Gehrmann et al., 2019), leading researchers to

consider automated detection methods that may identify signals dif-ficult for humans to recognize. Such methods might give teachers and news-readers more confidence in the human origin of the text that they consume.

As in prior work (Jawahar et al., 2020), we study the machine-generated text detection problem as a binary clas-sification problem. Specifically, we aim to classify whether

Figure 1. We aim to determine whether a piece of text was generated by a particular LLM p, such as GPT-3. To classify a candidate passage x, DetectGPT first generates minor **perturbations** of thepassage $\tilde{x}_i$ using a generic pre-trained model such as T5. Then DetectGPT **compares** the log probability under p of the original sample x with each perturbed sample $\tilde{x}_i$. If the average log ratio is high, the sample is likely from the source model.

2022), the articulate nature of such generated text may stillmake LLMs attractive for replacing human labor in some contexts, notably student essay writing and journalism. At least one major news source has released AI-written content with limited human review, leading to substantial factual er-rors in some articles (Christian, 2023). Such applications ofLLMs are problematic for a variety of reasons, making fairstudent assessment difficult, impairing student learning, and proliferating convincing-but-inaccurate news articles. Un- fortunately, humans perform only slightly better than chance when classifying machine-generated vs human-written text(Gehrmann et al., 2019), leading researchers to consider automated detection methods that may identify signals dif-ficult for humans to recognize. Such methods might give teachers and news-readers more confidence in the human origin of the text that they consume.

As in prior work (Jawahar et al., 2020), we study the machine-generated text detection problem as a binary clas-sification problem. Specifically, we aim to classify whether