Generational Debt: The Degeneration of Self-Consuming Generative Language Models

Madison Thantu
mgt2143@columbia.edu

Abstract

The data that is used to train large language models (LLMs) influences the decisions that it makes. Furthermore, the rate at which such models can produce content dramatically outweighs that which a human being can. The present research examines the effects of training LLMs on synthetic data, where that synthetic data was generated by a previous iteration of that model, and evaluates on the task of abstractive text summarization, although the synthetic data pipeline is relevant in nearly all of the applications for which these models are used. This research is intersted in both the quantitative and qualitative dimensions along which generated text evolves. Specifically, I consider the text characteristics of toxicity, formality, and emotion intensity, using three different abstractive summarization datasets, the set of which were selected with the intention of representing different text characteristics to differing degrees. Results support existing evidence that recursive training can reduce output quality. The GitHub repo can be found at -> https://github.com/madisonthantu/recursive_LLMs/tree/main

1 Introduction

State of the art large language models (LLMs) are trained on enormous amounts of data and billions of parameters, and the scaling laws for neural language models (Kaplan et al., 2020) indicates that this size plays a crucial role in model performance. Given the vast amounts of training data required, these datasets are typically procured via web scraping, such as the Common Crawl dataset[1], which is comprised of nearly 20TB of text data and was used to train GPT-3 (Brown et al., 2020). The deployment of LLMs is already changing the information landscape—the European Union Agency for Law Enforcement Cooperation estimates that, by 2026, up to 90% of online content may be synthetically generated[2]. Given the necessity of web scraping in LLM dataset curation, AI-generated content will not only saturate the Internet, but also the data that is used to train those very models. Although LLMs demonstrate striking capabilities, they do not replicate human use of language and are associated with numerous limitations, such as hallucinations, bias and toxicity, and lack of transparency.

The amount of AI-generated content on the Internet will continue its dramatic increase. And the amount of data required to train state-of-the-art generative models necessities the scraping of said Internet. Logically, if both of these trends continue, it follows that AI-generated data will, intentionally or unintentionally, percolate into the training process. However, there is limited research into the effects of this compounded algorithmic confounding.

Consequently, understanding the effects of training these models on data generated by their own species is crucial—and is the motivation behind this present research project. This research evaluates the qualitative and quantitative effects that generational LLM self-consumption has on text output for three different base models, where each generation is fine-tuned on synthetic data generated by the previous generation. The qualitative metrics of interest include emotion intensity, bias and toxicity, and formality. Existing research on this topic is scarce and even more so in the domain of generative *language* models. Given the generally negative flavor with which AI-generated content is discussed in the mainstream media, I am interested in seeing whether this hype is real and if so then to what extent. This is why I am particularly interested in the qualitative evolution of generated text.

[1] http://commoncrawl.org/

[2] https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes

2 Related Work

2.1 Summarization

Zhang et al. conducted some of the earliest work that leveraged pre-trained LMs in the abstractive summarization task. Input sequence encoding were obtained via BERT (Devlin et al., 2019), and underwent a two-stage decoding process: a Transformer generated an initial output sequence, which then underwent masked language modeling to predict a refined token for each masked position. There is a growing body of research on domain-specific text summarization, including the biomedical (Goff and Loehfelm, 2018; Moradi and Ghadiri, 2019), e-commerce (Hanni et al., 2016), and multi-modal domains (Emad et al., 2021). While there is a growing body of research focused on text summarization, controllability (Goyal et al., 2022) and the evaluation of quality and factuality of generated summaries remains an active area of interest (Aharoni et al., 2023; Tang et al., 2022).

2.2 Emotion Intensity Detection

Sentiment analysis aims to identify the general sentiment of a given text, whereas emotion analysis intends to detect specific emotions that are present. This domain has diverse applications, ranging from business (Kang et al., 2020) to public health (Sosea et al., 2022; Nijhawan et al., 2022).

Analyzing the intensity of emotions is both more difficult and less explored. One approach to emotion intensity analysis is lexicon-based, which uses corpus of emotion-associated words paired with intensity annotations (Mohammad and Bravo-Marquez, 2017) Gupta and Yang developed an affective lexicon Emotion Intensity for the four emotions of fear, anger, sadness, and joy, where each lexicon item is coded with an intensity score in the range [-3,3]. They also presented CrystalFeel, a model for predicting emotion intensity in Tweets, using a combination of affective lexicon-based and non-affective lexicon-based features.

2.3 Toxicity and Bias

Detection of toxicity and bias is a natural application of NLP tools — the 4.9 billion people that use social media generate content that requires moderation [3]. However, w.r.t. generative LMs, there is an even greater need to prevent such models from producing toxic or bias content. Wen et al. use an adversarial approach to identify the risks of implicit toxicity, which is conveyed via linguistic features rather than overtly toxic words. They observed impressive zero-shot performance of LLMs in generating implicitly toxic test, and further tested these bounds using reinforcement learning to optimize for an implicit toxicity preference in a GPT-3.5-turbo base model. Ousidhoum et al. use a probing technique to study toxic content against social groups in pre-trained LMs (PTLMs), considering French, English, and Arabic. Their findings indicate a propensity for bias against minorities, with the highest score of 46.36% being associated with BERT and the social group of refugees. In addition to LLM-generated toxicity and bias, research has indicated that existing toxicity detection models are insufficient when analyzing LM-generated text, such that the nature of flagged content is itself systematically biased (Baldini et al., 2022; Rosenblatt et al., 2022).

2.4 Synthetic Data

In NLP, and ML more broadly, there is a direct relationship between the volume of training data and a model's capability. As such, despite the existing limitations of LLMs, synthetic data is increasingly incorporated into the training process, with numerous studies showing improved results in the text summarization task (Karn et al., 2021; Siledar et al., 2023; Magooda and Litman, 2021). In particular, for resource scarce languages and applications, the use of synthetic data can expand model capabilities and increase fairness by way of increased access. For example, Dutta Chowdhury et al. use synthetic data to train a multi-modal machine translation model for the low-resource language pair of Hindi and English.

2.5 Model Degeneration

Existing work on the topic of self-consumption in the synthetic data pipeline is extremely limited. Shumailov et al.'s work is most similar to the present research. Evaluating the effects of training on generated data using Variational Autoencoders, Gaussian Mixture Models, and the the OPT-125m causal language model, they find that recursive training leads to irreversible defects that distort the original token distributions and yield increasingly degraded coherence and fluency. They term this deterioration model collapse and find that the models are essentially reinforcing misinformation that it initially constructed, albeit by way of statistical and

[3] https://www.forbes.com/advisor/business/social-media-statistics/#source

functional approximation errors. In a similar vein, Alemohammad et al. and Martínez et al. study the degradation that ensues from different recursive training regimes in generative image models. Their results indicate that sufficient amounts of fresh real data in each training iteration is able to mitigate the degenerative effects of data confounding. This finding lends inspiration to the research objectives of the current study.

3 Data

In order to explore qualitative differences in text summarization performance and generation, three different datasets are being used to fine-tune each of the three base models separately. This study is interested in the text properties of emotion intensity, toxicity, and formality, and the datasets were selected with this motivation in mind, given the task of abstractive summarization.

The first dataset is the *SAMSum* corpus (Gliwa et al., 2019), which consists of approximately 16K chat dialogues paired with manually annotated summaries. Second, is the *Reddit TIFU* dataset (Kim et al., 2019). This corpus contains approximately 120K posts that were obtained from the online discussion forum Reddit via a web-crawl from 2013-2018 in the *TIFU* subreddit, which imposes strict rules with respect to posting guidelines such that all posts must be terminated with a *TL;DR* summary. This dataset comes in two forms, *TIFU-short* and *TIFU-long*, where the post's title or the *TL;DR* summary are used as the target summary in the abstractive summarization task, respectively. For this study, *TIFU-long* was used. The third summarization dataset is the *CNN/DailyMail* corpus (Nallapati et al., 2016), which contains approximately 312K samples. Samples were obtained via human-generated abstractive summary bullets collected from the *CNN* and *Daily Mail* online platforms.

Due to the large number of samples in *TIFU* and *CNN/DailyMail*, these datasets were sharded in order to obtain more manageable dataset sizes. *TIFU* and *CNN/DailyMail* were reduced by factors of 2 and 8, respectively.

This study is interested in whether, and if so how, the aforementioned text characteristics vary according to the abstractive summarization dataset on which they are fine-tuned. Table 1 presents summary statistics on these three summarization task datasets, including both the reduced and complete sets for *TIFU* and *CNN/DailyMail*.

A lexicon-based approach is used to evaluate emotion intensity in the summaries generated by the three language models, as well as those supplied in the initial datasets, further discussed in section 4. The NRC Emotion Intensity Lexicon (NRC-EIL), introduced by Mohammad consists of approximately 10,000 words, where each word is associated with one of eight emotions—anger, anticipation, disgust, fear, joy, sadness, surprise, and trust—and a real-valued score of intensity in the range [0,1]. These real-valued scores were obtained using best-worst scaling., such that scores closer to 1 indicate higher associations with an emotion e, and vice versa.

Additionally, the RealToxicityPrompts dataset (Gehman et al., 2020) and Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018) are used to evaluate the evolution of toxicity and bias and text style (w.r.t. formality), respectively. Further details on how these dimensions are evaluated on are discussed in section 4.

4 Methods

The effects of generational self-consumption are evaluated using three different base models: BART (Lewis et al., 2019), T5 (Raffel et al., 2023), and GPT-2 (Radford et al., 2019). The former two are encoder-decoder models, whereas the latter is a decoder-only model.

Figure 1 illustrates the synthetic data pipeline. A model is fine-tuned on a given dataset, where the inputs consist of documents and the outputs consist of target summaries. After this fine-tuning process, the model is then used to generate, or predict, summaries for the entire set of input documents, spanning both the training and validation datasets. The documents from the initial, *real* dataset are then paired with the *synthetic* summaries, constituting the new *synthetic* dataset, which is then used to fine-tune the succeeding generation.

The models used were obtained from Huggingface[4] The BART-base model uses the `BartForConditionalGeneration` architecture and contains 139M parameters. The `facebook/bart-base` checkpoint was used to initialize $\texttt{Gen-0}_{\texttt{BART}}$. The T5-base model uses the `T5ForConditionalGeneration` architecture, initialized with the `t5-base` checkpoint with 223M parameters. Lastly, the GPT-2-base model used the `GPT2LMHeadModel` architecture, the `gpt2`

[4]https://huggingface.co/

	SAMSum	TIFU-long		CNN/DailyMail	
		Complete	Reduced	Complete	Reduced
Total no. samples	16,369	42,140	21,070	311,971	38,997
Training no. samples	11,458	29,498	14,749	218,379	27,297
Validation no. samples	4,911	12,642	6,321	93,592	11,700
Avg. document length	96	396	397	698	699
Avg. summary length	21	21	21	50	50

Table 1: Summary statistics for the three abstractive summarization datasets: SAMSum, TIFU-long, and CNN/DailyMail.

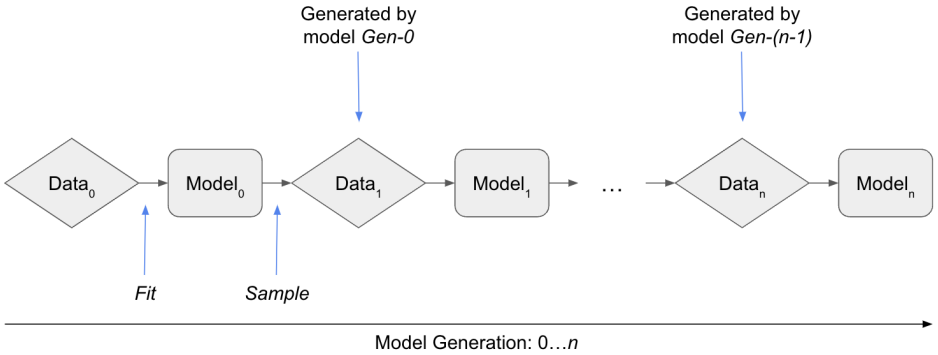


Figure 1: A conceptual diagram of the self-consuming mechanism that occurs during the fine-tuning process, adapted from Shumailov et al.. This is the intended protocol for the first set of experiments described in section 4, where each generation of the model is trained on data that is sampled from the corresponding previous generation only.

checkpoint with 137M parameters for initialization. Additionally, each generation was trained for one epoch.

All three models use cross-entropy loss during fine-tuning (Eq.2) and the `AdamW` optimizer with betas $= (0.9, 0.999)$, epsilon $=$ 1e-08.

$$L_{CE} = -\sum_{i=1}^{n} t_i \log(p_i) \quad (1)$$

For the encoder-decoder models, the inputs for training, validation, and generation consisted of the tokenized documents (model inputs) and the tokenized summaries (labels). For T5, the summarization task prefix `"summarize: "` was added to the beginning of each sample. For the decoder-only model, the inputs during training and validation consisted of the tokenized concatenation of the document and summary pairs, separated by `" TL;DR "`. For generation, the inputs for GPT-2 consisted of the document text and the separator. After fine-tuning and validation of each generator, the model then performed generation on the tokenized document inputs.

Experiments with GPT-2 were only conducted for the *TIFU* dataset and only up through generation 3. This was due to poor quality and incoherence of generated summaries.

5 Experiments

The initial datasets were obtained HuggingFace. The *TIFU* and *CNN/DailyMail* datasets were reduced by factors of 2 and 8, respectively. For data preprocessing, rows with no value for either the document or summary columns were dropped. No additional removals were made for BART or T5; however, for GPT2, the synthetic datasets produced by each successive generation of model contained additional rows with no value, which were then dropped. Each dataset was then split into training (70%) and validation (30%) sets. The test set consisted of each dataset in its entirety.

The ROUGE metrics (ROUGE-1, ROUGE-2,

ROUGE-L, and ROUGE-L*sum*) are used to evaluate the generated summaries (Lin, 2004). Qualitative evaluation was performed on each dataset, including the initial, real dataset, as well as the iteratively generated synthetic datasets, . The real datasets served as a baseline for the synthetic datasets to be evaluated against.

Several different metrics are used to evaluate text quality. To assess abstractiveness, the metric of *coverage* is used, which is computed as the fraction of words present in a summary that are also present in the document. The length metrics of compression ratio and average length for documents and summaries are also recorded, where compression ratio is computed as the ratio between the absolute length (i.e., number of words) of a summary and its input document.

In addition to the afforementioned abstractiveness and length metrics, generated summaries were evaluated on the "stylistic" dimensions of toxicity, formality, and emotion intensity. PerspectiveAPI [5], a popular toxicity detection system is used to evaluate toxicity, which is accessed via their public API.

A RoBERTa-based classifier (Babakov et al., 2023), trained on GYAFC and Online Formality, is used to evaluate style formality. I access this tool via the Hugging Face API. Due to rate limiting parameters, scores for toxicity and formality are computed for a sample size of 1,000 summaries and their average is taken.

Emotion intensity is evaluated via a lexicon-based approach. For each of the eight emotions present in the NRC Emotion Intensity Lexicon (NRC-EIL) corpus, the corresponding tokens that are present in the output summaries are identified. For each summary, the number of lexeme-specific tokens that are present and the sum of those tokens' corresponding emotion intensity scores are computed. For each summary, an emotion intensity ratio is computed as the sum of emotion intensity scores across all emotions weighted by the number of tokens in the sentence:

$$\text{Ratio}_{\text{EI}} = \frac{1}{|S|} \sum_{e \in E, s \in S} s_e, \qquad (2)$$

where S is a sentence, $|S|$ is the number of tokens in the sentence, e is an emotion in the set of eight emotions E, and s is the score of a token relative to the emotion e, such that tokens which are not present in the lexicon for a given emotion have a score of 0.

Lastly, summary token distributions are computed for each dataset.

6 Results

Five generations of the T5- and BART-based models were trained for the *SAMSum* and *TIFU* datasets, and four generations were trained for the *CNN/DailyMail* dataset. *GPT-2* was trained for three generations for the *TIFU* dataset only. Table 2 presents ROUGE metrics for each generation of synthetic summaries, evaluated against the real, "ground truth" summaries. Based on the ROUGE scores, the T5 base model generally performs the better than the BART base model, and the there is a small decrease in ROUGE scores with each successive generation, with two exceptions: $\texttt{Gen-0}_{\texttt{T5}}$ v. $\texttt{Gen-4}_{\texttt{T5}}$ for the *SAMSum* dataset and $\texttt{Gen-0}_{\texttt{BART}}$ v. $\texttt{Gen-1}_{\texttt{BART}}$ for the *TIFU* dataset. The summary length statistics of compression ratio and coverage can be found in table 10 of the Appendix section. The qualitative evaluations of toxicity, formality, and emotion intensity for the baseline and synthetic datasets are presented in table 3. These results indicate that very limited variation for the *SAMSum* datasets along all three dimensions. For the *TIFU* dataset, the T5 and GPT-2 base models exhibit decreases in average toxicity scores and average intensity ratios. The BART-based models also indicate a decrease in average emotion intensity ratio, as well as an increase in average formality. For the *CNN/DailyMail* dataset, both the T5 and BART-based models indicate an increase in formality scores, with T5-based increasing with each successive generation. Violin plots of the distribution of scores for these qualitative text characteristics in the *TIFU* dataset can be found in figure 2, and such plots for the *SAMSum* and *CNN/DailyMail* datasets can be found in tables 3 and 4, respectively, of the Appendix

7 Error Analysis

Text generation samples are included for the BART base model using the *TIFU* dataset for the extreme ends of the spectrum with respect to the qualitative characteristics of minimum and maximum toxicity scores (Tables **??**), minimum and maximum formality scores (Tables **??**, and minimum and maximum emotion intensity ratios (Tables **??**. Looking at the samples for minimum and maximum

[5] https://www.perspectiveapi.com/

		SAMSum				TIFU-long				CNN/DailyMail			
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
T5	Gen_0	9.3993	0.49	7.71	8.80	12.095	0.51	9.859	10.16	41.41	19.90	30.33	38.83
	Gen_1	9.3789	0.49	7.67	8.79	11.81	0.46	9.60	9.90	40.62	19.21	29.62	37.98
	Gen_2	9.3945	0.48	7.67	8.80	11.81	0.44	9.58	9.88	40.07	18.69	29.07	37.37
	Gen_3	9.3925	0.48	7.67	8.80	11.76	0.44	9.49	9.79	39.62	18.30	28.65	36.91
	Gen_4	9.4187	0.48	7.68	8.83	11.79	0.42	9.46	9.76	NA	NA	NA	NA
BART	Gen_0	9.1317	0.47	7.58	8.55	11.56	0.69	9.36	9.65	38.91	17.27	27.18	36.06
	Gen_1	8.9758	0.45	7.46	8.41	11.41	0.70	9.25	9.53	38.35	16.69	26.77	35.35
	Gen_2	8.9562	0.45	7.41	8.38	11.37	0.67	9.24	9.51	37.94	16.41	26.51	34.93
	Gen_3	8.886	0.44	7.36	8.32	11.36	0.66	9.24	9.49	37.68	16.21	26.40	34.66
	Gen_4	8.876	0.42	7.35	8.29	11.32	0.63	9.18	9.45	NA	NA	NA	NA
GPT-2	Gen_0	NA	NA	NA	NA	5.57	0.33	5.15	4.79	NA	NA	NA	NA
	Gen_1	NA	NA	NA	NA	4.93	0.28	4.67	4.20	NA	NA	NA	NA
	Gen_2	NA	NA	NA	NA	5.20	0.30	4.89	4.40	NA	NA	NA	NA

Table 2: ROUGE scores of predicted synthetic summaries evaluated against real (ground truth) summaries.

		SAMSum			TIFU-long			CNN/DailyMail		
		Avg. toxicity score	Avg. formality score	Avg. $RATIO_{EI}$	Avg. toxicity score	Avg. formality score	Avg. $RATIO_{EI}$	Avg. toxicity score	Avg. formality score	Avg. $RATIO_{EI}$
Baseline		0.05	0.84	0.07	0.33	0.36	0.12	0.07	0.69	0.10
T5	Gen_0	0.06	0.85	0.06	0.28	0.36	0.09	0.06	0.73	0.09
	Gen_1	0.06	0.85	0.06	0.27	0.34	0.09	0.06	0.74	0.09
	Gen_2	0.06	0.84	0.06	0.27	0.34	0.09	0.06	0.74	0.09
	Gen_3	0.06	0.83	0.06	0.27	0.35	0.09	0.06	0.75	0.09
	Gen_4	0.06	0.83	0.06	0.27	0.37	0.09	NA	NA	NA
BART	Gen_0	0.05	0.87	0.07	0.36	0.44	0.10	0.07	0.75	0.09
	Gen_1	0.06	0.87	0.07	0.36	0.43	0.09	0.07	0.75	0.09
	Gen_2	0.06	0.86	0.07	0.35	0.43	0.09	0.07	0.76	0.09
	Gen_3	0.06	0.86	0.07	0.32	0.44	0.09	0.07	0.75	0.09
	Gen_4	0.06	0.86	0.07	0.32	0.44	0.09	NA	NA	NA
GPT-2	Gen_0	NA	NA	NA	0.23	0.58	0.02	NA	NA	NA
	Gen_1	NA	NA	NA	0.21	0.56	0.01	NA	NA	NA
	Gen_2	NA	NA	NA	0.21	0.58	0.02	NA	NA	NA

Table 3: Qualitative evaluation of the datasets - both synthetic and real. Metrics include average toxicity score, average formality score, and average emotion intensity ratio ($RATIO_{EI}$) for all combinations of base models, generations, and datasets.

$RATIO_{EI}$, the metric appears to be biased w.r.t. the number of tokens, which makes sense since this was used to weight the computation of this metric. Additionally, this bias and the unintuitive correlation between emotion intensity ratios and the associated summary indicates shortcomings in both the present implementation as well as possibly lexicon-based approaches in general.

8 Conclusions, Limitations, and Future Work

Large language models (LLMs) are increasingly influencing the way that society works, communicates, and learns on a daily basis. Given the far-reaching capabilities and capacities of said models, increasing scrutiny is being given to the transparency, or lack thereof, of how and from what sources these models learn. And, in the digital age, where information is saturated in amount and often subpar in quality, studying how large language models can affect the information landscape is crucial.

This research investigated the effects of training large language models recursively on synthetic data generated by previous iterations of the same model, with a focus on the task of abstractive text summarization. Results indicate that recursive training can lead to a decrease in output quality, aligning with existing evidence in the field.

One limitation of this study includes the potential biases of the tools used for qualitative analysis, such as the PerspectiveAPI for toxicity detection. Although a popular tool for commercial toxicity detection, existing research (Rosenblatt et al., 2022; Sap et al., 2022), has observed biases in PerspectiveAPI's evaluation, such that ratings of toxicity were more aligned with certain demographic groups more than others. While this is an important issue in the field of toxicity and bias detection, in the scope of this present research this tool suffices, as the present experiments focus on the qualitative differences between datasets. Additionally, the lexicon-based approach for measuring emotion intensity presents limitations more generally. Multiple corpora should be used in the evaluation of emotion intensity, if a lexicon-based approach must be used at all.

Future work should is needed to identify alternative methods for qualitative analysis that reduce bias and provide more nuanced insights. Additionally, investigating the impact of recursive training

	ID	Max toxicity score	Summary
Baseline	2881	0.9607	fuck telemarketing, fuck cheap ass owners, fuck!!! totally gonna fucking rage at crossfit tonight. yes i have been doing crossfit and any idiot to comment how dangerous it is grow a fucking sack and try it out unless your too pussy or busy with other athletic tasks that may be more related to functionality. !
Gen-0_{BART}.	9357	0.9607	i got fucked in the ass by my asian friend.
Gen-4_{BART}.	11163	0.9564	i said "fuck you, i pay my taxes motherfucker" to a black dude pushing his shitty cd's onto socially awkward nerds.

Table 4: The summaries and corresponding scores for the BART-base generated text summaries associated with the highest toxicity scores for the *TIFU* dataset.

	ID	Min toxicity score	Summary
Baseline	12922	0.00886	a 165 euro second baggage fee is now turning into a $1000+ ticket back home from paris, france to dfw, tx.
Gen-0_{BART}.	7205	0.00738	took a cab to get work, didn't check the bus schedule, got late for work and now have to leave 2 hours early for a 4 hour shift.
Gen-4_{BART}.	7645	0.00792	didn't check the trip counter before leaving my motorcycle, ended up being two hours late for work.

Table 5: The summaries and corresponding scores for the BART-base generated text summaries associated with the lowest toxicity scores for the *TIFU* dataset.

	ID	Max RATIO_{EI}	Summary
Baseline	2990	1.33	i hate bananas
Gen-0_{BART}	5045	1.48	i'm a terrible liar.
Gen-4_{BART}	472	1.28	i annihilated a bullfrog.

Table 6: The summaries and corresponding scores for the BART-base generated text summaries associated with the highest ratio of emotion intensity for the *TIFU* dataset. The implementation of this method exhibits clear bias towards short summaries.

on other aspects of language generation, such as factual accuracy, coherence, and the presence of hallucinations, is crucial — given that LLMs are not only generating an increasing amount of content, but they are also participating in an increasing number of social interactions with real human beings.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2023. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*.

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *Natural Language Processing and Information Systems*, pages 47–61, Cham. Springer Nature Switzerland.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

	ID	Min RATIO$_{EI}$	Summary
Baseline	21050	0.00165	i was at the urinal, our vp walked in and stood at the urinal next to me, i glanced over out of the corner of my eye and peed all over the floor right in front of his foot. splashes were made and his shoe was wet with urine.
Gen-0$_{\text{BART}}$.	8972	0.00146	i signed up for classes on tuesday and wednesday mornings, didn't check my schedule to figure out when the first day of class was, and now i can't take the required class next semester.
Gen-4$_{\text{BART}}$.	18291	0.00141	tried to decorate the kitchen with helium filled mylar balloons, ended up hitting the transformer at the end of the power lines.

Table 7: The summaries and corresponding scores for the BART-base generated text summaries associated with the lowest ratio of emotion intensity for the \textit{TIFU} dataset.

	ID	Max formality score	Summary
Baseline	11359	0.997	i ruined a very important night for my girlfriend at a cat power concert. something dumb i said will go unredeemable unless i can find a way to fix it. thank you for your sincere answers.
Gen-0$_{\text{BART}}$.	3109	0.995	i accidentally sent my girlfriend a text that said "x acts of affection" while i was asleep listening to steve harvey on "family feud". momma was irate and sent it to my mom.
Gen-4$_{\text{BART}}$.	8784	0.996	i introduced my younger cousin to anime, and now he is showing signs of depression.

Table 8: The summaries and corresponding scores for the BART-base generated text summaries associated with the highest formality scores for the *TIFU* dataset.

bidirectional transformers for language understanding.

Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. 2018. Multimodal neural machine translation for low-resource language pairs using synthetic data. Association for Computational Linguistics (ACL).

Ahmed Emad, Fady Bassel, Mark Refaat, Mohamed Abdelhamed, Nada Shorim, and Ashraf AbdelRaouf. 2021. Automatic video summarization with timestamps using natural language processing text fusion. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0060–0066.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Daniel J Goff and Thomas W Loehfelm. 2018. Automated radiology report summarization using an open-source natural language processing pipeline. *Journal of digital imaging*, 31:185–192.

Tanya Goyal, Nazneen Rajani, Wenhao Liu, and Wojciech Kryściński. 2022. Hydrasum: Disentangling style features in text summarization with multi-decoder models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 464–479.

Raj Kumar Gupta and Yinping Yang. 2018. CrystalFeel at SemEval-2018 task 1: Understanding and detecting emotion intensity using affective lexicons. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 256–263, New Orleans, Louisiana. Association for Computational Linguistics.

Akkamahadevi R Hanni, Mayur M Patil, and Priyadarshini M Patil. 2016. Summarization of customer reviews for a product on a website using natural language processing. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2280–2285.

Yue Kang, Zhao Cai, Chee-Wee Tan, Qian Huang, and Hefu Liu. 2020. Natural language processing (nlp) in management research: A literature review. *Journal of Management Analytics*, 7(2):139–172.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Sanjeev Kumar Karn, Francine Chen, Yan-Ying Chen, Ulli Waltinger, and Hinrich Schütze. 2021. Few-shot learning of an interleaved text summarization model by pretraining with synthetic data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 245–254, Kyiv, Ukraine. Association for Computational Linguistics.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of reddit posts with multi-level memory networks.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

	ID	Min formality score	Summary
Baseline	8137	0.00626	dont fucking trust none unless it shows u pussy over webcam. thats it guys.
Gen-0_{BART}.	3987	0.00544	gave my gf a new years eve kiss and now she thinks i don't want her.
Gen-4_{BART}.	4608	0.00527	i thought a girl was wearing the same hair like where do y'all shop for your weaves smh.

Table 9: The summaries and corresponding scores for the BART-base generated text summaries associated with the lowest formality scores for the *TIFU* dataset.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ahmed Magooda and Diane Litman. 2021. Mitigating data scarceness through data synthesis, augmentation and curriculum for abstractive summarization.

Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. 2023. Combining generative artificial intelligence (ai) and the internet: Heading towards evolution or degradation?

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. *arXiv preprint arXiv:1708.03696*.

Milad Moradi and Nasser Ghadiri. 2019. Text summarization in the biomedical domain. *arXiv preprint arXiv:1908.02285*.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond.

Tanya Nijhawan, Girija Attigeri, and T Ananthakrishna. 2022. Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, 9(1):1–24.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Lucas Rosenblatt, Lorena Piedras, and Julia Wilkins. 2022. Critical perspectives: A benchmark revealing pitfalls in PerspectiveAPI. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 15–24, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget.(may 2023).

Tejpalsingh Siledar, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2023. Synthesize, if you do not have: Effective synthetic dataset creation strategies for self-supervised opinion summarization in E-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13480–13491, Singapore. Association for Computational Linguistics.

Tiberiu Sosea, Chau Pham, Alexander Tekle, Cornelia Caragea, and Junyi Jessy Li. 2022. Emotion analysis and detection during COVID-19. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6938–6947, Marseille, France. European Language Resources Association.

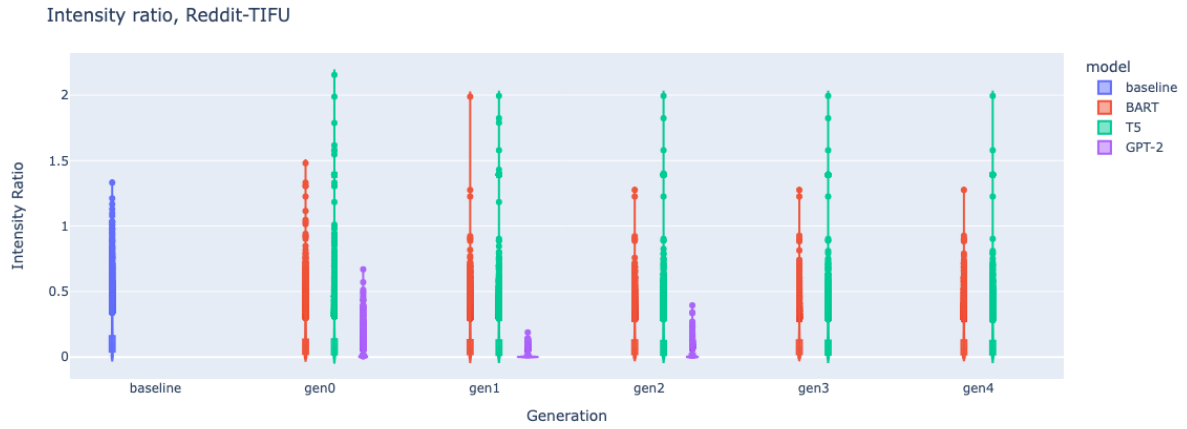
Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models.

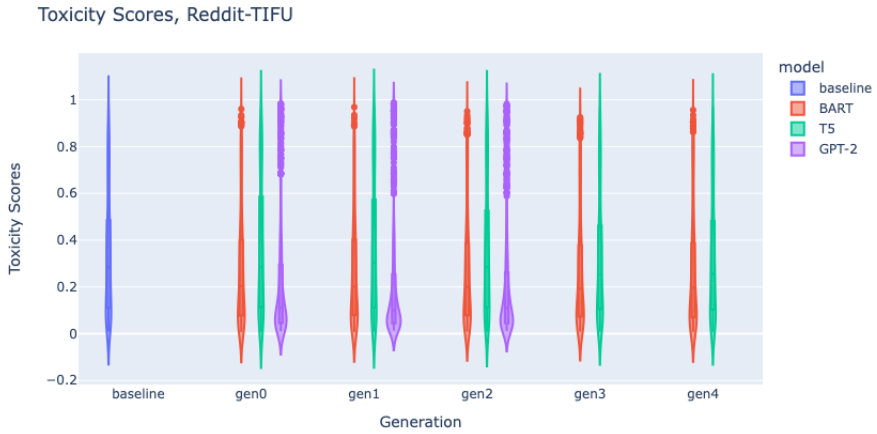
Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization.

		SAMSum		TIFU-long		CNN/DailyMail	
		Coverage	Compression ratio	Coverage	Compression ratio	Coverage	Compression ratio
Baseline		0.24	0.34	0.07	0.09	0.13	0.09
T5	Gen_0	0.26	0.27	0.09	0.07	0.13	0.08
	Gen_1	0.27	0.27	0.09	0.08	0.14	0.08
	Gen_2	0.27	0.27	0.10	0.08	0.14	0.08
	Gen_3	0.27	0.27	0.10	0.08	0.15	0.09
	Gen_4	0.27	0.27	0.10	0.08	NA	NA
BART	Gen_0	0.24	0.30	0.08	0.06	0.16	0.09
	Gen_1	0.25	0.30	0.09	0.06	0.16	0.09
	Gen_2	0.25	0.31	0.09	0.06	0.16	0.09
	Gen_3	0.25	0.31	0.09	0.06	0.15	0.08
	Gen_4	0.26	0.31	0.09	0.06	NA	NA
GPT-2	Gen_0	NA	NA	0.06	0.33	NA	NA
	Gen_1	NA	NA	0.06	0.34	NA	NA
	Gen_2	NA	NA	0.07	0.34	NA	NA

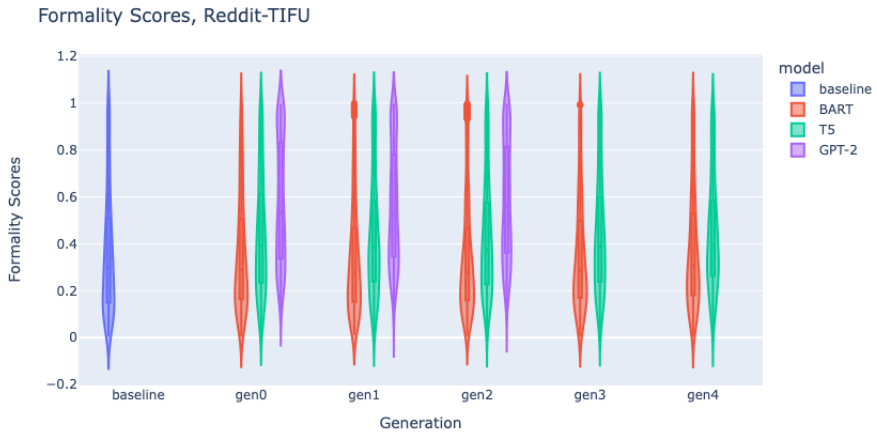
Table 10: Summary length metrics of compression ratio and coverage for the different combinations of base model, generation, and dataset.



(a) Distribution of emotion intensity scores for entire set of summaries in the real and synthetic *TIFU* datasets, by base model and generation number.

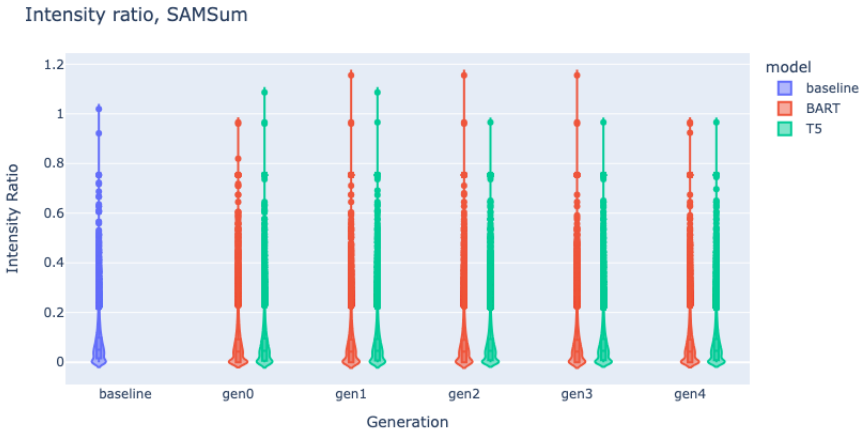


(b) Distribution of toxicity scores for a random sample of 1,000 summaries in the real and synthetic *TIFU* datasets, by base model and generation number.

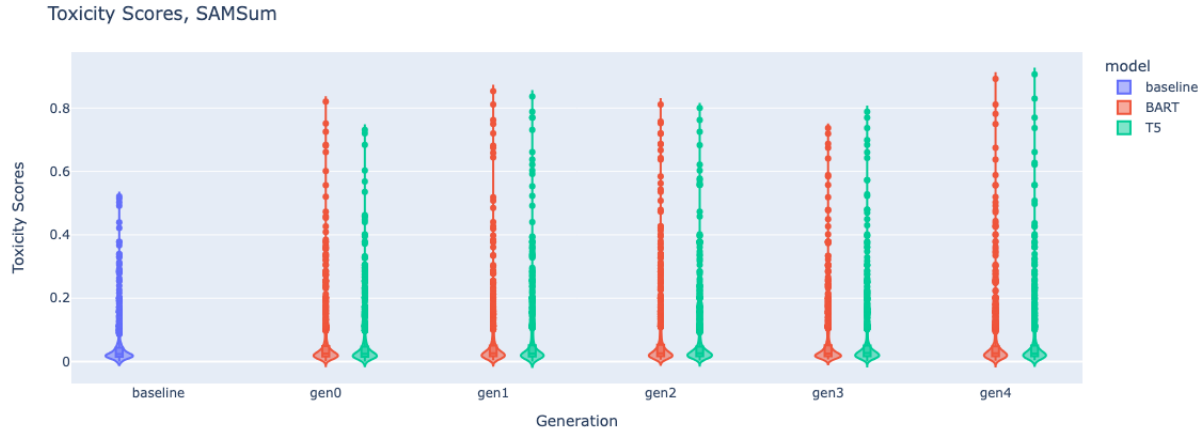


(c) Distribution of formality scores for a random sample of 1,000 summaries in the real and synthetic *TIFU* datasets, by base model and generation number.

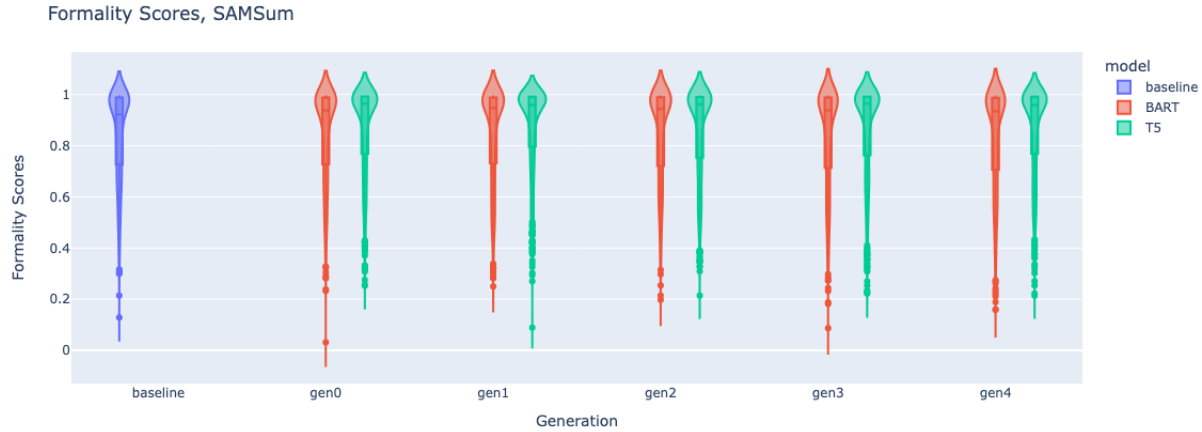
Figure 2: Violin plots illustrating the distribution of scores for the qualitative text characteristics of emotion intensity (a), toxicity (b), and formality (c) for the *TIFU* dataset.



(a) Distribution of emotion intensity scores for entire set of summaries in the real and synthetic *SAMSum* datasets, by base model and generation number.

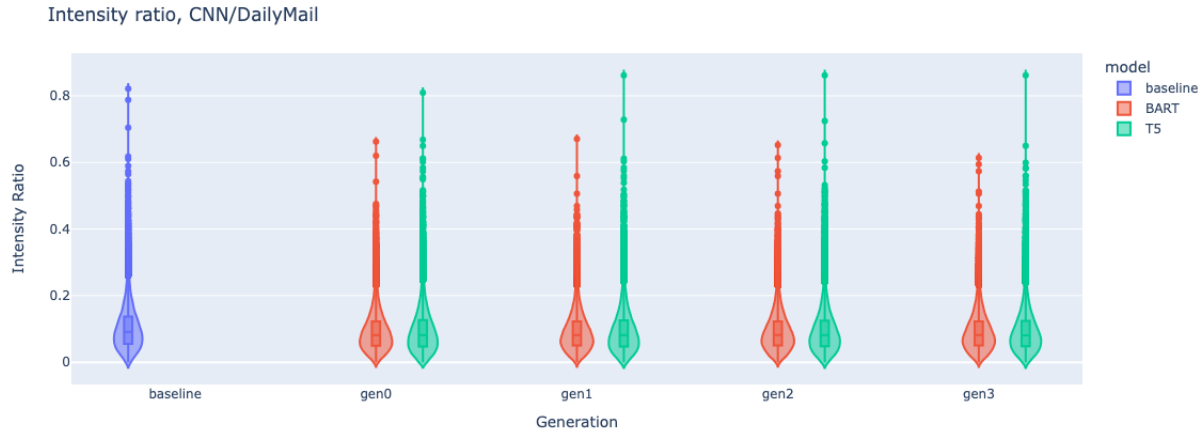


(b) Distribution of toxicity scores for a random sample of 1,000 summaries in the real and synthetic *SAMSum* datasets, by base model and generation number.

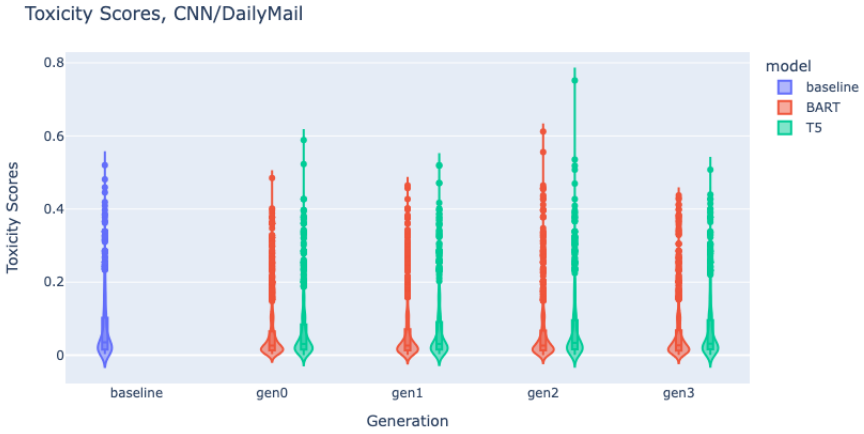


(c) Distribution of formality scores for a random sample of 1,000 summaries in the real and synthetic *SAMSum* datasets, by base model and generation number.

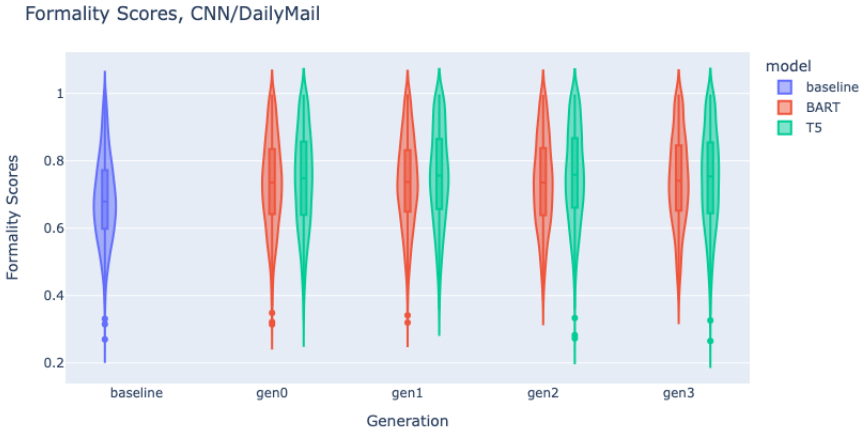
Figure 3: Violin plots illustrating the distribution of scores for the qualitative text characteristics of emotion intensity (a), toxicity (b), and formality (c) for the *SAMSum* dataset.



(a) Distribution of emotion intensity scores for entire set of summaries in the real and synthetic *SAMSum* datasets, by base model and generation number.



(b) Distribution of toxicity scores for a random sample of 1,000 summaries in the real and synthetic *CNN/DailyMail* datasets, by base model and generation number.



(c) Distribution of formality scores for a random sample of 1,000 summaries in the real and synthetic *CNN/DailyMail* datasets, by base model and generation number.

Figure 4: Violin plots illustrating the distribution of scores for the qualitative text characteristics of emotion intensity (a), toxicity (b), and formality (c) for the *CNN/DailyMail* dataset.