DIFFUSION DB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models

Zijie J. Wang  Evan Montoya  David Munechika
Haoyang Yang  Benjamin Hoover  Duen Horng Chau
College of Computing, Georgia Tech
{jayw|emontoya30|david.munechika|alexanderyang|bhoov|polo}@gatech.edu

Abstract

With recent advancements in diffusion models, users can generate high-quality images by writing text prompts in natural language. However, generating images with desired details requires proper prompts, and it is often unclear how a model reacts to different prompts or what the best prompts are. To help researchers tackle these critical challenges, we introduce DIFFUSION DB, the first large-scale text-to-image prompt dataset totaling 6.5TB, containing 14 million images generated by Stable Diffusion, 1.8 million unique prompts, and hyperparameters specified by real users. We analyze the syntactic and semantic characteristics of prompts. We pinpoint specific hyperparameter values and prompt styles that can lead to model errors and present evidence of potentially harmful model usage, such as the generation of misinformation. The unprecedented scale and diversity of this human-actuated dataset provide exciting research opportunities in prompt engineering, deepfake detection, and understanding large generative models.

DIFFUSION DB is publicly available at: https://poloclub.github.io/diffusiondb.

1 Introduction

Recent diffusion models have gained immense popularity by enabling high-quality and controllable image generation based on text prompts written in natural language (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022). Since the release of these models, people from different domains have quickly applied them to create award-winning artworks (Roose, 2022), synthetic radiology images (Chambon et al., 2022), and even hyper-realistic videos (Ho et al., 2022).

However, generating images with desired details is difficult, as it requires users to write proper prompts specifying the exact expected results. Developing such prompts requires trial and error, and can often feel random and unprincipled (Liu and Chilton, 2022). Willison et al. (2022) analogize writing prompts to wizards learning “magical spells”: users do not understand why some prompts work, but they will add these prompts to their “spell book.” For example, to generate highly-detailed images, it has become a common practice to add special keywords such as “trending on artstation” and “unreal engine” in the prompt.

Prompt engineering has become a field of study in the context of text-to-text generation, where researchers systematically investigate how to construct prompts to effectively solve different downstream tasks (Branwen, 2020; Reynolds and McDonell, 2021). As large text-to-image models are relatively new, there is a pressing need to understand how these models react to prompts, how to write effective prompts, and how to design tools to help users generate images (Liu and Chilton, 2022). Our work helps researchers tackle these critical challenges, through three major contributions:

- DIFFUSION DB (Fig. 1), the first large-scale prompt dataset totaling 6.5TB, containing 14 million images generated by Stable Diffusion (Rombach et al., 2022) using 1.8 million unique prompts.
unique prompts and hyperparameters specified by real users. We construct this dataset by collecting images shared on the Stable Diffusion public Discord server (§ 2). We release DIFFUSIONDB with a CC0 1.0 license, allowing users to flexibly share and adapt the dataset for their use. In addition, we open-source our code\(^1\) that collects, processes, and analyzes the images and prompts.

- **Revealing prompt patterns and model errors.** The unprecedented scale of DIFFUSIONDB paves the path for researchers to systematically investigate diverse prompts and associated images that were previously not possible. By characterizing prompts and images, we discover common prompt patterns and find different distributions of the semantic representations of prompts and images. Our error analysis highlights particular hyperparameters and prompt styles that lead to model errors. Finally, we provide evidence of image generative models being used for potentially harmful purposes such as generating misinformation and nonconsensual pornography (§ 3).

- **Highlighting new research directions.** As the first-of-its-kind text-to-image prompt dataset, DIFFUSIONDB opens up unique opportunities for researchers from natural language processing (NLP), computer vision, and human-computer interaction (HCI) communities. The scale and diversity of this human-actuated dataset will provide new research opportunities in better tooling for prompt engineering, explaining large generative models, and detecting deepfakes (§ 4).

We believe DIFFUSIONDB will serve as an important resource for researchers to study the roles of prompts in text-to-image generation and design next-generation human-AI interaction tools.

\(^1\)Code: [https://github.com/poloclub/diffusiondb](https://github.com/poloclub/diffusiondb)

### 2 Constructing DIFFUSIONDB

We construct DIFFUSIONDB (Fig. 2) by scraping user-generated images from the official Stable Diffusion Discord server. We choose Stable Diffusion as it is currently the only open-source large text-to-image generative model, and all generated images have a CC0 1.0 license that allows uses for any purpose (StabilityAI, 2022b). We choose the official public Discord server as it has strict rules against generating illegal, hateful, or NSFW (not suitable for work, such as sexual and violent content) images, and it prohibits sharing prompts with personal information (StabilityAI, 2022a).

Our construction process includes collecting images (§ 2.1), linking them to prompts and hyperparameters (§ 2.2), applying NSFW detectors (§ 2.3), creating a flexible file structure (§ 2.4), and distributing the dataset (§ 2.5). We discuss DIFFUSIONDB’s limitations and broader impacts in § 7, § 8, and a Data Sheet (Gebru et al., 2020) (‡ A).

#### 2.1 Collecting User Generated Images

We download chat messages from the Stable Diffusion Discord channels with DiscordChatExporter (Holub, 2017), saving them as HTML files. We focus on channels where users can command a bot to run Stable Diffusion Version 1 to generate images by typing a prompt, hyperparameters, and the number of images. The bot then replies with the generated images and used random seeds.

#### 2.2 Extracting Image Metadata

We use Beautiful Soup (Richardson, 2007) to parse HTML files, mapping generated images with their prompts, hyperparameters, and image size. Each image also has a unique filename, a hash of its creator’s Discord username, and a creation timestamp. To help researchers filter out potentially unsafe or harmful content, we employ state-of-the-art models to compute an NSFW score for each image and prompt.
images as a grid (e.g., a $3 \times 3$ grid of $n = 9$ images); these images have the same prompt and hyperparameters but different seeds. We use Pillow (Clark, 2015) to split a collage into $n$ individual images and assign them with the correct metadata and unique filenames. Finally, we compress all images in DIFFUSIONDB using lossless WebP (Google, 2010).

2.3 Identifying NSFW Content

The Stable Diffusion Discord server prohibits generating NSFW images (StabilityAI, 2022a). Also, Stable Diffusion has a built-in NSFW filter that automatically blurs generated images if it detects NSFW content. However, we find DIFFUSIONDB still includes NSFW images that were not detected by the built-in filter or removed by server moderators. To help researchers filter these images, we apply state-of-the-art NSFW classifiers to compute NSFW scores for each prompt and image. Researchers can determine a suitable threshold to filter out potentially unsafe data for their tasks.

**NSFW Prompts.** We use a pre-trained multilingual toxicity prediction model to detect unsafe prompts (Hanu and Unitary team, 2020). This model outputs the probabilities of a sentence being toxic, obscene, threat, insult, identity attack, and sexually explicit. We compute the text NSFW score by taking the maximum of the probabilities of being toxic and sexually explicit (Fig. 3 Top).

**NSFW Images.** We use a pre-trained EfficientNet classifier to detect images with sexual content (Schuhmann et al., 2022). This model predicts the probabilities of five image types: drawing, hentai, neutral, sexual, or porn. We compute the image NSFW score by summing the probabilities of hentai, sexual, and porn. We use a Laplacian convolution kernel with a threshold of 10 to detect images that have already been blurred by Stable Diffusion and assign them a score of 2.0 (Fig. 3 Bottom). As Stable Diffusion’s blur effect is strong, our blurred image detector has high precision and recall (both 100% on 50k randomly sampled images).

**NSFW Detector Accuracy.** To access the accuracy of these two pre-trained state-of-the-art NSFW detectors, we randomly sample 5k images and 2k prompt texts and manually annotate them with two binary NSFW labels (one for image and one for prompt) and analyze the results. As the percentage of samples predicted as NSFW (score $> 0.5$) is small, we up-sample positive samples for annotation, where we have an equal number of positive and negative examples in our annotation sample. After annotation, we compute the precisions and recalls. Because we have up-sampled positive predictions, we adjust the recalls by multiplying false negatives by a scalar to adjust the sampling bias. The up-sampling does not affect precisions. Finally, the precisions, recalls and adjusted recalls are $0.3604$, $0.9565$, and $0.6661$ for the prompt NSFW detector, and $0.315$, $0.9722$, and $0.3037$ for the image NSFW detector. Our results suggest two detectors are progressive classifiers. The lower adjusted recall of the prompt NSFW detector can be attributed to several potential factors, including the use of a fixed binary threshold and the potential discrepancy in the definition of NSFW prompts between the detector and our annotation process.

2.4 Organizing DIFFUSIONDB

We organize DIFFUSIONDB using a flexible file structure. We first give each image a unique filename using Universally Unique Identifier (UUID, Version 4) (Leach et al., 2005). Then, we organize images into 14,000 sub-folders—each includes 1,000 images. Each sub-folder also includes a JSON file that contains 1,000 key-value pairs mapping an image name to its metadata. An example of this image-prompt pair can be seen in Fig. 2. This modular file structure enables researchers to flexibly use a subset of DIFFUSIONDB.

We create a metadata table in Apache Parquet format (Apache, 2013) with 13 columns: unique image name, image path, prompt, seed, CFG scale, sampler, width, height, username hash, timestamp, image NSFW score, and prompt NSFW score.
We store the table in a column-based format for efficient querying of individual columns.

2.5 Distributing DIFFUSIONDB

We distribute DIFFUSIONDB by bundling each image sub-folder as a Zip file. We collect Discord usernames of image creators (§ 2.2), but only include their SHA256 hashes in the distribution—as some prompts may include sensitive information, and explicitly linking them to their creators can cause harm. We host our dataset on a publicly accessible repository\(^2\) under a CC0 1.0 license. We provide scripts that allow users to download and load DIFFUSIONDB by writing two lines of code.

We discuss the broader impacts of our distribution in § 7, § 8, and the Data Sheet (‡ A). To mitigate the potential harms, we provide a form for people to report harmful content for removal. Image creators can also use this form to remove their images.

3 Data Analysis

To gain a comprehensive understanding of the dataset, we analyze it from different perspectives. We examine prompt length (§ 3.1), language (§ 3.2), characteristics of both prompts (§ 3.3) and images (§ 3.4). We conduct an error analysis on misaligned prompt-image pairs (§ 3.5) and provide empirical evidence of potentially harmful uses of image generative models (§ 3.6).

3.1 Prompt Length

We collect prompts from Discord, where users can submit one prompt to generate multiple images and experiment with different hyperparameters. Our dataset contains 1,819,808 unique prompts. We tokenize prompts using the same tokenizer as used in Stable Diffusion (Platen et al., 2022). This tokenizer truncates tokenized prompts at 75 tokens, excluding special tokens <|startoftext|> and <|endoftext|>. We measure the length of prompts by their tokenized length. The prompt length distribution (Fig. 4) indicates that shorter prompts (e.g., around 6 to 12 tokens) are the most popular. The spike at 75 suggests many users submitted prompts longer than the model’s limit, highlighting the need for user interfaces guiding users to write prompts within the token limit.

3.2 Prompt Language

We use a pre-trained language detector (Joulin et al., 2017) to identify the languages used in prompts. 98.3% of the unique prompts in our dataset are written in English. However, we also find a large number of non-English languages, with the top four being German (5.2k unique prompts), French (4.6k), Italian (3.2k), and Spanish (3k). The language detector identifies 34 languages with at least 100 unique prompts in total. Stable Diffusion is trained on LAION-2B(en) (Schuhmann et al., 2022) that primarily includes images with English descriptions, thus our findings suggest that expanding the training data’s language coverage to improve the user experience for non-English communities.

3.3 Characterizing Prompts

In this section, we explore the characteristics of prompts in DIFFUSIONDB. We examine the syntactic (§ 3.3.1) and semantic (§ 3.3.2) features of prompt text via interactive data visualizations. Lastly, We discuss the implications of our findings and suggest future research directions.

3.3.1 Prompt Syntactic Features

To characterize the composition of prompts, we parse phrases from all 1.8M unique prompts. We split each prompt by commas and then extract named entities (NE) and noun phrases (NP) from each separated component using use Spacy (Honnibal et al., 2020). If there is no noun phrase in a comma-separated component, we extract the whole component (C) as a phrase. We keep track of each NP’s root to create a hierarchy of noun phrases.

For example, for the prompt “draw baby yoda in a loading screen for grand theft auto 5, highly detailed, digital art, concept art,” we extract six phrases: “baby yoda” (NE), “a loading screen” (NP with root “screen”), “grand theft auto 5” (NE), “highly detailed” (C), “digital art” (NP with root “art”), and “concept art” (NP with root “art”). We group

---

\(^2\)Public dataset repository: https://huggingface.co/datasets/poloclub/diffusiondb
“digital art” and “concept art” into the same hierarchy as they share the same NP root “art.”

Visualizing Prompt Phrases. We create an interactive circle packing visualization to gain an understanding of the distribution and relationships between different phrases (Fig. 5). Circle packing (Wang et al., 2006) is a technique to visualize hierarchical data, and each phrase is represented as a circle whose size encodes the phrase’s frequency in the dataset. We position sibling noun phrases (e.g., phrases sharing the same NP root) inside their parent phrase’s circle through a front-chain packing algorithm (Wang et al., 2006). Viewers can hover over a circle to see the corresponding phrase and its frequency. Viewers can also click a circle (Fig. 5A) to zoom into that sub-tree to see more details about a phrase (Fig. 5-B1) or a sub-phrase (Fig. 5-B2).

Insights and implications. Our interactive visualization reveals that key phrases such as “highly detailed,” “intricate,” and “greg rutkowski” are commonly used in prompts (Fig. 5A). The hierarchical visualization also surfaces popular image styles specified by users, such as “digital painting,” “oil painting,” and “portrait painting” for painting styles (Fig. 5-B1) and “studio lighting,” “volumetric lighting”, and “atmospheric lighting” for lighting. These phrases can be unfamiliar to Stable Diffusion users, especially beginners, which highlights the importance of helping users develop prompting vocabularies. Researchers can leverage DIFFUSIONDB and our visualization to design tutorials and user interfaces that integrate exemplar prompts to guide users in describing their desired images.

3.3.2 Prompt Semantic Features

In addition to analyzing the syntactic characteristics of prompts, we also analyze their semantic features. We use a pre-trained CLIP model (Radford et al., 2021) to extract semantic features (Ramesh et al., 2022). We use a frozen CLIP ViT-L/14 text encoder (the same model used in Stable Diffusion) to convert prompts into 768-dimension vectors.
Visualizing Prompt Embeddings. To study the distribution of prompts in high-dimensional space, we use UMAP (McInnes et al., 2020) to project 768-dimensional vectors into 2-D vectors for easy visualization. UMAP is a popular dimensionality reduction technique that is better at preserving the global structure of data and more scalable to large datasets compared to t-SNE (van der Maaten and Hinton, 2008) and PCA (Hotelling, 1936). We use grid search to fine-tune hyperparameters $n_{neighbors}$ (60) and $min\_dist$ (0.1) so that prompts are more spread out in a 2-D space.

We develop an interactive visualization tool\(^4\) to explore prompts’ semantic embeddings (Fig. 6). We use Kernel Density Estimation (KDE) (Rosenblatt, 1956) with a standard multivariate Gaussian kernel and Silverman bandwidth (Silverman, 2018) to estimate the distribution of prompts’ UMAP representations. Then, we visualize the estimated distribution as a contour plot. To summarize prompts that are in the same region, we create four grids with varying granularity and pre-compute keywords for each grid tile, by treating all prompts in the tile as a document and selecting the top 4 keywords with the highest TF-IDF scores.

Interactions. Our visualization shows keywords of tiles that are close to high-density regions and prompt clusters by default. Viewers can hover over a tile to see its keywords, pan and zoom in to see more details of specific regions, and click a button to display each prompt as a small dot that viewers can hover over to read its prompt text.

Insights and implications. Our semantic embedding visualization (Fig. 6) highlights two popular prompt categories: art-related prompts (left in the plot) and photography-related prompts (dark blue regions on the right). These two groups appear distant from each other in the UMAP space, suggesting that the prompts for art and photography typically have distinct semantic representations. Interestingly, photography prompts appear to contain two clusters: one for non-human objects (top right) and another for celebrities (bottom right). Small prompt clusters outside the central area often feature artist names. Our findings suggest that future researchers can leverage the prompt usage distribution to fine-tune generative models to tailor to specific popular prompt categories.

3.4 Characterizing Images

We visualize\(^5\) the CLIP embedding distribution of 2 million unique image instances randomly sampled from DIFFUSIONDB (Fig. 7) by defining the unique key as the combination of the image’s prompt and hyperparameters $CFG\_scale$, $step$, $size$, and $seed$. We use the UMAP model that was previously trained on the prompt embeddings to project the image embeddings into the same 2-D space. Finally, we apply the same method we used for our prompt embedding visualization (§ 3.3.2) to generate a contour plot and grid label overlays.

Insights and implications. Our image embedding visualization reveals that generated images have a different distribution from their prompts in the CLIP embedding space. For example, the

---

\(^4\)Prompt embedding visualization: https://poloclub.github.io/diffusiondb/explorer/#prompt-embedding

\(^5\)Image embedding visualization: https://poloclub.github.io/diffusiondb/explorer/#image-embedding
“movie” cluster in the prompt embedding has been replaced by the “portrait” cluster in the image embedding. This suggests the semantic representations of prompts and their generated images may not be perfectly aligned. One hypothesis is that large image generative models face limitations when generating photorealistic human faces (Borji, 2022), and therefore some images generated with movie-related prompts appear to be closer to art and portrait regions in the embedding space.

3.5 Stable Diffusion Error Analysis

We leverage DIFFUSIONDB to discover Stable Diffusion generation failure cases and examine potential causes. To surface poor image generations, we compute CLIP embeddings for all prompts and images in DIFFUSIONDB. We then select prompt-image pairs with a large cosine distance (d) between their embeddings. The cosine distances have a normal distribution (N(0.7123, 0.04132)). In this analysis, we focus on 13,411 “bad” prompt-image pairs (1) with a distance that is larger than 4 standard deviations from the mean and (2) the image was not blurred by Stable Diffusion (§ 2.3).

Impacts of hyperparameters. We conduct a logistic regression test to analyze the relationship between Stable Diffusion hyperparameter values (e.g., CFG scale, step, width, and height) and the likelihood of generating an image that is semantically different from its prompt. The results reveal that all four hyperparameters are negatively correlated with the likelihood of generating a bad image. The correlation is statistically significant with a p-value of less than 0.0001 for all four variables. Furthermore, we find the distribution of selected sampler options when generating bad images is significantly different from the overall distribution ($X^2 = 40873.11, p < 0.0001$).

CFG scale controls how much the generated image looks like the prompt. We find some users specify negative CFG scales that make images look different from their prompts (large cosine distance d). In the example shown on the right, a user generates an image using a prompt about “superman” with all default hyperparameters values, except for setting CFG scale to -1. This results in an image featuring a bowl of soup instead of “superman”.

A small step could also generate under-developed images that look different from the specified prompts. As demonstrated in the example on the right, a user generates an image about “plague doctor” with all default hyperparameter values, except for setting step to 2, which leads to a blurry image.

Stable Diffusion struggles with generating images with a small size or large aspect ratios. The dissimilar image shown on the right is generated with default hyperparameters except for a size of (64, 512).

Impacts of prompts. Despite controlling all hyperparameters to be close to default values, we still find 1.1k unique bad image-prompt pairs. Most of these instances have non-English prompts, very short prompts, or prompts consisting primarily of emojis (see an example on the right). The token lengths of these instances are significantly lower than the overall token length (one-tailed $t = -23.7203, p < 0.0001$). The English prompt frequency among these instances is also significantly lower than the overall frequency ($X^2 = 1024.56, p < 0.0001$). Interestingly, we also find that Stable Diffusion sometimes generates unexpected images even when prompts are meaningful English sentences. Future researchers can use our error analysis and failure cases to check potentially mislabeled training data.

Implications. Our study reveals Stable Diffusion can make mistakes when generating images with certain hyperparameter values or prompt styles. Negative CFG scales, small steps, or small sizes contribute to generating images dissimilar to prompts. Short and non-English prompts can also lead to errors. To improve the quality of future generative models, researchers can expand the training data to cover these edge cases. There are opportunities for researchers to design user interfaces that can help users understand the impact of different hyperparameters and guide them in choosing values that fit their specific use cases.

3.6 Potentially Harmful Uses

To identify potentially malicious uses of Stable Diffusion, we use named entity recognition to analyze prompts. We find that many prompts include names
of influential politicians, such as over 65k images generated with a prompt including “Donald Trump” and over 48k images with “Joe Biden.” Some prompts portray these politicians in negative lights, ranging from depicting them “as Gollum with hair” to “arrested in handcuffs.” Additionally, we find female celebrities are frequently used in prompts, with a high frequency after artists and influential politicians. Some of these prompts are presented in a sexual context that could be considered nonconsensual pornography.

Through keyword search, we discover prompts generating misinformation that could cause harm. For example, the prompt "scientists putting microchips into a vaccine" may harm public trust in medical institutions by potentially validating conspiracy theories. Similarly, the prompt "Russian soldiers in gas masks found the last surviving Ukrainian after a nuclear war to liberate Ukraine" depicts false images of the Russo-Ukrainian War and could lead to new forms of propaganda. Our findings highlight the crucial need for further research on the broader impacts of large generative models and ways to regulate and mitigate their harms.

4 Enabling New Research Directions

The unprecedented scale and diversity of DIFFUSIONDB bring new exciting research opportunities to help users generate images more effectively and efficiently, and enable researchers to improve, explain, and safeguard generative models.

Prompt Autocomplete. With DIFFUSIONDB, researchers can develop an autocomplete system to help users construct prompts. For example, one can use the prompt corpus to train an n-gram model to predict likely words following a prompt part. Alternatively, researchers can use semantic auto-complete (Hyvönen and Mäkelä, 2006) by categorizing prompt keywords into ontological categories such as subject, style, quality, repetition, and magic terms (Oppenlaender, 2022). This allows the system to suggest related keywords from unspecified categories, for example suggesting style keyword “depth of field” and a magic keyword “award-winning” to improve the quality of generated images. Additionally, researchers can also use DIFFUSIONDB to study prompt auto-replace by distilling effective prompt patterns and creating a “translation” model that replaces weaker prompt keywords with more effective ones.

Generation through Search. As DIFFUSIONDB contains 14 million images, this dataset might have already included images with a user’s desired effects. Thus, a user can quickly search images in DIFFUSIONDB instead of running Stable Diffusion, which can be slow and costly. Lexica (Shameem, 2022), an AI start-up, provides such a search engine, where users can search Stable Diffusion images by natural language or images. Researchers can also construct a structured index of images and prompts, such as building a semantivisual image hierarchy of images (Li et al., 2010) or a hierarchical topic model of prompts (Griffiths et al., 2003), to help users easily discover and explore images and prompts with similar styles.

Improving Generative Models. With DIFFUSIONDB, a large and diverse collection of Stable Diffusion usage logs, researchers not only can identify weak points and failure modes of Stable Diffusion but also gain insights into user preferences. For example, we demonstrate that researchers can use joint text-image embeddings between prompts and images to detect generation misalignments (§ 3.5). Additionally, DIFFUSIONDB provides important metadata such as username hash and timestamp for each generated image. By analyzing these metadata fields, researchers can trace the evolution chain of prompts, parameters, and images, which offers valuable insights into how users develop mental models of large generative models and their preferences of generated images. This understanding can inform future researchers to enhance generative models and design interfaces that facilitate better image-generation experiences.

Explainable Generation. As generative models have been gaining immense popularity, there is a call for explainable creativity (Llano et al., 2022). Many explanation techniques use input permutation that computes feature attribution scores by running a model on slightly-modified input values (Lundberg and Lee, 2017). DIFFUSIONDB contains 14 million prompt-image pairs including similar prompts with minor differences, such as “a happy dog” and “a sad dog”, allowing researchers to investigate how individual keywords affect the generation process.

Deepfake Detection. Breakthroughs in generative models raise concerns about deepfakes—fake images of real individuals for unethical purposes (Wiggers, 2022). DIFFUSIONDB is valu-
able for detecting deepfakes, as it contains a large-scale collection of model-generated images and their metadata. Researchers can use this collection to train ML models to identify synthetic artifacts and train classifiers that classify synthetic images from real images (Mirsky and Lee, 2022).

5 Related Work

Text-to-text Prompting. Researchers have been studying prompt engineering for text-to-text generation (e.g., Liu et al., 2022; Lu et al., 2022; Rubin et al., 2022). To facilitate this line of research, researchers develop PromptSource (Bach et al., 2022), a dataset of 2k text prompts along with a framework to create and share prompts. In contrast, our work focuses on text-to-image prompting, and DIFFUSIONDB has an unprecedented scale of 14 million real prompt-image pairs.

Text-to-image Prompting. There is a growing interest in text-to-image prompt engineering research from NLP, Computer Vision, and HCI communities (e.g., Qiao et al., 2022; Pavlichenko and Ustalov, 2022). For example, Oppenlaender (2022) identifies six types of prompt modifiers through an ethnographic study, and Liu and Chilton (2022) proposes design guidelines for text-to-image prompt engineering by experimenting with 1,296 prompts. Closest in spirit to DIFFUSIONDB is Lexica (Shameem, 2022) which allows users to search over 5 million Stable Diffusion images with their prompts, but it does not release its internal database. In comparison, DIFFUSIONDB is open-source and publicly available to everyone.

6 Conclusion

We present DIFFUSIONDB, the first large-scale text-to-image prompt dataset, containing 14 million images with their prompts and hyperparameters collected from the Stable Diffusion Discord server. We release the dataset with a CC0 1.0 license and open source all collection and analysis code, broadening the public’s access to cutting-edge AI technologies. We discuss findings on prompt and image patterns. We hope our work will serve as a cornerstone for the future development of large generative modes and tools that help users use these modes.

7 Limitations

We discuss four limitations of our work: the inclusion of unsafe content, potential biases in data sources, a limited measure of image quality and generalizability to different generative models.

- **Inclusion of unsafe images and prompts.** We collect images and their prompts from the Stable Diffusion Discord server (§ 2). The Discord server has rules against users generating or sharing harmful or NSFW (not suitable for work, such as sexual and violent content) images. The Stable Diffusion model used in the server also has an NSFW filter that blurs the generated images if it detects NSFW content. However, we observe that DIFFUSIONDB includes some NSFW images that were not detected by the NSFW filter or removed by the server moderators. To mitigate the potential harm, we compute and share the likelihood of an image or a prompt containing unsafe content using the state-of-the-art NSFW detectors (§ 2.3). In addition, we provide a Google Form on the DIFFUSIONDB website where users can report harmful or inappropriate images and prompts. We will closely monitor this form and remove reported images and prompts from DIFFUSIONDB.

- **Potential biases of the data source.** The 14 million images in DIFFUSIONDB have diverse styles and categories. However, Discord can be a biased data source. Our images come from channels where early users could use a bot to use Stable Diffusion before release. As these users had started using Stable Diffusion before the model was public, we hypothesize that they are AI art enthusiasts and are likely to have experience with other text-to-image generative models. Therefore, the prompting style in DIFFUSIONDB might not represent novice users. Similarly, the prompts in DIFFUSIONDB might not generalize to domains that require specific knowledge, such as medical images (Chambon et al., 2022).

- **Limited measure of image quality.** We use joint text-image CLIP embeddings between prompts and images to detect generation misalignments (§ 3.5). While the CLIP embedding distance can indicate the degree of alignment between the prompts and generated images, it does not provide a measure of the overall image quality. When constructing our dataset, we have considered including image properties such as entropy, variance, and the most common colors to help users gauge image qualities. However, these metrics do not provide a good measure of the overall image quality as well. To better mea-
sure image quality, future researchers can recruit annotators to rate images in DIFFUSION DB.

- **Generalizability.** Previous research has shown a prompt that works well on one generative model might not give the optimal result when used in other models (Borji, 2022). Therefore, different models can need users to write different prompts. For example, many Stable Diffusion prompts use commas to separate keywords, while this pattern is less seen in prompts for DALL-E 2 (Ramesh et al., 2022) or Midjourney (Holz, 2022). Thus, we caution researchers that some research findings from DIFFUSIONDB might not be generalizable to other text-to-image generative models.

8 Ethics Statement

In this section, we discuss two main ethical considerations of DIFFUSIONDB.

- **Copyright.** By using the Stable Diffusion Discord server, all users agree to the entirety of CC0 1.0 Universal Public Domain Dedication. This includes waiving any intellectual property rights related to any content shared on the server (StabilityAI, 2022b). All prompts and images in the Discord server are considered to be public domain and can be used by anyone for any purpose. Also, we release DIFFUSIONDB under the CC0 1.0 license (§ 2.5).

- **Privacy.** While it is possible that some prompts may contain sensitive information, this is not common because the Stable Diffusion Discord has strict rules against writing personal information in the prompts and has moderators in place to remove violative messages. To further protect user privacy, we have anonymized the usernames of all users in our dataset (§ 2.4). Users also have the option to remove their prompts and images from our dataset through an online form (§ 2.5).

We provide a thorough discussion on the limitations and broader impacts of DIFFUSIONDB in its Data Sheet (Gebru et al., 2020) (‡ A).

Acknowledgements

We thank Stability AI for releasing Stable Diffusion and hosting the Stable Diffusion Discord server. We especially appreciate the Stable Diffusion Discord moderators and users for creating an open and friendly online community that makes our work possible. We also extend our appreciation to Hugging Face for hosting our dataset. Lastly, we would like to acknowledge the anonymous reviewers for their valuable feedback and insightful comments that helped improve our paper. This work was supported in part by a J.P. Morgan PhD Fellowship, NSF grants IIS-1563816, DARPA GARD, gifts from Cisco, Bosch, and NVIDIA. Use, duplication, or disclosure is subject to the restrictions as stated in Agreement number HR00112030001 between the Government and the Performer.

References

Apache. 2013. Apache Parquet: Open Source, Column-oriented Data File Format Designed for Efficient Data Storage and Retrieval.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Weber, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Safful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshibani, Shanya Sharma, Urmish Thakker, Khalid Alnubar, Xiangru Tang, Dragomir Radiev, Mike Tian-jian Jiang, and Alexander Rush. 2022. Prompt-Source: An Integrated Development Environment and Repository for Natural Language Prompts. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Ali Borji. 2022. Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2. arXiv 2210.00586.

Gwern Branwen. 2020. GPT-3 Creative Fiction.

Pierre Chambon, Christian Bluethegen, Curtis P. Langlotz, and Akshay Chaudhari. 2022. Adapting Pre-trained Vision-Language Foundational Models to Medical Imaging Domains. arXiv 2210.04133.

Alex Clark. 2015. Pillow: Python Imaging Library (Fork).

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. arXiv:1803.09010 [cs].

Google. 2010. Comparative Study of WebP, JPEG and JPEG 2000.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. In Advances in Neural Information Processing Systems, volume 16.

Laura Hanu and Unitary team. 2020. Detoxify: Toxic Comment Classification with Pytorch Lightning and Transformers.
Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasempour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv 2205.11487.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarszyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv 2210.08402.

Sharif Shameem. 2022. Lexica: Building a Creative Tool for the Future.

Bernard W Silverman. 2018. Density Estimation for Statistics and Data Analysis.

StabilityAI. 2022a. Stable Diffusion Discord Server Rules.

StabilityAI. 2022b. Stable Diffusion Dream Studio beta Terms of Service.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9.

Weixin Wang, Hui Wang, Guozhong Dai, and Hongan Wang. 2006. Visualization of large hierarchical data by circle packing. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Kyle Wiggers. 2022. Deepfakes for all: Uncensored AI art model prompts ethics questions.

Simon Willison, Adam Stacoviak, and Jerod Stacoviak. 2022. Stable Diffusion Breaks the Internet.
A Data Sheet for DIFFUSION DB

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The DIFFUSIONDB project was inspired by important needs in research focused on diffusion models and prompt engineering. As large text-to-image models are relatively new, there is a pressing need to understand how these models work, how to write effective prompts, and how to design tools to help users generate images. To tackle these critical challenges, we present DIFFUSIONDB, the first large-scale prompt dataset with 14 million real prompt-image pairs.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau at the Georgia Institute of Technology.

Who funded the creation of the dataset?

If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funded in part by J.P. Morgan PhD Fellowship, NSF grants IIS-1563816, DARPA GARD, and gifts from Cisco, Bosch, and NVIDIA.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance consists of an image generated by the Stable Diffusion model and the prompt as well as parameters that were input into the model to generate the image. The input parameters include seed, CFG scale, sampler, width, height, username hash, timestamp, image NSFW score and prompt NSFW score.

How many instances are there in total (of each type, if appropriate)?

There are 14 million instances in total.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified.

The dataset is a sample of instances. It represents a sample of images from the Stable Diffusion discord server. No tests were run to determine representativeness.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of the image generated by the Stable Diffusion model (with a unique id), along with the prompt used to generate the image and the model parameters as a JSON file.

Is there a label or target associated with each instance? If so, please provide a description.

The labels associated with each image are the prompt and other input parameters.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Not applicable.

Are there recommended data splits (e.g., training, development/validation, testing)? If
so, please provide a description of these splits, explaining the rationale behind them.

No. This dataset is not for ML model benchmarking. Researchers can use any subsets of it.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No. All images and prompts are extracted as is from the Discord chat log.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is entirely self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ nonpublic communications)? If so, please provide a description. Unknown to the authors of the datasheet.

It is possible that some prompts contain sensitive information. However, it would be rare, as the Stable Diffusion Discord has rules against writing personal information in the prompts, and there are moderators removing messages that violate the Discord rules.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

We collect images and their prompts from the Stable Diffusion discord server. Even though the discord server has rules against users sharing any NSFW (not suitable for work, such as sexual and violent content) and illegal images, DIFFUSIONDB still contains some NSFW images and prompts that were not removed by the server moderators.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Any other comments?

None.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was directly observed from the Stable Diffusion Discord Channel. It was gathered from channels where users can generate images by interacting with a bot, which consisted of messages of user generated images and the prompts used to generate those images.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

The data was gathered using a DiscordChatExporter (Holub, 2017), which collected images and chat messages from each channel specified. We then extracted and linked prompts to images using Beautiful Soup (Richardson, 2007). Random images and prompts were selected and manually verified to validate the prompt-image mapping.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

DIFFUSIONDB does not sample from a larger set. However, DIFFUSIONDB-2M is a sample from a larger set. For certain messages, there would exist a collage of \( n \) images (e.g., \( n = 2, 4, 9 \)) with identical prompts consolidated into a single image. These images were split and a single image would be randomly selected to include in DIFFUSIONDB-2M from \( n \) images with equal probability of any image being selected. This saved space and prioritized unique prompts.
Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Students conducted the data collection process and were compensated with stipend or course credits.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?

All messages were generated in August 2022 and messages were collected between October 18th and 24th 2022. DIFFUSIONDB includes the generation timestamps of all images.

Were any ethical review processes conducted (e.g., by an institutional review board)?

There were no ethical review processes conducted.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was directly obtained from individual messages in the Discord server.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to any supporting documentation.

Users of the channel were not notified about this specific gathering of data but agree to forfeit any intellectual property rights claims by using Stable Diffusion. In addition, users are instructed that the images are public domain and can be used by anyone for any purpose. The exact language is as follows (StabilityAI, 2022b):

Note, that while users have forfeited copyright (and any/all intellectual property right claims) on these images, they are still public domain and can be used by anyone for any purpose, including by the user. Feel free to use images from DreamStudio Beta and the Stable Diffusion beta Discord service for anything, including commercial purposes.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

This message is contained in the rules and terms of service section of the Stable Diffusion Discord (StabilityAI, 2022a,b). In conjunction with the previous statement about images being public domain (CC0 1.0 license), it is established that the images made by using Stable Diffusion can be used for other purposes.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Users will have the option to report harmful content or withdraw images they created through a Google Form listed on the DIFFUSIONDB website: https://github.com/poloclub/diffusiondb.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting...
documentation. No analysis has been conducted.

Any other comments?
None.

Preprocessing

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

The Discord chat logs include collage images, where each collage contains a grid of images that share the same prompt but have different seeds. We use Pillow (Clark, 2015) to split a collage into individual images. For DiffusionDB, we include all split images. However, for DiffusionDB-2M, we only include one randomly selected split image to save space and prioritize unique prompts.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

Raw data was not saved.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

All our data collection and preprocessing code is available at: https://github.com/poloclub/diffusiondb.

Any other comments?
None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No.

What (other) tasks could the dataset be used for?
This dataset can be used for (1) prompt autocomplete, (2) generating images through search, (3) detecting deepfake, (4) debugging image generation, (5) explaining image generation, and more.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description.

There is minimal risk for harm: the data were already public. Personally identifiable data (e.g., discord usernames) were removed during the collection/preprocessing phases.

Are there tasks for which the dataset should not be used? If so, please provide a description.
All tasks that utilize this dataset should follow the licensing policies and the regulations (StabilityAI, 2022b) posed by Stability AI, the company that both made Stable Diffusion and runs the official Discord server.

Any other comments?
None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes, the dataset is publicly available on the internet.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is distributed on the project website: https://poloclub.github.io/diffusiondb. The dataset shares the same DOI as this paper.

When will the dataset be distributed?
The dataset is released on October 25th, 2022.
Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

All images generated by stable diffusion discord services are under the CC0 1.0 License, and therefore so are images in this dataset. In addition, the distribution of the dataset is under the Terms of Use (StabilityAI, 2022b) posed by Stability AI, the company that both made Stable Diffusion and runs the official Discord server.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

All images in this dataset have a CC0 1.0 License and follows the Stability AI’s Terms of Use (StabilityAI, 2022b).

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?
None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The authors of this paper will be supporting and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The contact information of the curators of the dataset is listed on the project website: https://poloclub.github.io/diffusiondb.

Is there an erratum? If so, please provide a link or other access point.
There is no erratum for our initial release. Errata will be documented in future releases on the dataset website.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
Yes, we will monitor the Google Form where users can report harmful images and creators can remove their images. We will update the dataset bimonthly. Updates will be posted on the project website https://poloclub.github.io/diffusiondb.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
People can use a Google Form linked on the project website to remove specific instances from DIFFUSIONDB.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
We will continue to support older versions of the dataset.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
Anyone can extend/augment/build on/contribute to DIFFUSIONDB. Potential collaborators can contact the dataset authors.

Any other comments?
None.
ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
  
  Section 7 and Appendix A

- A2. Did you discuss any potential risks of your work?
  
  Section 8 and Appendix A

- A3. Do the abstract and introduction summarize the paper’s main claims?
  
  Section 1

- A4. Have you used AI writing assistants when working on this paper?
  
  Left blank.

B ✔ Did you use or create scientific artifacts?

- We created a new dataset, described in Section 2.

- B1. Did you cite the creators of artifacts you used?
  
  Section 2

- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
  
  Section 2.5.

- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
  
  Section 2.

- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
  
  Section 2.5.

- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
  
  Abstract, section 2

- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
  
  Section 2 and 3

C ✗ Did you run computational experiments?

- Left blank.

  □ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
  
  Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.
C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants’ demographic (e.g., country of residence)?
Not applicable. Left blank.

D3. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.