Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

The Limitations and Ethical Considerations of ChatGPT

Shangying Hua, Shuangci Jin, Shengyi Jiang[†]

Guangdong University of Foreign Studies , Guangzhou 510000, China

Keywords: ChatGPT; ChatGPT Technology; Limitations; Ethical considerations; Artificial Intelligence

Citation: Shangying Hua, Shuangci Jin, Shengyi Jiang.: The Limitations and Ethical Considerations of ChatGPT. Data Intelligence XX(XX),XX–XX (2023). doi: 10.1162/dint_a_XXXX
Submitted: March 15, 2023; Revised: October 8, 2023; Accepted: November 30, 2023

ABSTRACT

With the advancements of artificial intelligence technology, ChatGPT, a new practice of artificial intelligence, holds immense potential across multiple fields. Its user-friendly human-machine interface, rapid response capabilities, and delivery of high-quality answers have attracted considerable attention and widespread usage. Regarded by many as a groundbreaking advancement in AI, ChatGPT represents a new milestone in the field. However, as with any technological evolution, the emergence of ChatGPT brings not only benefits, but also inevitable security risks and ethical issues. This paper provides specific information about ChatGPT, including its technology, limitations, ethical issues, governance paths and future directions. Specifically, we firstly offered a thorough exploration of the technical implementation details of GPT series models. Next, we provided an intricate analysis elucidating the reasons for limitations and scrutinized the consequential impacts, such as malicious misuse, privacy violation, and so on. Finally, we explore diverse governance paths to mitigate the impacts of ChatGPT and present future directions. This review aims to equip users with crucial knowledge, facilitating well-informed decision-making, effectively handling of potential challenges in employing ChatGPT, and staying abreast with the rapidly evolving landscape of this technology.

1. INTRODUCTION

ChatGPT (Chat Generative Pre-Trained Transformer), developed by OpenAI (Open Artificial Intelligence), made a stunning debut and quickly swept the world with its unprecedented text generation capabilities In November 2022. While people are excited about ChatGPT's excellent capabilities of language understanding and question answering, they are also concerned about whether this "human-like" language communication bot will pose a threat to human life, as shown in science fiction movies.

† Corresponding author: Shengyi Jiang (E-mail: jiangshengyi@163.com; ORCID: 0000-0002-6753-474X)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

The development of ChatGPT has gone through several iterations. In 2018, GPT-1 was published and opened the era of pre-trained large models. Compared to previous natural language models based on supervised learning, GPT-1 used a new "semi-supervised" training method which first trained a pre-trained model on unlabeled data, and then used a small size of labeled data to fine-tune the model, thereby gaining generalization ability. GPT-2 then came out next year, with unsupervised pre-trained and bigger scale of training datasets, its text generation and generalization capability were significantly improved. GPT-3, published in 2020, was developed to be capable of performing most natural language processing tasks. Two years later, a model with less harmful output was published called InstructGPT. It is a fine-tuned version of GPT-3 using Reinforcement Learning from Human Feedback (RLHF). In November of the same year, a chatbot named ChatGPT surprised people all over the world. It was improved on the basis of InstructGPT, having a new ability of multi-round conversation.

From GPT-1 to ChatGPT, the capabilities of the model are gradually expanded. ChatGPT now can effectively complete much work, such as text summarization, story generation, error diagnostics, paper writing, knowledge questioning and role playing, etc. Besides, ChatGPT is a multilingual model that can not only understand low-resource languages but also computer languages. In addition, ChatGPT has almost all fields of knowledge like healthcare, finance, biology, business, mathematics, and so on. Some people have high hopes for ChatGPT to promote social development for its outstanding abilities of text comprehension, text processing, and text generation. However, as ChatGPT was used by more and more people, its emerging drawbacks and limitations make people nervous about the negative effects that ChatGPT may bring. Particularly, the main limitations demonstrated by ChatGPT and other LLMs are as follow:

- **Hallucination**: ChatGPT generated text that looks semantically or grammatically correct but actually unfaithful and meaningless.
- **Originality**: The sentence or main idea of the text generated by ChatGPT is a copy or combination of the training data.
- **Toxicity**: ChatGPT may produce harmful content which contain biased or discriminatory, or speech that is aggressive, insulting or misleading.
- **Privacy**: ChatGPT is trained on large-scale datasets and interacts with countless messages, which may lead to some privacy and security risks.
- **Sustainability**: The training and maintenance costs of ChatGPT are high, including but not limited to cost in money, environmental and manpower, etc.

Scholars are generally focused on the above limitations and the ethical issues as well as negative effects the limitations bring. For example, the hallucination of ChatGPT can produce misleading erroneous text, leading to the risk of spreading misinformation. Also, non-original text violates copyright and the right to know, as well as leading to discussions about ChatGPT's attribution. The biased and discriminatory text generated by ChatGPT has an impact on social fairness. ChatGPT also have problems like data privacy and sustainable development, having an impact on social development.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

Contributions: We explore the reasons for five limitations the ChatGPT has. We present the ethical issues and social impacts ChatGPT may bring. We give some suggestions for the management and development of ChatGPT. We show the promising future research directions of ChatGPT.

Organization: The rest of this article is organized as follows. In Section 2, we give an introduction of the key technology used in ChatGPT. In Section 3, we firstly point out the possible reasons for five limitations listed above and then conclude the corresponding ethical issues based on its application scenarios. Recommendations to mitigating the impacts are provided in Section 4. We present several promising directions of the ChatGPT in Section 5. Finally, we conclude this paper in Section 6.

2. KEY COMPONENTS OF CHATGPT

In reality, OpenAI hasn't extensively disclosed the technical intricacies of ChatGPT in a dedicated paper. Alternatively, a succinct overview can be accessed via their official website. As it describes, ChatGPT is a sibling model to InstructGPT, belong to GPT-3.5 series models, which is trained to follow an instruction in a prompt and provide a detailed response. To figure out the implementation of ChatGPT, conducted a detailed review of GPT series models including GPT-1 [1], GPT-2 [2], GPT-3 [3], InstructGPT [4], GPT3.5, and ChatGPT①. This chapter presents the technical implementation of ChatGPT in four parts: model's architecture, training datasets, model training methods, the evolution and improvements of GPT series models.

2.1 Generative Pre-Trained Transformer

GPT (Generative Pre-trained Transformer) is a self-regressive generative pre-trained model developed exclusively using the decoder within Transformer. The Transformer architecture [5] demonstrates exceptional capability in capturing extensive textual dependencies.

GPT, relying exclusively on the Transformer's decoder, employs a multi-layer stacked architecture of transformer decoders for text generation. In contrast to the original transformer decoder, the structure utilized by GPT omits the Encoder-Decoder Attention, retaining solely the Multi-Head Attention layer and the Feedforward layer. In GPT-2, the structure underwent slight modifications: Layer normalization was moved to the input of each sub-block, and an extra layer of normalization was added after the final self-attention block. Besides, the initialization of residual layer weights was also scaled down.

Later iterations of GPT models have not explicitly described architectural alterations in published papers. Figure 1 provides reference to the foundational architecture of GPT series models as well as the original transformer decoder. Various GPT models differ in the number of transformer layers, multi-head attention, input vector dimensions, etc. Table 1 shows the comparisons of different versions of GPT models.

① https://openai.com/blog/chatgpt

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

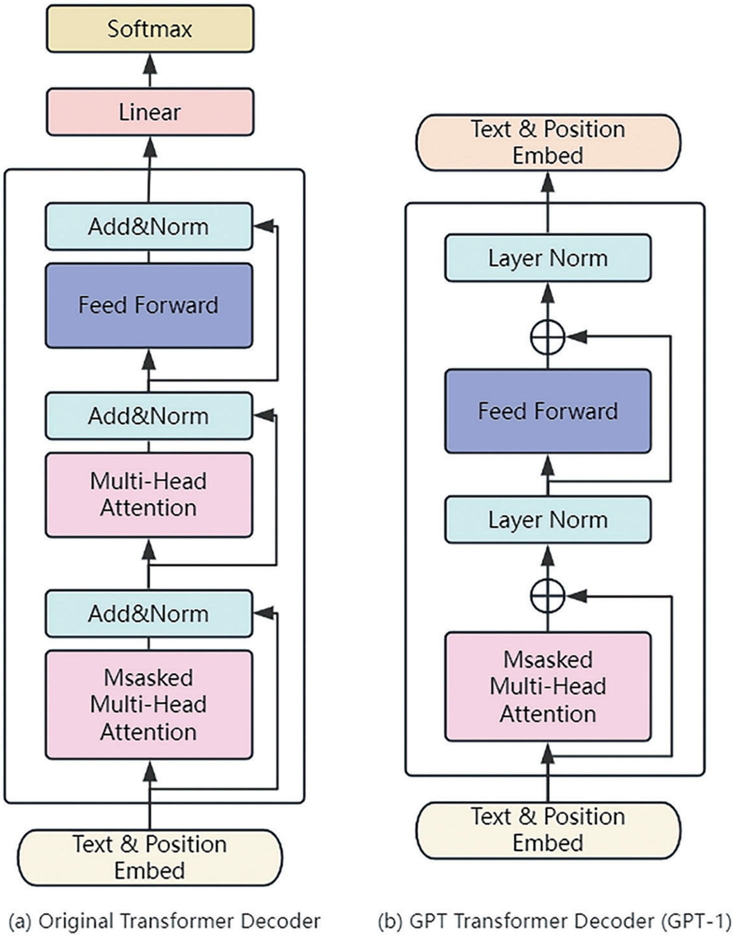

Figure 1. the foundational architecture of GPT series models.

Table 1. the comparison of different versions of GPT models.

Year	Model Name	Trained Tokens	Layers	Parameters	Multi-Head Attention	Training Algorithm
2018	GPT-1	512	12	117M	12	Pretrained+SFT
2019	GPT-2	1024	12~48	1.5B	-	Pretrained
2020	GPT-3	2048	12~96	175B	64~128	Pretrained
2022	InstructGPT	2K	12~96	175B	64~128	Pretrained+RLHF

2.2 Training Datasets

A notable aspect of ChatGPT is its utilization of vast datasets for training. To address the limitations posed by insufficient labeled training data, OpenAI sought to augment the training datasets significantly, enabling the model to develop a more comprehensive understanding of language itself.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

2.2.1 Pre-trained Datasets

(1) BooksCorpus

BooksCorpus [6] comprises over 7,000 unique unpublished books spanning various genres, including adventure, fantasy, and romance. Texts within the datasets typically exhibit a long length, making them conducive for model training aimed at comprehending extensive dependencies. GPT-1 was pretrained using the datasets.

(2) WebText

WebText [2], an innovative web scraping datasets focusing on document quality, consists of the textual subset from 45 million links, amounting to over 8 million documents and a cumulative text size of 40 GB. The datasets exclusively gathered outbound links from Reddit, a social media platform, that received at least 3 karma and underwent human filtration. GPT-2 was pretrained using the datasets.

(3) Collective datasets

The Collective datasets was derived from the original CommonCrawl (CC)[②]. To ensure high quality, this version of CC was created by comparing similarity to various high-quality reference corpora and conducting fuzzy deduplication at the document level. Additionally, for enriching CommonCrawl and enhance its diversity, high-quality reference corpora such as WebText [2], two internet-based book corpora (Books1 and Books2), and the English-language Wikipedia were incorporated into the training amalgamation.

The datasets encompass documents published between 2016 and 2019, totaling 570GB, roughly equating to 400 billion byte-pair-encoded tokens. Figure 2 illustrates the composition of this datasets. In fact, GPT models after GPT-3 were fine-tuning using different datasets from GPT-3. Therefore, these datasets can be seen as ChatGPT's pre-trained datasets.

2.2.2 Fine-Tuning Datasets

In the GPT series models, only GPT-2 and GPT-3 were not explicitly fine-tuned. GPT-1 conducted Supervised Fine-tuning (SFT) across various downstream tasks. InstructGPT and subsequent GPT models employed fine-tuning methods including Instruction Fine-Tuning (IFT), Supervised Fine-Tuning (SFT), and Reinforcement Learning from Human Feedback (RLHF). The fine-tuning datasets required for especially manual collection and construction.

(1) Supervised Fine-tuning (SFT)

GPT-1 utilized datasets related to natural language tasks like Classification, Entailment, Similarity and Multiple choice, but did not specify the datasets used.

② https://commoncrawl.org/the-data/

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

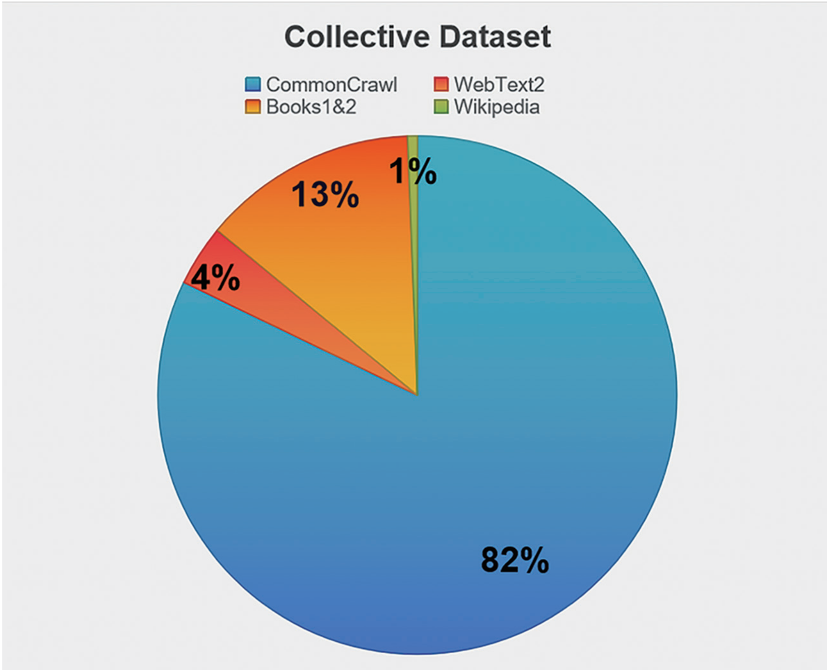

Figure 2. ChatGPT's pre-trained datasets.

(2) Reinforcement Learning from Human Feedback (RLHF)

Three kinds of Prompts: InstructGPT use Reinforcement Learning from Human Feedback (RLHF) to fine-tuning model. An initial source of instruction-like prompts is necessary to bootstrap the InstructGPT fine-tuning process. As these kinds of prompts weren't often submitted to the regular GPT-3 models on the API, labelers were asked to write the prompts.

- Plain: an arbitrary task.
- Few-shot: an instruction, and multiple query/response pairs for the instruction.
- User-based: use-cases stated in waitlist applications to the OpenAI API. Labelers should come up with prompts corresponding to these use cases.

The Prompts datasets consist of 10 tasks, such as generation, OpenQA, brainstorming, etc. Prompts containing personally identifiable information will be excluded.

Labeler: OpenAI hired a team of about 40 constructors on Upwork and through ScaleAI to work well on labelling.

SFT (Supervised Fine-Tuning) datasets: Prompts datasets with labeler demonstrations used to train Supervised Fine-Tuning (SFT) models.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

RM (Reward Model) datasets: Collect for different outputs of SFT model based on the inputs from the Prompts datasets. And the output should be ranked by labelers. The datasets are used to train reward model (RM).

PPO (Proximal policy optimization) datasets: Prompts datasets without any human labels, which are used as inputs for RLHF fine-tuning.

(3) Dialogue datasets for Reinforcement Learning from Human Feedback

It's widely recognized that ChatGPT excels in multi-turn conversations, owing in part to the specially curated dialogue data collected by OpenAI. ChatGPT uses the same fine-tuning method RLHF as InstructGPT, but slightly different in datasets.

About SFT (Super Fine-Tuning) datasets for ChatGPT, OpenAI let human AI trainers provided conversations in which they played both sides—the user and an AI assistant. And then this new dialogue datasets were mixed in the InstructGPT datasets. The whole datasets were transformed into a dialogue format to fine-tuning the base model using supervised learning.

As for RM (Reward Model) datasets for training reward model. OpenAI collected comparison data, which consisted of two or more model responses ranked by quality. OpenAI gathered conversations between AI trainers and the chatbot. Then, multiple alternative completions were sampled from randomly selected chatbot outputs, and AI trainers were tasked with ranking them.

2.3 Training Methods

The GPT series models universally employ a two-stage training method: pre-trained (unsupervised) and fine-tuning (supervised). In practice, both GPT-2 and GPT-3 were not fine-tuning. However, subsequent researches have demonstrated the pivotal role of fine-tuning in enhancing model performance. Since then, GPT models have maintained the fine-tuning stage and some new tuning methods were introduced like Instruction Fine-Tuning(IFT) [7]、Chain-of-thought (COT) [8].

2.3.1 Unsupervised Pre-Training

During pre-training, GPT learns from extensive unannotated text data to understand words relationships and their context, acquiring general language patterns. The choice of optimization objective function affects the pre-trained model's applicability to downstream tasks. GPT employs a standard unsupervised language modeling task, predicting the next word based on the context of preceding words. Unlike BERT [9], which uses a masked language model requiring the prediction of masked words in context, GPT's 'predict the future' task is more demanding, fostering models with enhanced potential for natural language understanding (NLU).

During the pre-training phase, GPT undergoes training on a vast corpus of unannotated text data, aiming to comprehend the relationships between words and their context within the training data, thereby

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

acquiring general features and patterns of natural language. The selection of the optimization objective function impacts the pre-trained model's generalizability to downstream tasks. GPT employs a standard language modeling task for unsupervised training, predicting the next word based on preceding words. In contrast to BERT [9], which utilizes a masked language model—predicting the masked word within the context, GPT's 'predict the future' training task is more challenging, thereby fostering models with greater potential for natural language understanding (NLU).

Furthermore, both the scale and quality of the training data significantly influence the efficacy of unsupervised pre-training. Throughout the iterations of the GPT series models, the pre-training corpora have substantially expanded while maintaining quality. This expansion has contributed positively to the model's performance in downstream tasks.

2.3.2 Fine-Tuning

(1) Supervised Fine-Tuning (SFT)

In this process, pre-trained model upgrades its parameters by training it on a task-specific annotated datasets, enabling it to better adapt to the specifics of a particular task or domain.

In the Supervised Fine-Tuning phase of GPT-1, datasets related to semantic similarity, classification, question answering and commonsense reasoning was used to fine-tuning the model. For fitting the pre-trained model training on contiguous sequences of text, task-specific input transformations should be made. Simply put, it involves converting structured inputs into an ordered sequence that a pre-trained model can process. Taking textual entailment as an example, the premise and hypothesis token sequences are concatenated with a delimiter token in between. As for Similarity, altering the input sequence to encompass both potential sentence orderings, and delineated sentences by a delimiter. This can help mitigate the influence of sentence order during training. Figure 3 illustrate the input transformations.

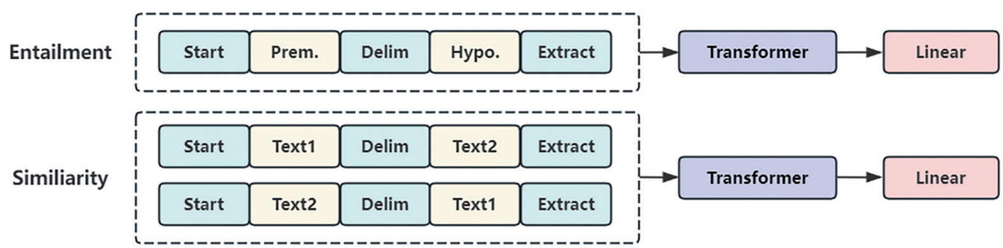

Figure 3. the input transformations.

(2) Reinforcement Learning from Human Feedback (RLHF)

To ensure GPT outputs more helpful, secure and user-instructive information, InstructGPT was fine-tuning from GPT-3 using RLHF. ChatGPT also used RLHF fine-tuning methods but differ in the training

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

datasets that we mentioned in 2.2.2. Below, take the RLHF method employed in InstructGPT as example, we introduce the three steps involved in the fine-tuning process and an illustration in Figure 4.

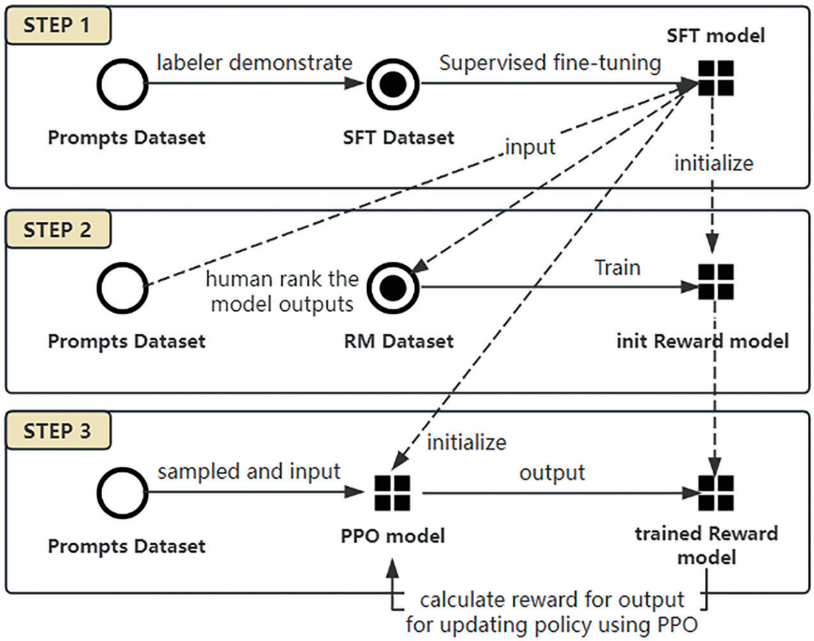

Figure 4. the RLHF method employed in InstructGPT.

Step1: Using SFT datasets to perform Supervised Fine-Tuning on GPT-3

Adopt the SFT method mentioned before to fine-tuning the GPT-3 model using the SFT datasets, and the trained GPT-3 is referred to as the SFT model.

Step 2: Training Reward Model

For reduced manual intervention and increased efficiency in subsequent reinforcement learning stages, the initial training involves a reward model (RM) that learns human response preferences using RM datasets. The RM model utilizes the SFT model with its last unembedding layer removed, initializing model's parameters from SFT datasets and training by RM datasets. The final trained model call Reward Model.

Step 3: Reinforcement learning

This phase aims to refine the model's strategy through the utilization of the Proximal Policy Optimization (PPO) algorithm [10].

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

Initially, the SFT model initializes a PPO model. Subsequently, the PPO model generates responses to the queries in PPO datasets which excluded human annotations.

Next, the trained reward model (RM) rates and ranks the predictions made by the PPO model, assessing if the model's results align with human preferences.

Finally, based on the reward rank, PPO model's generative policy can be optimized by PPO algorithm. The ranking and optimization process iterate until an optimal strategy is achieved.

2.3.3 In-context Learning

GPT-3 [3] does not utilize conventional fine-tuning methods for updating model parameters, instead employing Implicit fine-tuning known as In-context Learning. GPT-3 was experimented with in zero-shot, one-shot, and few-shot scenarios. In these context, zero-shot involves directly inputting the task description into GPT-3, one-shot means feeding the task description along with a single sample into the model, and few-shot entails an increase in the number of sample instances.

With both a large model parameter size and extensive training data, GPT-3 demonstrates exceptional performance in n-shot scenarios, surpassing state-of-the-art models. This proved its robust learning capacity during the pretraining phase.

2.3.4 Instruction Fine-Tuning

Instruction Fine-Tuning (IFT) [8] is a new fine-tuning paradigm introduced by the Google team in 2021. IFT data typically consist of a collection of manually crafted instructions and instruction instances, comprising three primary components: instructions, inputs and outputs. The IFT process involves, when faced with a given task A, initially fine-tuning the model on several diverse tasks unrelated to task A. This fine-tuning process involves concatenating task instructions with data and feeding them into the model. Subsequently, the instructions for task A are directly provided for inference.

FLAN model trained using IFT outperformed GPT-3 on most datasets, even with smaller model sizes. Experiments demonstrated that the more tasks involved in instruction tuning, the better the model performs. Instruction tuning shows positive impact on cross-task generalization. Moreover, task-specific training aids in enhancing the efficacy of general language models.

2.3.5 Chain-of-Thought

In 2022, Google introduced a new prompting mechanism named Chain of Thought (CoT) to mitigate the deficiency in mathematical reasoning abilities observed in large language models.

Essentially, this approach transfers the human thought process into the model's reasoning, guiding it step by step towards satisfactory answers. For instance, by incorporating mathematical problem-solving steps into the model's input instead of solely providing questions and answers,

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

the model's outputs significantly improve when prompted with reasoning inputs compared to those without.

Experiments demonstrate that after fine-tuning with CoT, LaMDA [11], GPT and PaLM [12] exhibit noticeable enhancements to varying degrees, surpassing even the results of optimally supervised fine-tuned models. Furthermore, regardless of whether it's a few-shot or zero-shot scenario, the integration of CoT technology enables these models to address certain mathematical reasoning problems that were previously unanswerable.

2.4 The Evolution of ChatGPT

Before the advent of GPT, most NLP models relied on annotated data for specific task completion. The model's performance was closely linked to the size and quality of annotated datasets. However, annotated datasets incurred high costs, and models trained on specific datasets can't generalize to other tasks. To address these limitations, OpenAI's GPT-1 emerged, employing a strategy of extensive corpus pre-trained followed by fine-tuning on smaller datasets. It showcased impressive performance in zero-shot scenarios, marking the initial stride in the development of universal text generation models.

To enhance the pre-trained model's comprehension capability, OpenAI began to augment the training data, creating larger and higher-quality pretraining datasets. They also scaled up the model size, increasing GPT-2's parameters by tenfold compared to GPT-1. In GPT-2, there was a departure from supervised fine-tuning, instead evaluating the model's performance directly in zero-shot scenarios. Evaluations across various downstream task datasets showed improved accuracy in recognizing long-distance relationships and predicting sentences. In specific tasks, it even surpassed supervised models, heralding the dawn of prompt tuning.

GPT-3 introduced the novel concept of prompt tuning. With an impressive 175 billion parameters, GPT-3 showcased remarkable natural language understanding, reasoning, extensive text generation abilities and adeptness across zero-shot, single-shot, and few-shot scenarios via In-context Learning. This laid a robust foundation for the groundbreaking ChatGPT that followed.

Amidst the utilization of large models like GPT-3, the emergence of biases, discrimination, and inaccuracies in the model's outputs became glaringly apparent. To address this significant concern, OpenAI introduced reinforcement learning from human feedback (RLHF). Engaging human feedback in supervising the model's training aimed to steer content generation closer to human preferences, thereby mitigating the generation of toxic texts and improving output accuracy. Although complete eradication remained unachievable, the utilization of RLHF in training InstructGPT significantly decreased the occurrence of generating harmful texts.

As the performance of generative large language models stabilized, OpenAI endeavored to introduce these models to the public sphere, aiming to integrate GPT into everyday life to augment human productivity. By augmenting code training datasets and employing diverse fine-tuning methods, models

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

like CodeX and ChatGPT were developed. Figure 5 depicts the evolution of ChatGPT from GPT-1 to its current state, highlighting the advancements across various models.

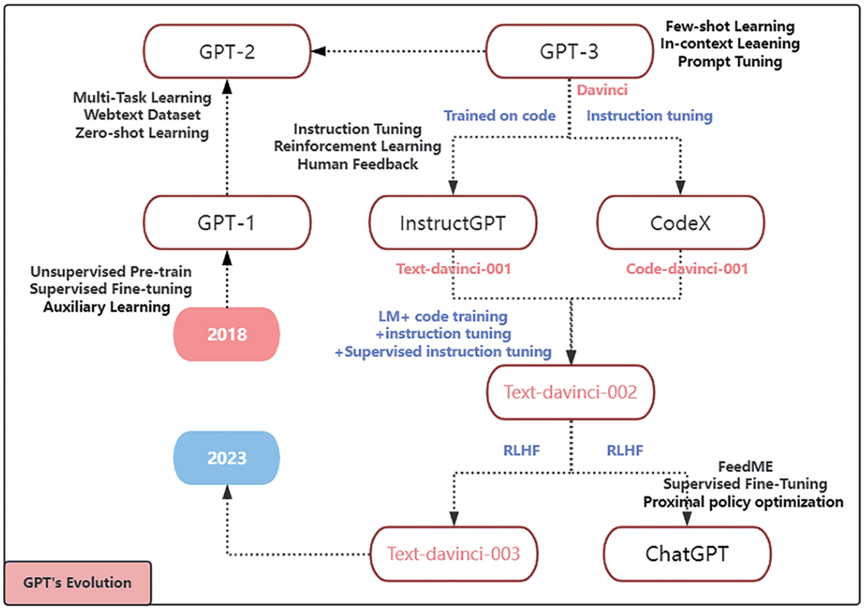

Figure 5. Evolution of ChatGPT.

3. LIMITATIONS AND ISSUES

3.1 Hallucination

Since released in November 2022, ChatGPT has gained over a one million users within five days and over a billion daily active users within two months, and its breakthrough performance in AI content generation has garnered widespread attention and discussion. Many researchers have tried to apply ChatGPT's powerful text generation capabilities to their respective domains, such as medical report writing [13], business writing [14], financial analysis [15], journalism and media education [16], etc. To some extent, ChatGPT performed well on those work, but incorrect and nonsensical texts were still found in the generated result [17]. Remarkably, dissemination of such harmful information may have adverse effects.

The phenomenon of ChatGPT generating text that seems semantically or grammatically reasonable but actually incorrect or meaningless is called "hallucination", which is a common problem in all LLM (large language models). Hallucination can be divided into two types [18]: Intrinsic Hallucinations, which refer to situations where output text contradicts input text, and Extrinsic Hallucinations, which refer to situations where output content includes knowledge that does not exist in or derive from input. The occurrence of hallucination is attributed to the quality of ChatGPT's training data, training inference algorithms and

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

prompt types [18] [19] [20]. The impressive conversational capability of ChatGPT was accredited to huge size of training data (45 TB of text data), from which the model acquires linguistic knowledge (syntactic and semantics) and world knowledge (fact and common sense) [21]. The generated text is determined by both the input source text and the prior knowledge in the language model. Assuming user's input has no factual error, then the culprit of hallucinatory output is the incorrect knowledge learned by model itself. Since the huge size of the data cannot be verified manually one by one, it is inevitable that training datasets contains misinformation. ChatGPT, which has no discriminative ability, will undifferentiated learn and store the wrong parameters. In addition, the training data of ChatGPT are all before December 31, 2021, so the model can't track the latest and accurate information. Besides, added to its "random parrot" feature [22], ChatGPT can easily generate false and outdated information and thus mislead users. Furthermore, the incompleteness of a domain dataset can cause the model learn the wrong knowledge structure and thus output incorrect answers.

In the training and inference process of ChatGPT, incorrect decoding is another reason causes model generated Hallucinatory text. the GPT-x models is a language model based on autoregressive decoder, and the randomness brought by decoding strategy naturally gives a certain probability of generating incorrect texts, such as top-p decoding, which makes model more randomness in exchange of improvement in content diversity [21].

Second, the exposure bias problem [23] arising from the inconsistency between training target and inferred target is also a major cause of Inaccurate text. The decoder is usually trained with maximum likelihood estimation in training process, and uses ground-truth as input for subsequent token prediction. However, in reference time, the decoder predicts the next token according to the historical sequence. The difference between ground-truth and historical sequence will give rise to error accumulation.

ChatGPT is essentially a probabilistic statistical model for generating texts that satisfy the statistical consistency with the hints, so the output may sometimes look logically correct but completely deviated from the facts in human's view. It was found in [24] that models pretrained with large datasets prioritize the parametric knowledge before input texts when producing results. This kind of parametric knowledge bias was the reason for Extrinsic Hallucinations.

ChatGPT's generated response also depends on users' prompts types. The expression form, syntax, and information veracity of prompts are all affect the quality and accuracy of output. A dataset of questions that humans are prone to answer incorrectly was constructed [25] as a benchmark for measuring the accuracy of generated content, and experiments showed that models such as GPT-3 achieved 50% accuracy, which is much smaller than human performance. Another investigation [26] used two types of prompts respectively containing correct and incorrect information as input to test the variability of answers, and the results showed that the accuracy dropped by nearly 30% when added wrong information into the prompts.

As the extension of application scope, the possible ethical issue brought by ChatGPT's hallucination should not be neglected. On the one hand, the dissemination and circulation of hallucinated texts will damage the communication ecosystem, especially in academic circles, and also hinder further

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

development and application of ChatGPT in some industries with low fault tolerance. On the other hand, the hallucination was found to be able to reproduce private information in training data, which poses the risk of privacy violation. In addition, the hallucination of ChatGPT may be abused for illegal purposes, affecting social security and order.

3.1.1 Disruption of Dissemination Ecology

ChatGPT may produce results that appear correct but are logically incorrect, and the dissemination of such texts in fields that require intellectual rigor, such as medicine and academia, can mislead learners with unimaginable consequences. In recent years, the number of published papers has grown rapidly, and there is no shortage of low-quality papers. ChatGPT makes this situation worse by spreading misinformation with its hallucinations that may distort scientific facts without being easily detected. However, the current review process does not yet address this problem [27]. Jürgen Wittmann [28] used ChatGPT to write an immunology review article and found its general wording convincing but many incorrect statements and citations in factual details and references, thus the authors were negative about such AI in aiding academic writing. In medical applications, if ChatGPT generates drug instructions that are hallucinatory, it may trigger life-threatening events for patients.

ChatGPT is also not yet complete in terms of code generation. First, it has a limited scope of application because its training data is biased towards programming languages such as Python, C++, and Java, and may not include all programming languages. Second, the code format needs to be manually optimized. Finally, there is no guarantee of the accuracy of the generated code, as it relies heavily on natural language input, and the understanding of the input may contain errors and ambiguities [29].

3.1.2 Privacy Violation

The hallucination of ChatGPT may lead to potential privacy violations. Carlini et al. [30] found experimentally that the GPT-2 model can inadvertently generate sensitive personal information in the training corpus, such as email addresses, phone/fax numbers, and addresses, among others. This is the second manifestation of ChatGPT hallucination, that is, generating detailed information that is not present in the source input. In addition, models not trained with privacy-preserving algorithms are vulnerable to similar privacy inference attacks [31], and some advanced LMs such as GPT-2 and GPT-3 do not use these privacy-preserving techniques and face the risk of privacy leakage due to the current limitations of LMs by training costs.

3.1.3 Malicious Abuse

ChatGPT can be used for malicious purposes, such as spreading misinformation and impersonating identities. Besides, sophisticated language models such as ChatGPT can be misused to create spam, fake news, deeply fake content, or engage in cyberbullying [32].

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

Once false information is published on the Internet, it will be widely disseminated and form a hot topic, and the false information will muddy the public opinion field and arouse social concern, which will control the direction of public opinion and cause incalculable impact. The spread of a large amount of false information makes people doubt the authenticity of the news, and over time people will think that the media reports are not credible, resulting in a decrease in the social trust of the news media industry. The combination of disinformation and social media can intensify social polarization and easily lead to large-scale social unrest, riots, and even ethnic confrontations, which undermine social security and affect social harmony and stability. Disinformation that is difficult to identify can easily be used by unscrupulous groups as a tool to discredit domestic political parties, incite violence, and provoke internal conflicts in society. Information involving ethnic conflicts and racial rivalries can also intensify internal conflicts and racial clashes, trigger social panic and unrest, and endanger national security. Furthermore, false information may provoke mistrust between countries and even trigger border conflicts, jeopardizing territorial security. ChatGPT's ability to generate human-like text also increases the risk of identity impersonation scams [33].

3.2 Originality

ChatGPT's powerful text generation capability makes it a hot tool for creative writing. Through detailed query to ChatGPT, it seems that we can get brand new texts in various types of writing, like marketing copy, review, script, novel, poem, song or even an academic paper with reference. Nonetheless, ChatGPT's creativity dealt a blow to previous idea of that artificial intelligence won't replace humans to do the creative work and the fear that ChatGPT will replace creative workers has begun to spread in society.

Many scholars have evaluated the originality and innovation of generated content. Ventayen [34] used ChatGPT to write papers on specific topics and then used Quillbot [35] to check the text similarity. Among the four texts provided by ChatGPT, the similarity ranged from 0% to 10%, and one of the most similar contents was originated from about 50 texts, where the longest similar token reached 46 words and the highest similarity reached 87%, which exceeded the standard required for academic originality. Barakat et al. [36] used plagiarism detection tools SmallSEOTools and Turnitin to detect assignment answers provided by ChatGPT, and found out that the generated answers had statistically significant text sources with a plagiarism rate of about 50%. Khalil et al [44] also used Turnitin and iThenticate for ChatGPT's originality detection experiments, and the results showed that 25% of the texts generated from 50 different topics had an unacceptable level of similarity, and the highest text similarity exceeded 40%.

In summary, the text generated by ChatGPT on different topics has varying levels of originality, but the innovation is generally low since the text is basically a combination of existing ideas. In fact, ChatGPT is destined to unable to generate innovative results like human do. Because its model output is limited by the scope covered by the training data [37]. Furthermore, strictly speaking, the originality exhibited by ChatGPT in novels, dramas, papers and other texts is actually the manifestation of the internal potential rules of relevant types of training data, rather than the generation of new rules. By the way, once the model learned a wrong rule, it is likely to output a hallucinated text that appears reasonable but is actually absurd.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

The originality of AI-generated texts such as ChatGPT has led to discussions on copyright, plagiarism, and thinking influences.

3.2.1 Plagiarism

ChatGPT has been widely used in academia, such as writing a qualified college admission essay within ten minutes③, composing literature reviews [38], and finishing a coursework that met the requirements for Wharton School's MBA degrees [39], etc. The writing ability of ChatGPT has raised concerns among academic and educational workers about academic plagiarism and other academic misconduct.

Plagiarism [40] is considered as the act of Intentionally or unintentionally stealing all or part of someone else's work and attempting to cover it up and claim it as one's own independent creative work. Studies have shown [41] that as artificial intelligence rapidly develops, academic misconduct has tripled in 20-21 from the previous year. The revolutionary ability of ChatGPT in solving almost any problem makes it easier for students to engage in plagiarism and disrupt the integrity of the academic community. If students don't think deeply about the generated text and unconsciously use the low-originality one, they are prone to be involved in academic plagiarism and disrupt the integrity of the academic community.

Although OpenAI has proposed a solution to plagiarism by adding watermark to ChatGPT's responses [42], this watermark can only be accessed by OpenAI's development team, and this watermark is also easily removed. Various of plagiarism detection systems were built for coping with this problem. However, some scholars worried about those systems can't not completely distinguish AI text or misjudge the original one, which may bring another kind of inequity. Gao et al. [43] used a RoBERTs-based AI generated detector and artificial judgment to test summaries generated by ChatGPT. Both methods can recognize around 70% of AI-generated abstracts. GPTZero, an AI-generated detector invented by Svrluga [44], detects text content based on randomness and burstiness, but suffers from the same problem of misclassification.

The ease of use of ChatGPT makes plagiarism more convenient, but as it stands now, the performance of plagiarism detection software still has room for improvement and cannot effectively prevent academic plagiarism. Currently, many schools can only manage this situation by explicitly banning the use of ChatGPT, such as schools in New York City, Seattle, Queensland, Australia, and Tasmania④.

3.2.2 Technical Dependency

Compared to previous chatbot, ChatGPT has more powerful contextual understanding, faster response speed, better response quality and wider types of problem-solving ability, which undoubtedly helps to improve productivity. However, some scholars [45] argued that long-term use of ChatGPT for problem solving is likely to hinder the development of critical thinking and the enhancement of problem-solving skills and creativity. ChatGPT can handle problems for almost all disciplines in a short time, whether for

③ https://www.forbes.com/sites/emmawhitford/2022/12/09/a-computer-can-now-write-your-college-essay---maybe-better-than-you-can/?sh=1d427781dd39
④ https://www.dailymail.co.uk/news/article-11688905/UNSW-student-fails-exam-using-OpenAIs-ChatGPT-write-essay.html

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

academic writing, homework quizzes, code writing, or data analysis. If students lack self-discipline and completely depend on ChatGPT to finish homework, in a long run, their corresponding academic research skills, writing skills and even problem-solving skills will not be sufficiently exercised and technical dependency will be formed. Even worse, it may be a blow to traditional education and new educational styles need to be proposed to adapt to the development of this new technology. Sandra Wachter [46] of the Oxford Internet Institute also believes that uncontrolled use of ChatGPT will take over student thinking.

3.2.3 Copyright Infringement and Attribution of Authorship

Regarding the copyright issue of ChatGPT, there are two problems that have sparked lively discussions among scholars. One is the risk of intellectual property infringement, and the other is the attribution of authorship.

ChatGPT's original generated content comes from a stitched-together combination of training data. Since OpenAI has not disclosed the source of the colossal size of datasets, some researchers reasonably suspect that some of the training data involved may be protected by copyright. If OpenAI has not reached agreements with these copyright owners, there may be a risk of copyright infringement. On January 13, 20231, Sarah Andersen et al filed a lawsuit against Stable Diffusion and Midjourney for this reason[⑤].

Another hot topic related to copyright is whether artificial intelligence such as ChatGPT should has authorship and copyright for their generated content. Research [47] has found that ChatGPT does provide significant assistance in generating high-quality papers. However, its contribution to the entire academic research is difficult to define because whether it is sufficient to be listed as a co-author. Nature [48] insisted that it does not accept any LLM tools as an attributed author because it cannot assume responsibility as an author of the paper.

The "ICMJE Recommendations for Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals" [49] published in May 2022 by the ICMJE suggests that authorship should meet the following four criteria: 1. Contribute significantly to the research ideas and design; 2. Make critical revisions of key information; 3. Approve the final version of paper to be published; 4. Be accountable for all aspects of the research work. Silva [50] analyzed whether ChatGPT could have co-authorship status based on the ICMJE recommended guidelines and the conclusion was that ChatGPT did not comply with any of these rules. In addition, in decided cases of training data infringement or cases involving AI as inventors of patents, courts have tended not to identify AI as authors or inventors [48].

3.3 Toxicity

As a chatbot, ChatGPT's abilities in semantic understanding and response accuracy surpass all previous chatbot systems. Many users have been eager to test ChatGPT on tough, sensitive, and taboo topics, to see if it can provide unprejudiced responses to questions that even humans cannot answer objectively. Not

⑤ https://www.vice.com/en/article/dy7b5y/artists-are-suing-over-stable-diffusion-stealing-their-work-for-ai-art

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

surprisingly, ChatGPT produces some toxic content with biases and discrimination. In this paper, the biased, discriminatory, and harmful content output by ChatGPT is collectively referred to as “toxicity”. The reason why ChatGPT exhibits toxicity are because of toxic training data and toxic guided prompts.

Similar to the hallucination, toxic content is rooted in biased and discriminatory training data and biased prompts datasets influenced by annotators’ unrighteous thoughts. If model learns with biased or offensive datasets, it is likely to generate toxic text. Apart from that, imbalanced class of training dataset can also result in toxic output. ChatGPT’s training data may mostly reflect the will of native English-speaking races, and the power structure that this skewed dataset represented will then be reflected in output [27], causing toxic text generation. In addition, the training method of reinforcement learning based on artificial feedback (RLHF) used by ChatGPT relies on human-supplied labeled prompts dataset, and the unbiased and truthfulness of the labeled data greatly affects the relevance, accuracy, and ethical correctness of ChatGPT’s response texts [51]. TIME once revealed in an article that OpenAI employed cheap labor from poor people in backward countries to carry out a large number of harmful data labeling tasks to support the ethical standardization work of ChatGPT and other AI systems. If the impartiality of the labeled data is adversely affected by subjective judgments of humans, there is no guarantee that the text generated by ChatGPT will be objective and impartial. More importantly, the bias caused by the training data is permanent [52].

The prompts accepted by ChatGPT also guide the model to produce toxic sentences. Since opening up to the public, ChatGPT has been subject to many adversarial attacks, which means deliberate manipulation of inputs to mislead the model’s predictions. Among ChatGPT’s diverse “user experiences,” many people have been keen to lure ChatGPT into generating unethical, false, malicious, and other offensive statements as a way to satirize ChatGPT’s text generation capabilities. Unfortunately, even if the developers set some basic ethical and moral rules and manually mark the taboo content as a way to regulate the information generated by ChatGPT, still can’t completely block the generation of toxic content. In addition, an interesting feature of ChatGPT is to perform role-playing to answer questions. Deshpande et al. [53] conducted an in-depth study on ChatGPT’s toxicity and found that different roles specified in the prompt would affect the degree of toxic content generated by ChatGPT, and it is easier to Induce ChatGPT to give toxic speech under role-playing mode.

ChatGPT-generated texts exhibit bias in four types [54]: cultural bias, linguistic bias, temporal bias and political bias. Cultural bias refers to the bias that the model exhibits towards specific gender, race, or social groups. Language bias refers to the different accuracy that model response to various language due to the lack of low-resource language in the training data. Temporal bias refers to the lagging bias of the model to time-sensitive events, trends, or updated knowledge due to a fixed time horizon of training data. Political bias refers to the cognitive bias related to political views or ideologies due to the lack of representative data in training data.

ChatGPT’s broad application scope make it easy to mislead users’ thinking and exacerbate social biases and discriminatory phenomena. It can also become a tool for illegal criminals to carry out information warfare.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

3.3.1 Intensification of Social Prejudice and Discrimination

ChatGPT's broad knowledge base allows it to perform well in a specific area with fine-tuning. Currently, researchers from various fields are actively exploring ChatGPT's potential applications in assisting development, such as healthcare, finance, academic research, business decision-making, and government policymaking. However, ChatGPT's toxicity poses a major challenge to its development because the biased and discriminatory ideas in models may lead to unfair treatment of relevant users. If biased models integrated into our daily lives, they will probably marginalize vulnerable groups and exacerbate social inequalities.

There are a few research on the toxicity of generated-AI models' output. Dahmen et al [55] analyzed the potential of ChatGPT to assist medical research and found bias in the generated abstracts of medical research papers, risking misleading readers. Ghosh et al. [56] found that ChatGPT exhibits gender bias and stereotyping in occupation for low-resource languages machine translation (Persian, Malay, Tagalog, Thai, and Turkish), such as replacing neutral pronouns with "he" or "she" in the context of a doctor or a nurse. Lucy et al. [57] also found that GPT-3 also exhibited gender bias in story creation. Religious bias can be found in LLM as well. Abid et al. [58] used a complementation experiment to find that Muslims are more likely to be described as "terrorists" by GPT-3 than Christians and Buddhists. Although OpenAI has set guardrails to mitigate the toxic behavior [59], it is still unable to completely avoid harmful text production. Such potential biases and discriminatory thoughts instilled in some teenagers who have not developed complete values are detrimental.

3.3.2 Political Manipulation

Texts are the core medium of politics. Political participants and theorists are highly concerned about the hidden politics driven by ideologies in texts [60]. It has been noted that ChatGPT trained using humans' own biases with texts containing malicious intent has an automated bias, and the widespread use of such AI systems may amplify hegemonic worldviews [22]. Potential political biases and biased ideologies in texts can subtly influence the nation. If ChatGPT is maliciously abused for political manipulation purposes, the consequences would be unimaginable. Scholars have pointed out that ChatGPT, which is trained on data that containing human biased and malicious thinking, has an inherent automation bias and the widespread use of such AI systems may amplify a hegemonic worldview.

Rozado et al. [61] conducted 15 different political orientation tests on ChatGPT, and 14 of them drew a conclusion that ChatGPT has leftist political bias. Rutinowski et al. [62] also investigated ChatGPT's political preferences using political questionnaire and standard criterion, and obtained the same result that ChatGPT belongs to the totalitarian left or liberal left quadrant. Except that, in personality tests, the most pronounced negative traits of ChatGPT appeared to be egoism and sadism to, both of which are below average (ranking 35% and 29.1%, respectively).

ChatGPT's usability, efficiency, and high performance make it a new generation of "communication tool". As the speed and breadth of information dissemination have greatly increased, people are able to access information faster and better, while having a risk of being poisoned by the malicious manipulation of ideas.

Therefore, ethical issues arising from ChatGPT's biases, discrimination, and other harmful features should not be ignored. Therefore, correcting the political biases in ChatGPT should be given serious attention.

3.4 Privacy

As the newest and most powerful generative AI language model in the LLM, ChatGPT's leap forward in multi-round conversational capabilities has been made possible by large-scale model training data. However, the privacy issues involved are one of the reasons why the commercial use of LLMs has not received sufficient attention [63]. Although the training data of ChatGPT has not been described in detail in relevant papers, the training data of its predecessor GPT series have been disclosed [3] [64] [65].

The types and sizes of training datasets for GPT-1, GPT-2 and GPT-3 models are shown in figure 6. Books1, used by GPT-1, is composed of unpublished and freely used books. All the books were collected from Smashwords which is an e-book website that describes itself as "the world's largest independent e-book distributor.". Books1 and Books2 used by GPT-3 are actually not detailed in the original work [61], but some related researchers speculated [66], that Books1 used by GPT-3 is the Gutenberg data from SPGC (Standardized Project Gutenberg Corpus) and Books2 may be a dataset composed from the Bibliotik e-book website resources. Common Crawl is an open data platform that has been crawling the information of Internet for years to build a large public dataset consisting of various web pages, including a broad range of text from different languages and domains. Moreover, GPT3.5 and InstructGPT [67] are based on GPT-3 fine-tuned using manually annotated datasets and user prompted datasets collected by the OpenAI API.

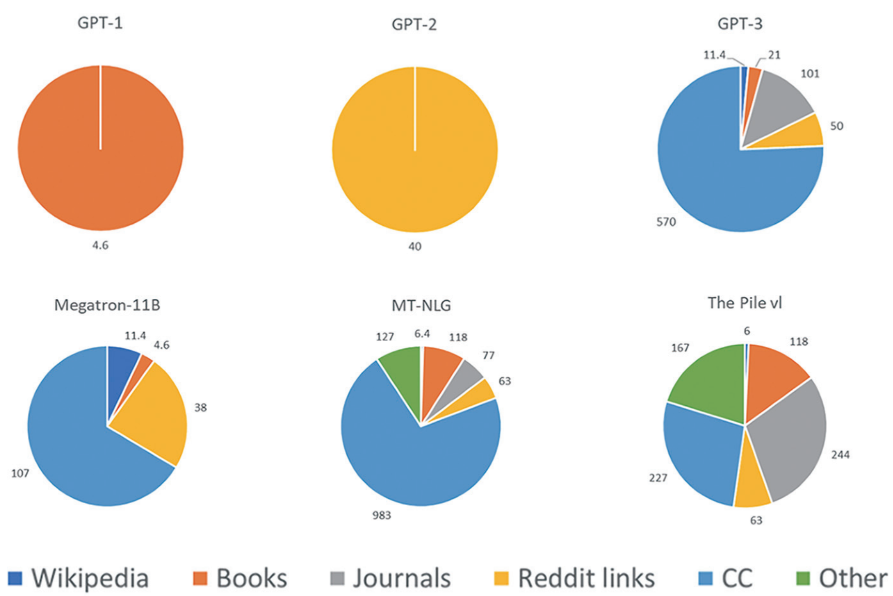

Figure 6. The figure shows the types of training data for six large language models, including Wikipedia, Books, Journals, Reddit links, Common Crawl, and Other. The size shown in the figure is in GB.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

The training data of artificial intelligence systems like ChatGPT is extremely large, and there are ethical issues such as invasion of privacy and data leakage.

3.4.1 Privacy Infringement

The training data of generative AI systems including ChatGPT are large in size, widely sourced, and rich in subject matter, which can easily create ethical issues of violating the privacy and right to know of data stakeholders. The sources of training data for generative AI models such as ChatGPT can be summarized as the following four: public datasets, crawling using data mining or crawling techniques, manual collection, and user data collected through APIs. Among them, manual collection and use of public datasets fully respect the scope of authorization of data subjects and are in line with the norms. But the use of crawling technology or interactive interfaces to acquire user data has the ethical pitfall that can't guarantee the right to know of authorized subjects. If there is personal sensitive data in the training data, it is undoubtedly an infringement on the privacy and right to know of the data related persons.

In fact, the training data used by ChatGPT and the method used to obtain the data are not fully disclosed. Alexander, a support member of European Data Protection Board (EDPB), has pointed out that if ChatGPT obtains its training data by crawling the Internet, it may not be legal. The above analysis of training data sources shows that more than half of the training data of similar large language models as ChatGPT are from web crawlers, which means that the training data may contain unauthorized data, and the risk of infringement is self-evident.

3.4.2 Data Breaches

ChatGPT's data interactions with hundreds of millions of users are subject to the security risk of data leakage. According to ChatGPT's privacy policy statement [68], users' personal information is collected for further research and development of new projects, and may be provided to third parties under special circumstances. Personal information includes, but is not limited to, personal account information, communication information, and personal interaction data with ChatGPT. If data used by ChatGPT for training purposes is hacked, leaked internally by the company, or data backups failure, the privacy and property rights of the individuals involved are vulnerable to infringement. Not only that, if the sensitive information provided by users to ChatGPT is used for subsequent research, whether user information will be further leaked in the form of model output is an aspect worthy of caution. Moreover, though the privacy policy promises that users have the right to delete personal information, it is not possible to investigate whether the information is completely deleted or used for further training of the model.

So far, many companies and countries have become wary of the data leakage and privacy security risks associated with AI tools like ChatGPT. On February 22, 2023, investment bank JPMorgan announced restrictions on employees' use of ChatGPT in the workplace. Microsoft and Amazon have announced a ban on company employees sharing to ChatGPT. Microsoft's internal engineers also warned against

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

sending sensitive data to OpenAI terminals, as OpenAI may use it for training future models. Then on March 31, the Italian Personal Data Protection Authority declared a ban on the use of ChatGPT citing ChatGPT's leakage of private user data and lack of an age verification system, and restricted its development company OpenAI from handling Italian user information. Subsequently, German regulators also announced a ban on ChatGPT, and European countries such as France, Ireland, and Spain ware beginning to consider stricter regulation of AI chatbots. On April 4, 2023, the Office of the Privacy Commissioner of Canada also launched an investigation into OpenAI, the developer of ChatGPT, regarding allegations that "OpenAI collected, used and disclosed personal information without consent"[⑥]. An initiative released by the China Payment Clearing Association on April 10 also pointed out that intelligent tools like ChatGPT have been exposed to cross-border data leakage and other risks. In addition, due to technical reasons, users using ChatGPT have briefly seen others' chat topics on private pages, which has also raised public concerns about privacy leaks[⑦].

3.5 Sustainability

Since ChatGPT has become a popular topic, most research has focused either on the future of the ChatGPT application process or on how ChatGPT can enable general artificial intelligence. Nevertheless, the sustainability of ChatGPT in terms of greenhouse gas and carbon emissions, which directly contribute to climate change, is rarely discussed. An in-depth study of the sustainability of ChatGPT development requires consideration of its training costs. The carbon emissions of the training model can be determined by the electric power consumption and other related consumption. Electricity consumption also assumes the use of hardware, while other production methods that consume as electricity, such as wind, solar, coal or nuclear [63].

3.5.1 The Training Process of LLMs

The training of LLMs used by ChatGPT is performed on large scale datasets and during the training process, the model is fed with a large amount of text to adjust the model weights. The training process is considered computationally intensive, which is the main reason for associating the training process with a carbon footprint. Most LLMs demand transformer architecture that are trained on large amounts of text data. Most LLMs demand transformer architecture that are trained on large amounts of text data. ChatGPT was created on top of GPT-3.5, which contains 175 billion parameters. The network is often trained several times until it produces satisfactory results. Although the user may only fine-tune the pre-trained network, it also requires multiple attempts to produce satisfactory results [69]. In recent years, several studies have analyzed the carbon emissions of LLMs. For example, Bannour et al. [70] conducted a carbon emissions analysis using six different tools. According to the study [71], AWS Canada (Central), Azure Canada (East), and GCP Europe (West6) produce the lowest carbon emissions. Canada uses hydroelectricity and Switzerland is based on the carbon neutrality initiative and therefore has a low carbon footprint. In contrast, Azure South Africa (West) and Azure South Africa (North)have higher carbon emissions because

⑥ https://www.cbc.ca/news/politics/privacy-commissioner-investigation-openai-chatgpt-1.6801296
⑦ ChatGPT bug raises privacy concerns (proactiveinvestors.co.uk)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

they use oil and coal as a source of electricity generation. Several studies have analyzed the carbon emissions of the GPT-3 and Meta's OPT training processes [72] [73] and found that the latter emits 75 metric tons, while the former emits 500 metric tons. However, this is only the training up to now, and as LLMs become more popular and develop, the carbon emission figures will rise more significantly.

3.5.2 Reasoning for Carbon Emissions from LLMs

In their recent paper, the artificial intelligence startup Hugging Faces has proposed an efficient method for counting carbon emissions. [74] This provides an opportunity not only for AI tech companies but also for governments, regulators and technology auditors to be able to assess the environmental impact of LLMs. The paper measured the carbon emissions of their own LLM BLOOM, the cost of electricity to manufacture the computer hardware, and the energy required to run the model. A recent paper by Patel and Ahmed [75] assumes that the number of active users of ChatGPT is 13 million and also assumes that each active user makes 15 queries per day. Based on the above information, multiplying 13 million by 15 requests will generate about 195 million requests per day. OpenAI does not provide GPT-3 power consumption, and while these figures are hypothetical and utilize information from existing studies, this hypothesis still provides a data reference that can be used as a basis for revising policies [63].

4. EXPLORING THE GOVERNANCE PATH OF GENERATIVE AI

With the rapid development and wide application of ChatGPT, one day it will be integrated into our life and work. The technical risks and ethical issues associated with it do not mean that the technology will immediately turn the world upside down. In order to ensure the safe, reliable and controllable development of ChatGPT, we need to be proactive and do our best to meet these challenges from various aspects. For example, through industry statutes and legal norms to effectively regulate and punish the application of new technologies. Adhere to the principle of human-centeredness and build a safe and reliable system. The related ethical issues brought by ChatGPT can also be solved through technological paths. In the following, we will specifically introduce the governance methods.

4.1 Practice

4.1.1 Regulation Practice

Following the release of the ChatGPT, AI legislation has accelerated around the world. In March 2023, the UK government released a policy paper, A pro-innovation approach to AI regulation, in which it sets out its AI governance principles and framework [76]. In addition, it focused on specific application scenarios in which AI is used and choosing to adopt proportionate, risk-based responses. The paper identified five principles, including safety, security and robustness, appropriate transparency and explain ability, fairness, accountability and governance, and contestability and redress. It further demonstrated the ability of regulators to adapt and apply the framework as needed to the requirements of different industries.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

In October 2022, the U.S. enacted the Blueprint for an Artificial Intelligence Bill of Rights, which set out a total of five principles for establishing safe and effective systems, avoiding algorithmic discrimination, focusing on data privacy, advocating for clear notice and explanation, and setting up alternatives and exit mechanisms [77]. In April 2023, with the quick progress of generative AI, the U.S. National Telecommunications and Information Administration issued the "AI Accountability Policy Request for Comment", publicly soliciting suggestions from relevant stakeholders on whether and how to regulate and account for tools such as generative AI [78]. Policies in both the UK and the US reflect an attitude that encourages innovation and promotes the development of AI. They focus on the regulation of the use of AI technology rather than the technology itself or the industry as a whole, and ensure that regulation is proportionate and adaptable [79].

As a global windsock in the development of digital governance regimes, Europe is attempting to take the initiative and lead in the development of global rules for AI governance. In 2019, the European Commission published a white paper on AI [80]. The white paper suggested that existing regimes are equally applicable to AI, that a horizontal extension approach will be taken, and that a risk-based approach will help to ensure that regulatory interventions are proportionally appropriate. In June 2023, the European Parliament adopted a draft Artificial Intelligence Act [81], which will now enter the final phase before regulation is initiated. The draft grouped AI systems according to their level of risk and categorized them for regulation. The more dangerous applications, the stricter the rules that apply. It includes four risk levels, namely unacceptable risk, high risk, limited risk and minimal risk. The act prohibits AI belonging to the unacceptable risk category, requires high-risk AI systems to complete market access and certification, and for limited risk requires the product to realize transparency and openness, while there are no specific constraints on the minimal risk act for the time being. The act has a number of merits, including a more granular categorization of AI risks, a clear indication of AI practices that need to be banned, and the addition of a European public database. However, the act is still deficient in the protection of fundamental rights, liability mechanisms, transparency obligations, AI definition and classification rationality [82].

China has also done some exploration in regulating AI as well. The most representative one is that, in September 2021, the Professional Committee on Governance of New Generation Artificial Intelligence issued the "Code of Management of New Generation Artificial Intelligence", which proposes six basic ethical requirements include "promoting human welfare, promoting fairness and justice, protecting privacy and security, ensuring control and trust, strengthening responsibility and enhancing ethical literacy", as well as 18 specific ethical requirements for specific activities such as management and research. The latest policy "draft for comments" proposes that generative AI should reflect core socialist values, avoid algorithmic bias and discrimination, respect equal competition, truthful content, and protect the rights of individuals [83]. In terms of AI regulation and governance, China has clarified the idea that development and safety in parallel, and innovation and ethics in parallel. While promoting the development and innovation of AI, it need guarantee the safety, reliability and controllability of technology applications. Effective governance of generative AI also requires the joint participation of all sectors of society. Utilizing the cooperative power of multi-party governance, we will build a safe and trustworthy AI applications [79].

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

4.1.2 Technology Practice

Countries are actively exploring technical and managerial safety controls to proactively address the potential risks of AI. These measures include interventions in training data, improvements in model architecture, censorship of model output content, and monitoring of user behavior. In terms of the internal mechanism of the model, OpenAI's GPT-4 model reduces harmful outputs by adding additional safety reward signals during the RLHF training phase. This approach generated better results, significantly increasing the difficulty of inducing the AI to produce malicious behaviors and harmful content, and improving the safety of the model. In addition, external mechanisms can be used to compensate for the model's own flaws and limitations. For example, the Azure OpenAI service launched by OpenAI can help us identify and filter various categories of prohibited content and realize the control of input and output data. The models can also be subjected to multiple probes, tests, and attacks to identify potential problems and solve them before they are released through Red Teaming [4]. In September 2023 OpenAI announced a brand-new project, the OpenAI Red Teaming Network. they will collaborate with experts around the globe to identify and address potential risks to models such as GPT-4. In June 2023, Google officially released the SAIF (Secure AI Framework), articulating its vision for creating secure AI. To support innovators in bringing new ideas to market, the UK government also funded the construction of sandboxes where companies can test how regulation can be applied to AI products and services.

4.2 Purpose of Generative AI Governance

4.2.1 Ethical AI

The design of all technologies should follow the concept of human-centeredness and insist on maintaining the power and dignity of human beings. Oppose the risks that technology abuse poses to people's privacy, copyright, liability, and more. Ensuring the privacy and security of user data is critical to preserving users' trust in technology. In addition, ethical considerations should be incorporated into the development process, such as developing guidelines for appropriate use and ensuring transparency in the deployment of the technology. Also do not set uniform standards for value judgments and avoid algorithmic bias against individuals or groups based on personal characteristics such as race, gender or religion. Artificial intelligence technology should have the ethical cornerstone of maximizing the benefits to all humans [83]. Informing the capabilities and limitations of AI and focusing on the guidance of users will lead to the development of AI technology in the right direction.

4.2.2 Trusty AI

With the rise of Artificial Intelligence, the trustworthiness of ChatGPT has become one of the most important social issues. Trust is a key aspect of how people choose and adopt an item. For an AI product like ChatGPT, gaining people's trust is even more important. There is a huge uncertainty in the development and deployment of AI, and this uncertainty leads to people's manifestations of caution, skepticism, and mistrust. People's distrust of AI is well-founded and manifests itself in many ways. Despite the great success of ChatGPT, it still suffered

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

failures in a number of areas. These failures could be the toxic text produced by the ChatGPT model that is biased and discriminatory, or the security risk of data leakage that comes with the use of ChatGPT. In fact, ChatGPT's success can likewise fuel fears and lead to mistrust. ChatGPT's excellence in certain areas can easily generate feelings of substitution and the fear of being reduced to AI inputs [84].

4.2.3 Responsible AI

Responsible AI focuses on adherence to ethical principles and human values, promotes fairness by reducing bias, and emphasizes transparent development and deployment. However, this responsibility includes not only ethical issues, but also needs to pay attention to legal, economic, and cultural issues to make AI technology beneficial to the whole society [85]. Especially in the direction of business decision-making, healthcare, financial risk control, and government, AI may give false predictions that can have a large impact. Responsible AI needs to take actions to serve the society widely. It is an intelligent system built on fundamental human principles and values, and the responsibility is to ensure that the results are beneficial to the majority, rather than being used as a profit-making tool by the minority. The ultimate goal of building AI technologies based on responsible principles is to avoid large negative impacts on human and societal well-being [85].

4.2.4 Secure AI

The security of AI involves not only security for the user, but also the security of the AI itself. In this era full of Internet applications, people touch many applications in their lives. The secure of the application system is crucial for data protection, privacy protection and corporate reputation. A secure system can protect our privacy and data from being leaked, while a system with security vulnerabilities may bring serious consequences such as property loss, legal liability, and reputation damage. Only a secure application system can gain people's favor and sustained development. Moreover, AI systems themselves face many challenges in the complex network environment. Moreover, AI systems themselves face many challenges in a complex network environment. For instance, diverse cyber attackers, continuously expanding attack scope, and so on. AI systems need to ensure protection from cyber-attacks and maintain system security and stability by accelerating threat detection and improving response speed.

4.2.5 Explainable AI

As AI becomes more advanced, the challenge for humans is to understand and trace how algorithms arrive at their results. Since ChatGPT is still a black-box model, it has not yet been possible to break down its intrinsic algorithmic logic. This leads to the fact that when ChatGPT makes and uses irrational decisions or generates biased speech, we simply do not have access to detailed explanations of its behavior, which can trigger mistrust. Humans are often reluctant to adopt techniques that cannot be directly explained and are not secure, and model explanation is an important bridge that connects algorithms to human cognition. When developing models, it would be useful to consider explain ability to fulfill a range of needs, such as being able to help improve the fairness of model decisions by detecting and correcting biases in the

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

training datasets, assisting with debugging and auditing in the management and development of model, and contributing to improved robustness [86].

4.3 Solution Path

The birth of ChatGPT is a change in global information dissemination and the rise of a new kind of intelligent communication beyond traditional communication methods [87]. Artificial intelligence technology is created to better serve human beings, there is no good or evil in technology, the key lies in how people use it. Large language models such as ChatGPT will eventually penetrate our work and life. To ensure a more long-term development of AI, we explore coping strategies from multiple perspectives and strive to eliminate potential risks and pitfalls.

4.3.1 Legal Regulation

For managing the use of ChatGPT technology, it is fundamental to establish ethical norms to delineate the boundaries of the application of the new technology. ChatGPT technology has numerous applications, and the benign impact of all applications cannot be denied because of the immature development of certain applications, nor can all applications be allowed to develop unregulated and unchecked. A reasonable legal regulation scheme can promote the enhancement of ChatGPT's orderliness and standardization. Meanwhile, the technology of ChatGPT is constantly and rapidly developing and its functions are becoming more and more powerful, and the corresponding laws and regulations also need to keep pace with the times. First of all, we need to legislate to solve some infringement problems of ChatGPT. Various countries have now begun to strengthen laws to restrict artificial intelligence. The European Union introduced the AI Act, which provides basic rules for the further development of AI technology by regulating it in a uniform manner. The United States issued the Blueprint for an AI Bill of Rights, which protects the safety and rights of citizens. The Code of Management of New Generation Artificial Intelligence issued by China emphasizes the requirement to enhance human well-being. However, these regulations are fragmented, lack top-level comprehensive legislation specifically regulating the application of AI, and are short of prior preventive strategies [88]. Future policy formulation should clarify which behaviors are illegal and need to be explicitly prohibited, which applications are legal, and what ethical and legal assessments are required for putting them into use. The scope and boundaries of ChatGPT in practical applications should be standardized, and foresight should be strengthened to avoid new risks that ChatGPT may bring. Secondly, it is also necessary to establish a review and supervision system. There is a need to clarify the scope of algorithmic review and the specialized subject of algorithmic review [89], which can also be based on ChatGPT's algorithms, datasets, and model training to eliminate the potential crisis of ChatGPT technology.

4.3.2 Collaborative Governance

Governing ChatGPT requires a concerted effort from all sides. Collaborative governance requires not only synergy between internal efforts in every country, but also cooperation between countries. The risk

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

analysis of ChatGPT shows the necessity of cross-sector and even cross-national governance cooperation [83]. For the country's internal, it requires the participation of the government, enterprises, society, and people. The government takes its lead and uses laws and regulations to limit the malicious misuse of technology. Enterprises actively research new technologies to improve the development of content filtering and detection technologies and continuously complete technical defects. Society plays a propaganda role to create a civilized and harmonious social atmosphere. Individuals take the initiative to improve their own quality and resolutely resist malicious behavior, while being vigilant, taking precautions and enhancing their sense of self-protection. If every member has a regulated attitude and behavior, then the order of the whole society can be easily established. In addition, the different governance approaches adopted by different countries may hinder the innovation and management of AI. AIGC is a "nuclear-weapon level" technology, and its creative and destructive effects are far beyond imagination. Countries should gradually form certain rules for the future development of general AI. Otherwise, the huge creative destructive effects of AIGC may completely subvert the achievements of human civilization [90]. Therefore, in addition to the need to strengthen multi-subject and multi-sector collaborative governance within each country, international collaborative cooperation in the governance of AI should be actively promoted.

4.3.3 Model Constraints

Guaranteeing the quality of the data generated by ChatGPT should focus on the accuracy of the raw data and the accuracy of the output results [83]. ChatGPT's current models are trained based on historical data (as of 2021) and thus lack real-time understanding of current events. In today's information explosion, this is a critical issue. Because the reliability of the prior knowledge bases is gradually decreasing, it may produce inaccurate results, especially in some rapidly evolving fields. Additionally, these models cannot review training data when the training data consists of content from various sources, some of which are unreliable and may lead to seemingly reasonable but meaningless results. Therefore, researchers should continue their efforts to improve model training methods while filtering pre-trained data to reduce the presence of misinformation in the knowledge base in order to obtain accurate responses [91]. Algorithm bias also mainly stems from unrepresentative data sets [92]. For example, the training datasets may be inadequate leading to under-sampling as well as not represent a random sample from the target population leading to sample selection bias [17]. Thus, guaranteeing data quality is key to training good algorithms. In response to the problem of information leakage during data storage and transmission, it is also essential to take corresponding measures to protect the confidentiality of data, such as using encryption technology for data processing. Furthermore, because ChatGPT is a black-box model and users cannot understand its internal mechanism, researchers could develop relevant tools to visualize the modeling process, increase the transparency of the model, help people understand the operation principle of the model, and improve the reliability of the model.

4.3.4 Application Restrictions

When organizations put ChatGPT into practical applications, they need to ensure that it is secure, reliable, controllable, and compliant in their business. As a result, some third parties can develop

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

supervisory applications with appropriate supervisory algorithms to help enterprises ensure the safety and compliance of AI. Currently, there are many supervisory platforms and services available. For example, the GRACE platform launched by 2021.AI is able to provide AI auditing capabilities, implement power controls to comply with relevant AI acts, identify and continuously monitor model bias, fairness, explainability, accuracy, security, and privacy. Datatron is a reliable AI platform built for the enterprise that obtains a overall health score of AI models, supports model bias monitoring, and ensures that models are reliable, interpretable, and accountable. Snowflake builds a data-centric platform for generative AI to help secure user data. Additionally, a user feedback system can be added. Users are often the first discoverers of some risks, and user feedback algorithms can be the first to receive user feedback and judge and filter the feedback, which can then be added to a new training dataset to further optimize the model.

5. FUTURE DIRECTIONS

5.1 Technology

5.1.1 Model Lightweighting

The GPT-3 model has many parameters, is computationally complex, requires powerful hardware support for pre-trained and inference, and costs tens of millions of dollars a day to run the model. This shows that training large language models is costly, expensive in monetary terms, difficult to implement, and environmentally expensive. Model lightweighting can reduce the number of parameters, computation, memory usage, speed up training, and save resources. Lightweighting methods include neural network pruning, quantization, knowledge distillation, neural structure search, and so on [93]. Model lightweighting may also lead to loss of model details and model deformation. Researchers also need to improve optimization algorithms to advance the lightweighting process in conjunction with the use of deep learning. Joint hardware, software, and algorithm research is also needed in the future.

5.1.2 Multimodal Technology Combination

Through the amazing ability of ChatGPT based on large pre-trained models, people have seen the dawn of developing Artificial general intelligence (AGI). Enabling to deal with more complex tasks, multi-modal AI became another promising research direction. Now, some researches have achieved a considerable progress, such as DALL-E 2, GPT-4, PaLM-E, Stable Diffusion and ERNIE-ViLG 2.0, etc. Multi-modal like text, images, and videos can be input and out by all of these models.

At present, the development of multi-modal pre-trained models is still in its early stages, and there are several aspects that deserve further research, such as (1) the design of multi-modal pre-trained objectives. Most of training objectives for multi-modal tasks are directly borrowed from single-modal models. Tailoring training objectives for multi-modal tasks may improve performance, which is also a challenging task. (2) Optimization of training methods and computing power support. The increase in need of processing modal types implies an increase in the difficulty of training effective models and the demand for computing power.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

5.1.3 Model Transparency and Interpretation

To protect the security of people's privacy and information and to help users understand AI actions and decisions, AI needs to improve transparency as well as interpretability. The uninterpretability of AI actually deepens human fears. In terms of transparency, there is a need for development companies to disclose the data used by the system, and the direction in which the data is used. In terms of interpretability, typically the highest performing methods (e.g., deep learning) are the most difficult to interpret, while the easiest to interpret methods (e.g., decision trees) are the least accurate. Interpretable methods are, before modeling, visualizing the data and, when modeling, using rule-based and individual feature-based models. For deep learning algorithms explanatory are implicit analysis methods [94], agent models [95]. Currently, there are not many methods for evaluating the interpretability of complex model metrics like deep learning, and there is a lot of room for development of interpretability research.

5.2 Applications

5.2.1 Intelligent Transformation of Society

ChatGPT has triggered an artificial intelligence change, and industries are starting to move towards widespread intelligence. ChatGPT's powerful generation capability can help us solve many problems. Since its generated content mainly depends on the sources and the model lacks interpretability, it is not responsible for the authenticity of the content and therefore requires users to make their own judgment. With the widespread use and popularity of ChatGPT and other AIGCs, it is bound to bring a large output of information, leaving many people drowned in a massive stream of information that is hard to distinguish the authenticity. Information detection techniques include content-based methods [96], social context-based methods [97], feature fusion methods [98], and deep learning methods [99].However information detection techniques into the awareness of false information, standard datasets, model adaptation, model resistance to attack these aspects are yet to be addressed.

5.2.2 Domain Specialization

ChatGPT, as a general model, can be applied with multiple domains, including education, media, healthcare, finance, etc. But it is not very targeted and specific. With the growing demand for expertise in various industries, we can expect more dedicated models for medical, financial, legal and scientific domains that require specific knowledge to provide more accurate, relevant and in-depth information to users in these domains. In addition, ChatGPT needs to develop more substantial applications. In the media industry, ChatGPT can be used in focusing on optimizing scripts and automatically editing to generate well-formatted news. In the e-commerce industry, ChatGPT can be used to automatically generate product graphic presentations and intelligent customer service. In the film and television industry, ChatGPT is used to analyze large amounts of script data to stimulate creativity, reduce creation time, and automatically generate promotional copy, posters, and trailers. Importantly, ChatGPT focuses on practical functions by analyzing specific problems in each field.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

5.2.3 Search Engine

Before the emergence of ChatGPT, people always find solutions through search engines, but there is still much room for improving the retrieval accuracy of search engines. ChatGPT's high accuracy in response to questions has shown its potential for improving the quality of retrieval. Microsoft and Google have respectively released conversational search engines Bard⑧ and New Bing⑨. However, the answers provided by both of them are not as accurate as imagined [100].

Conversational search engines still face the following issues: (1) The completeness of answers still needs to be improved: answers provided may sometimes miss some key information contained in web pages, which does not achieve the purpose of Improving retrieval's quality; (2) Answers lack credibility: from the demonstration video of New Bing⑩, it can be seen that some queries do not have reference sources provided, or there may be conflicting information between the answer and the reference source. (3) Query understanding bias: large Language models like ChatGPT are pre-trained models, and words that are not exist in training datasets maybe be misunderstood by conversational search engines, influencing the accuracy of query. Furthermore, different query prompts may lead to different query results.

Research on how to better integrate chatbots with search engines to provide faster and more accurate search services is crucial in improving the efficiency of obtaining information. This is also a meaningful direction for generative AI.

5.3 Ethics

5.3.1 Human-AI Collaboration

Much of the opposition to generative AI like ChatGPT comes from people who have a conflict of interest with them, such as the producers of original works. They fear of losing their benefits from the use of their original works in AI training datasets and the potential threat from AI-generated works taking over the original market. If a new mode or rule of human-machine collaboration can be found to help utilize the advantages of both while protecting human interests, more and better original works can be created, and artificial intelligence will become a strong ally of humanity.

AI-generated content requires human participation. An appropriate prompt can achieve satisfactory results while bad one cannot. Take ChatGPT as example, if the generated novel will be used for profit, can we see the prompts provided as people's work and have copyright and pricing, right? A new revenue mode may reconcile the conflicts of interest between humans and machines.

All in all, new ethical rules of copyright and the principle of income from AI-generated works need to be formulated by professionals, so that generative AI can be developed in the direction of benefiting the majority of people.

⑧ Microsoft. 2023. New bing demo page. https://www.bing.com/
⑨ Sundar Pichai. 2023. An important next step on our ai journey.
⑩ https://www.youtube.com/watch?v=rOeRWRJ16yY

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

5.3.2 Data Issues

Data is one of the key factors to support good performance of large language models such as ChatGPT. However, large-scale training data carries the risk of privacy data leakage, and unbalanced or poor-quality training data can affect the accuracy and unbiasedness of the model output content. Therefore, how to address the limitations imposed by the data while ensuring model performance is a direction worth investigating in the future.

Measures to protect sensitive data in training datasets include but are not limited to data deidentification, data access restrictions, and accountability [101]. In addition, using synthetic data instead of real data is a promising method to solve data problems. The use of synthetic data can not only avoid problems such as copyright infringement and sensitive data leakage, but also reduce the cost of data collection and data annotation. High-quality synthetic data for training is able to improve model training efficiency and performance.

5.3.3 Automatic Content Review

Language models such as ChatGPT often generate answers that are rich and accurate in response to questions, but still cannot achieve 100% accuracy and impartiality. The answers generated by large language models such as ChatGPT for questions are often rich and accurate, but still not 100% correct and impartiality. So far, there is no effective algorithm to evaluate the correctness of generated answers in real time or provide alert about harmful texts. If these erroneous, biased, and unethical texts are spread on the Internet, they may affect people's values and have significant negative impacts on society. Therefore, reviewing the content generated by AI models such as ChatGPT and minimize the impact of harmful outputs on humans is research that needs to be explored in the future.

6. CONCLUSION

By observing and reviewing the current status and development of ChatGPT applications, we can see the avid interest in ChatGPT capabilities and its huge potential for application in various fields in society. However, the rapid development of LLM-based generative AI systems such as ChatGPT has led to the emergence of potential ethical issues. If the corresponding regulatory policies and approaches do not catch up with the speed of technological development, serious consequences may result. Therefore, in order to explore the potential ethical issues of generative AI, led by ChatGPT, and address the challenges posed by these advanced AI systems, we present a more detailed analysis and summary of the technologies, limitations and corresponding ethical issues of ChatGPT, and give some suggestions for the development and regulation of ChatGPT.

Most of the limitations of ChatGPT are found in the generated texts, such as hallucinations, originality, and toxicity. Although the quality of ChatGPT responses has improved significantly over previous chatbots, it is still inevitably to produces erroneous, biased and discriminatory, offensive and

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

Plagiarized texts. With ChatGPT's widespread application in various fields, it is highly likely to integrate into every aspect of our lives in the future. Therefore, the impact of these limitations cannot be underestimated, such as the disruption of the knowledge dissemination ecosystem caused by erroneous texts, the influence of biased and discriminatory texts on social fairness, and so on. In addition, ChatGPT's limitations in privacy data and sustainability determine whether this technology can safely develop in real life, increasing people's productivity rather than bring new troubles. Besides, we have explored potential research directions for ChatGPT in terms of technology, applications, and ethics.

We believe that the most important issue for ChatGPT now is to resolve the conflicts arising from its potential ethical issues. It is important to fully understand underlying technology and limitations of ChatGPT, so that we can take full advantage of it without hurting by its shortcomings. Only in a secure development environment, artificial intelligence models such as ChatGPT can provide maximum convenience for human life.

We hope that this article can help researchers and users of ChatGPT to have a comprehensive look at the limitations of ChatGPT, reducing the negative impacts in the process of using it. In addition, we hope that the future research directions of ChatGPT mentioned can give some inspiration to whom are intend to do research on it. In the future, we will keep abreast of the latest developments of ChatGPT and investigate related work on alleviating the limitations of ChatGPT to evaluate the safety of ChatGPT in practical applications, promoting the integration of ChatGPT into human production and life with minimal harm.

ACKNOWLEDGEMENTS

The paper is grateful to Professor Shengyi Jiang of Guangdong University of Foreign Studies for his guidance.

AUTHOR CONTRIBUTION STATEMENT

Shangying Hua (email: syhua13@foxmail.com) Theme design, data collection, paper writing. Shuangci Jin (email: cc17837342836@126.com) Theme design, data collection, paper writing. Shengyi Jiang (email: jiangshengyi@163.com) served as the corresponding author and contributed significantly to the conception, design, and drafting of the manuscript.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no data sets were generated or analyzed during the current study.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

REFERENCES

[1] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-trained. OpenAI blog, (2018)

[2] Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9 (2019)

[3] Brown, T., et al.: Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901 (2020)

[4] Ouyang, L., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35, 27730–27744 (2022)

[5] Vaswani, A., et al.: Attention is all you need. Advances in neural information processing systems 30, (2017)

[6] Zhu, Y., et al.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pp. 19–27 (2015)

[7] Wei, J., et al.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)

[8] Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35, 24824–24837 (2022)

[9] Devlin, J., et al.: Bert: Pre-trained of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[10] Schulman J., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[11] Thoppilan R., et al.: Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)

[12] Chowdhery A., et al.: Palm: Scaling language modeling with pathways [J]. arXiv preprint arXiv:2204.02311 (2022)

[13] Biswas, S.: ChatGPT and the future of medical writing. Radiology 307.2 (2023): e223312

[14] AlAfnan, M. A., et al.: Chatgpt as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses. In: Journal of Artificial Intelligence and Technology, vol. 3.2, pp. 60–68 (2023)

[15] Dowling, M., Brian L.: ChatGPT for (finance) research: The Bananarama conjecture. Finance Research Letters 53, 1544–6123 (2023)

[16] Pavlik, John V.: Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. Journalism & Mass Communication Educator 78(1), 84–93 (2023)

[17] Akter, S., et al.: Algorithmic bias in data-driven innovation in the age of AI. International Journal of Information Management 60, (2021)

[18] Ji, Z., et al.: Survey of hallucination in natural language generation. In: ACM Computing Surveys, vol. 55.12, pp.1–38 (2023)

[19] Lee, N., et al.: Towards few-shot fact-checking via perplexity. arXiv preprint arXiv:2103.09535 (2021)

[20] Lee, N., et al.: Factuality enhanced language models for open-ended text generation. In: Advances in Neural Information Processing Systems, vol.35, pp. 34586–34599 (2022)

[21] Zhang, Y., et al.: When do you need billions of words of pretraining data?. arXiv preprint arXiv:2011.04946 (2020)

[22] Bender, Emily M., et al.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. (2021)

[23] Wang, C., Rico S.: On exposure bias, hallucination and domain shift in neural machine translation. arXiv preprint arXiv:2005.03642 (2020)

[24] Longpre, S., et al.: Entity-based knowledge conflicts in question answering. arXiv preprint arXiv:2109.05052 (2021)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

[25] Lin, S., Jacob H., Owain E.: Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958 (2021)
[26] Zuccon, G., Bevan K.: Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness. arXiv preprint arXiv:2302.13793 (2023)
[27] Dwivedi, Y. K., et al.: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management 71, (2023)
[28] Wittmann, J.: Science fact vs science fiction: A ChatGPT immunological review experiment gone awry. Immunology Letters 256, 42–47 (2023)
[29] Liu, Y., et al.: Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852 (2023)
[30] Carlini, N., et al.: Extracting Training Data from Large Language Models. In: USENIX Security Symposium, vol. 6. (2021)
[31] Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE symposium on security and privacy (SP). IEEE, pp. 739–753 (2019)
[32] Ray, P. P.: ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, (2023)
[33] Deng, J., Lin, Y.: The Benefits and Challenges of ChatGPT: An Overview. Frontiers in Computing and Intelligent Systems 2.2, 81–83 (2022)
[34] Ventayen, R. J. M.: OpenAI ChatGPT generated results: similarity index of artificial intelligence-based contents. SSRN 4332664, (2023)
[35] Fitria, T. N.: QuillBot as an online tool: Students' alternative in paraphrasing and rewriting of English writing. Englisia: Journal of Language, Education, and Humanities 9.1, 183–196 (2021)
[36] Steponenaite, A., Basel, B.: Plagiarism in AI empowered world. arXiv (2023)
[37] Rudolph, J., Samson T., Shannon, T.: ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. Journal of Applied Learning and Teaching 6.1, (2023)
[38] Golan, R., et al.: Artificial intelligence in academic writing: a paradigm-shifting technological advance. Nature Reviews Urology, 1–2 (2023)
[39] Terwiesch, C.: Would Chat GPT get a Wharton MBA? A prediction based on its performance in the operations management course. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania (2023)
[40] Kleebayoon, A., Wiwanitkit, V.: Artificial intelligence, chatbots, plagiarism and basic honesty: comment. Cellular and Molecular Bioengineering 16.2, 173–174 (2023)
[41] Tatzel, A., Mael, D.: 'Write a paper on AI Plagiarism': An Analysis on ChatGPT and its impact on Academic Dishonesty in Higher Education, (2023)
[42] Wiggers, K.: OpenAI's attempts to watermark AI text hit limits. TechCrunch, December 10 (2022). Available at: https://techcrunch.com/2022/12/10/openais-attempts-to-watermark-ai-text-hit-limits/
[43] Gao, C. A., et al.: Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. BioRxiv 2022–12, (2022)
[44] Svrluga, S.: Princeton student builds app to detect essays written by a popular AI bot. The Washington Post (2023)
[45] O'Connor, S.: Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?. Nurse Education in Practice 66, 103537–103537 (2022)
[46] Stokel-Walker, C.: AI bot ChatGPT writes smart essays-should academics worry?. Nature, (2022)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

[47] Dowling, M., and Brian, L.: ChatGPT for (finance) research: The Bananarama conjecture. Finance Research Letters 53, (2023)
[48] Editorials, N.: Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 613.612, 10–1038 (2023)
[49] COPE (Committee on Publication Ethics) (2023). Authorship and contributorship. Available at: https://publicationethics.org/authorship
[50] da Silva, J. A. T.: Is ChatGPT a valid author?. Nurse Education in Practice 68, (2023)
[51] Floridi, L.: AI as Agency without Intelligence: On ChatGPT, large language models, and other generative models. Philosophy & Technology 36.1, (2023)
[52] Temsah, O., et al.: Overview of early ChatGPT's presence in medical literature: insights from a hybrid literature review by ChatGPT and human experts. Cureus 15.4, (2023)
[53] Deshpande, A., et al.: Toxicity in chatgpt: Analyzing persona-assigned language models. arXiv preprint arXiv:2304.05335 (2023)
[54] Ferrara, E.: Should ChatGPT be biased? Challenges and risks of bias in large language models. arXiv preprint arXiv:2304.03738 (2023)
[55] Dahmen, J., et al.: Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. Knee Surgery, Sports Traumatology, Arthroscopy 31.4, 1187–1189 (2023)
[56] Ghosh, S., Aylin, C.: ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. arXiv preprint arXiv:2305.10510 (2023)
[57] Lucy, L., Bamman, D.: Gender and representation bias in GPT-3 generated stories. In Proceedings of the Third Workshop on Narrative Understanding, pp. 48–55 (2021)
[58] Abid, A., Maheen, F., James, Z.: Persistent anti-muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 298–306 (2021)
[59] Prates, M. O., Avelar, P. H., Lamb, L. C.: Assessing gender bias in machine translation: a case study with google translate. Neural Computing and Applications 32, 6363–6381 (2020)
[60] Mitrani, M., Adams, T., Noy, I.: Can We Algorithmize Politics? The Promise and Perils of Computerized Text Analysis in Political Research. PS: Political Science & Politics 55.4, 809–814 (2022)
[61] Rozado, D.: The political biases of chatgpt. Social Sciences 12.3 (2023)
[62] Rutinowski, J., et al.: The Self-Perception and Political Biases of ChatGPT. arXiv preprint arXiv:2304.07333 (2023)
[63] Khowaja, S. A., Khuwaja, P., Dev, K.: ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation: A Review. arXiv preprint arXiv:2305.03123 (2023)
[64] Radford, A., Wu, J., Child, R., et al.: Language models are unsupervised multitask learners [J]. OpenAI blog 1(8): 9, (2019)
[65] Radford, A., et al.: Language models are unsupervised multitask learners OpenAI blog 1.8, (2019)
[66] Thompson, A. D.: What's in my ai. A comprehensive analysis of datasets used to train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher. LifeArchitect. ai Report (2022)
[67] Ouyang, L., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35, 27730–27744 (2022)
[68] Open AI (no date) Privacy policy. Open AI. Available at: https://openai.com/policies/privacy-policy. Accessed: 5 Nov 2023
[69] Azadi, M., et al.: Transparency on greenhouse gas emissions from mining to enable climate change mitigation. Nature Geoscience 13.2, 100–104 (2020)
[70] Bannour, N., et al.: Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, pp. 11–21 (2021)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

[71] Writer, S.: Carbon footprint of training GPT-3 and large language models, Shrink That Footprint. (2023)
[72] Zhang, S., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
[73] Patterson, D., et al.: Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350 (2021)
[74] Luccioni, A. S., Viguier, S., Ligozat, A. L.: Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. arXiv preprint arXiv:2211.02001 (2022)
[75] Patel, D., Ahmad, A.: The inference cost of search disruption–large language model cost analysis, SemiAnalysis. (2023)
[76] Natasha L.: UK to Avoid Fixed Rules for AI – in Favor of 'Context-Specific Guidance'. Available at: https://techcrunch.com/2023/03/29/uk-ai-white-paper/. Accessed 20 Nov 2023
[77] The White House: Blueprint for an AI bill of rights: making automated systems work for the American people. Available at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/. Accessed 20 Nov 2023
[78] National Telecommunications and Information Administration: AI accountability policy request for comment. Available at: https://www.ntia.gov/issues/artificial-intelligence/request-for-comments. Accessed 20 Nov 2023
[79] Cao, J.: Towards Trustworthy AI: The Governance Challenges and Responses for Generative AI like ChatGPT. Journal of Shanghai University of Political Science and Law(The Rule of Law Forum) 38(04), 28–42 (2023)
[80] European Commission White Paper on Artificial Intelligence: A European Approach to Excellence and Trust, European Commission. Available at: https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf. Accessed 20 Nov 2023
[81] European Commission: The AI A. Available at: https://artificialintelligenceact.eu/the-act/. Accessed 20 Nov 2023
[82] Yu, P., Liu, Q.: Review of the EU Artificial Intelligence Act and Implications. Hainan Finance (06), 45–53 (2023)
[83] Shang, J.: On the Meta-rules for Risk Governance of Generative Artificial Intelligence. Oriental Law, 1–14 (2023)
[84] Lukyanenko, R., Maass, W., Storey, V. C.: Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. Electronic Markets 32(4), 1993–2020 (2022)
[85] Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer (2019)
[86] Arrieta, A. B., et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion 58, 82–115 (2020)
[87] Zhong, X., Fang, X., Gu, Y.: Governance of ChatGPT: Challenges and Countermeasures. Media Observer 3, 25–35 (2023)
[88] Yu, S., Fan, D. Z.: The Main Characteristics, Social Risks and Governance Paths of the New Generation of Artificial Intelligence (ChatGPT). Journal of Dalian University of Technology (Social Sciences) 44(05), 28–34 (2023)
[89] Zou, K. L., Liu, Z. B.: On ChatGPT-like General Artificial Intelligence Governance: Based on the Perspective of Algorithmic Security Review. Journal of Hohai University (Philosophy and Social Sciences), 1–13 (2023)
[90] Gao, Q. Q..: GPT Technology and the Modernization of National Governance: A Framework Based on Order, Empowerment and Innovation. Journal of Shandong University (Philosophy and Social Sciences), 1–10 (2023)
[91] Liu, Y. H., et al.: Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852 (2023)
[92] Israeli A, Ascarza E. Algorithmic bias in marketing [R]. Harvard Business School Technical Note 521–020, (2020)
[93] Wang, C. H., et al.: Lightweight deep learning: an overview. IEEE Consumer Electronics Magazine (2022)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00243

[94] Alain, G., and Yoshua, B.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016)

[95] Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154 (2017)

[96] Potthast, M., et al.: A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638 (2017)

[97] Jing, M., Gao, W., Wong, K. F.: Detect rumors in microblog posts using propagation structure via kernel learning. In: Association for Computational Linguistics, pp. 708–717 (2017)

[98] Volkova, S., Jang, J. Y.: Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In: Companion Proceedings of the The Web Conference 2018, pp. 575–583 (2018)

[99] Monti, F., et al.: Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673 (2019)

[100] Zhao, R. C., et al.: Can chatgpt-like generative models guarantee factual accuracy? on the mistakes of new generation search engines. arXiv preprint arXiv:2304.11076 (2023)

[101] Monti, F. et al.: Big data privacy: a technological perspective and review. Journal of Big Data 3, 1–25 (2016)