Exploring Challenges in Applying Foundation and Generative Models in AI

Ayse Kok Arslan

Abstract:
This study provides a comprehensive review on generative models, and basic components both from the perspective of unimodality and multimodality. The analysis aims to distinguish contemporary generative AI models from their predecessors. After providing a brief historical background the study discusses the recent applications of generative AI models, commonly used techniques in AIGC, and addresses concerns surrounding trustworthiness and responsibility in the field. Finally, it explores open problems and future directions for AIGC, highlighting potential avenues for innovation and progress.

INTRODUCTION

In recent years, Artificial Intelligence Generated Content (AIGC) has gained much attention beyond the computer science community.

AIGC refers to content that is generated using advanced Generative AI (GAI) techniques, as opposed to being created by human authors, which can automate the creation of large amounts of content in a short amount of time.

Generally, GAI models can be categorized into two types: unimodal models and multimodal models (Fig. 1.0). Unimodal models receive instructions from the same modality as the generated content modality, whereas multimodal models accept cross-modal instructions and produce results of different modalities.

Technically, AIGC refers to, given human instructions which could help teach and guide the model to complete the task, utilizing GAI algorithms to generate content that satisfies the instruction.

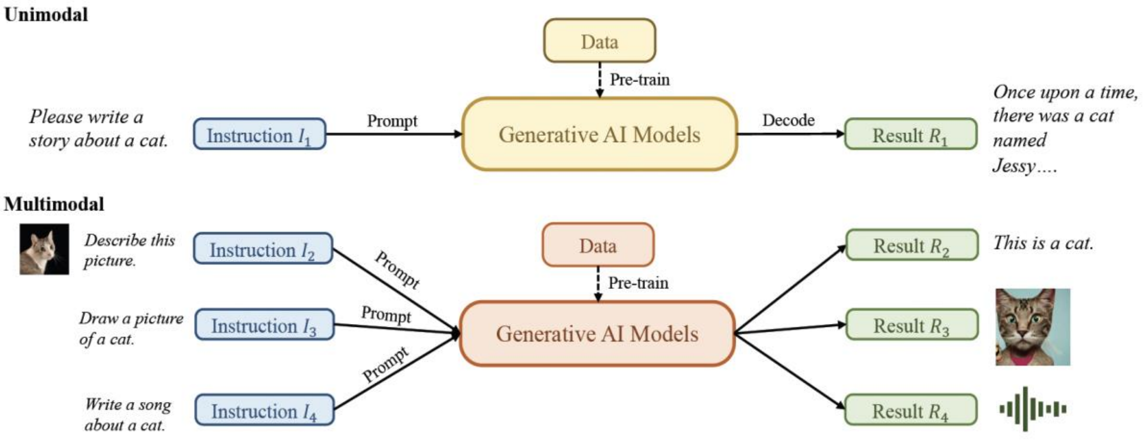


Fig. 1. Overview of AIGC (Credit: Yenala et al (2019))

By combining these advancements, models have made significant progress in AIGC tasks and have been adopted in various industries, including art [14], advertising [15], and education [16]. In the near future, AIGC will continue to be a significant area of research in machine learning. It is therefore crucial to conduct an extensive review of past research and identify the open problems in this field.

This study focuses on the core technologies and applications in the field of AIGC. The primary objective is to provide readers with a comprehensive understanding of recent developments and future challenges in generative AI.

BACKGROUND OF GENERATIVE AI

Generative models have a long history in artificial intelligence, dating back to the 1950s with the development of Hidden Markov Models (HMMs) [20] and Gaussian Mixture Models (GMMs) [21]. These models generated sequential data such as speech and time series. However, it wasn't until the advent of deep learning that generative models saw significant improvements in performance.

In early years of deep generative models, different areas do not have much overlap in general. In natural language processing (NLP), a traditional method to generate sentences is to learn word distribution using N-gram language modeling [22] and then search for the best sequence. However, this method cannot effectively adapt to long sentences. To solve this problem, recurrent neural networks (RNNs) [23] were later introduced for language modeling tasks, allowing for modeling relatively long dependency.

This was followed by the development of Long Short-Term Memory (LSTM) [24] and Gated Recurrent Unit (GRU) [25], which leveraged gating mechanism to control memory during training. These methods are capable of attending to around 200 tokens in a sample [26], which marks a significant improvement compared to N-gram language models.

In recent years, researchers have also begun to introduce new techniques based on these models. For instance, in NLP, instead of fine-tuning, people sometimes prefer few-shot prompting [38], which refers to including a few examples selected from the dataset in the prompt, to help the model better understand task requirements. In visual language, researchers often combine modality-specific models with self-supervised contrastive learning objectives to provide more robust representations.

In the future, as AIGC becomes increasingly important, more and more technologies shall be introduced, empowering this area with vitality.

FOUNDATIONS FOR AIGC

This section introduces important models such as foundation and generative models in AIGC.

Foundation Model

Transformer:

Transformer is the backbone architecture for many state-of-the-art models and is mainly based on a self-attention mechanism that allows the model to attend to different parts in an input sequence. Transformer consists of an encoder and a decoder. The encoder takes in the input sequence and generates hidden representations, while the decoder takes in the hidden representation and generates output sequence.

Each layer of the encoder and decoder consists of a multi-head attention and a feed-forward neural network. The multi-head attention is the core component of Transformer, which learns to assign different weights to tokens according their relevance.

***Pre-trained Language Models*:**

Generally, these transformers based pre-trained language models can be commonly classified into two types based on their training tasks: autoregressive language modeling and masked language modeling [41].

Given a sentence, which is composed of several tokens, the objective of masked language modeling, e.g., BERT [42] and RoBERTa [43], refers to predicting the probability of a masked token given context information. The most notable example of masked language modeling is BERT [42], which includes masked language modeling and next sentence prediction tasks.

Reinforcement Learning from Human Feedback:

Despite being trained on large-scale data, the AIGC may not always produce output that aligns with the user's intent, which includes considerations of usefulness and truthfulness. In order to better align AIGC output with human preferences, reinforcement learning from human feedback (RLHF) has been applied to fine-tune models in various applications such as Sparrow, InstructGPT, and ChatGPT [10, 46]. Typically, the whole pipeline of RLHF includes the following three steps: pre-training, reward learning, and fine-tuning with reinforcement learning.

Generative Models

Unimodal Models:

Generative Language Models:

Generative language models (GLMs) are a type of NLP models that are trained to generate readable human language based on patterns and structures in input data that they have been exposed to. These models can be used for a wide range of NLP tasks such as dialogue systems [58], translation [59] and question answering [60].

Recently, the use of pre-trained language models has emerged as the prevailing technique in the domain of NLP. Generally, current state-of-the-art pre-trained language models could be categorized as masked language models (encoders), autoregressive language models (decoders) and encoder-decoder language models.

Decoder models are widely used for text generation, while encoder models are mainly applied to classification tasks. By combining the strengths of both structures, encoder-decoder models can leverage both context information and autoregressive properties to improve performance across a variety of tasks.

Multimodal Models:

The goal of multimodal generation is to learn a model that generates raw modalities by learning the multimodal connection and interaction from data [7]. This connection and interaction between modalities can sometimes be very intricate, which makes the multimodal representation space hard to learn compared to the unimodal one. However, with the emergence of the powerful modality-specific foundation architectures mentioned in previous sections, a growing number of methods are proposed in response to this challenge.

Vision Language Generation:

The encoder-decoder architecture is a widely used framework for solving unimodal generation problems in computer vision and natural language processing. The encoder is responsible for learning a contextualized representation of the input data, while the decoder is used to generate raw modalities that reflect cross-modal interactions, structure, and coherence in the representation.

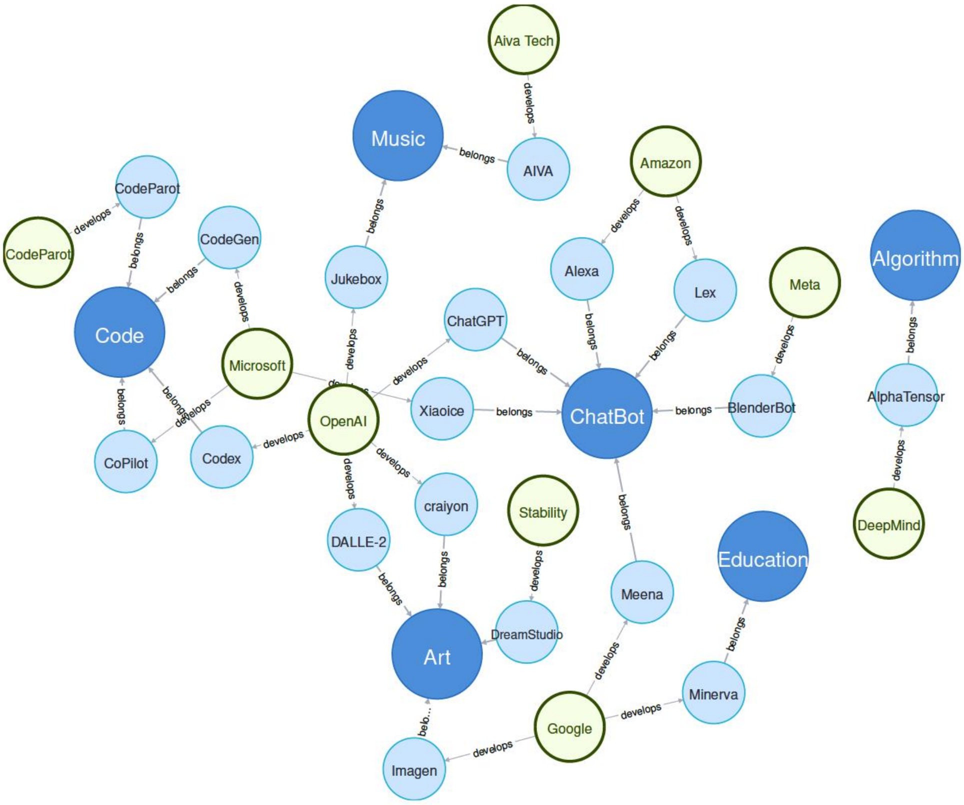
Vision Language Encoders:

Recently, the development of encoders for single modalities has advanced significantly, leading to the question of how to learn contextualized representations from multiple modalities. A common way to do this is to combine modality-specific encoders using a fusion function and then leverage multiple pre-training tasks to align the representation space [37, 134, 135]. Generally. these encoder models could be separated into two categories, concatenated encoders and cross-aligned encoders [7].

APPLICATIONS

ChatBot

A chatbot is a computer program designed to simulate conversation with human users through text-based interfaces. Chatbots normally use language models to understand and respond to user queries and inputs in a conversational manner. They can be programmed to perform a wide range of tasks, for example, providing customer support and answering frequently asked questions.



Art

AI art generation refers to using computer algorithms to create original works of art. These algorithms are trained on large datasets of existing artwork and use machine learning techniques to generate new pieces that mimic the styles and techniques of famous artists or explore new artistic styles.

Music

Deep music generation refers to the use of deep learning techniques and artificial intelligence algorithms to generate novel and original pieces of music. A prominent approach is to produce a symbolic representation of the music in the form of a piano roll. This approach entails specifying the timing, pitch, velocity, and instrument for each note to be played.

Code

AI-based programming systems generally aim for tasks including code completion, source code to pseudo-code mapping, program repair, API sequence prediction, user feedback, and natural language to code generation.

It can be fine-tuned for various code generation tasks such as code completion, summary, or translation based on a vast amount of source code data.

One unique feature is the scaffolding strategy which splits complicated tasks into smaller and manageable steps to help students gradually build their coding skills.

Education

AIGC has the potential to achieve significant advancements in education by leveraging multimodality data, for example, tutorial videos, academic papers, and other high-quality information, thereby improving the personalized education experience.

On the academic side, Google Research introduced Minerva [207], which is built upon PaLM general language models [209] and an additional science-and-math-focused dataset, to solve college-level multi-step quantitative tasks, covering algebra, probability, physics, number theory, precalculus, geometry, biology, electric engineering, chemistry, astronomy, and machine learning.

Given these applications, the increasing model footprint and complexity, as well as the cost and resources required for training and deployment, pose challenges for practical deployment in the real world. The core challenge is efficiency, which can be broken it down as follows:

- ***Inference Efficiency:*** This is concerned with the practical considerations of deploying a model for inference, i.e., computing the model's outputs for a given input. Inference efficiency is mostly related to the model's size, speed, and resource consumption (e.g., disk and RAM usage) during inference.
- ***Training Efficiency:*** This covers factors that affect the speed and resource requirements of training a model, such as training time, memory footprint, and scalability across multiple applications.

An important technique to overcome issues in efficiency is 'prompt learning' which is a relatively new concept that proposed in recent years within the context of pre-trained large language models. Previously, to make a prediction y given input x, the goal of traditional supervised learning is to find a language model that predicts the probability $P\ (y|x)$. With prompt learning, the goal becomes finding a template x' that directly predicts the probability $P\ (y|x')$ [211].

Normally, prompt learning will freeze the language model and directly perform few-shot or zero-shot learning on it. This enables the language models to be pre-trained on large amount of raw text data and be adapted to new domains without tuning it again. Hence, prompt learning could help save much time and efforts.

Traditional Prompt Learning

The process of utilizing prompt learning with a language model can be divided into two main stages: prompt engineering and answer engineering.

- ***Prompt Engineering:*** In general, there are two commonly used forms of prompt engineering: Discrete prompt and continuous prompt. Discrete prompts are typically manually designed by humans for specific tasks, while continuous prompts are added to the input embeddings to convey task-specific information.
- ***Answer Engineering:*** After the task has been reformulated, the answer generated by the language model based on the provided prompt needs to be mapped to the ground truth space. There are different paradigms for answering engineering, including discrete search space and continuous search space.

In addition to single-prompt learning methods, there are also multi-prompt methods. These approaches primarily focus on ensembling multiple prompts together as input during inference to improve prediction robustness, which is more effective than relying on a single prompt.

Another approach to multi-prompt learning is prompt augmentation, which aims to assist the model in answering questions by providing additional prompts that have already been answered.

In-Context Learning

This approach is a subset of prompt learning and involves using a pre-trained language model as the backbone, along with adding a few input-label demonstration pairs and instructions to the prompt.

SECURITY AND PRIVACY IN AIGC

While AIGC has the potential to be incredibly useful in many different applications, it also raises significant concerns about security and privacy.

Security

Factuality:

Systematic definitions of truthfulness standards and approaches for governing AI-generated content were proposed in Truthful AI [24]. The standard proposed by Truthful AI aims to avoid "negligent falsehoods" and explicitly train AI systems to be truthful via curated datasets and human interaction.

Based on GPT-3, WebGPT [25] proposed a humanoid prototype that models the AI answering process into web searching and evidence-composing phrases. Since the model is trained to cite its sources, the factual accuracy of AI-generated content is significantly improved in multiple benchmark datasets [26, 27].

Toxicity:

Besides utility, it is important for AI-generated content (AIGC) to be helpful, harmless, unbiased, and non-toxic. Extensive research has been conducted on the potential harm caused by deployed models [229–231], which can include biased outputs [232, 233], stereotypes [234], and misinformation [25].

To address this issue of toxicity in the language domain, OpenAI proposes InstructGPT [10], which aligns language models with human preferences by using human feedback as a reward signal to fine-tune the models, ensuring more relevant and safe responses. Concurrently, Google proposes LaMDA [26], a family of neural language models specialized for safe and factual dialog by leveraging fine-tuning and external knowledge sources.

Privacy

Membership Inference:

The goal of the membership inference attack (MIA) is to determine whether an image x belongs to the set of training data. Wu et al. [238] investigated the membership leakage in text-to-image (diffusion-based and sequence-to-sequence-based) generation models under realistic black-box settings. Specifically, three kinds of intuitions including quality, reconstruction error, and faithfulness are considered to design the attack algorithms.

Data Extraction:

The objective of a data extraction attack is to retrieve an image from the set of training data, denoted as $x \in D$. The attack can be considered a success if the attacker is able to obtain an image ˆx that closely resembles image $x \in D$.

Compared to the membership inference attack, the data extraction attack poses stronger privacy risks to the model. The feasibility of such an attack might be due to the memorization property of large-scale models [243], in which they turn to memorize parts of their training data.

OPEN PROBLEMS AND FUTURE DIRECTIONS

Many fundamental challenges to developing a high-quality model capable of performing well in real world applications still exist. For example, it is now increasingly well-understood that large language models trained on unlabeled datasets will learn to imitate patterns and biases inherent in their training sets [10]. Such biases can be hard to detect since they manifest in a wide variety of subtle ways. For example, the axes of marginalization differ greatly across geo-cultural contexts, and how they manifest in pre-trained language models is an under-studied area [11].

Known approaches to mitigate undesirable statistical biases in generative language models include attempts to filter pre-training data, train separate filtering models, create control codes to condition generation, and fine-tuning models. While these efforts are important, it is critical to also consider the downstream applications and the socio-technical ecosystems where they will be deployed when measuring the impact of these efforts in mitigating harm. For example, bias mitigations in certain contexts might have counter-intuitive impacts in other geo-cultural contexts [10].

The field of algorithmic bias measurement and mitigation is still growing and evolving rapidly, so it will be important to continue to explore novel avenues of research to ensure the safety of dialog agents Future work should explore the benefits of greater coordination across the research community and civil society in the creation of benchmarks and canonical evaluation datasets to test for harmful and unsafe content.

Another potential area of exploration is to study how different applications may warrant distinct levels of safety, quality, and groundedness based on the risk/benefit tradeoffs of these individual applications.

It should also be taken into account that various traits measured for safety objectives depend heavily on socio-cultural contexts. Therefore, any meaningful measure of safety should take into account the societal context where the system will be used, employing a "participatory finetuning" approach that brings relevant communities into the human-centered data collection and curation processes.

Another challenge in GAI models relates to reasoning which is a crucial component of human intelligence that enables us to draw inferences, make decisions, and solve complex problems. However, even trained with large scale dataset, sometimes GAI models could still fail at common sense reasoning tasks [256, 257]. Recently, more and more researchers began to focus on this problem.

Chain-of-thought (CoT) prompting [256] is a promising solution to the challenge of reasoning in generative AI models. It is designed to enhance the ability of large language models to learn about logical reasoning in the context of question answering. By explaining the logical reasoning process that human-beings use to arrive at answers to models, they can follow the same road that humans take in processing their reasoning.

Model training is always limited by compute budget, available dataset and model size. As the size of pretraining models increases, the time and resources required for training also increases significantly. This poses a challenge for researchers and organizations that seek to utilize large-scale pretraining for various tasks, such as natural language understanding, computer vision, and speech recognition.

Another issue pertains to the efficacy of pretraining with large-scale datasets, which may not yield optimal results if experimental hyperparameters, such as model size and data volume, are not thoughtfully designed. As such, suboptimal hyperparameters can result in wasteful resource consumption and the failure to achieve desired outcomes through further training.

AI models can inadvertently perpetuate or amplify existing societal biases, particularly if the training data used to develop the models are themselves biased. This can have significant negative consequences, such as perpetuating discrimination and inequities in areas such as hiring, loan approvals, and criminal justice.

Overall, while AI-generated content holds significant promise in various domains, it is crucial to address these concerns to ensure that its use is responsible and beneficial for society as a whole.

CONCLUSION

This study provides a comprehensive overview of the history and recent advancements in AIGC, with a particular focus on both unimodality and multimodality generative models.

The primary objective is to provide readers with a comprehensive understanding of recent developments and future challenges in generative AI. The analysis of the general framework of AI generation aims to distinguish contemporary generative AI models from their predecessors.

Hopefully, this study will aid readers in gaining deeper insights into this field.

REFERENCES

1. Akhtar: Google defends its search engine against charges it favors Clinton, ‖ USA Today (10 June) https://www.usatoday.com/story/tech/news/2016/06/10/google-says-search-isntbiased-toward-hillaryclinton/85725014/, accessed 14 July 2020 (2016)

2. Arentz, W and B. Olstad: Classifying offensive sites based on image content, ‖ Computer Vision and Image Understanding, volume 94, numbers 1–3, pp 295– 310.doi: https://doi.org/10.1016/j.cviu.2003.10.007, accessed 14 July 2020. (2016).

3. Gulli, A: A deeper look at Autosuggest, ‖ Microsoft Bing Blogs (25 March), at https://blogs.bing.com/search/2013/03/25/a-deeper-look-at-autosuggest/, accessed 14 July 2020. (2013)

4. McGuffie and A. Newhouse: The radicalization risks of GPT-3 and advanced neural language models, ‖ arXiv:2009.06807v1 (15 September), at https://arxiv.org/abs/2009.06807, accessed 9 April 2021. (2020)

5. Miller and I. Record, M: Responsible epistemic technologies: A social-epistemological analysis of autocompleted Web search, ‖ New Media & Society, volume 19, number 12, pp. 1,945–1,963. doi: https://doi.org/10.1177/1461444816644805, accessed 14 July 2020. (2017)

6. Olteanu, C. Castillo, J. Boy, and K. Varshey: The effect of extremist violence on hateful speech online, ‖ Proceedings of the Twelfth International AAAI Conference on Web and social media, at https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17908/17013, accessed 14 July 2020 (2018)

7. Olteanu, K. Talamadupula, and K. Varshney: The limits of abstract evaluation metrics: The case of hate speech detection, ‖ WebSci '17: Proceedings of the 2017 ACM on Web Science Conference, pp. 405–406.doi: https://doi.org/10.1145/3091478.3098871, accessed 30 January 2022. (2017)

8. Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil: Learning semantic representations using convolutional neural networks for Web search, ‖ WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web, pp. 373– 374.doi: https://doi.org/10.1145/2567948.2577348, accessed 14 July 2017.

9. Olteanu, C. Castillo, F. Diaz, and E. Kcman: Social data: Biases, methodological pitfalls, and ethical boundaries, ‖ Frontiers in Big Data (11 July).doi: https://doi.org/10.3389/fdata.2019.00013, accessed 14 July 2020. (2019)

10. H. Yenala, M. Chinnakotla, and J. Goyal: Convolutional bi-directional LSTM for detecting inappropriate query suggestions in Web search, ‖ In: J. Kim, K. Shim, L. Cao, J.G. Lee, X. Lin, and Y.S. Moon (editors). Advances in knowledge discovery and data mining. Lecture Notes in Computer Science, volume 10234. Cham, Switzerland: Springer, pp. 3–16.doi: https://doi.org/10.1007/978-3-319-57454- 7_1, accessed 14 July 2020. (2017)