Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 5458–5468

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Generative AI and Large Language Models - Benefits, Drawbacks, Future and Recommendations

Anne Håkansson[a,*], Gloria Phillips-Wren[b]

[a]UiT The Arctic University of Norway, Hansine Hansens vei 54, Tromsø, 9037 Norway
[b]Loyola University Maryland, 4501 N. Charles Street, Baltimore, MD 21210 USA

Abstract

Natural language processing, with parsing and generation, has a long tradition. Parsing has been easier to perform than a generation but with generative artificial intelligence (a.k.a Gen AI) and large language models (abbr. LLMs), this has changed. Generative artificial intelligence is a type of artificial intelligence that uses a large data set to create something in the genre of that data set. It can generate different outputs ranging from texts, audio, objects, pictures, and paintings to videos, but also synthetic data. LLMs use deep learning and deep neural networks to train on large text corpora for recognizing and generating texts. These models are based on massive data sets, collected from databases and the web. They use transformer models to detect how elements in sequences relate to each other. This provides context support. Two well-known large language models are the Generative Pre-trained Transformer, GPT, used in ChatGPT and Bidirectional Encoder Representations from Transformers, BERT. Although LLMs have advantages, they have problems. This paper presents generative artificial intelligence and LLMs with benefits and drawbacks. Results from applying these models have shown that they can work well for accuracy in specificity, user personalization and human-computer communication but they may not provide acceptable, reliable and truthful results. For example, ethics, hallucinations and incorrect information, or misjudgments, are some major problems. The paper ends with future directions, research questions on LLMs, and recommendations.

Keywords: Natural Language Processing; Generative AI; Large Language Models;

1. Introduction

Natural language processing (NLP) has been around since the 1960s, i.e., when artificial intelligence, AI, was coined [9]. NLP has two parts: parsing and generation. Parsing examines texts for correctness by churning through the corpora whereas generation produces new texts from a given data set.

* Corresponding author.
E-mail address: annehak@kth.se

10.1016/j.procs.2024.09.689

Since the beginning of NLP, parsing has been seen as easier to perform than generating well-formed, viable texts. Parsing for, e.g., text control and editing, with syntactical, and semantic correctnesses, is fairly straightforward whereas parsing for pragmatic correctness is much more difficult. Generating syntactically correct sentences is easier than generating semantically and pragmatically correct sentences, even though this has changed with the support of deep learning technology.

Syntactically, parsing analyses the correctness of structures of sentences. With simple means, such as parse trees (also called parsing trees) showing the structure of the sentence, it is possible to parse all commonly structured sentences [9]. These so-called parsing trees, provided as part-of-speech (POS) tagging, check the syntactic correctness and semantic correctness. Syntactic structures, S, can be represented as a noun phrase, NP and a verb phrase, VP, S= < NP, VP >. This can be further broken into < N, V, N > and < Det, N, V, Det, Adj, N > (Det=Determination, N=Noun, V=Verb, Adj=Adjective). For example, the parse tree < Det, N, Conj, Adj, N > can check sentences such as < The, boy, and, the, happy, girl >. With conjunction complex texts can be built, i.e., < S, Conj, S > [9].

Semantics is the linguistic meaning and concerns the analysis of meanings of words and their relationships to other words. Here methods, such as parse trees, lexicons, and ontologies are applied to provide semantically correct sentences [9].. Semantics use information about nouns and verbs to avoid accepting nonsense sentences. Lexicons provide the information about the words and the parse trees are applied to check the semantic correctness. For example ”the” < determiner >, ”boy” < noun, male, singular >, ”and” < conjunction >, ”happy” < adjective, feeling pleasure and enjoyment > and ”girl” < noun, female, singular >. This can prevent sentences, such as < The house likes the boy >, which is syntactically correct but semantically wrong. Ontologies are formal descriptions that provide information about the relationships of words [9]. Ontologies structure pieces of information by linking the different words in the ontology. This facility makes ontologies good for semantics analysis, word translations and domain concept linkages [28].

Pragmatics is context-dependent and requires analyses of text bodies. The sentences are interpreted in a certain situation to deal with the sensibility and realism of each sentence in the text. The pragmatics provide contextual information to support interpreting the sentences in their right contexts. For example, the sentence ”I saw her duck” has more than one meaning, i.e., ”she ducked” or ”she has a duck”. The precedes and posterior sentences provide the necessary contexts for the text. The pragmatics can be handled with machine learning, e.g., sentiment analysis and intent analysis and deep learning, e.g., transformer models.

Sentiment analysis is a contextual mining tool that analyses texts to determine the attitude of the underlying sentiment, i.e., syntactic features, such as positive, negative or neutral. To determine the attitude, the analysis can use a lexicon-based approach but the most common is machine learning classification. The lexicon-based approach uses, besides lexicon, syntactic features of the text, rules and heuristics; the machine learning uses algorithms like decision trees, support vector machines (SVMs), and neural networks.

Neural networks can, besides attitudes, also identify intents, such as complaints, suggestions and appreciation. The process of determining the underlying intention behind the text is called Intent analysis. The analysis captures intents by classification and intent recognition.

Deep neural networks and deep learning are effective techniques for in-depth conversation analysis. Deep-learning-based approaches, like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), use pre-trained language models [14] for text parsing and generation to handle, e.g., sentiment classification and text summarization of long text bodies, documents, articles and papers.

Instead of word-by-word using a lexicon, deep learning such as BERT trains bi-directionally, to analyse texts. Bi-directional implies that sentences are analysed from two directions, left-to-right and right-to-left. GPT is commonly a unidirectional approach and trains in one direction, i.e., left-to-right. Generation of semantically correct sentences is difficult and pragmatically correct has almost been impossible [9]. However, with newly developed techniques, especially generation is improved. For example, Generative artificial intelligence (Generative AI) and Large Language Models (LLM) have provided some excellent results for both semantically and pragmatically correct sentences.

Technically the solutions have not been easy to develop and the results have shown that generated texts do not always correspond to real-world knowledge [17]. Various techniques have been applied, ranging from finite-state machines to deep learning, which is considered to be the most powerful learning technique today. Enablers are the advancements of hardware and software moving from character-to-character computation to corpora generations and images, videos and audio creations.

This paper presents Generative AI, Gen AI, and LLMs with benefits and drawbacks. Results from using these large language models have shown that the models can work well for quantitative data sets and accuracy in specificity, user personalization and human-computer communication. However, when generating texts with qualitative data using small data sets, the large language models may provide unacceptable, unreliable and untruthful results. Especially, the context-dependent results are problematic, which can be hallucinations [17] with source-reference divergence, biased training data, and privacy concerns. Results can also be incorrect information, misjudgments, discriminatory content, and privacy violations. Besides benefits and drawbacks, the paper sheds light on future directions and research questions on LLMs and provides recommendations.

2. Generative AI

Generative AI or Gen AI for short, creates content based on natural language input. In NLP, Gen AI uses deep learning to generate all sorts of texts like books, poems, product descriptions and other material. Generating texts in NLP is essentially Gen AI although the term was not used in the beginning i.e., 1950-1960. The term became popular in 2014-2020 starting with generative adversarial networks (GANs) and their powerful possibilities of creating output quickly.

Besides generating texts, Gen AI handles other media, such as images, videos, audio and synthetic data. Gen AI identifies, and encodes patterns and relationships in huge data sets and then uses the data to generate or create new content [20]. Generating solutions involves training, tuning, generating, evaluating, and re-tuning to continuously improve the quality and accuracy of the generated content. The results can be customized by providing feedback about, e.g., style, and tone but also other elements, like personalization, that the content shall reflect.

Gen AI uses different techniques to generate content. The most common are variational autoencoders (VAE) and autoencoders, GANs and diffusion models, and transformers [20].

2.1. Variational autoencoders

Variational autoencoders, VAEs, as the name suggests provide ”variations” of the output. Initially, the VAEs trained with unsupervised learning [21] but they can train with supervised learning [23], and semi-supervised learning [22] as well. A VAE compresses input data to features (encode), like autoencoders, and constructs (decode), while training, several new variations of the content to provide more accurate and high-quality content. To encode and decode, autoencoders incorporate two connected neural networks [20]. They efficiently encode by compressing input data to essential features. From the compressed representation, they decode by reconstructing the input. Autoencoders use unsupervised learning to discover latent variables, that are hidden or random variables. VAEs have been used for anomaly detection, facial recognition and natural language generation. The variational autoencoders and autoencoders handle domain knowledge and are special cases of encoder-decoder neural networks. The encoder-decoder neural networks are often Long Short-Term Memory Networks, LSTM, that handle sequence-to-sequence, seq2seq, and predictions, e.g., translations. The encoder-decoder LSTMs can consist of two recurrent neural networks (RNN). RNN Encoder–Decoders can score a pair of sequences, i.e., input sequence and output sequence, and generate target sequences given a source sequence.

2.2. GANs and diffusion models

Generative adversarial networks (GANs) is a deep-learning generative technique. GANs train with supervised learning, unsupervised learning and semi-supervised learning [19] to create good, realistic texts, images and videos [19]. As VAE, GANs also train two neural networks for learning the models: a generator and a discriminator. The generator model captures the data distribution and generates samples in the domain (from an input vector). The discriminator model classifies or predicts labels of samples that either came from the training data (true) or the generator model (false). Thus, the generator model generates new examples and the discriminator model discriminates the examples (real and fake) [24]. The models train together until the generator model generates plausible examples that can be considered to be true. Thus, GANs generate e.g., images with one part to evaluate how realistic they are with the other part [18]. Technically, for MNIST and CIFAR-10 image datasets, the models are fully connected layers

networks where the generator model uses ReLU activations and the discriminator uses maxout activations. GANs have had problems with generating texts due to the texts' discrete nature. The text is composed of distinct, separate words handled as tokens and the generation of texts involves producing sequences of discrete tokens. To generate texts, GANs train with backpropagation, which does not handle structures well. In addition, the dependencies between words, tokens, correct structure and context of words are hard to capture.

Another Gen AI technique is diffusion models. These diffusion models create high-quality, photorealistic images, videos and animations by adding or removing noise in the output image. It starts with a part of the image and then slowly fills the gaps by defining a Markov chain of diffusion steps, creating a diffusion process. The diffusion models learn by using the diffusion process: *forward* alters data by adding noise until the data becomes identically to pure noise and *reverse* generates by reversing the forward process, i.e., "denoising" it to reconstruct the original data. Thus, reverse removes noise to generate new data samples that resemble the original data distribution.

Stable diffusion is a specific and optimized implementation of diffusion models. Stable diffusion uses a text-to-image technique enabling the transformation of text into visual representations by blending VAE (output variation) with diffusion models (forward and reverse process). They generate an image adhering to the text prompt, with text and image representations. From a user-given text request, a.k.a. text prompt, the stable diffusion converts the text into numerical values (tokenizing) and provides a vector for each token (text encoding). Representing the text prompt as numerical values is the text representation, a. k. a. embeddings. Then, the stable diffusion generates an image representation, i.e., a vector that numerically summarizes an image from the text. This image representation is refined and upscaled meaning removing noise from the current image representation (refine) and upscaling the image representation into a high-resolution image. The stable diffusion is used for generating detailed images conditioned on text description using diffusion techniques. However, it is not suitable for sentence generation with creating coherent and contextually appropriate text.

2.3. Transformers

A transformer is a network architecture [15]. It relies entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution[15]. Through self-attention mechanisms, the transformers focus on learning dependencies, i.e., drawing global dependencies between input and output. Self-attention mechanism relates different positions of a single sequence to each other, to compute a representation of the sequence.

Transformers are designed to handle sequential data. They utilize an encoder and a decoder to process and generate sequential data. The transformer architecture (often an encoder-decoder architecture) consists of an encoder to map an input sequence of symbol representations and a decoder that generates an output sequence. The input is the entire text as a sequence of words, which is converted into tokens, i.e., numerical representations required by the neural network. These tokens are embedded into vectors and processed in a feed-forward neural network [15]. Transformers use the attention mechanism to provide weights for words' significance in a sentence, allowing the model to focus more on what is relevant in the text. Some information about the relative or absolute position of the tokens in the sequence must be provided since no recurrence or convolution is applied.

The transformer architectures can handle the context of each word in a sentence by considering it concerning every other word in the sentence. During training, the model learns to predict the next word in the sentence. The model adjusts its internal parameters and can, thereby, reduce the differences between predictions and the actual outcomes.

Examples of approaches that use transformer architectures to capture dependencies and attention in the corpora are Generative Pre-trained Transformer, GPT, [15] and Bidirectional Encoder Representations from Transformers, BERT. The difference between GPT and BERT is the training. GPT trains unidirectional and BERT bidirectional.

GPT is a transformer model that uses multi-head attention to allow the model to handle information from different subspaces at different positions [15]. Commonly, GPTs use minimal task-specific parameters and train by fine-tuning all pre-trained parameters. GPTs apply generative pre-training of a language model on a diverse corpus of unlabeled text, using semi-supervised learning, which is followed by discriminative fine-tuning on each specific task [26]. Hence, GPT can generate coherent and contextually relevant texts for various topics and styles.

GPT uses a left-to-right architecture, i.e., unidirectional. Every token attends to previous tokens in the self-attention layers of the transformer. Thus, GPT transformer's self-attention only looks at dependencies to the left meaning that

it misses the right side dependencies, which can be regarded as constrained [25]. A solution to this problem is a bidirectional approach like BERT that is based on a bidirectional self-attention transformer.

BERT [25] is designed to pre-train deep bidirectional representations from unlabeled data. It trains in both directions in all layers, hence bidirectional. BERT uses two steps solution, i.e., pre-training and fine-tuning, to provide a pre-trained BERT model that can be fine-tuned with labelled data from the downstream tasks [25] . BERT's model architecture is a multi-layer bidirectional transformer encoder [25] based on the original implementation of transformers, like GPT.

For pre-training, BERT uses a masked language model, MLM, to randomly mask some of the tokens from the input and to predict the masked word from the context[25]. MLM fuses the left and the right context, which supports pre-training a deep bidirectional transformer[25]. BERT also uses next-sentence prediction that pre-trains text-pair representations. In deep learning NLP, BERT can be seen as a predecessor of generative AI.

A difference between BERT and GPT is that BERT generates texts by answering questions (so-called question answering) and fills in the blanks missing in the text, i.e., BERT works by predicting masked words or missing words in the input text. GPT generates texts from scratch by using keywords, phrases, or sentences.

Some autoencoders combine BERT and GPT, for example, BART. BART is an autoencoder for pre-training sequence-to-sequence models [27]. It pre-trains by combining bidirectional and auto-regressive transformers (bidirectional encoder, like BERT, with a left-to-right decoder, like GPT) [27]. BART uses a standard transformer-based neural machine translation architecture with a GeLU activation function, just like BERT. GPT uses the ReLU activation function.

2.4. Benefits and Drawbacks of Generative AI

Generative AI has positive and negative impacts. Positives are instant language translations, voice and text, supporting communication between people not speaking the same language. It also provides movie dubbing and enriched educational content to support humans in tedious labour work. Negatives are fake news and deepfakes with digitally forged images or videos [19] pretending to be a person who can then fool other people, as in the case where Generative AI was used to fool a parent into sending money to the child. Another problem is harmful cybersecurity attacks on businesses, including requests that realistically mimic an employee's boss[19].

Table 1. Comparison of VAE, GAN, and Transformers

Aspects	Differences Between VAE, GAN, and Transformers
Purpose:	VAE and GAN: Generative models for creating new data samples Transformer: Sequence modeling and representation learning
Architecture Components:	VAE: Encoder, Decoder, Latent Space GAN: Generator, Discriminator Transformer: Encoder, Decoder, Self-Attention Mechanism
Key Concept:	VAE: Learn a latent space representation GAN: Learn to generate realistic data samples through adversarial training Transformer: Learn dependencies in sequential data using self-attention
Applications, examples:	VAE: Image generation, anomaly detection GAN: Text-to-image generation, Image generation, video generation Transformer: Text generation, image recognition, Machine translation
Advantages (+) / Disadvantages (-):	VAE: (+) Probabilistic modelling (-) Generated samples can be blurry GAN: (+) High-quality image generation (+) adversarial training provides sharp outputs (-) Training instability Transformer: (+) Handles long-range dependencies well (+) are scalable (-) Requires large amounts of data (-) computationally intensive

3. Large Language Models

Large Language Models are models handling enormous amounts of text data. The term LLM grew from the need to handle language models that require large quantities to provide plausible results. Language models are probabilistic models that handle the effects of random events or actions by using probabilistic modelling to predict the potential occurrence of future outcomes.

Training generative AI models, like GANs, is computationally intensive, time-consuming and expensive [20]. No matter the generative AI technique, all of them require thousands of clustered GPUs and weeks of processing[20] to perform well. To make the processing more efficient, foundation models, FMs, and large language models, LLM, are developed and applied.

FMs and LLMs can achieve general-purpose generation. FMs refer to any model trained on broad data and adapted to a wide range of generative tasks (movies,) whereas LLMs are solely used for natural language processing.

Foundation models (FM): The foundation models are deep learning neural networks that function as platforms to empower Gen AI generating, e.g., realistic texts, images and videos, and sounds and music applications used in computer vision, synthetic music and virtual reality. The FM can also support several kinds of content generation, such as multimodal FMs that are trained simultaneously with multiple modalities.

A characteristic of FMs is the training on massive datasets with e.g., terabytes of data and billions of parameters [20]. The FM trains the deep learning algorithm on huge volumes of raw, unstructured, unlabeled data. The algorithm evaluates millions of blank spots and predicts the next element in a sequence[20]. To give an idea of sizes, the early BERT model was trained using 340 million parameters and a 16 GB training dataset. GPT-4 is trained using 170 trillion parameters and a 45 GB training dataset [20].

The FMs are often used as a starting point to develop AI models for faster and more cost-effective applications. The most common FMs are BERT, and GPT created for text generation applications. GPT can be regarded as the start of LLMs with the invention of transformers. BERT can be perceived as one of the first LLMs with thousands of open-source, free, and pre-trained BERT models available for specific use cases, such as sentiment analysis. As BERT uses MLM, the FM can fill in the next word in a sentence or the next element in an image[20]. It can also fill in the next command in a line of code and adjust itself continually to minimize the differences between predictions and the actual result [20].

Large Language Models: Large Language Models, LLMs, can be considered as a subset of FMs since LLM generates texts but not images, and sounds. LLMs are designed to process and generate natural language. LLMs are trained using deep learning algorithms and they learn language patterns and structures. Besides natural language comprehension, generation and writing texts, creatively, LLMs handle text completeness and predictions, provide machine translation and sentiment analysis, and chatbots and virtual assistants. In addition, LLMs create content by generating, translating, or detecting errors in programming code.

Commonly, LLMs are based on transformer architecture with encoder-only, decoder-only, or encoder-decoder configurations. They are characterized by their large number of parameters and extensive training on diverse text corpora of petabytes. The generated outputs are probabilities of various possibilities and combinations creating an extensive list of possible words and their probabilities of being used in a text.

LLMs can be seen as ”self-supervising”. They generate labels and, hence, do not require any manually provided labels. Instead, LLMs utilise the fundamental structure and data properties to enable the model to learn useful representations and features.

There are several different LLMs: GPT-4 (OpenAI with 1.76 T parameters), LLaMA (Meta, 70 B parameters), PaLM2 (Google, 540 B parameters) and BART / BART-large (Facebook AI/Meta Platforms, 406 M parameters). These models can be considered to ”understand” or ”comprehend” text. The models are rather symbolically treating or adequately administering a simulation of understanding languages and contexts. ”Understand” context comes from capturing long-range dependencies between parts of words, words and sentences, supported by an attention mechanism. Dependencies are the relationships learned by vast amounts from hundreds of thousands of texts collected from, e.g., news articles, social media posts, and web pages.

Popular LLM interfaces for texts are, e.g., ChatGPT and Gemini[19]. ChatGPT (OpenAI) is a chatbot tool pre-trained on data up to a certain year, currently 2021. ChatGPT can analyse data, create charts, present information using images and create images. Gemini (Google’s AI), earlier Bard and successor to LaMDA and PaLM 2 is a

chatbot handling multimodal reasoning by interpreting and combining texts, codes, images, audio, and videos. It also handles complex coding and mathematical reasoning. Gemini can draw conclusions, analyze the context, make decisions, suggest courses of action, draw parallels between different concepts and scenarios, and symbolically treat and maintain context in conversations.

Conclusively, LLMs can generate contextually relevant, grammatically correct, and semantically meaningful texts, as well as demonstrate an advanced comprehension with symbolically treating languages by multi-layer neural networks, diverse and extensive datasets, pertaining (often unsupervised learning) and fine-tuning (adjusting the learned representations to specific tasks or domains), text representation (tokenization and embeddings), transformers with analysing the relationships between tokens in a sequence and attention mechanisms focusing on relevant parts of the text.

4. Benefits and Drawbacks of LLMs

LLMs offer both potential benefits and drawbacks [31]. One way to increase the likelihood of realizing the benefits is a well-tailored prompt. To interact with a LLM, the user provides a prompt in either written or verbal form so that the system can respond. Prompts are "instructions given to an LLM to enforce rules, automate processes, and ensure specific qualities (or quantities) of generated output" that program "the LLM by customizing it and/or enhancing or refining its capabilities" [7](p. 1). The term 'prompt engineering' is a phrase used to describe the tailoring of a prompt to achieve the desired outcome and is a type of programming. The structuring of prompts as prompt patterns or templates is similar in format to classic software patterns with modifications to address the context. Domain-specific pattern catalogs are useful to address problem solving in specific domains such as medical and legal fields. However, general classifications of prompt patterns have been identified including input semantics (i.e., how the LLM understands the input), output customization (i.e., constraining or tailoring output), error identification (i.e., resolving errors), prompt improvement (i.e., improving the quality of input and output), interaction (i.e., user-LLM interaction), and context control (i.e., controlling the context) [7]. In the process of interacting with the user, the LLM will begin to tailor its responses to the prompts, becoming more accurate in terms of specificity and more personal in terms of responding to the user. However, the LLM can also become confused or inaccurate if not guided appropriately by the user, increasing the risk of error propagation and poor decision making.

Specific tasks for the LLM can be evaluated to determine the expected benefit compared to existing methods or a human processor. A number of studies have evaluated natural language processing including sentiment analysis, text classification, and natural language inference [1]. Sentiment analysis, also called opinion analysis, is a subfield of natural language processing that analyzes textual data to attempt to interpret its emotive nature or polarity [6]. In traditional ML this is often done by analyzing individual words within the text and applying sentiment dictionaries to arrive at a conclusion about the document, sentence or phrase. LLMs group words to determine sentiment and show good results superior to single-word traditional analysis. Text classification is related to sentiment analysis as it attempts to process a body of text. For example, text classification methods can produce credibility ratings for news outlets. Studies have shown that the results are moderately correlated with a human expert [1]. Semantic understanding of language and associated concepts such as meaning and intent is poor compared to humans at the current time, and LLMs often classify nonsense phrases as meaningful [1].

Reasoning tasks using evidence, arguments and logic are some of the most anticipated benefits of LLMs as an intelligent AI system, and they exhibit an ability to reason when they have enough data [2]. LLMs show potential and continual improvement in reasoning dependent on the specific task and specific LLM. For example, some tasks require complex multi-step reasoning while other tasks are simple step problems. Some LLMs perform well on analogical reasoning but lack spatial reasoning. Performance is dependent on the specific LLM since models have been trained on tailored datasets for a specific task.

Natural language and dialogue offer improvements in human-computer interaction, and these tasks are most associated with LLMs. To interact successfully, the model needs to understand context and answer questions appropriately. Fine-tuned models can be used for specialized domains such as healthcare and give reasonable responses. Question answering is an increasing application area for intelligent customer service and search engines. Many LLMs are able to reasonably answer questions, summarize text, generate code, solve math problems, translate language, and provide commonsense answers [1]. In responding to human dialogue, LLMs are able to provide various styles such as infor-

mal, professional, argumentative, and creative. However, in multilingual contexts, particularly non-Latin languages, research is progressing slowly.

LLMs display variation in robustness, ethics, bias, and trustworthiness [1] [8]. Robustness considers the stability of the system when facing unexpected inputs, and studies have shown that LLMs are vulnerable to adversarial prompts and "internalize, spread and potentially magnify harmful information ... [using] ... toxic languages, like offensiveness, hate speech, and insults as well as social biases against people with a particular demographic identity" [1](p. 39:14). Besides the societal impact, LLMs can introduce bias though the training data, the algorithm, and the output[3]. Bias can be reinforced through viewpoints of individual actors who seek to sway public opinion on topics such as complex environmental issues [3]. Since LLMs seem to be authoritative, they may have undue impact on users similar to the negative implications of social media by amplifying misinformation [5], creating fake news [4], and creating digital echo chambers. Ethical issues represent a serious challenge and risk to society as their use spreads.

A major known issue in LLMs is hallucinations, responses that appear reliable and logical but are factually inaccurate or do not represent reality. The false information may appear highly plausible, making it challenging for models or humans to detect [8]. Hallucinations in LLMs have been characterized as input-conflicting (i.e., content that deviates from the source input or prompt from the user), context-conflicting (i.e., content that conflicts with its own previously generated information, or fact-conflicting (i.e., content that is inconsistent with established or accepted world knowledge). In addition, LLMs can be inconsistent and imprecise since they are relating words statistically without considering meaning [30]. In their current form, LLMs cannot replace human judgement, so a human must be part of the use of LLMs.

Some domains show promise for the use of LLMs. For example, LLMs can act as medical assistants providing clinical information or summarization of research. Educational applications are emerging such as assisting students with improving their writing, explanations of complex ideas, delivering useful feedback, and providing personalized attention and feedback. Although they are not sentient, LLMs can appear as coaches or mentors to humans by generating reasonable contextual speech. However, they may also provide incorrect information or misjudgment, so LLMs require monitoring by a human decision maker. Information retrieval and incorporation into recommender systems is also emerging, although there are potential risks when the human does not exercise independent judgment.

5. Future Research for LLMs

LLMs have attracted the attention of the research community toward exploring benefits and drawbacks of these models. At a societal level, a persistent question is the effect of using LLMs over the long term. Will people begin to substitute LLM-generated text for the effort to write their own documents, and will that result in less creativity and expertise on the part of the human user? LLMs depend on data, so will they simply reiterate and retread the same data without meaningful new input, and will the net result be a lowering of human expertise?

A major known issue with LLMs is hallucinations and tokenization-related errors referred to as 'glitch tokens' [32]. Designers of these systems are working to mitigate against propagation of errors [32] and disinformation. Repeated often enough, misinformation can be perceived as factual. Thus, teaching people how to discriminate between 'information' that is reliable versus 'information' that is false is a persistent problem that is exacerbated by social media and the seemingly authoritative voice of AI. At all levels of education, critical thinking skills become essential in combating misinformation and disinformation.

Since LLMs are based on vast quantities of human language, they can repeat and reinforce stereotypes, pick up biased language, amplify hate speech, present socially unacceptable ideas, and sow social discord. Appropriate guard rails are needed that respect free speech and yet stay within the norms of society. Since the phrase 'socially acceptable' is a societal judgement call, government legislation is needed to find a balance between competing ideas. In addition, as with social media, external efforts to influence viewpoints or sow discord are ever-present challenges, and methods need to be developed to combat them.

LLMs can be improved by taking into account the knowledge or expertise level of the end user. For example, the LLM could modify its output if the user identifies themselves in terms of general characteristics. Other approaches to specificity are AI frameworks such as Retrieval-Augmented Generation, or RAG, that have been shown to overcome imprecision for some prompts, update world knowledge, provide provenance for decisions, and generate more tailored and factual knowledge [30].

A sample of potential research questions are shown in Table 1. As a relatively new technology, research on LLMs is likely to increase, and the list of research questions appears endless. There is a need for research on all aspects of LLMs including the growth of capabilities, expanding use cases, and societal response to the LLM technology.

Table 2. Sample of potential research questions on LLMs.

Topic	Research Questions
Societal Impact	Does persistent LLM use reduce human expertise?
	Over time, will LLM use reduce the amount of creativity and new ideas in society?
	Are LLMs actually generating anything new that is meaningful?
	What security mechanisms are needed to combat external influences considered a threat?
	What are the far-term impacts of LLM use and how can we minimize negative outcomes?
	What laws are needed to govern LLMs and protect society?
	How to prevent Generative AI and LLMs from steeling identities with face and voice?
	What will society do with Generative AI and LLMs' almost real-time performances?
	Can Generative AI and LLMs' make society unintelligent and blasé due to self-learning wrong produced outcomes?
Benefits of LLMs	How can an LLM be tailored to assist specific fields?
	What metrics should be used to benchmark LLM functionality?
	What is the appropriate level of human interaction with an LLM for a desired output?
	How can LLMs provide more accurate and context-aware interpretations and results?
Hallucinations	How can LLMs be fact-checked, and what sources of information should be considered authoritative as checks on LLM output?
	Should LLMs filter their input data? If so, against what norms?
	How can LLMs combat the propagation of errors and disinformation?
	How can the internal consistency and stability of LLMs be improved?
Bias	How can bias be identified in LLMs?
	What filters or guardrails are appropriate to modify LLM outputs?
	How can social discord topics such as hate-speech be identified in LLMs and filtered out?
Decision making	What fields can be assisted by LLM use during decision making?
	How is medical decision making influenced by LLM use?
	How can reasoning be improved in LLMs?
Education	Can LLMs serve as an effective tutor for students?
	Do students interact with an LLM differently depending on age?
	How can we teach students to discriminate factual information from misinformation during LLM use?
Communication	Are LLMs useful in patient-physician communication such as explaining medical diagnosis?
	Can an LLM modify its behavior to account for the expertise of the user?
	How accurate are LLMs for text classification?
	How well do LLMs perform with different languages?
Prompt engineering	How can humans effectively communicate with an LLM to achieve a desired output?
	What types of formats or templates are useful in communicating with an LLM?
	How should we evaluate the response from an LLM?
	How many more advances are needed in Generative AI and LLMs before it sees an AI winter?
Energy efficiency	How can LLM train on massive datasets with minimal energy consumption?
	How can infrastructures [28] be consolidated to achieve energy efficiency?

6. Recommendations

Since LLMs are infiltrating our daily lives in forms such as digital assistants on mobile phones and copilots for work tasks, we need to be aware of the potential for harm resulting from inaccurate or biased guidance. A major source of error is the lack of control of the underlying data on which the system is based. LLMs have been called 'statistical parrots', indicating that both the quantity and quality of the input data are fundamentally important in training the system. Although LLMs are trained on a massive amount of data, there is no assurance that these are the right data to answer the question. The data may contain errors or biases, be wrong and distorted, diverse and scattered, and contain contradictions. In addition. text generation with GPT-3 and BERT may be limited by challenges such as data availability and quality, model complexity and scalability, and ethical and social implications. Although the quantity

of data in LLMs may not considered problematic, the well-known issue of hallucinations shows that the designers must be vigilant to ensure that they have enough data from the right sources to produce accurate results.

Secondly, pre-trained LLMs are useful for models that are very complex and require a lot of computational resources to train and run. However, due to their size and complexity, robustness and stability can be problematic. For example, overfitting or underfitting with algorithms such as BERT and GPT may impact their robustness and stability and also reduce their ability to generalize. Questions or tasks related to the same problem but worded in different ways should produce equivalent correct results. Thus, similar to any analysis, the designer must be aware of the outcome and evaluate the likelihood of correctness in the result. The designer should use the same model with different training sets and evaluate the comparability of the results.

The rise of synthetic data has been driven by lack of data due to privacy restrictions, sensitive information, data quality issues, and a lack of labelled data. Although synthetic data is generated from original data and on the trained model that reproduces characteristics and structure of the original, synthetic data can be skewed, incomplete, or discriminating in a way that skews the output more than the original results. Results from the system should be analysed and interpreted from this point of view. In some domains synthetic data may not be problematic and even be good to use, but in other domains it may be a disaster.

Federated learning, where the system collaboratively learns a shared model while keeping all the training data decentralized, is a positive for privacy reasons. Another interrelated concept is federated data. Federated data refer to distributed data sources that are used without sharing raw data. These techniques are helpful in enhancing privacy, security, and data governance. However, both federated learning and federated data have issues with handling the complexity of the exchange. Handling remote, decentralized data can be costly in communication and runtime if data come from many different areas.

An area that has received considerable attention is the effect of generative AI and LLMs on society. These technologies are being used to spread misinformation, disinformation and malinformation. Deepfakes create realistic-looking, believable videos by manipulating images of real people using their voices without consent. Intellectual property is a concern as these systems emulate source data that can stylistically sound and look like it was created by the original author or artist. Society will need to decide when the results tip over into plagiarism, violate privacy, violate intellectual property, and run afoul of laws, regulations and rules of conduct. Moreover, they may have significant impacts on human communication, information, and knowledge that need to be taken into account with respect to ethical and social norms and values. Certainly our current laws need to be updated to provide guidance on who is the responsible party and to provide ways to ensure accountability and reduce harm.

7. Conclusions

This paper presents Generative AI and LLMs with their benefits and drawbacks. Results from applying these LLMS show that the models work well for accuracy in specificity, user personalization, and human-computer communication but may provide unacceptable, unreliable and untruthful results. Ethics with biases, hallucinations and incorrect information, or misjudgments, are some of the problems. We also shed a light on future directions, research questions on LLMs and recommendations.

Generative AI and LLMs are powerful technologies but come with significant drawbacks. Some questions concern the techniques themselves, but the societal impact is also not known. We have suggested several research questions that are but a small selection of ones that need to be investigated. Similarly, our recommendations are provided from some of the problems which the authors have seen during research in the area. Neither the research questions nor the recommendations are comprehensive, but we hope that we have highlighted some of primary ones to further the discussion in this emerging area.

References

[1] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, U., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2023). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and TechnologyVolume 15Issue 3Article No.: 39pp 1–45https://doi.org/10.1145/3641289.

[2] Huang, J., and Chang, K. C. C. (2022). Towards reasoning in large language models: A survey. arXiv preprint arXiv:2212.10403.

[3] Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., and Sauerland, U. (2023). Risks and benefits of large language models for the environment. *Environmental Science & Technology*, *57*(9) 3464-3466.

[4] Shin, J., Jian, L., Driscoll, K., and Bar, F. (2018). The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, *83*, 278-287.

[5] Wang, Y., McKee, M., Torbica, A., and Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, *240*, 112552.

[6] Wankhade, M., Rao, A. C. S., and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780.

[7] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

[8] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Tuan Luu, A., Bi, W., Shi, F., and Shi, S. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

[9] Håkansson, A., andHartung, R. L. (2020) *Artificial Intelligence: concepts, areas, techniques and applications*, Studentlitteratur, Sweden, ISBN: 9789144125992

[10] Filippini, M., and Lester C. H. (2011) "Energy demand and energy efficiency in the OECD countries: a stochastic demand frontier approach." *Energy Journal* **32** (2): 59–80.

[11] Filippini, M., and Lester C. H. (2012) "US residential energy demand and energy efficiency: A stochastic demand frontier approach." *Energy Economics* **34** (5): 1484–1491.

[12] Weyman-Jones, T, Mendonça Boucinha, J., and Feteira Inácio, C. (2015) "Measuring electric energy efficiency in Portuguese households: a tool for energy policy." *Management of Environmental Quality: An International Journal* **26** (3): 407–422.

[13] Sorrell, S. (2009) "The Rebound Effect: definition and estimation", in Joanne Evans and Lester Hunt (eds) *International Handbook on the Economics of Energy*, Cheltenham, Edward Elgar.

[14] Xu H., Zhengyan Z., Ning D., Yuxian G., Xiao L., Yuqi H.,Jiezhong Q., Yuan Y., Ao Z., Liang Z., Wentao H., Minlie H., Qin J., Yanyan L., Yang L., Zhiyuan L., Zhiwu L., Xipeng Q., Ruihua S., Jie T., Ji-Rong W., Jinhui Y., Wayne X. Z., and Jun Z. (2021) Pre-trained models: Past, present and future, AI Open, Volume 2, 2021, Pages 225-250, ISSN 2666-6510, https://doi.org/10.1016/j.aiopen.2021.08.002.

[15] Vaswani, A., Shazeer N., Parmar, N., Uszkoreit, J., Jones L., Gomez, A. N., Kaiser L., and Polosukhin I. (2017) "Attention is All You Need", https://arxiv.org/pdf/1706.03762.pdf.

[16] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014), Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, https://doi.org/10.48550/arXiv.1406.1078.

[17] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2023), A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, https://doi.org/10.48550/arXiv.2311.05232.

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio Y. (2014). Generative Adversarial Nets. Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014), pp. 2672–2680.

[19] Lawton G. (2024). What is generative AI? Everything you need to know. Techtarget Enterprise AI. https://www.techtarget.com/searchenterpriseai/definition/generative-AI.

[20] Stryker, C., and Scapicchio M. (2024). What is generative AI? https://www.ibm.com/topics/generative-ai, (Accessed 2024-06-01).

[21] Kingma, D. P., and Welling M. (2019). "An Introduction to Variational Autoencoders, Foundations and Trends® in Machine Learning: Vol. 12: No. 4, pp 307-392. http://dx.doi.org/10.1561/2200000056.

[22] Xu, W., Sun, H., Deng, C., and Tan, Y. (2017). Variational Autoencoder for Semi-Supervised Text Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). https://doi.org/10.1609/aaai.v31i1.10966.

[23] Kameoka, H., Li, L., Inoue, S., and Makino, S. (2019). Supervised Determined Source Separation with Multichannel Variational Autoencoder, in Neural Computation, vol. 31, no. 9, pp. 1891-1914, https://doi.org/10.1162/neco_a_01217.

[24] Brownlee, J. (2019). A Gentle Introduction to Generative Adversarial Networks (GANs). https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/, (Accessed 2024-06-04).

[25] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics (NAACL).

[26] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. (Accessed 2024-06-04).

[27] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871-7880.

[28] Håkansson, A., Hartung, R., Moradian, E., and Wu, D. (2010). Comparing Ontologies Using Multi-Agent System and Knowledge Base. Proceedings of the 14th international conference on KES: Part IV, DOI: 10.1007/978-3-642-15384-6_14

[29] Håkansson, A., (2018). Ipsum – An Approach to Smart Volatile ICT-Infrastructures for Smart Cities and Communities. Procedia Computer Science 126:2107-2116, DOI: 10.1016/j.procs.2018.07.241.

[30] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. https://arxiv.org/abs/2005.11401.

[31] Robert J. E., and Schmidt D. (2024, Jan 10). 10 Benefits and 10 Challenges of Applying Large Language Models to DoD Software Acquisition https://insights.sei.cmu.edu/blog/10-benefits-and-10-challenges-of-applying-large-language-models-to-dod-software-acquisition/.

[32] Li, Y., Liu, Y., Deng, G., Zhang, Y., Song, W., Shi, L., Wang, K., Li, Y., Liu, Y., and Wang, H. (2024). Glitch Tokens in Large Language Models: Categorization Taxonomy and Effective Detection. https://arxiv.org/abs/2404.09894.