Natural Language Processing Journal 6 (2024) 100048

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

A survey of GPT-3 family large language models including ChatGPT and GPT-4

Katikapalli Subramanyam Kalyan[1]

Akmmus AI, Trichy, Tamil Nadu, 620015, India

ARTICLE INFO

Keywords:
Large language models
LLMs
GPT-3
ChatGPT
GPT-4
Transformers
LLM survey

ABSTRACT

Large language models (LLMs) are a special class of pretrained language models (PLMs) obtained by scaling model size, pretraining corpus and computation. LLMs, because of their large size and pretraining on large volumes of text data, exhibit special abilities which allow them to achieve remarkable performances without any task-specific training in many of the natural language processing tasks. The era of LLMs started with OpenAI's GPT-3 model, and the popularity of LLMs has increased exponentially after the introduction of models like ChatGPT and GPT4. We refer to GPT-3 and its successor OpenAI models, including ChatGPT and GPT4, as GPT-3 family large language models (GLLMs). With the ever-rising popularity of GLLMs, especially in the research community, there is a strong need for a comprehensive survey which summarizes the recent research progress in multiple dimensions and can guide the research community with insightful future research directions. We start the survey paper with foundation concepts like transformers, transfer learning, self-supervised learning, pretrained language models and large language models. We then present a brief overview of GLLMs and discuss the performances of GLLMs in various downstream tasks, specific domains and multiple languages. We also discuss the data labelling and data augmentation abilities of GLLMs, the robustness of GLLMs, the effectiveness of GLLMs as evaluators, and finally, conclude with multiple insightful future research directions. To summarize, this comprehensive survey paper will serve as a good resource for both academic and industry people to stay updated with the latest research related to GLLMs.

1. Introduction

Large Language Models (LLMs), the recent buzz in Artificial Intelligence, have garnered a lot of attention in both academic and industry circles with their remarkable performances in most of the natural language processing (NLP) tasks. These models are essentially deep learning models, specifically transformer-based, pretrained on large volumes of text data and then aligned to human preferences using meta-training. Pretraining provides universal language knowledge to the model (Kalyan et al., 2021), while meta-training aligns the model to act based on the user's intentions. Here user's intention includes both explicit intentions, like following instructions, and implicit intentions, like maintaining truthfulness and avoiding bias, toxicity, or any harmful behaviour (Ouyang et al., 2022). Large language models (LLMs) are a special class of pretrained language models obtained by scaling model size, pretraining corpus and computation. For downstream task usage, PLMs leverage supervised learning paradigm, which involves task-specific fine-tuning and hundreds or thousands of labelled instances (Kalyan et al., 2021, 2022). LLMs leverage in-context learning (ICL), a new learning paradigm which does not require task-specific fine-tuning and a large number of labelled instances (Brown et al., 2020). LLMs treat any NLP task as a conditional text generation problem and generate the desired text output just by conditioning on the input prompt, which includes task description, test input and optionally, a few examples. Fig. 1 shows the evolution of artificial intelligence from machine learning to LLMs.

In the beginning, NLP systems are predominantly rule-based. These rule-based models are built on top of domain expert-framed rules. As manual rule framing is a laborious, expensive process and also requires frequent changes, rules-based models are gradually replaced by machine models, which learn the rules automatically from the training data and completely avoid manual rule framing (Kalyan et al., 2021). However, machine learning models require human intervention in the form of domain experts for feature engineering. The evolution of dense text vector representation models like Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), FastText (Bojanowski et al., 2017) and the advancement of computer hardware like GPUs, NLP systems are built using traditional deep learning models like CNN (Kalchbrenner et al., 2014), RNN (Salehinejad et al., 2017), LSTM (Hochreiter and Schmidhuber, 1997), GRU (Chung et al., 2014), Seq2Seq (Sutskever et al., 2014) and Attention-based Seq2Seq models (Bahdanau et al.,

E-mail address: kalyan@akmmusai.pro.
URLs: https://www.akmmusai.pro/kalyanksnlp, https://www.akmmusai.pro.
[1] NLP Researcher and Founder.

https://doi.org/10.1016/j.nlp.2023.100048
Received 16 October 2023; Received in revised form 5 December 2023; Accepted 10 December 2023

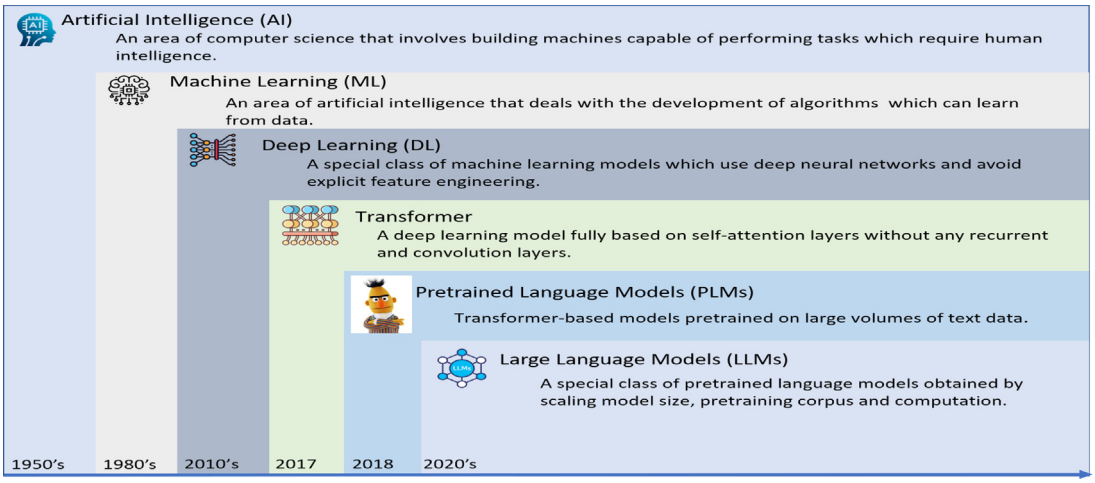

Fig. 1. Evolution of artificial intelligence from machine learning to LLMs.

2015; Luong et al., 2015). However, the drawbacks of these models like the inability to (i) capture long-term dependencies and (ii) leverage GPUs fully because of sequential processing (except in the case of CNN), resulted in the evolution of advanced deep learning models like Transformers (Vaswani et al., 2017), which are fully attention based without any recurrent and convolution layers.

Inspired by the success of image-pretrained models (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; Szegedy et al., 2015) built on top of transfer learning and large convolution models, the research community focused on building pretrained language models (PLMs) like BERT (Devlin et al., 2018) and GPT-1 (Radford et al., 2018) with transformers as the backbone and pretrained based on a new learning paradigm called self-supervised learning (Kalyan et al., 2021; Liu et al., 2021c; Gui et al., 2023). Unlike traditional deep learning models and vanilla transformers, which require training from scratch for downstream usage, PLMs can be easily adapted to downstream tasks with fine-tuning. The huge success of BERT and GPT-1 models triggered the development of other PLMs like RoBERTa, XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2019), ALBERT (Lan et al., 2019), DeBERTa (He et al., 2022a, 2020), GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020), BART (Lewis et al., 2020) etc.

Although PLMs have many advantages compared to traditional deep learning and vanilla transformer models, they still suffer from drawbacks like the inability to generalize to unseen tasks without task-specific training. So, the research community focused on developing more advanced models like LLMs which can generalize to unseen tasks without any task-specific training. The era of LLMs started with GPT-3 (Brown et al., 2020), and the success of GPT-3 inspired the development of other LLMs like PaLM (Chowdhery et al., 2022), Chinchilla (Hoffmann et al., 2022), GLaM (Du et al., 2022), LaMDA (Thoppilan et al., 2022), Gopher (Rae et al., 2021), Megatron–Turing NLG (Smith et al., 2022; Du and Cardie, 2020), BLOOM (Scao et al., 2022), Galactica (Taylor et al., 2022), OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023a,b) etc. The popularity of LLMs is increasing exponentially after the recent launch of Open AI's models like ChatGPT and GPT-4 (OpenAI, 2023). For example, ChatGPT has garnered millions of users within a few weeks of its launch. Because of the ability to generalize to unseen tasks based on the task description and a few examples without requiring any task-specific training, just like humans, LLMs can be considered as a baby step towards Artificial General Intelligence (Bubeck et al., 2023). In this survey paper, we mainly focus on Open AI LLMs like GPT-3 models, GPT-3.5 models (InstructGPT, ChatGPT etc.) and GPT-4, which we refer to as GPT-3 family large language models (GLLMs). This survey paper provides a comprehensive review of research works related to GLLMs in multiple dimensions.

Contributions. The key contributions of this survey paper are

- First survey paper to present a comprehensive review of GPT-3 family large language models (GLLMs) in multiple dimensions covering more than 350 recent research papers.
- We discuss various foundation concepts like transformers, transfer learning, self-supervised learning, pretrained language models and large language models.
- We discuss GPT-3 family large language models in detail, starting from GPT-3 to the latest ChatGPT and GPT-4.
- We discuss the performances of GLLMs in various downstream tasks and present a thorough discussion on the data labelling, and data augmentation abilities of GLLMs.
- We discuss the robustness and the evaluation abilities of GLLMs.
- We present multiple insightful future research directions which will guide the research community to improve the performances of GLLMs further.

Comparison with existing surveys. The existing survey papers provide a review of LLMs (Zhao et al., 2023d) and the relevant concepts like in-context learning (Dong et al., 2022), evaluation (Chang et al., 2023; Zhuang et al., 2023), alignment with human values (Wang et al., 2023o; Liu et al., 2023k), safety and trustworthiness (Huang et al., 2023c), reasoning (Huang and Chang, 2022), challenges and applications (Kaddour et al., 2023), LLM compression (Zhu et al., 2023a), prompting frameworks (Liu et al., 2023h), security risks (Derner et al., 2023), chain-of-thought prompting (Zhang et al., 2023i), open-source LLMs (Chen et al., 2023c) and multi-modal LLMs (Yin et al., 2023). For example, Zhao et al. (2023d) are the first to provide a comprehensive of LLMs. Unlike Zhao et al. (2023d), the other existing survey papers focus on specific concepts of LLMs. For example, the survey papers written by Dong et al. (2022), Chang et al. (2023), Wang et al. (2023o) and Huang and Chang (2022) focus on in-context learning, evaluation of LLMs, alignment of LLMs with human values and reasoning ability of LLMs respectively. Similarly, the survey papers written by Yin et al. (2023) and Huang et al. (2023c) provide a review of multi-modal LLMs and the safety and trustworthiness of LLMs, respectively. However, there is no existing survey paper which provides a comprehensive survey of GPT-3 family LLMs. With the ever-rising popularity of GPT-3 family LLMs like GPT-3, InstructGPT, ChatGPT, GPT-4 etc. and a lot of research works using these models, there is a strong need for a survey paper which focuses exclusively on GPT-3 family LLMs.

Papers collection. For this survey paper, we gathered over 350 research papers that appeared online in the period of June 2020 to September 2023. Initially, we selected GLLMs like GPT-3, InstructGPT, Codex and GPT-4 papers as seed papers and collected all the citing papers. We also collected papers from popular venues like ACL, EMNLP, COLING, AAAI, ICML, ICLR, NeurIPS etc. and popular databases like

Google Scholar and ScienceDirect using the keywords GPT-3, ChatGPT, GPT-3.5, InstructGPT, Codex and GPT-4. After removing the duplicate papers, we did a manual review to arrive at a final set of over 350 relevant research papers.

Survey paper organization. The survey paper is organized as follows: Section 2 presents a brief overview of various foundation concepts like transformers, transfer learning, self-supervised learning, pretrained language models and large language models. Section 3 presents GPT-3 family LLMs in detail, starting from GPT-3 to the latest ChatGPT and GPT-4. Sections 4, 5, and 6 discuss the performances of GLLMs in various downstream tasks, specific domains and multilingual scenarios, respectively. Section 7 presents the data labelling and data augmentation abilities of GLLMs. Section 8 discusses various research works presenting approaches to detect text generated by GLLMs. Sections 9, 10 and 11 discuss the evaluation, robustness and evaluation abilities of GLLMs, respectively. Section 12 presents multiple insightful future research directions.

2. Foundation concepts

2.1. Transformer

2.1.1. Traditional deep learning models

Before the evolution of the transformer model, most of the research in natural language processing involved deep learning models like multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM) network, gated recurrent unit (GRU), sequence-to-sequence and attention-based sequence-to-sequence (Young et al., 2018). MLP is a feed-forward neural network with three or more layers (input layer, one or more hidden layers, and output layer), and the neurons in these layers are fully connected. MLPs are easy to understand and simple to implement. However, as MLPs ignore the sequence information and struggle to capture the semantic relationships, these models are subsequently replaced by advanced models like CNN and RNN. CNN, originally developed to process images, is also explored for natural language processing tasks by treating text as a one-dimensional image (Kalchbrenner et al., 2014; Kim, 2014). CNNs can learn local features (n-grams) effectively using convolution layers but struggle to capture long-term dependencies. RNNs evolved as a deep learning model exclusively to process sequential data like text, time series, etc. (Salehinejad et al., 2017). RNNs can handle input with varying lengths and process sequential data by maintaining a hidden state to capture the context from previous inputs. However, RNNs suffer from vanishing gradients problems and struggle to capture long-term dependencies. LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Chung et al., 2014; Cho et al., 2014) evolved as advanced RNN variants to address the issues with the vanilla RNN model. The gating mechanism in these models helps to regulate the flow of information along the sequence and retain the most important information. Compared to LSTM, which includes three gates (input, forget and output gates), GRU is more parameter efficient as it includes only two gates, namely the input and the reset gates.

RNN and its variants like LSTM and GRU expect the input and output sequences to be the same length. However, in the case of natural language generation tasks like machine translation, text summarization, etc., the input and output sequences can be of different lengths. So, the researchers introduced the sequence-to-sequence (Seq2Seq) model to handle tasks with different input and output sequence lengths (Sutskever et al., 2014). The Seq2Seq model is originally developed for machine translation and later explored for other NLP tasks. The Seq2Seq model consists of an encoder and decoder based on RNN, LSTM or GRU to process the input sequence and generate the output sequence. The encoder processes the input sequence to generate a fixed-size context vector based on which the decoder generates the output sequence. However, the fixed-size context vector fails to encode the entire information in the input sequence, especially when the input sequence is long (Bahdanau et al., 2015). The attention mechanism is introduced to address this issue, allowing the decoder to focus on the relevant input tokens at each decoding step (Bahdanau et al., 2015; Luong et al., 2015). However, as the encoder and decoder of the Seq2Seq model are based on RNN and its variants, the Seq2Seq model suffers from vanishing gradients and struggles to capture long-term dependencies.

2.1.2. Drawbacks of traditional deep learning models

Here are the drawbacks of traditional deep learning models

- *Lack of sequence and semantic understanding* - MLPs ignore sequence information, treating all input tokens as independent. Moreover, MLPs can learn statistical patterns but struggle to capture semantic information in the input sequence.
- *Computationally expensive* - CNNs require a large number of parameters to achieve good results. Although LSTM and GRU address the limitations of vanilla RNNs to some extent, these models include a gating mechanism which significantly increases the number of model parameters. The large number of parameters makes these models computationally expensive to train and use.
- *Vanishing gradients* - RNN suffer from vanishing gradients problem. Although LSTM and GRU address this problem to some extent, these models also suffer from vanishing gradient problem and have difficulties in capturing long-term dependencies (Qiu et al., 2020; Kalyan et al., 2021).
- *Sequential Computation* - RNN and its variants process the input sequence token by token, i.e. sequentially . This sequential computation is a bottleneck for these models to leverage parallel computing capability in advanced computing hardware like GPUs and TPUs (Vaswani et al., 2017; Kalyan et al., 2021). This sequential computation also slows down training and inference processes, especially for long sequences.

2.1.3. Transformer description

The transformer model evolved as an effective alternative to traditional deep learning models and addressed most associated issues (Vaswani et al., 2017). In no time, the transformer model, with its novel and efficient architecture, gained a lot of popularity and became a de facto choice for building PLMs and LLMs using self-supervised learning paradigm (Kalyan et al., 2021; Zhao et al., 2023d). The key ingredient behind the massive success of the transformer model is its self-attention mechanism. The self-attention mechanism allows the transformer model to process the input sequence without using recurrent or convolution layers. When compared to convolution and recurrent layers, the self-attention mechanism can better capture long-range dependencies in the input sequence which makes the transformer model highly effective for natural language understanding and generation tasks. Although the self-attention mechanism is comparatively better, still it has difficulties in capturing long-range dependencies because of the quadratic complexity (both time and memory) (Tay et al., 2022; Li et al., 2023e). This drawback is later addressed in efficient transformer variants like Linformer (Wang et al., 2020a), Performer (Choromanski et al., 2020), Longformer (Beltagy et al., 2020) etc. For a detailed discussion on efficient transformer variants, refer to the survey paper by Tay et al. (2022).

In this paper, we present the description of the vanilla transformer model (Vaswani et al., 2017). The transformer consists of encoder and decoder components. The encoder processes the input text using a stack of encoder layers and then produces rich contextualized vector representations for each token in the input sequence, which are later used by the decoder. Each encoder layer consists of a self-attention mechanism and a feedforward neural network. The self-attention mechanism adds contextual information to the token vectors by allowing each token to attend to all other input tokens, and this helps the model capture long-term dependencies better. After the self-attention

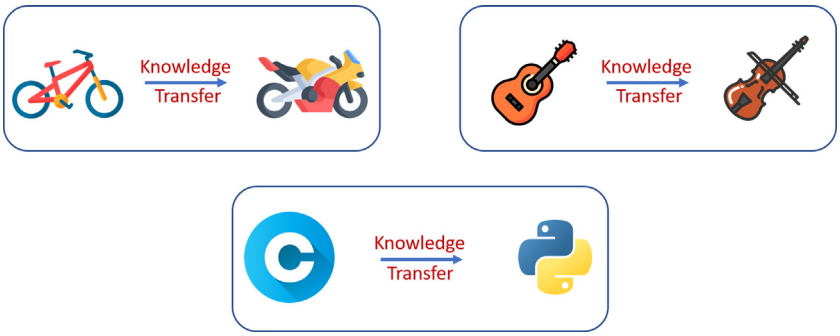

Fig. 2. Real-life examples of knowledge transfer (transfer learning).
Source: Examples are inspired from Zhuang et al. (2020).

mechanism, the token vectors are passed through a feedforward neural network, which introduces non-linearity and further transforms the representations. In this way, each encoder layer applies self-mechanism and feed-forward network to add more contextual information to the token vector representations.

The decoder receives the output from the last encoder layer and processes it sequentially by applying a stack of layers, with each decoder layer having masked self-attention, encoder–decoder self-attention and feed-forward neural network. The masked self-attention allows each token to attend to the previously generated tokens only and prevents the model from attending to future tokens. The encoder–decoder self-attention allows the decoder to attend to the encoded input sequence and helps the decoder focus on relevant input sequence tokens to generate the output tokens.

The self-attention mechanism in the Transformer uses multiple attention heads, which allow the model to learn different aspects of relationships between tokens and encode more contextual information in the token representations. The encoder and decoder layers also include the embedding layer, residual connections (He et al., 2016) and layer normalization (Ba et al., 2016). The embedding layer transforms input tokens into vector representations where each vector representation encodes both the meaning and position information. The residual connections and layer normalization are applied after the self-attention mechanism and feed-forward network. Residual connection (He et al., 2016) avoids vanishing gradients and ensures a smooth flow of gradients, while layer normalization (Ba et al., 2016) is applied to normalize the token representations and stabilize training. Apart from the embedding layer and stack of decoder layers, the decoder also includes an output layer. The output layer is nothing but a softmax layer that assigns probabilities to each token in the vocabulary, indicating the likelihood of each token being the next word in the generated sequence.

2.2. Transfer learning

2.2.1. Why transfer learning?

Although machine learning models tasted some success, these models require feature engineering, which is a laborious and expensive process involving human intervention in the form of domain experts (Kalyan et al., 2021). Deep learning models, essentially a subset of machine learning, do not require feature engineering as deep learning models learn features during training. Over the years, deep learning witnessed the evolution of various models like multi-layer perceptron (MLP), convolution neural networks (CNN), recurrent neural networks (RNN), long short-term memory networks (LSTM), gated recurrent unit networks (GRU), encoder–decoder networks, encoder–decoder with attention networks and recently transformers (Young et al., 2018; Otter et al., 2020). Even though deep learning models eliminated the requirement of manual feature engineering and achieved significant progress, the main drawback with these models is the requirement of a large amount of labelled data to achieve good results. Along with developing various deep learning models, the research community also focused on developing high-quality datasets for various tasks (Han et al., 2021). However, manual data annotation is a time-consuming, expensive and laborious process. Additionally, when there is a change in the data distribution, it is essential to re-train deep learning models with new labelled data to maintain good performances (Pan and Yang, 2009). To reduce the costs, the research community focused on how to effectively train deep learning models with limited labelled data. Transfer learning evolved as one of the effective solutions to train deep learning models with limited labelled data (Zhuang et al., 2020; Pan and Yang, 2009).

2.2.2. What is transfer learning?

Transfer Learning in the context of artificial intelligence involves existing knowledge transfer from one task (or domain) to another different but related task (or domain) (Zhuang et al., 2020; Pan and Yang, 2009). Transfer learning avoids training a model from scratch and helps improve the model's performance on the target task (or domain) by leveraging already existing knowledge. Transfer learning is largely based on the idea that when two tasks (or domains) are similar, the knowledge from the source task (or domain) with sufficient data can be used to enhance the performance of the target task (or domain) with limited data. For example, consider the task of sentiment analysis of reviews of different products. It is highly expensive to annotate large data separately for each product. In such cases, transfer learning helps to adapt the model trained on one product reviews to perform well on other product reviews without requiring large labelled data (Blitzer et al., 2007).

Transfer learning draws inspiration from human beings, i.e., human beings can do new tasks without or with few examples just by reusing previously gained knowledge (Han et al., 2021). Fig. 2 illustrates real-life examples of knowledge transfer (transfer learning). For example, a person who can cycle can learn to ride a bike quickly with less effort. This is because riding a cycle and a bike involves a lot of common things like handling the balance, etc. Similarly, a person familiar with C programming language can learn Python programming language easily. This is because both C and Python are programming languages and share many common concepts. So, due to the ability to reuse the existing knowledge and train the target models with limited data, transfer learning evolved as a promising learning paradigm and eventually played a crucial role in the evolution of advanced deep learning models like PLMs (Kalyan et al., 2021, 2022) and the recent LLMs. Overall, the advantages of transfer learning are

- Transfer learning helps to reduce the requirement of labelled data. (Data efficiency)
- Transfer learning avoids training models from scratch by providing a good initialization from existing related models. (Faster training and development)
- Transfer learning helps to enhance the performance on the target task (or domain) by reusing existing knowledge. (Enhance target task performance)

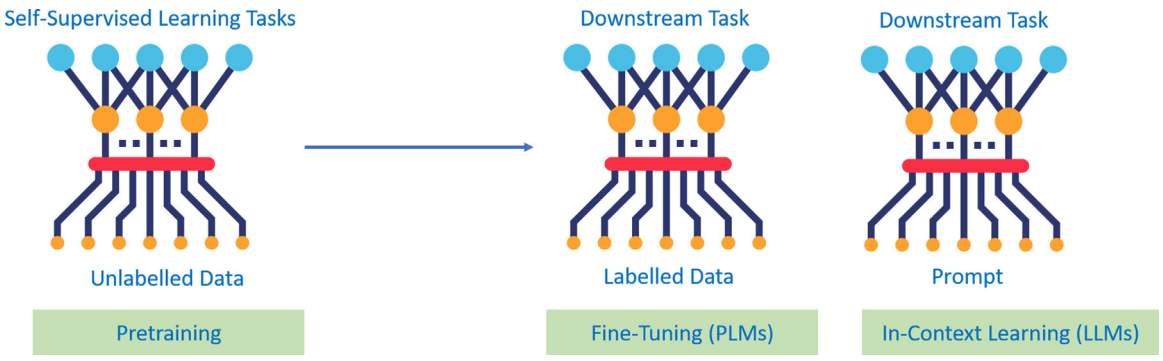

Fig. 3. Illustration of self-supervised learning paradigm.

- Transfer learning is explored across AI areas like computer vision, natural language processing, and speech processing. (Versatile)

In conclusion, transfer learning is a powerful learning paradigm in artificial intelligence that has benefits regarding data efficiency, speed, performance, adaptability, and real-world practicality.

2.2.3. Transfer learning vs. other learning paradigms

Along with transfer learning, the other learning paradigms that evolved to address large labelled data requirements are semi-supervised learning (Van Engelen and Hoos, 2020) and multi-task learning (Zhang and Yang, 2021). Semi-supervised learning is a learning paradigm in artificial intelligence that uses labelled and unlabelled data to train models (Van Engelen and Hoos, 2020). As semi-supervised learning uses labelled and unlabelled data, it lies between unsupervised and supervised learning paradigms. As semi-supervised learning uses only a small amount of labelled data, it reduces the amount of labelled data required, like transfer learning. However, unlike transfer learning, where the distribution of source and target tasks can be different, in semi-supervised, the distribution of labelled and unlabelled data should be the same (Zhuang et al., 2020). Multi-task learning is a learning paradigm which focuses on enhancing the performance of a group of tasks by leveraging the interconnections between the tasks and learning them simultaneously (Van Engelen and Hoos, 2020). Unlike multi-task learning, which simultaneously learns all the tasks, transfer learning first learns the source task and then transfers the knowledge to the target task. In multi-task learning, the focus is generally on all the tasks, while transfer learning focuses more on the target task (Pan and Yang, 2009).

2.3. Self-supervised learning (SSL)

2.3.1. Why self-supervised learning?

The main drawback with traditional deep learning models like CNN is the requirement of training from scratch. Training from scratch requires a large amount of labelled data. Data labelling is not only expensive but also a time-consuming and laborious process, which eventually makes the model development expensive. To reduce the requirement of labelled data and make the model development process less expensive, the computer vision research community focused on developing models like VGGNet (Simonyan and Zisserman, 2015), AlexNet (Krizhevsky et al., 2012) and GoogleNet (Szegedy et al., 2015) on top of large CNNs, transfer learning and supervised learning. These models are pretrained on a large number of labelled images from ImageNet dataset (Deng et al., 2009) using supervised learning, and then adapted to downstream tasks. These pretrained models avoid training downstream models from scratch by providing a good initialization. Moreover, downstream models initialized from pretrained models converge faster and achieve good results even with limited labelled data (Han et al., 2021).

Inspired by the huge success of pretrained image models, the NLP research community focused on developing PLMs (Han et al., 2021; Kalyan et al., 2021, 2022). However, the main challenge here is the use of supervised learning at scale to pretrain language models. This is because supervised learning at scale requires huge volumes of labelled data, which is almost impossible to obtain in many cases because of highly expensive annotation costs. Besides high annotation costs, supervised learning also suffers from generalization errors and spurious correlations (Kalyan et al., 2021; Gui et al., 2023). Self-supervised learning with the ability to automatically generate the labels and make use of unlabelled data evolved as an effective alternative to supervised learning to pretrain language models at scale (Liu et al., 2021c; Gui et al., 2023; Kalyan et al., 2021).

2.3.2. What is self-supervised learning?

Self-supervised learning, a promising learning paradigm in artificial intelligence, helps models from different modalities like language, speech or image to learn background knowledge from large volumes of unlabelled data (Liu et al., 2021c; Gui et al., 2023). Unlike supervised learning, which relies on large volumes of labelled data, SSL pretrains the models at scale based on the pseudo supervision offered by one or more pretraining tasks. Here, the pseudo supervision stems from the labels, which are automatically generated without human intervention based on the description of the pretraining task. In general, SSL involves one or more pretraining tasks (Kalyan et al., 2021, 2022). Moreover, the efficiency of SSL is heavily influenced by the choice of pretraining task (Kalyan et al., 2021; Clark et al., 2019; He et al., 2022a).

Fig. 3 presents the SSL paradigm. In the pretraining phase, the labels are automatically generated based on the description of pretraining tasks, and the models learn universal knowledge using the pseudo supervision offered by one or more pretraining tasks. Pretraining helps the models to gain strong background knowledge, which allows the models to provide a good initialization to downstream models. The initialization from pretrained models enhances the downstream models in terms of generalization, performance, and robustness and makes them data efficient. After pretraining, PLMs can be easily adapted to downstream tasks with limited labelled data, and LLMs can be used to solve downstream tasks using in-context learning without any task-specific fine-tuning.

2.3.3. Evolution of self-supervised learning

Fig. 4 shows the evolution of SSL in natural language processing from embedding models to the recent LLMs. The evolution of SSL in natural language processing happened in three stages, namely embedding models, PLMs and LLMs. Initially, SSL is explored to develop non-contextual embedding models (e.g. Word2Vec Mikolov et al., 2013, FastText Bojanowski et al., 2017), followed by sentence embedding (e.g. Sent2Vec Pagliardini et al., 2018) and contextual embedding models (e.g. ELMo Peters et al., 2018). The quest to develop pretrained models motivated NLP researchers to explore SSL to develop

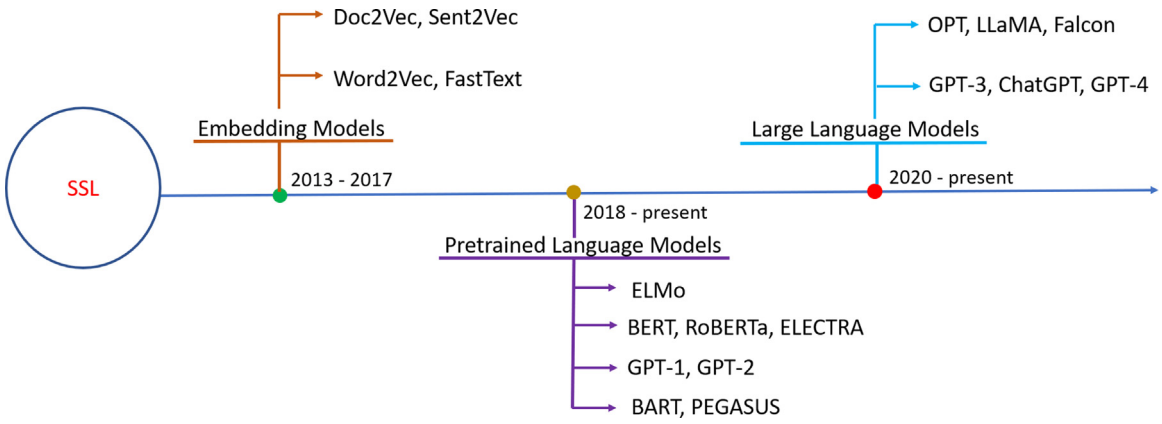

Fig. 4. Evolution of self-supervised learning in natural language processing.

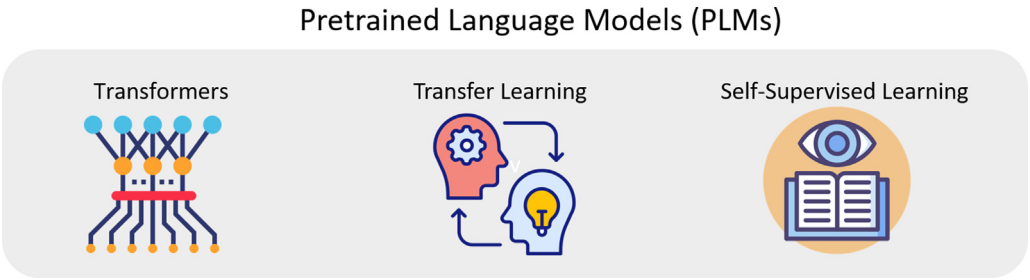

Fig. 5. Key ingredients in the evolution and success of PLMs.

PLMs (Kalyan et al., 2021, 2022; Han et al., 2021). As PLMs cannot generalize to NLP tasks without fine-tuning, the NLP research community focused on developing LLMs using SSL at a large scale (Brown et al., 2020; Touvron et al., 2023a,b; Anil et al., 2023; OpenAI, 2023). To summarize, self-supervised is undergoing a rapid evolution and is also treated as a significant element in achieving near human-level intelligence (Gui et al., 2023).

2.3.4. Self-supervised learning vs. other learning paradigms

Self-supervised learning, with its exceptional ability to make use of unlabelled data at scale, evolved as an alternative to supervised learning to pretrain models. However, SSL has similarities and dissimilarities with supervised learning (Kalyan et al., 2021). Both self-supervised and supervised provide supervision. However, unlike supervised learning, which offers supervision based on human-labelled data, SSL offers supervision based on automatically generated data. Supervised learning is mostly used to train downstream models with task-specific data, while SSL is used to train pretrained models to offer good initialization to downstream models. Similarly, SSL has similarities and dissimilarities with unsupervised learning (Kalyan et al., 2021). Both SSL and unsupervised learning make use of unlabelled data without requiring any labelled data. However, unlike SSL, which focuses on learning rich data representations using pseudo supervision, the main focus of unsupervised learning is to identify the hidden patterns in the data without any supervision.

2.4. Pretrained language models (PLMs)

2.4.1. Overview

Deep learning witnessed the evolution of several models, from convolution neural networks to the latest transformers (Young et al., 2018; Otter et al., 2020). Transformer addressed drawbacks of traditional deep learning models like convolutional neural network, recurrent neural network and its variants and achieved significant progress (Vaswani et al., 2017; Lin et al., 2022b). However, transformer and traditional deep learning models suffer from one major drawback: training from scratch, which requires large volumes of labelled data and makes model development expensive. Inspired by the success of pretrained image models like VGGNet (Simonyan and Zisserman, 2015), AlexNet (Krizhevsky et al., 2012) and GoogleNet (Szegedy et al., 2015) in computer vision, NLP researchers focused on developing pretrained models for natural language processing based on transformers and self-supervised learning (Kalyan et al., 2021, 2022; Han et al., 2021; Qiu et al., 2020). Pretrained language models are advanced deep learning models essentially transformer-based, pretrained on large volumes of text data and can be adapted to downstream tasks with limited labelled data. Along with transformer model, self-supervised learning and transfer learning are key concepts which make PLMs possible (Kalyan et al., 2021) (refer Fig. 5). The era of PLMs started with GPT-1 (Radford et al., 2018) and BERT (Devlin et al., 2018) models. The massive success of BERT and GPT-1 models triggered the development of other PLMs like RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2019), ALBERT (Lan et al., 2019), DeBERTa (He et al., 2022a, 2020), GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020), BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020) etc.

2.4.2. Evolution of pretrained language models

The evolution of PLMs happened along three dimensions: encoder-based models, decoder-based models and encoder–decoder based models (Kalyan et al., 2021). Encoder-based models consist of an embedding layer and stack of encoder layers, with each encoder layer having self-attention and feed-forward networks. Encoder-based models are primarily used for natural language understanding tasks like text classification, entity extraction, relation extraction, etc. Some of the popular encoder-based PLMs are BERT, RoBERTa, XLNet, ALBERT, ELECTRA, DeBERTa, etc. Decoder-based models consist of an embedding layer and a stack of decoder layers, with each decoder layer having self-attention, masked self-attention and feed-forward networks. Decoder-based models are used for both natural language understanding and

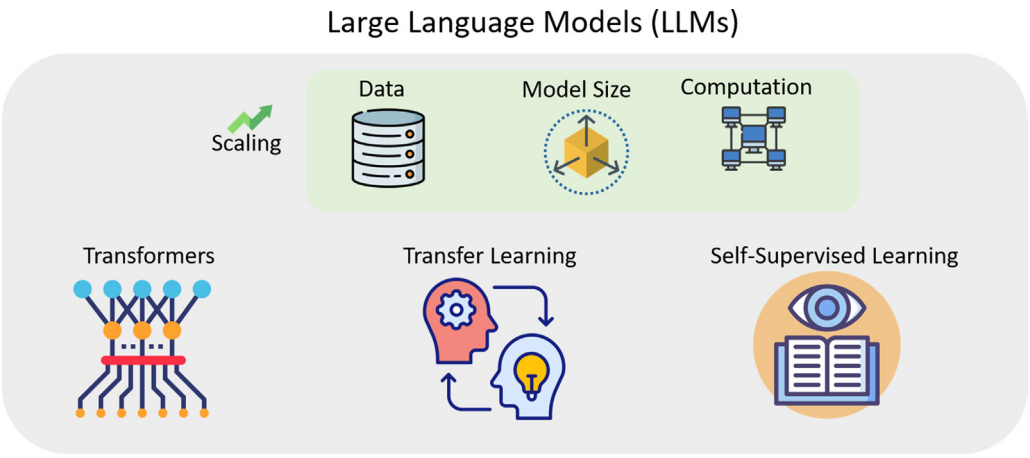

Fig. 6. Key ingredients in the evolution and success of LLMs.

generation tasks. Some of the popular decoder-based PLMs are GPT-1, GPT-2 etc. Encoder–decoder based models consist of both encoder and decoder modules. In general, encoder–decoder based models are used for natural language generation tasks like machine translation, text summarization, etc., while some are explored for both natural language understanding and generation tasks. Some of the popular encoder–decoder based models are T5, BART, PEGASUS, M2M100, NLLB, etc.

After the massive success of PLMs in the English language, the research community started to develop multilingual PLMs (Doddapaneni et al., 2021) and PLMs for non-English languages (Kalyan et al., 2021). Some of the popular multilingual PLMs are mBERT (Devlin et al., 2018), mT5 (Xue et al., 2021), mBART (Liu et al., 2020a), IndicBERT (Kakwani et al., 2020), XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), mDeBERTa (He et al., 2022a) etc. As the performance of general domain PLMs is limited in domain-specific tasks (Kalyan et al., 2021, 2022), the research community focused on developing PLMs for specific domains like social media (Nguyen et al., 2020; Barbieri et al., 2020), finance (Yang et al., 2020a; Araci, 2019; Liu et al., 2021a), legal (Chalkidis et al., 2020; Leivaditi et al., 2020), coding (Feng et al., 2020; Wang et al., 2021d, 2023d), healthcare (Lee et al., 2020; Gu et al., 2020; raj Kanakarajan et al., 2021) etc., As PLMs have millions of parameters which make model fine-tuning and deployment expensive, compact PLMs like DistilBERT (Sanh et al., 2019), TinyBERT (Jiao et al., 2020), MobileBERT (Sun et al., 2020), MiniLM (Wang et al., 2020b)etc., are developed. As PLMs have a limited context length which limits the performance on long sequences, long-sequence PLMs like LongFormer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020) etc., are developed. PLMs encode only the universal language knowledge available in the pretraining corpus and lack valuable knowledge available in ontologies. So, the research community developed ontology-enriched models like SapBERT (Liu et al., 2021), UmlsBERT (Michalopoulos et al., 2020), etc.

2.5. Large language models (LLMs)

2.5.1. Overview

The pretrained language models, starting from GPT-1 (Radford et al., 2018), BERT (Devlin et al., 2018) models to the latest DeBERTa (He et al., 2022a, 2020), achieved significant progress and also reduced the amount of labelled data required to train the task-specific models (Kalyan et al., 2021, 2022). Pretrained language models follow the paradigm "pretrain then fine-tune", i.e., the model is pretrained first and then adapted to downstream tasks by fine-tuning. As task-specific fine-tuning is mandatory to adapt the pretrained language model to downstream tasks, PLMs cannot generalize to unseen downstream tasks without task-specific fine-tuning. Moreover, task-specific fine-tuning requires labelled data and creates a separate copy of the pretrained language model for each downstream NLP task, increasing the model development and deployment costs (Kalyan et al., 2021).

Pretrained language models are treated as narrow AI systems as they are adapted through fine-tuning and then used for specific downstream tasks. However, the main focus of the research community is to develop artificial general intelligence systems (Goertzel, 2014; Bubeck et al., 2023) which are not narrowly focused on specific tasks but have the ability for general problem-solving and can handle even the unseen tasks by utilizing the existing knowledge like human beings. The NLP researchers observed that the performance of PLMs can be enhanced further through scaling along three dimensions: pretraining computation, pretraining data and model size (Liu et al., 2019; Radford et al., 2019; Raffel et al., 2020). Large size allows the models to capture more nuanced language patterns, which in turn enhances their ability to understand and generate text, while large pretraining data helps the model to learn from a wider range of text. The promising results from scaling and the quest to build artificial general intelligence systems motivated NLP researchers to build much bigger and bigger models, which eventually resulted in the evolution of GPT-3 and its successor models (Brown et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022; Du et al., 2022). Learning paradigms like transfer learning and self-supervised learning make LLMs possible, but scaling makes these models powerful.

The research community coined a new phrase, "large language models", to refer to GPT-3 and its successor large models to differentiate these models from small PLMs (Zhao et al., 2023d). Large language models (LLMs) are a special class of pretrained language models obtained by scaling model size, pretraining corpus and computation as shown in Fig. 6. LLMs are essentially deep learning models, specifically transformer-based, pretrained on large volumes of text data and aligned to human preferences using meta-training. Pretraining provides universal language knowledge to the model (Kalyan et al., 2021), while meta-training aligns the model to act based on the user's intentions. Here, the user's intention includes explicit intentions, like following instructions, and implicit intentions, like maintaining truthfulness and avoiding bias, toxicity, or harmful behaviour (Ouyang et al., 2022).

Because of their large size and pretraining on large volumes of text data, LLMs exhibit special abilities referred to as emerging abilities (Wei et al., 2022a; Schaeffer et al., 2023), allowing them to achieve remarkable performances without any task-specific training in many natural language processing tasks. For downstream task usage, PLMs leverage supervised learning paradigm, which involves task-specific fine-tuning and hundreds or thousands of labelled instances (Kalyan et al., 2021, 2022). LLMs leverage in-context learning (ICL), a new learning paradigm that does not require task-specific fine-tuning and many labelled instances (Brown et al., 2020; Dong et al., 2022).

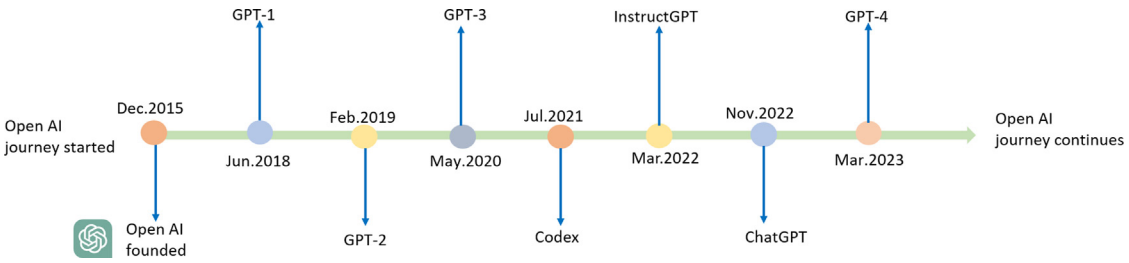

Fig. 7. Open AI journey starting from GPT-1 to the latest GPT-4.

LLMs treat any NLP task as a conditional text generation problem and generate the desired text output by conditioning on the input prompt, including task description, test input and optionally, a few examples.

2.5.2. Evolution of large language models

The evolution of LLMs happened along two dimensions: closed-source LLMs and open-source LLMs. The era of LLMs roughly started with GPT-3. Following the success of GPT-3, Open AI developed successor models like InstructGPT (Ouyang et al., 2022), Codex (Chen et al., 2021b), ChatGPT and GPT-4 (OpenAI, 2023). Google introduced models like GLaM (Du et al., 2022), PaLM (Chowdhery et al., 2022), PaLM2 (Anil et al., 2023), LaMDA (Thoppilan et al., 2022) and Bard. DeepMind developed models like Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022), AlphaCode (Li et al., 2022a) and Sparrow (Glaese et al., 2022). Companies like Baidu, AI21 labs and Amazon developed the models Ernie 3.0 Titan (Wang et al., 2021c), Jurassic-1 (Lieber et al., 2021) and AlexaTM (Soltan et al., 2022), respectively. Although the performances of closed-source LLMs are impressive, the main drawback with these models is that they are behind the paywalls, i.e., their weights are not publicly available, only some of them are accessible only through the APIs offered by the respective companies, and the model usage is charged based on the tokens processed and generated.

To address this issue, the research community focused on developing open-source LLMs with publicly available weights. Some of the popular open-source LLMs are OPT (Zhang et al., 2022), OPT-IML (Iyer et al., 2022), Galactica (Taylor et al., 2022), LLaMA (Touvron et al., 2023a), LLaMA2 (Touvron et al., 2023b) and Falcon. The performances of these open-source LLMs are on par with closed-source LLMs. Moreover, in some cases, open-source LLMs outperform closed-source LLMs. For example, Galactica beats closed-source LLMs like GPT-3, Chinchilla and PaLM. Inspired by the success of open-source LLMs in the English language, the research community focused on developing multilingual and bilingual LLMs. BLOOM (Scao et al., 2022) and BLOOMZ (Muennighoff et al., 2022) are examples of multilingual LLMs, JAIS (Sengupta et al., 2023) (English and Arabic), GLM (Zeng et al., 2022) (English and Chinese) and FLM-101B (Li et al., 2023j) (English and Chinese) are examples of bilingual LLMs.

The success of closed and open-source LLMs in the general domain triggered the development of domain-specific LLMs like FinGPT (Yang et al., 2023g) and BloombergGPT (Wu et al., 2023a) in the finance domain, MedPaLM (Singhal et al., 2023a) and MedPaLM2 (Singhal et al., 2023b) in the healthcare domain and StarCoder (Li et al., 2023a), CodeLlaMa (Rozière et al., 2023), CodeGen (Nijkamp et al., 2022) and CodeGen2 (Nijkamp et al., 2023) in the coding domains. For example, Bloomberg developed BloombergGPT, an exclusive LLM for the finance domain. Similarly, Google developed MedPaLM and MedPaLM2 LLMs exclusively for the healthcare domain based on PaLM and PaLM2 models respectively. Similarly, HuggingFace developed StarCoder, MetaAI developed Code LlaMA, and SalesForce developed CodeGen and CodeGen2 LLMs exclusively for coding tasks.

3. GPT-3 family large language models

3.1. Overview

Open AI, an AI company established in 2015, focused on building generative models. The Open AI researchers initially explored RNNs for developing generative language models (Radford et al., 2017). Inspired by the huge success of the transformer model and its ability to capture long-term dependencies, Open AI researchers leveraged the transformer decoder to build GPT-1 (117M parameters), the first-ever transformer-based pretrained language model (Radford et al., 2018). GPT-1 introduced a new paradigm, "pretrain and fine-tune", to develop downstream task models effectively. Originally, the "pretrain and fine-tune" paradigm was introduced by Dai and Le (2015) and then explored by Howard and Ruder (Howard and Ruder, 2018) to build language models for text classification. However, unlike Radford et al. (2018) work, these research works build language models based on LSTM, which lacks parallelization ability and has difficulties in capturing long-term dependencies. Radford et al. (2018) used casual language modelling as a pretraining task to pretrain the GPT-1 model. The casual language modelling pretraining task involves generating the next token based on the previous tokens. GPT-1 achieved SOTA results in 9 out of 12 NLP tasks (Radford et al., 2018).

Inspired by the success of GPT-1, Open AI researchers introduced the GPT-2 model to push the results further (Radford et al., 2019). The GPT-2 model is pretrained on the WebText corpus (40B text), which is much larger than the Books corpus used to pretrain the GPT-1 model. The authors developed four versions of the GPT-2 model with varying parameters: 117M, 345M, 762M and 1.5B. The authors observed that the perplexity decreases with an increase in the model's size, and even for the largest version of 1.5B, the decrease in perplexity did not exhibit saturation. This revealed that GPT-2 underfitted the pretraining dataset, and extending the training duration could have further reduced perplexity. This observation triggered the insight that "developing even larger language models will decrease the perplexity further and enhance natural language understanding and generation capabilities". The insights gained from the GPT-1 and GPT-2 models laid a strong foundation for the evolution of the GPT-3 family LLMs, including the latest models like ChatGPT and GPT-4. Fig. 7 shows the journey of Open AI starting from GPT-1 to the latest GPT-4 and Fig. 8 shows the GPT-3 family LLMs starting from GPT-3 series to the latest GPT-4.

3.2. GPT-3 models

The experiment results of GPT-2 showed that increasing the model size further reduces the perplexity, and the model with more parameters achieves better results than the models with fewer parameters. This observation motivated Open AI researchers to train much bigger GPT models, which eventually resulted in the introduction of the GPT-3 model (Brown et al., 2020). GPT-3 model contains 175B parameters and is 100 times bigger than its predecessor model, GPT-2. Moreover, the GPT-3 model is trained over a corpus with the text from multiple sources like webpages, Wikipedia and books, unlike GPT-1 and GPT-2

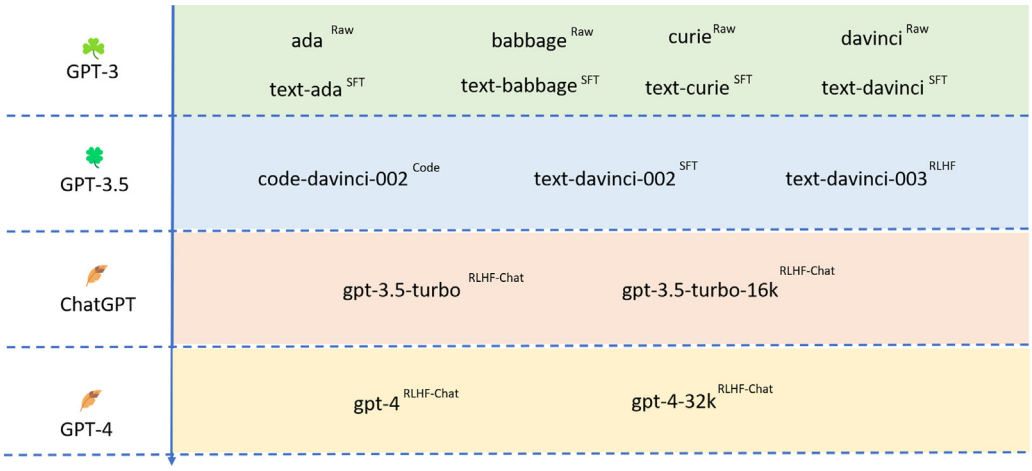

Fig. 8. GPT-3 family large language models (GLLMs) starting from GPT-3 series to the latest GPT-4. Here, SFT stands for supervised fine-tuning, and RLHF stands for reinforcement learning from human feedback. Here, raw represents that the model is just pretrained and is not aligned using SFT or RLHF. Here, RLHF-Chat represents that the model is aligned using RLHF and optimized for chat.

models, which are pretrained over corpora with the text from books and webpages, respectively. Scaling in three dimensions: pretraining data, model size, and pretraining computation allows the GPT-3 model to learn more from large volumes of texts from different sources, which eventually empowers the model to handle unseen tasks without any task-specific training. Unlike GPT-1 and GPT-2 models, which leverage supervised learning to do downstream tasks, GPT-3 leverages training-free in-context learning. In-context learning is a new learning paradigm that is training-free and solves the downstream tasks by using knowledge encoded in the model parameters (Dong et al., 2022). In-context learning accepts prompts as input where the input prompt consists of task descriptions, optimally few examples and other instructions.

3.3. GPT-3.5 models

Two main drawbacks of the GPT-3 model are (i) GPT-3 is not trained over code data, and hence, it lacks complex reasoning abilities like solving math problems (Zhao et al., 2023d), and (ii) GPT-3 model struggles to follow user instructions and sometimes generate harmful text (Ouyang et al., 2022). These two drawbacks are addressed by GPT-3.5 models. Brown et al. (2020) observed that GPT-3 can generate simple programs, although it is not specifically trained for generating code. The Open AI researchers triggered by this observation introduced Codex (Chen et al., 2021b), an exclusive GLLM for coding tasks. Codex is developed by fine-tuning a GPT model with 12B parameters over publicly available Github code. Moreover, it is observed that GPT models explicitly trained over code data exhibit better reasoning capabilities.

During pretraining, the GPT-3 model is optimized based on the casual language modelling objective, which involves predicting the next word based on the previous words. In-context learning during inference can be viewed as conditional text generation, where the model generates the output by conditioning on the given prompt. The model performs text generation during pretraining and inference, but it does vanilla text generation during pretraining and conditional text generation during inference. During pretraining, the model conditions on the previous words and generates the next word, i.e., vanilla text generation. However, during in-context learning, the model conditions on the prompt and generates the answer rather than generating the next words, i.e., conditional text generation. So, there is a gap between pretraining and in-context learning at inference. Due to this, in many cases during inference, the GPT-3 model fails to understand the given prompt and tends to generate the next words.

The pretraining corpus of the GPT-3 model includes some amount of text with undesired qualities like misinformation, abuse, hate, sexism, etc., due to which the model sometimes generates harmful text. To enhance complex reasoning ability, the instruction following ability and reduce the harmful text generation, GPT-3.5 models are developed by fine-tuning GPT-3 models over code data and then aligned using supervised fine-tuning (SFT) or reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). For example, the text-davinci-002 model is developed by fine-tuning the GPT-3 model (text-davinci) over code data to get code-davinci-002, which is further aligned using SFT.

3.4. ChatGPT and GPT-4

GPT-3 models are capable of understanding and generating natural language, while GPT-3.5 models are capable of understanding and generating both natural language and code. However, both GPT-3 and GPT-3.5 models are not chat optimized. This drawback is addressed by ChatGPT (GPT-3.5-turbo) and GPT-4 (OpenAI, 2023) models. Open AI introduced ChatGPT in November 2022. With extraordinary conversational abilities, ChatGPT, ChatGPT has garnered millions of users within a few weeks of its launch. Following ChatGPT, Open AI released the GPT-4 model in March 2023, which can handle both text and image inputs. Apart from generating text with human-like fluency, these models further pushed the results in many natural language processing tasks. The performance of these models in downstream tasks and specific domains is discussed in detail in Sections 5 and 6.

4. Performance of GLLMs in downstream tasks

4.1. Text classification

Overview. Text Classification is one of the fundamental tasks in natural language processing (Li et al., 2022c). It involves assigning label(s) from a predefined set of labels to a given piece of text. Here, the piece of text can be a phrase, sentence, paragraph or even a document. Many of the natural language processing problems, like offensive language identification, stance detection, sentiment analysis, hate speech detection, etc., are approached as text classification. Text Classification can be binary, multi-class or multi-label.

In the case of text classification, the large language model is prompted with a task description, a predefined set of labels, examples (optional) and the test input. Here, task description, a predefined set of labels and examples constitute the context. The model understands what actually the task is from the context and then assigns the most appropriate label(s) to the given test input. The additional inputs, like examples in the context, enrich the prompt with more information

Table 1
Summary of research works exploring GLLMs for various text classification problems. Here ZS represents zero-shot, and FS represents few-shot.

Paper	Task(s)	GLLMs explored	Prompt settings	Domain(s)	Language(s)	SOTA results
Zhang et al. (2023c)	Stance detection	ChatGPT	ZS, FS	Social media	English	No
Lamichhane (2023)	Stress detection, depression detection, suicidal detection	ChatGPT	ZS	Social media	English	No
Yang et al. (2023c)	Mental health analysis tasks	ChatGPT	ZS	Social media	English	No
Wang et al. (2023l)	Sentiment analysis	ChatGPT	ZS, FS	Social media	English, Chinese	No
Lopez-Lira and Tang (2023)	Stock prediction based on sentiment analysis	ChatGPT	ZS	Finance	English	No
Ziems et al. (2023a)	Computational social science tasks	GPT-3, ChatGPT	ZS	Social media	English	No
Kuzman et al. (2023)	Genre identification	ChatGPT	ZS	General	English, Slovenian	No
Bang et al. (2023)	Sentiment analysis, misinformation detection	ChatGPT	ZS	Social media	English, Indonesian, Javanese, Buginese	No
Kocoń et al. (2023)	Nine NLU tasks including sentiment analysis and natural language inference	ChatGPT	ZS	General, social media	English	No
Zhong et al. (2023)	Paraphrase detection, Sentiment analysis, Natural language inference	ChatGPT	ZS, FS	General	English	No
Ye et al. (2023)	Sentiment analysis, natural language inference	GPT-3, GPT-3.5, ChatGPT	ZS, FS	General, social media	English	No
Li et al. (2023k)	Financial news classification, sentiment analysis	ChatGPT, GPT-4	ZS	Finance	English	No
Wu et al. (2023c)	Natural language inference	ChatGPT, GPT4	ZS, FS	Healthcare	English	No
Wang et al. (2023n)	Natural language inference, document classification	GPT3.5, GPT4, Bard	ZS, FS	Healthcare	English	No
Chiu et al. (2021)	Hate Speech Detection	GPT-3	ZS, FS	Social media	English	No
Huang et al. (2023a)	Implicit hate speech detection	ChatGPT	ZS	Social media	English	No
Chen et al. (2023e)	Clinical text classification	GPT-3, ChatGPT, GPT-4	ZS, FS	Healthcare	English	No
Amin et al. (2023)	Sentiment analysis, suicide tendency detection, personality prediction	ChatGPT	ZS	Social media	English	No
Parikh et al. (2023)	Intent classification	GPT-3	ZS	Social media	English	No
Sun et al. (2023b)	News classification, sentiment analysis	InstructGPT	ZS, FS	General, social media	English	Yes

which allows the model to understand the task better and then perform better.

Research works exploring GLLMs for text classification. The recent works explored GLLMs like GPT-3, GPT-3.5 ChatGPT and GPT-4 for various text classification problems like sentiment analysis (Wang et al., 2023l; Lopez-Lira and Tang, 2023; Bang et al., 2023; Zhong et al., 2023; Li et al., 2023k; Amin et al., 2023; Sun et al., 2023b), stance detection (Zhang et al., 2023c), intent classification (Parikh et al., 2023), mental health analysis (Lamichhane, 2023; Yang et al., 2023c), hate speech detection (Chiu et al., 2021; Huang et al., 2023a), misinformation detection (Bang et al., 2023), paraphrase detection (Zhong et al., 2023), news classification (Li et al., 2023k), natural language inference (Zhong et al., 2023; Wu et al., 2023c; Wang et al., 2023n)etc. The evaluation is done in zero and few-shot settings using different prompting strategies like chain-of-thought (CoT) (Zhang et al., 2023c; Yang et al., 2023c; Zhong et al., 2023; Wu et al., 2023c; Wang et al., 2023n; Chen et al., 2023e; Sun et al., 2023b), self-question prompting (SQP) (Wang et al., 2023n), clue and reasoning prompting (CARP) (Sun et al., 2023b) etc. Most of the research works focused on English datasets, except a few research works focused on other languages like Chinese (Wang et al., 2023l), Slovenian (Kuzman et al., 2023), Indonesian (Bang et al., 2023), Javanese (Bang et al., 2023), and Buginese (Bang et al., 2023). A brief summary of research works exploring GLLMs for various text classification problems is presented in Table 1.

Most of the research works showed that compared to direct prompting, advanced prompting strategies help the model to achieve better results. This is because advanced prompting involves generating intermediate outputs, which in turn guide the model in generating the correct final output. Zhang et al. (2023c) explored the ChatGPT model with direct and chain-of-thought prompting for stance detection in tweets in zero and few-shot settings. Experiment results on three datasets showed that one-shot chain of thought prompting outperforms zero-shot direct prompting and also achieves near state-of-the-art results. Yang et al. (2023c) designed emotion-enhanced CoT prompting to combine emotion information with the power of CoT prompting for mental health analysis tasks. Experiments on five different mental health analysis tasks showed that ChatGPT with emotion-enhanced CoT outperforms other prompting strategies. Overall, ChatGPT outperforms traditional deep learning models like CNN and RNN but still lags behind task-specific fine-tuned models. Wu et al. (2023c) explored models like GPT-4 and ChatGPT for radiology natural language inference task. The authors reported that GPT-4 with IRSA prompting strategy outperforms ChatGPT in both zero and few-shot settings. IRSA stands for Instruction

Response Semantic Alignment. IRSA prompting strategy is almost the same as direct prompting except that in the case of IRSA prompting, the model is instructed to give the labels "contain" and "not contain" instead of "entailment" and "not entailment", just to reduce the complexity. Wang et al. (2023n) evaluated the performances of the latest LLMs like GPT-3.5, GPT-4, and Bard models on text classification tasks like natural language inference and document classification in the healthcare domain. The GPT-4 model with the newly designed self-question prompting (SQP) outperforms other models in both zero and few-shot settings. The SQP strategy involves identifying the key elements of input, generating questions and answers related to the key elements, and then using them to generate the final output. Parikh et al. (2023) showed that the performance of the GPT-3 model for intent classification in zero-shot settings can be enhanced by including intent class descriptions in the prompt.

Some of the research works demonstrated that GPT-3 family LLMs can outperform task-specific fine-tuned models (Kuzman et al., 2023; Zhong et al., 2023) and domain-specific LLMs (Li et al., 2023k). Kuzman et al. (2023) showed that ChatGPT outperforms fine-tuned XLM-R model in the task of automatic genre identification in the English language. Zhong et al. (2023) compared the performances of ChatGPT and fine-tuned models based on base and large versions of BERT and RoBERTa models on tasks like natural language inference, sentiment analysis and paraphrase identification. The results showed that ChatGPT outperforms both base and large fine-tuned models by a large margin in the case of natural language inference task. Li et al. (2023k) evaluated the performances of general LLMs like ChatGPT and GPT-4 and domain-specific LLMs like BloombergGPT on tasks like finance news classification and sentiment analysis. In the case of finance news classification, GPT-4 outperforms all other LLMs, including the domain-specific BloombergGPT model.

In all the above discussed research works, the performance of GLLMs is impressive but still lags behind SOTA results. Sun et al. (2023b) showed that it is possible to achieve SOTA results in text classification tasks with the newly designed clue And reasoning prompting (CARP) prompting strategy. CARP involves a progressive reasoning approach for handling complex linguistic phenomena, and it involves three steps: finding clues based on input, generating reasoning steps based on the input and the generated clues, and then arriving at the final output based on the input, generated clues and reasoning steps. Experiment results showed that the results are impressive as InstructGPT with CARP prompting strategy using just 16 examples achieves SOTA results on four text classification datasets.

4.2. Information extraction

Overview. Information Extraction (IE) in natural language processing involves extracting structured data like entities, relationships and events from unstructured text data (Lu et al., 2022b). Transforming unstructured text data into structured data enables efficient data processing, knowledge discovery, decision making and enhances information retrieval and search. Information extraction involves a number of tasks like entity typing, entity extraction, relation classification, relation extraction, event detection, event argument extraction and event extraction (Li et al., 2023d). Entity typing (ET) involves classifying identified named entity mentions into one of the predefined entity types (Chen et al., 2022). Named Entity Recognition (NER) or Entity Extraction (EE) involves identifying entity mentions and then assigning them to appropriate entity types (Das et al., 2022). Relation classification (RC) involves identifying the semantic relationship between the given two target entities in a sentence (Wu and He, 2019). Relation Extraction (RE) involves extracting the entities and then classifying the semantic relationship between the two target entities, i.e., involves entity extraction followed by relation classification (Ye et al., 2022). Event Detection (ED) aims to identify and categorize words or phrases that trigger events (Zhao et al., 2022a). Event Argument Extraction (EAE) involves identifying event arguments, i.e., entities involved in the event and then classifying their roles (Ma et al., 2022). Event Extraction (EE) aims to extract both the events and the involved entities, i.e., it involves event detection followed by event argument extraction (Du and Cardie, 2020).

Research works exploring GLLMs for information extraction tasks The recent works explored GPT-3 family LLMs for various information extraction tasks like entity typing (Li et al., 2023d), entity extraction (González-Gallardo et al., 2023; Hu et al., 2023a; Wei et al., 2023; Gutiérrez et al., 2022; Li et al., 2023d; Ma et al., 2023a; Wang et al., 2023i,n; Stammbach et al., 2022; Li et al., 2023k,g), relation classification (Gutiérrez et al., 2022; Li et al., 2023d; Chan et al., 2023; Xu et al., 2023e; Wan et al., 2023; Wang et al., 2023n; Zhang et al., 2023e), relation extraction (Wei et al., 2023; Rehana et al., 2023; Yuan et al., 2023b; Li et al., 2023d; Ma et al., 2023a; Wadhwa et al., 2023; Li et al., 2023g), event classification (Li et al., 2023d), event argument extraction (Li et al., 2023d) and event extraction (Wei et al., 2023; Gao et al., 2023e; Li et al., 2023d; Ma et al., 2023a). The evaluation is done in zero and few-shot settings using different prompting strategies like chain-of-thought (CoT) (Yuan et al., 2023b; Wan et al., 2023; Wang et al., 2023n; Wadhwa et al., 2023), self-verification (Wang et al., 2023i), self-question prompting (SQP) (Wang et al., 2023n), event ranking (ER) (Yuan et al., 2023b) etc. Most of the research works focused on English datasets, except a few research works focused on other languages like Chinese (Wei et al., 2023). A brief summary of research works exploring GLLMs for various information extraction tasks is presented in Table 2.

Hu et al. (2023a) demonstrated the performance of ChatGPT in extracting clinical entities like problem, treatment, and test can be enhanced by including additional information about entity types like synonyms and subtypes in the prompt. Wei et al. (2023) proposed ChatIE, a two-stage framework for information extraction, with each stage implemented as a multi-turn question answering. This two-stage framework helps the model break complex IE tasks into sub-tasks which allows the model to perform better. Results showed that ChatGPT used with the ChatIE framework outperforms vanilla ChatGPT by a large margin of more than 18 points. Gutiérrez et al. (2022) enhanced the performance of the GPT-3 model for entity extraction and relation classification by using techniques like contextual calibration (Zhao et al., 2021) to reduce bias and kNN-based demonstration selection. Gao et al. (2023e) examined the performance of ChatGPT for event extraction in few-shot settings. The model is prompted with task descriptions, definitions of event types, positive and negative examples, and test input. The authors reported that including negative examples decreases the performance of the model, which is in line with other existing works (Wang et al., 2022). The possible reason for this is that the model misunderstands negative examples as positive examples. Rehana et al. (2023) explored GPT-3 family models like GPT-3, ChatGPT and GPT-4 for protein–protein interaction extraction. It is reported that including normalized protein names in the prompt enhances the performance of the model. However, fine-tuned PubMedBERT model outperforms GPT-4 model with an F1-score of 86.47.

Yuan et al. (2023b) demonstrated that advanced prompting strategies like event ranking and chain-of-thought improve the performance of ChatGPT compared to vanilla prompting in temporal relation extraction. However, ChatGPT lags behind traditional neural networks like LSTM and fine-tuned pre-trained language models, which indicates the toughness of the temporal relation extraction task. Wang et al. (2023n) evaluated the performances of the latest LLMs like GPT-3.5, GPT-4, and Bard models on entity extraction and relation classification in the clinical domain. Experiment results showed that GPT-4 with self-question prompting outperforms other LLMs on most of the datasets. Li et al. (2023g) compared the performances of both natural language and code LLMs like GPT-3 and Codex using natural language and code style prompts. Experiment results showed that (i) Codex outperforms GPT-3 model and moderately sized fine-tuned models and (ii) Codex model

Table 2
Summary of research works exploring GLLMs for information extraction tasks. Here ZS represents zero-shot, and FS represents few-shot.

Paper	Task(s)	GLLMs explored	Prompt settings	Domain(s)	Language(s)	SOTA results
González-Gallardo et al. (2023)	Entity extraction	ChatGPT	ZS	General	English	No
Hu et al. (2023a)	Entity extraction	GPT-3, ChatGPT	ZS	Healthcare	English	No
Wei et al. (2023)	Entity extraction, event extraction, relation extraction	ChatGPT	ZS	General	English, Chinese	No
Gutiérrez et al. (2022)	Entity extraction, relation classification	GPT-3	FS	Healthcare	English	No
Gao et al. (2023e)	Event extraction	ChatGPT	FS	General	English	No
Rehana et al. (2023)	Protein–protein interaction extraction	GPT-3, ChatGPT and GPT-4	ZS	Healthcare	English	No
Yuan et al. (2023b)	Temporal relation extraction	ChatGPT	ZS	General	English	No
Li et al. (2023d)	Entity typing, entity extraction, relation classification, relation extraction, event detection, event argument extraction, event extraction	ChatGPT	ZS	General	English	No
Chan et al. (2023)	Temporal relation classification, causal relation classification, discourse relation classification	ChatGPT	ZS, FS	General	English	No
Xu et al. (2023e)	Relation classification	GPT-3.5	FS	General, scientific literature	English	Yes
Wan et al. (2023)	Relation classification	GPT-3.5	FS	General, scientific literature	English	Yes
Qin et al. (2023)	Entity extraction	GPT-3.5, ChatGPT	ZS	General	English	No
Ye et al. (2023)	Entity extraction, relation extraction	GPT-3, GPT-3.5, ChatGPT	ZS, FS	General, social eedia	English	No
Ma et al. (2023a)	Entity extraction, relation extraction and event detection	InstructGPT	FS	General	English	Yes
Wang et al. (2023i)	Entity extraction	GPT-3	FS	General	English	No
Wang et al. (2023n)	Entity extraction, relation classification	GPT-3.5, GPT-4	ZS, FS	Healthcare	English	No
Stammbach et al. (2022)	Entity extraction	GPT-3	ZS	General	English	No
Wadhwa et al. (2023)	Relation extraction	GPT-3	FS	General, healthcare	English	No
Li et al. (2023k)	Entity extraction	ChatGPT, GPT-4	FS	Finance	English	No
Li et al. (2023g)	Entity extraction, relation extraction	GPT-3, Codex	FS	General, scientific literature	English	No
Zhang et al. (2023e)	Relation classification	GPT-3.5, ChatGPT	ZS	General	English	No

with natural language or code style prompt outperforms GPT-3 model (iii) Code style prompts achieves better results in case of both Codex and GPT-3 models. The possible explanation for this is Codex which is pretrained over large volumes of code, encode structured code information which is useful for IE tasks as IE tasks involve structured outputs. Zhang et al. (2023e) proposed the QA4RE framework, which frames relation extraction as a question-answering problem. In the QA4RE framework, the sentence serves as context, and the relation types serve as options from which the LLMs choose. Experiment results showed that the proposed approach improves the performance of ChatGPT and GPT-3.5 models by a good margin in relation extraction.

Some of the research works (Xu et al., 2023e; Wan et al., 2023; Ma et al., 2023a) demonstrated that GPT-3 family models can achieve SOTA results in information extraction tasks. Wan et al. (2023) achieved SOTA results in relation extraction with the GPT-RE framework. GPT-RE framework overcomes the drawbacks in existing works using entity-aware demonstration retrieval based on fine-tuned model and gold label-induced reasoning. The use of representations from fine-tuned relation model for demonstration selection is more effective as they naturally include entity and relation information. Ma et al. (2023a) proposed a "filter then rerank" approach to use both fine-tuned models and LLMs to take advantage of the strengths of both models for few-shot information extraction. Here fine-tuned model acts as a filter while LLM acts as a re-ranker. The proposed approach achieves SOTA results with an average improvement of over 2 points in the F1 score.

4.3. Question answering

Overview. Question Answering (QA) is an important natural language processing task which deals with the development of algorithms to understand and interpret user queries in natural language and then deliver accurate responses (Zaib et al., 2022; Chali et al., 2011). The main aim of question answering systems is to enhance human–computer interaction, i.e., QA systems avoid the use of complex commands

and allow the user to interact with machines in a more natural way through natural language queries. For example, popular AI assistants like Amazon Alexa,[2] Google Assistant[3] and Apple Siri[4] rely on QA to provide accurate answers to user queries. The option of interaction through natural language queries enhances the reach of technology to a broader audience. QA can be treated as a fine-grained version of information retrieval (Torfi et al., 2020), and the demand for QA systems is increasing day by day because of the ability to generate answers which are accurate, relevant and short.

Research works exploring GLLMs for question answering tasks. The NLP research community explored GLLMs for question answering in various domains like education (Nunes et al., 2023; Joshi et al., 2023), news (Srivastava et al., 2022), healthcare (Samaan et al., 2023; Holmes et al., 2023a; Nori et al., 2023a; Hamidi and Roberts, 2023; Gupta et al., 2023; Tanaka et al., 2023b; Wang et al., 2023n; Weng et al., 2023; Kasai et al., 2023), social media (Ye et al., 2023), coding (Savelka et al., 2023), legal (Bommarito and Katz, 2022; Lin et al., 2022a), finance (Li et al., 2023k) and scientific literature (Pereira et al., 2023). Most of the research works focused on the English language, except a few research works focusing on languages like Portuguese (Nunes et al., 2023), Japanese (Tanaka et al., 2023b; Kasai et al., 2023) and Chinese (Weng et al., 2023). As advanced prompting methods allow GLLMs to perform well, some of the research works investigated the effectiveness of advanced prompting strategies like chain-of-thought (Nunes et al., 2023; Tan et al., 2023; Holmes et al., 2023a; Pereira et al., 2023; Wang et al., 2023n; Kasai et al., 2023), self-question prompting (Wang et al., 2023n; Weng et al., 2023) and holistically thought (Weng et al., 2023) for question answering. Table 3 presents a summary of research works exploring GLLMs for question answering across various domains and languages.

Zheng et al. (2023b) studied the shortcomings of ChatGPT in answering complex open-domain questions and found errors related to understanding, factual accuracy, specificity, and logical reasoning. They also analysed the importance of knowledge memorization, recall, and reasoning abilities in addressing these failures. The authors demonstrated that providing the model with external knowledge, cues for knowledge recall, and guidance for logical reasoning can enhance its ability to provide more accurate answers. Samaan et al. (2023) examined the accuracy of ChatGPT in answering questions related to Bariatric surgery. The authors reported that ChatGPT correctly answered 131 questions from 151 questions, i.e., ChatGPT achieves an accuracy of 86.8%. The impressive performance of ChatGPT shows that it can serve as an additional information resource in addition to healthcare professionals and reduce their burden in answering patient questions. Holmes et al. (2023a) compared the performances of GLLMs like ChatGPT, GPT-4 with other LLMs like Bard, BLOOMZ and medical physicists in answering related questions to Radiation Oncology Physics. The performance of GPT-4 is very impressive as the model outperforms medical physicists and other LLMs like ChatGPT, Bard and BLOOMZ. The performance of GPT-4 is further enhanced using CoT prompting, i.e., the model is prompted to arrive at the answer after step-by-step reasoning. Nori et al. (2023a) performed a comprehensive evaluation of the GPT-4 model on medical question answering in zero and few-shot settings. For evaluation, the authors used six datasets: two related to the United States Medical License Examination (USMLE) exam and four from the MultiMedQA benchmark (Singhal et al., 2023a). The performance of GPT-4 is very impressive as it outperforms not only general LLM like GPT-3.5 but also medical domain-specific LLM like Med-PaLM (Singhal et al., 2023a). Moreover, on USMLE exam datasets, GPT-4 model score is 20 points more than the passing score.

Hamidi and Roberts (2023) evaluated ChatGPT and Claude in answering patient-specific medical questions from MIMIC-III clinical notes. Experiment results demonstrated that the performances of both models are promising as these models display significant levels of coherence, accuracy, coverage and relevance in their answers. Li et al. (2023k) demonstrated that GPT4 achieves the best results for question answering in the finance domain and outperforms ChatGPT, domain-specific models like BloombergGPT, FinQANet and general LLMs like OPT (66B), and BLOOM (176B). Although the performance of GLLMs is impressive in zero and few-shot settings in multiple choice question answering, these models still lag behind SOTA results. The main reason for this is the use of cloze prompts. In cloze prompts, the model is prompted with only question without answer options, so the model generates the answers just by conditioning on the question. Robinson and Wingate (2022) proposed a new prompting strategy called multiple choice prompt which prompts the model with question and answer options so that the model generates the answer by conditioning on both question and answer options. Evaluation on 20 datasets showed that multiple-choice prompt helps GLLMs to achieve near SOTA results.

Some of the research works explored the effectiveness of GLLMs in answering exam questions from various domains. Nunes et al. (2023) investigated the performances of GLLMs like GPT-3.5, ChatGPT and GPT-4 in answering questions from the Brazilian university admission exam. Here all the questions are in Brazilian Portuguese language. The authors explored different prompting strategies like vanilla (zero-shot and few-shot) and CoT (few-shot). The authors observed that GPT-4 outperforms all other models by a large margin of over 11 points and achieves the best results with CoT prompting in few-shot settings. Joshi et al. (2023) evaluated ChatGPT in answering undergraduate-level computer science exam questions. For the evaluation, the authors gathered (i) questions from various computer science subjects like data structures, operating systems, machine learning and database management systems, (ii) questions from the GATE exam and (iii) programming questions from the Leetcode website. The results showed that ChatGPT is inconsistent in answering the questions, so students are not advised to rely on ChatGPT completely for their assignments and exams. Bommarito and Katz (2022) examined the ability of OpenAI's text-davinci-003 (GPT-3.5) model in answering multiple choice questions from the Bar Exam. Interestingly, human participants with extensive education and specialized training achieved a 68% accuracy rate, while the GPT-3.5 model achieved a lower accuracy rate of 50.3%. Gupta et al. (2023) evaluated how effective ChatGPT is in answering questions from plastic surgery inservice training examination. The authors reported that ChatGPT achieves an accuracy of 54.96% by correctly answering 242 questions. Tanaka et al. (2023b) evaluated the performances of GLLMs like GPT-3.5 and GPT-4 in answering questions from the Japanese National Medical Licensing Examination (NMLE). Here the input includes sample examples, instructions to translate the question into English, and then summarizing the question before answering. The authors reported that GPT-4 achieves a score better than the minimum passing score, and further analysis showed that the incorrect answers are due to insufficient medical knowledge and insufficient information about the Japanese-specific medical system. Kasai et al. (2023) reported that GPT-4 outperforms other models and passes the Japanese national medical licensing exam in the last six years. Moreover, ChatGPT with English-translated prompts achieves better results than ChatGPT with Japanese prompts. This is because ChatGPT is predominantly trained over the English text corpus.

Some of the research works explored GLLMs for more challenging tasks in question answering like tabular question answering (Srivastava et al., 2022), knowledge-based complex question answering (Tan et al., 2023), multiple choice code question answering (Savelka et al., 2023), multi-document question answering (Pereira et al., 2023) and conversational question answering (Weng et al., 2023). Srivastava et al. (2022) evaluated the effectiveness of GPT-3 for question answering on tabular data in zero and few-shot settings. Here the model is prompted with

[2] https://alexa.amazon.com.
[3] https://assistant.google.com.
[4] https://www.apple.com/in/siri/.

Table 3
Summary of research works exploring GLLMs for question answering tasks. Here ZS represents zero-shot, and FS represents few-shot.

Paper	Task(s)	GLLMs explored	Prompt settings	Domain(s)	Language(s)	SOTA results
Nunes et al. (2023)	Admission exam question answering	GPT-3.5, ChatGPT, GPT-4	ZS, FS	Education	Brazilian Portuguese	No
Tan et al. (2023)	Knowledge-based complex question answering	GPT-3, GPT-3.5, ChatGPT	ZS	General	Multiple languages	No
Yang et al. (2022)	Knowledge-based visual question answering	GPT-3	ZS	General	English	Yes
Srivastava et al. (2022)	Tabular question answering	GPT-3	ZS, FS	News	English	No
Zheng et al. (2023b)	Open domain question answering	ChatGPT	ZS	General	English	No
Samaan et al. (2023)	Bariatric surgery question answering	ChatGPT	ZS	Healthcare	English	No
Holmes et al. (2023a)	Radiation oncology physics question answering	ChatGPT, GPT-4	ZS	Healthcare	English	No
Joshi et al. (2023)	Computer science question answering	ChatGPT	ZS	Education	English	No
Nori et al. (2023a)	Medical question answering	GPT-3.5, GPT-4	ZS, FS	Healthcare	English	No
Hamidi and Roberts (2023)	Patient-specific question answering	ChatGPT	ZS	Healthcare	English	No
Bang et al. (2023)	Question answering	ChatGPT	ZS	General	English	Yes
Qin et al. (2023)	Boolean question answering	ChatGPT	ZS	General	English	No
Kocoń et al. (2023)	Multiple choice question answering	ChatGPT	ZS	General, social media	English	No
Ye et al. (2023)	Question answering	GPT-3, GPT-3.5, ChatGPT	ZS, FS	General	English	No
Savelka et al. (2023)	Multiple choice code question answering	GPT-3.5	ZS	Coding	English	No
Bommarito and Katz (2022)	Bar exam question answering	GPT-3.5	ZS	Legal	English	No
Pereira et al. (2023)	Multi-document question answering	GPT-3.5	FS	General, scientific literature	English	No
Gupta et al. (2023)	Plastic survey exam question answering	ChatGPT	ZS	Healthcare	English	No
Tanaka et al. (2023b)	Japanese medical exam question answering	GPT-3.5, GPT-4	FS	Healthcare	Japanese	No
Li et al. (2023k)	Financial question answering	ChatGPT, GPT-4	ZS	Finance	English	No
Wang et al. (2023n)	Medical question answering	GPT-3.5, GPT4	ZS, FS	Healthcare	English	No
Robinson and Wingate (2022)	Multiple choice question answering	GPT-3, Codex, InstructGPT	ZS	General	English	No
Weng et al. (2023)	Medical conversational question answering	GPT-3, InstructGPT	ZS	Healthcare	English, Chinese	No
Lin et al. (2022a)	Question answering	GPT-3	ZS	Multiple domains including Legal and Health	English	No
Kasai et al. (2023)	Japanese medical exam question answering	GPT-3, ChatGPT, GPT-4	FS	Healthcare	Japanese	No

unstructured passage text, tabular data in JSON format, examples (in the case of few-shot) and the question. The authors reported that GPT-3 displayed its ability to successfully locate the table, comprehend its structure, and accurately access the relevant cells or passages of text in order to provide answers to the given questions. Savelka et al. (2023) evaluated the effectiveness of GPT-3.5 models in answering multiple-choice questions (MCQs), particularly those involving code snippets from programming courses. Experiment results showed that MCQs with code snippets have lower success rates compared to those without code, indicating a challenge in answering multiple-choice questions with code snippets. Pereira et al. (2023) presented Visconde, a novel framework based on the GPT-3.5 model to tackle multi-document question answering. Visconde follows a three-step process involving decomposition, retrieval, and aggregation. The decomposition phase uses the GPT-3.5 model in few-shot settings for question simplification, the retrieval stage uses the SOTA model to select the relevant text chunks, and the final aggregation phase uses the GPT-3.5 with few-shot CoT prompting to get the answer. The authors observed that CoT prompting, i.e., generating reasoning steps before generating the final answer, enhances the performance. Weng et al. (2023) enhanced the

Table 4
Summary of research works exploring GLLMs for machine translation. Here ZS represents zero-shot, and FS represents few-shot.

Paper	GLLMs explored	Prompt settings	Domain(s)	Language(s)	Granularity	Outperforms Commercial Systems
Gu (2023)	ChatGPT	ZS	General	Japanese, Chinese	Sentence	No
Peng et al. (2023a)	ChatGPT	ZS	General, news, healthcare	English, Chinese, German, Romanian	Sentence	No
Jiao et al. (2023)	ChatGPT, GPT-4	ZS	General, healthcare,social media	English, Chinese, German, Romanian	Sentence	Yes
Hendy et al. (2023)	InstructGPT, ChatGPT, GPT-4	ZS, FS	News, social media, E-Commerce, dialogue	English, German, Chinese	Sentence, Document	Yes
Gao et al. (2023d)	ChatGPT	ZS, FS	General, news, social media, dialogue, E-Commerce	English, French, Spanish	Sentence	Yes
Wang et al. (2023h)	ChatGPT, GPT-4	ZS	General, social media, news, dialogue	English, German, Russian	Document	Yes
Zhu et al. (2023b)	ChatGPT	ZS, FS	General	102 languages in 202 directions	Sentence	No
Lyu et al. (2023b)	ChatGPT	ZS	General	English, Chinese, French	Paragraph	No
Bang et al. (2023)	ChatGPT	ZS	General	Twelve languages, including four low-resource languages	Sentence	No
Karpinska and Iyyer (2023)	GPT-3.5	ZS	General	18 language Pairs, including Japanese, English and Polish	Sentence, Paragraph	Yes
Moslem et al. (2023)	GPT-3.5	ZS, FS	General	English, Arabic, Chinese, German, Spanish	Sentence	Yes
He et al. (2023a)	GPT-3.5	ZS, FS	General	English, Chinese, Japanese, German, French	Sentence	No
Raunak et al. (2023b)	GPT-3.5, GPT-4	ZS	General	English, German, Chinese	Sentence	Yes
Raunak et al. (2023a)	GPT-3.5	ZS	General	English, German, Russian	Sentence	Yes

performance of GLLMs in answering medical conversational questions in English and Chinese using a novel prompt strategy called Holistically Thought (HoT). The HoT prompting strategy involves diffused thinking and focused thinking strategies to generate high-quality responses. Diffused thinking helps to generate various responses through diversified decoding, focused thinking generates a concise medical summary based on the dialogues and the final response is generated based on the dialogues, outputs of diffused thinking and focused thinking.

Unlike all the above discussed research works where the performances of GLLMs are just satisfactory but not SOTA, some of the research works (Yang et al., 2022; Bang et al., 2023) demonstrated that it is possible to achieve SOTA results for question answering task using GLLMs. For example, Yang et al. (2022) explored GPT-3 model for knowledge-based visual question answering. Knowledge-based visual question answering involves answering questions which require information which is not available in the input images. The authors propose a novel approach which uses GPT-3 as a knowledge source which is implicit and unstructured. Experiment results showed that the proposed approach achieves new SOTA results by outperforming existing approaches with a large margin of over 8 points.

4.4. Machine translation

Overview. Machine Translation (MT), an important task of natural language processing, deals with the development of models which can translate input text from the source language to the target language (Stahlberg, 2020; Yang et al., 2020b; Tan et al., 2020). MT models receive the input text in the source language, understand the syntax and semantics of the input text and then generate the translation in the target language. So, a good machine translation model should possess strong natural language understanding and generation skills to generate quality translations. The main objective of MT systems is to enhance cross-lingual communication by reducing the gap between individuals from different linguistic communities. The evolution of MT systems started with rule-based models followed by statistical and neural models (Tan et al., 2020). Rule-based MT systems are built on top of manually crafted syntactic and grammatical rules. As manually framing rules is heavily laborious and expensive, these systems are later replaced by statistical MT systems. Statistical MT systems use statistical models trained on bilingual data. With the evolution of deep learning models, the research community started to build neural machine translation (NMT) systems with the help of neural models (Sutskever et al., 2014; Bahdanau et al., 2014; Luong et al., 2015). These neural models are essentially based on the encoder–decoder architecture, where the encoder understands the input sequence and encodes it into a vector, and the decoder, based on the encoder output, generates the output sequence auto-regressively. Some of the recent neural models used for translation are mBART-50 (Tang et al., 2020), M2M100 (Fan et al., 2020), NLLB200 (Costa-jussà et al., 2022) etc.

Research works exploring GLLMs for machine translation. In recent times, GLLMs like ChatGPT and GPT-4 demonstrated remarkable performances in both natural language understanding and generation tasks. A good machine translation system requires strong natural language understanding and generation skills. As ChatGPT and GPT-4 possess strong natural language understanding and generation skills, the research community investigated the effectiveness of these models for machine translation across various domains like news (Peng et al., 2023a; Hendy et al., 2023; Gao et al., 2023d; Wang et al., 2023h), healthcare (Peng et al., 2023a; Jiao et al., 2023), social media (Jiao

Table 5
Summary of research works exploring GLLMs for keyphrase generation task. Here ZS represents zero-shot, and FS represents few-shot.

Paper	GLLMs explored	Prompt settings	Domain(s)	Language(s)	SOTA results
Martínez-Cruz et al. (2023)	ChatGPT	ZS	News, scientific literature	English	Yes
Song et al. (2023)	ChatGPT	ZS	Scientific literature	English	No

et al., 2023; Hendy et al., 2023; Gao et al., 2023d; Wang et al., 2023h), dialogue (Hendy et al., 2023; Wang et al., 2023h; Gao et al., 2023d) and e-commerce (Hendy et al., 2023; Gao et al., 2023d). Most of the research works focused on sentence-level machine translation (Gu, 2023; Peng et al., 2023a; Jiao et al., 2023; Hendy et al., 2023; Gao et al., 2023d; Zhu et al., 2023b; Bang et al., 2023; Karpinska and Iyyer, 2023; Moslem et al., 2023; He et al., 2023a; Raunak et al., 2023b,a), except a few research works focused on paragraph-level machine translation (Lyu et al., 2023b; Karpinska and Iyyer, 2023) and document-level machine translation (Hendy et al., 2023; Wang et al., 2023h). As advanced prompting methods allow GLLMs to perform well, some of the research works investigated the effectiveness of advanced prompting strategies like pivot (Jiao et al., 2023), chain-of-thought (Raunak et al., 2023b) and multi-aspect prompting and selection (He et al., 2023a). Table 4 presents a summary of research works exploring GLLMs for machine translation across various domains and languages.

Gu (2023) proposed a novel approach based on ChatGPT to enhance the quality of translation from Japanese to Chinese by effectively handling attribute clauses using a pre-edit scheme. The proposed approach, which integrates the pre-edit scheme with a novel two-step prompting strategy, enhances the translation quality by more than 35%. Peng et al. (2023a) explored the impact of temperature, task and domain information on the translation performance of ChatGPT. The authors showed that (i) ChatGPT performance degrades with an increase in temperature, and hence it is recommended to use a lower temperature (recommended is 0). and (ii) including task and domain information in the prompt enhances the performance of ChatGPT consistently for both high and low language translations. Zhu et al. (2023b) evaluated the performance of ChatGPT and other LLMs like OPT, BLOOM and XGLM on 102 languages in 202 translation directions. The authors reported that ChatGPT comprehensively outperforms other LLMs but still lags behind neural machine translation models like NLLB in the majority of the translation directions. Further analysis showed three errors, namely hallucination, monotonic translation and off-target translation. Lyu et al. (2023b) presented some interesting research directions with respect to using LLMs for machine translation. The presented interesting research directions include stylized machine translation, interactive machine translation and translation memory-based machine translation. Neural machine translation systems just focus on source–target text mapping, which results in a lot of errors. Unlike neural machine translation systems, the human translation process involves intermediate steps to ensure high translation quality. Inspired by the human translation process, He et al. (2023a) proposed MAPS, which involves three steps: knowledge mining, knowledge integration and knowledge selection to generate quality translations. Extension evaluation of the WMT22 test set shows that MAPS improves the performance of models like GPT-3.5 and Alpaca and also addresses the hallucination issue by resolving 59% of hallucination errors.

In all the above discussed research works, the performances of GLLMs are just satisfactory but not on par or beyond the performances of commercial machine translation systems. Some of the research works (Jiao et al., 2023; Hendy et al., 2023; Gao et al., 2023d; Wang et al., 2023h; Karpinska and Iyyer, 2023; Moslem et al., 2023; Raunak et al., 2023b,a) showed that it is possible to outperform commercial machine translation systems using GLLMs. For example, Jiao et al. (2023) investigated the translation capabilities of GLLMs like ChatGPT and GPT-4 and compared the performance with commercial systems like Google Translate, DeepL Translate and Tencent TranSmart. Extensive evaluation of multiple datasets showed that (i) the performance of GLLMs is on par with commercial systems in the case of high resources languages only, and (ii) the translation quality of low-resource languages can be enhanced using a novel pivot prompting strategy, which involves translating into high resource language before translating into the target low resource language. The naive prompts are unable to elicit the translation ability of ChatGPT fully. So, Gao et al. (2023d) focused on developing advanced prompting strategies by including additional information like task information, domain information and syntactic information like PoS (parts of speech) tags. The authors showed that ChatGPT, with the proposed advanced prompting strategy, achieves promising results and even outperforms commercial systems like Google Translate and DeepL Translate. Wang et al. (2023h) examined the performances of ChatGPT and GPT-4 for document-level machine translation and also compared the results with commercial systems from Google, DeepL and Tencent. The authors reported that GLLMs do well when the sentences in the document are combined and given at once to the model. Moreover, with this prompting strategy, both the GLLMs exhibit better performances than commercial machine translation systems according to human evaluation and also outperform most document-level neural machine translation methods in terms of d-BLEU scores. Karpinska and Iyyer (2023) explored the GPT-3.5 model for paragraph-level machine translation. The authors experimented with three different prompting strategies, namely translating sentence by sentence in isolation, translating sentence by sentence in the presence of the rest of the paragraph and translating the entire paragraph at once. After extensive evaluation of 18 language pairs, including English and Japanese, the authors report that translating the entire paragraph at once outperforms other strategies and commercial systems like Google Translate. Raunak et al. (2023a) examined the differences between the translations generated by GLLMs like GPT-3.5 and NMT systems like Microsoft Translator. The authors reported that GLLM generated translations are less literal, with better scores.

4.5. Keyphrase generation

Overview. Keyphrase generation (KPG) involves generating a set of phrases that capture the main ideas of a document (Meng et al., 2021). The primary advantage of KPG over keyphrase extraction is the ability to generate both extractive and abstractive keyphrases. Keyphrase generation is approached as a sequence-to-sequence generation task (Sutskever et al., 2014; Yuan et al., 2020; Kulkarni et al., 2022) in the existing works. The current state-of-the-art model for keyphrase generation is, KeyBART (Kulkarni et al., 2022), which is based on BART and trained using the text-to-text generation paradigm. Table 5 presents a summary of research works exploring GLLMs for keyphrase generation.

Research works exploring GLLMs for keyphrase generation. Martínez-Cruz et al. (2023) performed a comprehensive evaluation of ChatGPT as a keyphrase generator by evaluating its performance on six datasets using six candidate prompts. The authors reported that the results are promising, but ChatGPT struggles in the case of generating absent keyphrases. Song et al. (2023) evaluated ChatGPT on multiple datasets from news and scientific literature domains having both short and long documents. Experiment results showed that ChatGPT outperforms KeyBART (Kulkarni et al., 2022), the SOTA model, on all the datasets.

4.6. Dialogue tasks

Overview. Dialogue tasks in natural language processing (NLP) deal with understanding and generating human-like conversations between

Table 6
Summary of research works exploring GLLMs for various dialogue tasks. Here ZS represents zero-shot, and FS represents few-shot.

Paper	Task(s)	GLLMs explored	Prompt settings	Domain(s)	Language(s)
Pan et al. (2023)	Spoken language understanding and dialogue state tracking	GPT-3.5, ChatGPT	ZS	General	English
Zhao et al. (2023b)	Emotion dialogue understanding and generation tasks	ChatGPT	ZS, FS	General	English
Chintagunta et al. (2021)	Dialogue summarization	GPT-3	ZS	Healthcare	English
Bang et al. (2023)	Dialogue generation	ChatGPT	ZS	General	English
Qin et al. (2023)	Dialogue summarization	ChatGPT	ZS	General	English
Prodan and Pelican (2022)	Dialogue summarization	GPT-3	FS	General	English
Huynh et al. (2023)	Dialogue evaluation	GPT-3	FS	General	English
Fan and Jiang (2023)	Dialogue discourse analysis	ChatGPT	ZS, FS	General	English, Chinese
Wang et al. (2023k)	Dialogue question answering	ChatGPT	ZS, FS	General	English, Chinese

Table 7
Summary of research works exploring GLLMs for information retrieval tasks. Here ZS represents zero-shot, and FS represents few-shot.

Paper	Task(s)	GLLMs explored	Prompt settings	Domain(s)	Language(s)	SOTA results
Sun et al. (2023c)	Passage re-ranking	GPT-3, GPT-3.5, ChatGPT, GPT-4	ZS, FS	General, news, healthcare, scientific literature	English, ten low resource languages	Yes
Ziems et al. (2023b)	Document retrieval	GPT-3.5	ZS, FS	General	English	Yes

machines and users (Serban et al., 2018). The main objective of these tasks is to enable machines to have conversations with humans in a natural way. These dialogue tasks are essential components of building effective conversational agents, which have a wide range of applications, including customer support (Serban et al., 2018; Larson and Leach, 2022).

Research works exploring GLLMs for dialogue tasks. The research community explored GLLMs like GPT-3, GPT-3.5 and ChatGPT for various dialogue tasks like dialogue summarization (Chintagunta et al., 2021; Qin et al., 2023; Prodan and Pelican, 2022) , dialogue question answering (Wang et al., 2023k), emotion dialogue understanding and generation (Zhao et al., 2023b), dialogue state tracking (Pan et al., 2023), dialogue generation (Bang et al., 2023), and dialogue discourse analysis (Fan and Jiang, 2023). Some of the research works explored LLMs for the evaluation of dialogue tasks (Huynh et al., 2023). Most of the research works focused on general domain and English language datasets, except a few research works which focused on the medical domain (Chintagunta et al., 2021) and languages like Chinese (Fan and Jiang, 2023; Wang et al., 2023k). Table 6 presents a summary of research works exploring GLLMs for various dialogue tasks.

Pan et al. (2023) reported that ChatGPT exhibits better performance in dialogue state tracking compared to spoken language understanding. Further, the authors showed that the performance of ChatGPT can be enhanced by (i) using a multi-turn interactive prompt for dialogue state tracking and (ii) providing additional details like slot names, examples and descriptions for slot filling in spoken language understanding. Zhao et al. (2023b) explored the emotion dialogue capabilities of ChatGPT by evaluating the model on five different tasks, namely emotion recognition, emotion cause recognition, dialogue act classification (emotion dialogue understanding), empathetic response generation and emotion support generation. It is reported that ChatGPT exhibits better performances in emotion dialogue generation compared to emotion dialogue understanding. Chintagunta et al. (2021) showed that the in-house model trained on GPT-3 generated summaries achieves performances comparable to when trained on human-generated summaries. Further, the in-house model trained on mixed summaries (human-generated and GPT-3 generated) achieves better performances than those trained on either one of the summaries.

Prodan and Pelican (2022) proposed a scoring system to choose the best examples for dialogue summarizing using few-shot GPT-3. The proposed scoring system enhances the quality of generated summaries with an 11% reduction in failures. Huynh et al. (2023) studied the impact of various aspects influencing the performance of LLMs as Dialogue evaluators. The authors reported that the performance as a dialogue evaluator largely depends on the diversity and relevance of the datasets used for instruction tuning. Fan and Jiang (2023) investigated the effectiveness of ChatGPT for dialogue discourse analysis by evaluating its performance on three tasks, namely topic segmentation, discourse parsing and discourse relation recognition. ChatGPT's performance is promising in the case of topic segmentation, and CoT prompting enhances the performance. Wang et al. (2023k) proposed a novel approach based on explicit CoT prompting and demonstration selection to answer dialogue questions in few-shot settings.

4.7. Information retrieval

Information retrieval (IR) involves accessing and retrieving relevant information from large volumes of data. Here, the main objective is to provide users with the most relevant information by matching their queries to the content of documents and ranking them based on relevance (Anand et al., 2022). The process includes indexing, query formulation, search and retrieval, ranking, and presentation. Information retrieval is utilized in a wide range of fields, such as web search engines, digital libraries, e-commerce, healthcare, and scientific research (Anand et al., 2022). It plays a vital role in facilitating efficient and effective access to information in the modern digital era. Table 7 presents a summary of research works exploring GLLMs for information retrieval.

Sun et al. (2023c) explored the effectiveness of GPT-3 family models like GPT-3, GPT-3.5, ChatGPT and GPT-4 for passage re-ranking in information retrieval. The results are promising as GPT-4 outperforms SOTA models like monoT5-3B (Nogueira et al., 2020) on multiple benchmarks. Moreover, the compact model trained on ChatGPT-generated data demonstrates superior performance compared to the monoT5-3B model when evaluated on the MS MARCO dataset in BEIR (Thakur et al., 2021) benchmark. The existing approaches for document retrieval employ dual dense encoders, which encode query and document independently, resulting in shallow interaction between query and document (Zhao et al., 2022b). To overcome this drawback, Ziems et al. (2023b) proposed a novel approach which involves generating URLs using LLMs for document retrieval. The authors reported that document retrieval by generating URLs outperforms existing approaches.

4.8. Recommendation systems

Overview. Recommendation systems aim to reduce information overload and enhance the user experience by making relevant recommendations related to products or content based on user preferences

Table 8
Summary of research works exploring GLLMs for recommendation systems. Here ZS represents zero-shot, and FS represents few-shot.

Paper	GLLMs explored	Prompt settings	Domain(s)	Language(s)	SOTA results
Wang and Lim (2023)	GPT-3.5	ZS	Movies	English	No
Dai et al. (2023c)	GPT-3.5, ChatGPT	ZS, FS	News, books, movies, music	English	No
Gao et al. (2023c)	GPT-3.5, ChatGPT	ZS	Movies	English	No
Mysore et al. (2023)	InstructGPT	FS	Social media	English	No
Kang et al. (2023b)	GPT-3.5, ChatGPT	ZS, FS	Movies, books	English	No
Zhang et al. (2023a)	ChatGPT	ZS	Music, Movies	English	No
Liu et al. (2023e)	ChatGPT	ZS, FS	Beauty	English	Yes
Hou et al. (2023a)	ChatGPT	ZS	Movies, games	English	No
Zhiyuli et al. (2023)	ChatGPT	ZS, FS	Books	English	No

and behaviour (Adomavicius and Tuzhilin, 2005). In recent times, recommendation systems have gained immense popularity and are extensively utilized across a range of fields, such as entertainment, e-commerce, social media etc. For example, popular platforms like YouTube and Netflix use recommendation systems to suggest relevant videos and platforms like Amazon use recommendation systems to suggest relevant products to the user (Peng, 2022). The commonly used approaches for recommendation systems are based on collaborative filtering (Rezaimehr and Dadkhah, 2021), content-based (Xie et al., 2023a) and knowledge-based (Dong et al., 2020). The performance of traditional recommendation systems is limited by a number of issues like cold-start problem, poor generalization across domains and lack of explainability (Gao et al., 2023c; Zhu et al., 2021).

To overcome these drawbacks in traditional recommendation systems, recent works explored GPT-3 family LLMs for various tasks in recommendation systems like next item prediction (Wang and Lim, 2023), rating prediction (Gao et al., 2023c; Zhiyuli et al., 2023), top-k predictions (Gao et al., 2023c), direct recommendation (Liu et al., 2023e), sequence recommendation (Liu et al., 2023e) and generating explanations (Liu et al., 2023e). The evaluation is done in a variety of domains like movies (Wang and Lim, 2023; Dai et al., 2023c; Gao et al., 2023c; Kang et al., 2023b; Zhang et al., 2023a; Hou et al., 2023a), news (Dai et al., 2023c), books (Dai et al., 2023c; Kang et al., 2023b; Zhiyuli et al., 2023), music (Dai et al., 2023c; Zhang et al., 2023a), social media (Mysore et al., 2023), beauty (Liu et al., 2023e), and games (Hou et al., 2023a). Table 8 presents a summary of research works exploring GLLMs for recommendation systems.

Research works exploring GLLMs for recommendation systems. Wang and Lim (2023) proposed a novel prompting strategy called "Next-Item Recommendation (NIR)" to recommend movies using GLLMs. The proposed prompting strategy involves a three-step process to capture the user's preferences, choose representative movies they have watched in the past, and provide a ranked list of ten recommended movies. Dai et al. (2023c) reported that ChatGPT outperforms other GLLMs and is more effective with pair-wise and list-wise ranking compared to point-wise ranking. When it comes to balancing cost and performance, ChatGPT with list-wise ranking outperforms both point-wise and pair-wise ranking approaches. ChatGPT demonstrates the potential for providing explanations for recommendations and addressing the challenges of the cold start problem. Gao et al. (2023c) proposed Chat-REC, which leverages GLLMs to build conversational recommendation systems. The authors reported that Chat-REC performs well in tasks like top-k recommendations and zero-shot rating prediction. Moreover, Chat-REC enhances the conversational recommendation systems by making them more interactive and providing clear explanations.

Mysore et al. (2023) explored GLLMs like InstructGPT to generate synthetic data, and the experiment results showed that narrative-driven recommendation models trained on augmented datasets outperform LLM baselines and other approaches. Kang et al. (2023b) evaluated GLLMs like GPT-3.5 and ChatGPT on user rating prediction in zero and few-shot settings. Based on the experimental findings on datasets from movies and book domains, the authors reported that traditional models that have access to user interaction data perform better than GLLMs. Zhang et al. (2023a) introduced FaiRLLM, a new benchmark having eight sensitive attributes from domains like movies and music, to investigate the fairness of GLLM recommendations. The authors reported that GLLM-based recommendation systems are not fair to certain sensitive attributes.

Liu et al. (2023e) evaluated the performance of ChatGPT in five recommendation tasks, which include predicting ratings, direct recommendation, sequence recommendation, generating explanations, and summarizing reviews. Based on the evaluation of Amazon beauty datasets, the authors reported that (i) ChatGPT is much better in rating prediction compared to other tasks like direct and sequence recommendation. and (ii) ChatGPT achieves new SOTA results in generating explanations based on human evaluation. Hou et al. (2023a) demonstrated that GLLMs possess strong potential for zero-shot ranking tasks, showcasing performance that is comparable to or even superior to traditional recommendation models. Here, the authors designed the prompts in a way that important information like candidate items, sequential interaction history and ranking instruction is included. Zhiyuli et al. (2023) proposed BookGPT, a novel framework which leverages GLLMs like ChatGPT for book recommendation. Specifically, the performance of BookGPT is evaluated on three sub-tasks, namely the book rating task, book summary recommendation task and user rating recommendation task. The performance of BookGPT is promising in all three sub-tasks, and the performance increases with an increase in prompt examples.

4.9. Coding tasks

Overview. Software engineering is a discipline which deals with designing, developing, testing, and maintaining software systems (Hou et al., 2023b). To create software systems, software engineers use a variety of programming languages, development tools, and technologies. To aid software engineers and enhance their productivity, the research community focused on automating a number of coding tasks like code generation from natural language descriptions, code repair, code explanation generation, code hints generation, code completion, code document generation, test cases generation, code vulnerability detection, code refactoring, etc. The evolution of pre-trained source code models has paved the way for achieving cutting-edge results across coding tasks (Shi et al., 2023). Some of the popular pretrained source code models are CodeBERT (Feng et al., 2020), CodeGPT (Lu et al., 2021), CoTexT (Phan et al., 2021), GraphCodeBERT (Guo et al., 2020), CodeT5 (Wang et al., 2021d), CodeT5+ (Wang et al., 2023d), PLBART (Ahmad et al., 2021), PyCodeGPT (Zan et al., 2022) etc. Inspired by the success of GLLMs in NLP tasks, the research community focused on assessing the performances of these models in coding tasks also.

Research works exploring GLLMs for various coding tasks. The research community explored GLLMs for coding tasks across various languages like Java (Xia and Zhang, 2023; Cheshkov et al., 2023; Liu et al., 2023a; Khan and Uddin, 2022; Prenner and Robbes, 2021; Siddiq et al., 2023; Geng et al., 2023; Kang et al., 2023a; Destefanis et al., 2023; Yuan et al., 2023a), Python (Yetiştiren et al., 2023; Li et al., 2023l; Poldrack et al., 2023; Liu et al., 2023j; Chen et al., 2023b; Khan

Table 9
Summary of research works exploring GLLMs for various coding tasks. Here ZS represents zero-shot, and FS represents few-shot.

Paper	GLLMs explored	Task(s)	Prompt settings	Language(s)	SOTA results
Xia and Zhang (2023)	ChatGPT	Code repair	ZS, FS	Java	Yes
Cheshkov et al. (2023)	GPT-3, ChatGPT	Code vulnerability detection	ZS	Java	No
Yetiştiren et al. (2023)	ChatGPT	Code generation	ZS	Python	No
Li et al. (2023l)	ChatGPT	Finding failure-inducing test cases	ZS	Python	Yes
Liu et al. (2023a)	ChatGPT	Code generation	ZS	Java, C#	No
Poldrack et al. (2023)	GPT-4	Code generation, code refactoring, test case generation	ZS	Python	No
Liu et al. (2023j)	ChatGPT, GPT-4	Code generation	ZS	Python	No
Chen et al. (2023b)	ChatGPT	Code explanation generation	ZS	Python	No
Nascimento et al. (2023)	ChatGPT	Code generation	ZS	C++	No
Khan and Uddin (2022)	Codex	Code documentation generation	ZS, FS	Java, Python, PHP, GO, Ruby, JS	Yes
Leinonen et al. (2023)	GPT-3	Code explanation generation	ZS	C	No
Li et al. (2023i)	ChatGPT	Code generation	ZS	Python	Yes
Prenner and Robbes (2021)	Codex	Automatic code repair	ZS, FS	Python, Java	No
Siddiq et al. (2023)	Codex, ChatGPT	Unit test generation	ZS	Java	No
Tian et al. (2023)	ChatGPT	Code generation, APR, Code explanation generation	ZS	Python	No
Geng et al. (2023)	Codex	Code documentation generation	ZS, FS	Java	Yes
Kang et al. (2023a)	Codex, ChatGPT	Automate program repair	ZS	Python, Java	No
Kashefi and Mukerji (2023)	ChatGPT	Code generation	ZS	C, C++, Python, Julia, MATLAB	No
Destefanis et al. (2023)	GPT-3.5	Code generation	ZS	Java	No
Yuan et al. (2023a)	ChatGPT	Unit test generation	ZS	Java	No
Phung et al. (2023)	ChatGPT, GPT-4	Code repair, code completion, code explanation generation, coding hints generation	ZS	Python	No

and Uddin, 2022; Li et al., 2023i; Prenner and Robbes, 2021; Tian et al., 2023; Kang et al., 2023a; Kashefi and Mukerji, 2023; Phung et al., 2023), PHP (Khan and Uddin, 2022), GO (Khan and Uddin, 2022), Ruby (Khan and Uddin, 2022), JavaScript (Khan and Uddin, 2022), C (Leinonen et al., 2023; Kashefi and Mukerji, 2023), C++ (Nascimento et al., 2023; Kashefi and Mukerji, 2023), Julia (Kashefi and Mukerji, 2023), and MATLAB (Kashefi and Mukerji, 2023). Most of the research works focused on Python and Java languages, while a few research works focused on other languages like GO, PHP, GO, Ruby, JavaScript, C, C++, Julia and MATLAB. The assessment is done in zero and few-shot settings using mostly direct prompts. Table 9 presents a summary of research works exploring GLLMs for various coding tasks.

Some of the research works (Yetiştiren et al., 2023; Liu et al., 2023j; Nascimento et al., 2023; Kashefi and Mukerji, 2023; Destefanis et al., 2023) explored GLLMs for code generation task. Yetiştiren et al. (2023) compared various AI-assisted code generation tools like ChatGPT, Amazon's Code Whisperer and Github's Copilot on the Human Eval (Chen et al., 2021b) dataset. ChatGPT outperforms other tools by generating correct code 65.2% of the time, while the other tools generate correct code for a maximum of 46.3% of the time only. The test cases in existing datasets for code generation evaluation are limited in terms of quality and quantity. So, Liu et al. (2023j) proposed EvaPlus, a new framework for automatic test case generation using ChatGPT and the traditional mutation approach. The authors use EvaPlus to develop HumanEvalPlus on the top of the HumanEval (Chen et al., 2021b) dataset. The authors reported that HumanEvalPlus can detect a lot of incorrectly generated code that was previously undetected. Nascimento et al. (2023) compared the quality of code generated by ChatGPT and software developers for competitive coding problems on the LeetCode platform using various evaluation metrics. The authors reported that ChatGPT exhibits better performance compared to novice programmers but is outperformed by experienced programmers. Kashefi and Mukerji (2023) explored how effective ChatGPT is for generating code for numerical methods in five different programming languages: C, C++, Python, MATLAB and Julia. The authors observed that the results are promising but have some limitations which require further investigation. Destefanis et al. (2023) assessed the code generation ability of LLMs like Bard and GPT-3.5 by evaluating their performances in generating Java language code given the natural language descriptions. The authors observed that GPT-3.5 outperforms the Bard model by a large margin of more than 37%.

Some of the research works (Prenner and Robbes, 2021; Tian et al., 2023; Kang et al., 2023a; Phung et al., 2023) explored GLLMs for code repair task. Prenner and Robbes (2021) explored the Codex model for automatic program repair in Python and Java programming languages. The authors observed that the performance of Codex is comparable to state-of-the-art methods. Moreover, the Codex model is slightly better at fixing errors in Python language compared to Java language. Kang et al. (2023a) developed AutoSD, a novel framework for automatic program repair using GLLMs. The authors reported that the evaluation on three standard datasets showed that the proposed framework is on par with the baselines.

Unit tests generated using traditional approaches suffer from low readability (Yuan et al., 2023a). To address this drawback, some of the research works (Siddiq et al., 2023; Yuan et al., 2023a) explored GLLMs for test case generation. Siddiq et al. (2023) evaluated models like Codex and ChatGPT for unit test generation for Java code. Experiment results showed that Codex performs better with 80% coverage for the HumanEval dataset. However, both models perform poorly in the case of the SF110 benchmark, with less than 2% coverage. Yuan et al. (2023a) designed a ChatGPT-based unit test generation framework called "Chat-Tester". The iterative test refiner helps Chat-Tester to generate better unit tests compared to vanilla ChatGPT.

In all the above discussed research works, the performance of GLLMs in various coding tasks is promising but still lags behind SOTA results. Some of the research works (Xia and Zhang, 2023; Li et al.,

2023l; Khan and Uddin, 2022; Li et al., 2023i; Geng et al., 2023) demonstrated that GLLMs can achieve SOTA results in coding tasks. Xia and Zhang (2023) proposed ChatRepair, an automatic program repair tool based on ChatGPT. ChatRepair achieves remarkable performance, surpassing all the existing methods. It successfully resolves 114 and 48 bugs on Defects4j 1.2 and 2.0 (Just et al., 2014), respectively, outperforming the previous best by 15 and 17 bugs, respectively. Khan and Uddin (2022) explored Codex, GPT-3 family model pretrained on natural and programming languages to automate code documentation generation. The evaluation results on six programming languages showed that Codex, with just one example, outperforms existing approaches by a large margin of 11.2%. Geng et al. (2023) explored Codex for code document generation and demonstrated that few-shot in-context learning with systematic demonstration selection helps the GPT-3 model to achieve new SOTA results on two standard datasets related to Java language.

Some of the research works (Li et al., 2023l; Liu et al., 2023a; Li et al., 2023i) explored advanced prompting like CoT, brainstorming, differential prompting, etc., for coding tasks. Liu et al. (2023a) evaluated the code generation capabilities of ChatGPT by evaluating its performances on text-to-code and code-to-code generation tasks on CodeXGLUE (Lu et al., 2021) datasets. The authors observed that advanced prompting strategies like CoT enhance the code generation capabilities of models like ChatGPT. Li et al. (2023i) proposed Brainstorm, a new framework for code generation. Brainstorm involves three steps: brainstorming to generate diverse thoughts, thoughts selection to select the best thought using a ranking model and writing code to generate the code based on the problem statement and the best thought. The authors reported that the proposed framework helps ChatGPT to increase its performance by more than 50% and achieve new SOTA results on the CodeContests (Li et al., 2022a) benchmark. Li et al. (2023l) showed that directly using ChatGPT to find failure-inducing test cases results in poor performances. So, the authors proposed a new prompting strategy called "Differential Prompting", which enables ChatGPT to achieve new SOTA results on the Quixbugs dataset (Lin et al., 2017). Differential Prompting involves program intention inference followed by two more steps: program generation and differential testing.

4.10. Multimodal AI tasks

Overview. Traditional AI systems are designed to handle data from a single modality such as text, image, audio or video. As real-world data is often multi-modal, researchers focused on developing multi-modal AI systems which can leverage input data from multiple modalities to generate more accurate results. Multi-modal AI systems leverage techniques from different areas of AI, like natural language processing, computer vision, speech processing etc., to process multi-modal input data effectively (Sundar and Heck, 2022; Xu et al., 2023d). Multi-Modal AI systems can perform a variety of understanding and generation tasks like visual question answering (Shao et al., 2023; Lin et al., 2022c; Yang et al., 2022; Gui et al., 2022), text-to-image generation (Lu et al., 2023b; Zhu et al., 2023c; Zhang et al., 2023k), text-to-video generation (Hong et al., 2023), text-to-speech synthesis (Huang et al., 2023b), speech-to-text synthesis (Huang et al., 2023b), image captioning (Ranjit et al., 2023) etc.

Research works exploring GLLMs for Multimodal AI tasks. After the huge success of LLMs in natural language generation and understanding tasks, the research community recently explored GPT-3 family models in multi-modal understanding and generation tasks in various combinations like image+language (Kalakonda et al., 2022; Wu et al., 2023b; Shao et al., 2023; Yang et al., 2023e; Ranjit et al., 2023; Lin et al., 2022c; Lu et al., 2023b; Zhu et al., 2023c; Li et al., 2023f; Hakimov and Schlangen, 2023; Yang et al., 2022; Feng et al., 2023; Zhang et al., 2023k; Fan et al., 2023; Li et al., 2023h; Gui et al., 2022), video+language (Bhattacharya et al., 2023; Hong et al., 2023), audio+language (Mei et al., 2023; Zhang et al., 2023h). Most of the research works focused on general domain datasets, which some of the research works focused on specific domains like healthcare (Ranjit et al., 2023; Li et al., 2023h). Table 10 presents a brief summary of research works exploring GLLMs for various multimodal AI tasks.

Some of the research works developed multi-model AI systems for a specific task like action generation (Kalakonda et al., 2022), knowledge-based visual question answering (Shao et al., 2023; Lin et al., 2022c; Yang et al., 2022; Gui et al., 2022), X-ray report generation (Ranjit et al., 2023), named entity recognition (Li et al., 2023f), text-to-video generation (Hong et al., 2023), layout generation (Feng et al., 2023), text-to-image generation (Zhang et al., 2023k). Kalakonda et al. (2022) proposed GPT-3 based plug-and-play framework called Action-GPT for text-based action generation. Here, the authors generated multiple detailed body movement descriptions from the action phrases and then used them to generate actions. Shao et al. (2023) proposed Prophet, which avoids using an external knowledge base by using GPT-3 as an implicit knowledge base and includes vanilla visual question answering to provide answer heuristics to GPT-3. The answer heuristics, along with caption and question information, provide rich task-specific information to the GPT-3 model, which results in much better performances. Ranjit et al. (2023) proposed automatic X-ray report generation based on contrastively pretrained vision-language encoder and GPT-3 family models like GPT-3.5, ChatGPT and GPT-4. The contrastively pretrained encoder is used to encode input X-ray image into image vector embedding based on which the most similar sentences from the radiology report corpus are retrieved. The retrieved similar sentences form the context and allow LLM to generate a quality X-ray report. Li et al. (2023f) proposed PGIM, a two-stage approach which utilizes ChatGPT as an implicit knowledge base for multi-modal NER task. In the first stage, ChatGPT, when prompted with text descriptions of the image, generates the auxiliary knowledge. In the second stage, the downstream model receives the raw text and ChatGPT-generated auxiliary knowledge as input. The authors reported that the proposed approach outperforms existing SOTA approaches based on text-text and text-image paradigms.

Hong et al. (2023) proposed DirecT2V for text-to-video generation, which leverages GPT-4 model as a frame-level director. Here, the GPT-4 model generates descriptions for each frame based on a single prompt, and then the Text-to-Image model is used to generate frames based on these descriptions. Feng et al. (2023) developed LayoutGPT, which leverages LLM and Layout-to-Image models to generate 2D and 3D planning layouts from text descriptions. Zhang et al. (2023k) proposed "Control-GPT" based on LLMs and diffusion models for controllable text-to-image generation. Here, GPT-4 generates sketches based on Tikz code based on the text instructions, and then diffusion model generates realistic images with generated sketches and the text instructions as input. Here, the generated sketches help diffusion models to get a better idea about spatial relationships.

Some of the research works focused on developing multi-model AI systems which can handle multiple tasks (Wu et al., 2023b; Yang et al., 2023e; Bhattacharya et al., 2023; Hakimov and Schlangen, 2023; Zhao et al., 2023a; Huang et al., 2023b). As ChatGPT is trained on one data modality i.e., text data, ChatGPT can only handle text inputs and training models from scratch for vision-language tasks, is not a feasible option as it involves huge computation. So, Wu et al. (2023b) developed Visual ChatGPT based on ChatGPT and various visual foundation models to handle 22 vision language tasks. Bhattacharya et al. (2023) proposed a novel three-stage approach to handle five video understanding tasks. The proposed approach involves transforming video into text stories and then using this text content for video understanding tasks. Hakimov and Schlangen (2023) explored GPT-3 model for five vision language tasks, including four classifications and one question answering. Here the model is prompted with text description of the input image along with other elements like task instruction and similar examples. Huang et al. (2023b) proposed AudioGPT, which allows ChatGPT to handle multiple audio understanding and generation tasks with the help of audio foundation models.

Table 10
Summary of research works exploring GLLMs for various multimodal AI tasks. Here ZS represents zero-shot, and FS represents few-shot.

Paper	GLLMs explored	Task(s)	Prompt settings	Multimodality	Domain
Kalakonda et al. (2022)	GPT-3	Text-based action generation	ZS	Image + Language	General
Wu et al. (2023b)	ChatGPT	Twenty two vision language tasks	ZS	Image + Language	General
Shao et al. (2023)	GPT-3	Knowledge-based visual question answering	FS	Image + Language	General
Mei et al. (2023)	ChatGPT	Audio labelling	ZS	Audio + Language	General
Yang et al. (2023e)	ChatGPT	Multi-image reasoning, multi-hop document understanding, open-world concept understanding, video summarization	ZS	Image + Language	General
Ranjit et al. (2023)	GPT-3.5, ChatGPT, GPT-4	Chest X-ray report generation	ZS	Image + Language	Healthcare
Lin et al. (2022c)	GPT-3	Knowledge-based visual question answering	FS	Image + Language	General
Bhattacharya et al. (2023)	GPT-3.5	Five video understanding tasks	ZS	Video + Language	General
Zhang et al. (2023h)	GPT-4	Generate instructions	ZS	Audio + Language	General
Lu et al. (2023b)	GPT-3.5, GPT-4	Evaluator for text-to-image generation	ZS	Image + Language	General
Zhu et al. (2023c)	GPT-3, GPT-3.5	Editing in text-to-image generation	FS	Image + Language	General
Li et al. (2023f)	ChatGPT	Multimodal named entity recognition	FS	Image + Language	General
Hakimov and Schlangen (2023)	GPT-3	Five vision language tasks (four classification tasks and one question answering task)	FS	Image + Language	General
Hong et al. (2023)	GPT-4	Text-to-video generation	ZS	Video + Language	General
Yang et al. (2022)	GPT-3	Knowledge-based visual question answering	FS	Image + Language	General
Feng et al. (2023)	GPT-3.5, ChatGPT, GPT-4	Layout generation	FS	Image + Language	General
Zhao et al. (2023a)	ChatGPT, GPT-4	Multimodal tasks covering text, video, audio and images	ZS	Multimodal covering text, video, audio and images	General
Zhang et al. (2023k)	GPT-3.5, ChatGPT, GPT-4	Controlled text-to-image generation	ZS	Image + Language	General
Fan et al. (2023)	ChatGPT	Paraphrasing	ZS	Image + Language	General
Huang et al. (2023b)	ChatGPT	Audio understanding and generation tasks	ZS	Multimodal covering text, audio and images	General
Li et al. (2023h)	GPT-4	Generate instruction tuning dataset	FS	Image + Language	Healthcare
Gui et al. (2022)	GPT-3	Knowledge-based visual question answering	FS	Image + Language	General

Some of the research works explored GPT-3 family models for other tasks like data labelling (Mei et al., 2023), generating instructions (Zhang et al., 2023h), data generation (Fan et al., 2023), prompt editing (Zhu et al., 2023c) and evaluation (Lu et al., 2023b) while developing multimodal AI systems. Mei et al. (2023) used ChatGPT to rewrite those noisy audio captions and developed WavCaps, an audio captions dataset of 400k instances. The authors reported that the models trained on WavCaps datasets achieve new SOTA results. Zhang et al. (2023h) developed SpeechGPT and then do cross-modal instruction tuning to enhance its multi-model instruction following ability. Here, the authors use GPT-4 to generate the instructions for diverse tasks. Fan et al. (2023) proposed LaCLIP (Language augmented Contrastive Language-Image Pretraining), an extended version of CLIP which applies data augmentation to both text and image data to ensure that the model gets exposed to diversified texts during training. Here the data augmentation is performed using the open-source LLaMA model in few-shot settings, and the examples for LLaMA ICL are generated using ChatGPT. Zhu et al. (2023c) explored GPT-3 and GPT-3.5 models for prompt editing in text-to-image generation. The authors observed a potential reduction of 20%–30% in the remaining edits required by implementing the prompt edits suggested by GPT-3 family models. Lu et al. (2023b) proposed LLMScore, a new metric which can effectively capture both image and object-level compositionality for text-to-image generation evaluation.

4.11. Machine learning tasks

Overview. Machine learning (ML) is an area of artificial intelligence (AI) that deals with the development of algorithms that can learn from data and make decisions (Zhang et al., 2023j). Even though machine learning algorithms are successfully used in various real-world applications, creating an effective ML solution for a new task can be difficult due to the numerous design choices involved. In recent times, AutoML has evolved as a solution to reduce the human effort involved in designing ML solutions (Hutter et al., 2019). However, AutoML algorithms suffer from various drawbacks (Zhang et al., 2023j), like (i) the requirement of multiple rounds of trial-and-error, resulting in significant time consumption, (ii) starting the search for a new task from scratch, ignoring past experience gained from the previous tasks and (iii) many AutoML methods lack interpretability because of their black-box nature.

Research works exploring GLLMs to automate machine learning tasks. Inspired by the success of GLLMs in other tasks, the research community explored GLLMs as an alternative to AutoML to automate machine learning tasks (Zheng et al., 2023c; Shen et al., 2023b; Zhang et al., 2023j,d). Table 11 presents a summary of research works exploring GLLMs to automate machine learning tasks. Zheng et al. (2023c) explored how effective is GPT-4 for neural architecture search, i.e., designing optimal neural network configurations. The proposed approach involves two steps, namely (i) GPT-4 generates the optimal neural architecture based on the given problem statement, (ii) the generated configuration is evaluated, and for further refinement, the evaluation results along with the problem statement are passed to the model. This two-step process is repeated for a certain number of iterations to achieve the optimal configuration. Shen et al. (2023b) proposed HuggingGPT to solve AI tasks with the help of GLLMs like ChatGPT and models in AI communities like Hugging Face. HuggingGPT involves four steps, namely task planning, model selection, task execution and

Table 11
Summary of research works exploring GLLMs to automate machine learning tasks. Here ZS represents zero-shot, and FS represents few-shot.

Paper	Task(s)	GLLMs explored	Prompt settings	Language(s)
Zheng et al. (2023c)	Neural architecture search	GPT-4	ZS	English
Shen et al. (2023b)	Multiple AI tasks in language, speech and vision areas	GPT-3.5, ChatGPT, GPT-4	FS	English
Zhang et al. (2023j)	Machine learning tasks	GPT-3.5	FS	English
Zhang et al. (2023d)	Machine learning tasks	GPT-4	FS	English

Table 12
Summary of research works exploring GLLMs for planning. Here ZS represents zero-shot, and FS represents few-shot.

Paper	Task(s)	GLLMs explored	Prompt settings	Language(s)	SOTA results
Olmo et al. (2021)	Plan extraction	GPT-3	FS	English	Yes
Zhang and Soh (2023)	Planning in human–robot interaction	GPT-3.5	ZS	English	No
Xie et al. (2023b)	Plan extraction	GPT-3.5	FS	English	No
Hu et al. (2023b)	Planning	InstructGPT, ChatGPT	FS	English	No

response generation. The authors reported that HuggingGPT achieves promising results in solving AI tasks in language, vision and speech.

Zhang et al. (2023j) proposed MLCopilot, which leverages the power of GLLMs to solve machine learning tasks. MLCopilot works in two stages, namely offline and online. The offline stage involves creating an experience pool from which GLLM is used to retrieve relevant knowledge. The online stage involves retrieving relevant examples from the experience pool, and then GLLM generates results based on the task description, relevant examples and knowledge. Zhang et al. (2023d) proposed AutoML-GPT, which leverages the advanced GPT-4 GLLM to automatic machine learning tasks and reduces human efforts in building machine learning models. AutoML-GPT involves two stages. The first stage involves composing a prompt paragraph based on the model and data cards. The second stage involves performing the four crucial steps from data processing to training log prediction.

4.12. Planning

Overview. Many important industries, like finance and banking, often involve repetitive sequential tasks. These workflows, despite their significance, are typically not fully automated or formally defined. Recently, due to strong reasoning capabilities, the research community explored GLLMs for planning. Some of the research works (Zhang and Soh, 2023; Hu et al., 2023b) directly used LLMs for planning, while some of them (Olmo et al., 2021; Xie et al., 2023b) explored LLMs for planning extraction, which can then be used by automated systems.

Research works exploring GLLMs for planning. Table 12 presents a summary of research works exploring GLLMs for planning. Human models are crucial in facilitating human–robot interaction (HRI), as they empower robots to plan their behaviour based on the impact of their actions on individuals. As it is difficult to craft good human labels, Zhang and Soh (2023) used the GPT-3.5 model (i) as zero-shot human models and also (ii) for planning in trust-related scenarios. Hu et al. (2023b) proposed a novel prompting strategy called "Chain of Symbol" prompting to elicit better the planning abilities of LLMs like InstructGPT and ChatGPT. Unlike CoT prompting, which uses natural language descriptions to represent complex environments, CoS prompting uses condensed symbols to represent them in intermediate reasoning steps. The authors reported that CoS prompting outperforms CoT prompting in both performance and efficiency.

There are usually natural language documents that describe the procedures for the company's employees. Plan extraction methods offer the opportunity to extract structured plans from these natural language descriptions of workflows (Araci, 2019; Chalkidis et al., 2020). These extracted plans can then be used by automated systems. Olmo et al. (2021) explored the GPT-3 model for plan extraction in few-shot settings from the natural language descriptions of workflows and showed that GPT-3 model outperforms existing SOTA models in some cases. Xie et al. (2023b) explored GPT-3.5 models to extract plans from natural language descriptions. The authors reported that the models are poor planners on their own, which is in line with the existing works (Valmeekam et al., 2022; Collins et al., 2022; Mahowald et al., 2023) and are better at extracting plans from natural language. However, these models are sensitive to prompts and also struggle in the case of tasks involving spatial or numerical reasoning.

5. Performance of GLLMs in specific domains

Apart from the general domain, natural language processing is also explored in specific domains like healthcare, finance, legal, social media, etc. Analysing domain-specific texts is more challenging because of domain-specific terminology and abbreviations, complex language structures, etc. In domains like healthcare, finance and legal, domain experts use many words and abbreviations that are specific to the domain and not commonly found in general domain texts. In domains like social media, the texts are mostly authored by the general public using informal language and slang words. Moreover, social media texts are noisy, with many misspelt words, emojis, irregular grammar and abbreviations (Kalyan and Sangeetha, 2020a,b).

Inspired by the success of PLMs like BERT, RoBERTa, ELECTRA, DeBERTa and T5 in the general domain, these models are also explored for domain-specific NLP tasks (Kalyan et al., 2021). However, the performance of general domain models is limited as these models are pretrained on general domain texts (Yang et al., 2020a; Lee et al., 2020), and fine-tuning alone cannot provide enough domain knowledge (Kalyan et al., 2021). So, the research community focused on developing domain-specific PLMs either by continual pretraining or pretraining from scratch (Kalyan et al., 2021, 2022). Currently, domain-specific PLMs achieve state-of-the-art results in most tasks in specific domains like healthcare, finance, legal, social media, etc.

GPT-3 family large language models achieve impressive performances in most NLP tasks in zero and few-shot settings in the general domain. Surprisingly, these models outperform fine-tuned PLMs in some tasks and achieve state-of-the-art results (Sun et al., 2023b; Xu et al., 2023e; Wan et al., 2023; Ma et al., 2023a). Inspired by the massive success of GLLMs in the general domain, the research community explored GLLMs in specific domains to assess how good these models are in domain-specific NLP tasks. Moreover, an extensive evaluation of these models in domain-specific tasks helps to arrive at valuable insights that will guide the research community to improve the performance further and increase the usage of these models in domain-specific NLP tasks.

5.1. Healthcare domain

The recent works explored GLLMs for a variety of clinical NLP tasks like question answering (Holmes et al., 2023b; Nori et al., 2023b; Tanaka et al., 2023a; Liu et al., 2023n; Kasai et al., 2023; Moradi et al., 2021; Singhal et al., 2023b; Wang et al., 2023a; Hernandez et al., 2023;

Table 13
Summary of research works exploring GLLMs for various NLP tasks in the healthcare domain. Here ZS represents zero-shot, and FS represents few-shot. Here '–' represents there is no comparison between GLLMs and domain-specific PLMs in the paper.

Paper	GLLMs explored	Task(s)	Prompt settings	Language(s)	Outperforms domain-specific models
Holmes et al. (2023b)	ChatGPT, GPT-4	Question answering	ZS	English	–
Liu et al. (2023l)	ChatGPT, GPT-4	Text de-identification	ZS	English	Yes
Giorgi et al. (2023)	GPT-4	Dialogue summarization	FS	English	Yes
Nori et al. (2023b)	GPT-3.5, ChatGPT, GPT-4	Question answering	ZS, FS	English	Yes
Chen et al. (2023a)	GPT-3.5, GPT-4	Named entity recognition, relation extraction, document classification and semantic similarity	ZS, FS	English	Yes
Tanaka et al. (2023a)	GPT-3.5, ChatGPT	Question answering	ZS	Japanese	–
Liu et al. (2023n)	GPT-3.5, GPT-4	Question answering, reasoning	ZS	Chinese	Yes
Yang et al. (2023b)	GPT-3	Text simplification	FS	English	–
Gutiérrez et al. (2022)	GPT-3	Entity extraction, relation classification	FS	English	No
Wu et al. (2023c)	ChatGPT, GPT-4	Natural language inference	ZS, FS	English	–
Ma et al. (2023b)	ChatGPT	Text summarization	FS	English	Yes
Wang et al. (2023n)	GPT3.5, GPT4	Natural language inference, document classification	ZS, FS	English	–
Kasai et al. (2023)	GPT-3, ChatGPT, GPT-4	Question answering	FS	Japanese	–
Moradi et al. (2021)	GPT-3	Natural language inference, relation classification, semantic similarity, question answering, text classification	FS	English	No
Jeblick et al. (2022)	ChatGPT	Text simplification	ZS	English	–
Tang et al. (2023c)	GPT-3, GPT-4	Dialogue summarization	FS	English	–
Agrawal et al. (2022)	GPT-3	Clinical sense disambiguation, biomedical evidence extraction, coreference resolution, medication status extraction, medication attribute extraction	ZS, FS	English	–
Nair et al. (2023)	GPT-3	Dialogue summarization	ZS, FS	English	–
Shaib et al. (2023)	GPT-3	Text summarization	ZS, FS	English	–
Xu et al. (2023a)	ChatGPT	Multi-turn medical dialogue	ZS	Chinese	No
Singhal et al. (2023b)	GPT-4	Question answering	FS	English	No
Wang et al. (2023a)	ChatGPT	Question answering	ZS	Chinese	–
Carpenter and Altman (2023)	GPT-3	Synonym generation	ZS	English	–
Hernandez et al. (2023)	GPT-3	Natural language inference, question answering, text classification	ZS	English	No
Rao et al. (2023)	ChatGPT	Clinical decision support	ZS	English	–
Kung et al. (2023)	ChatGPT	Question answering	ZS	English	–
Hulman et al. (2023)	ChatGPT	Question answering	ZS	English	–
Hirosawa et al. (2023)	ChatGPT	Diagnosis lists generation	ZS	English	–
Liu et al. (2023i)	ChatGPT	Clinical decision support	ZS	English	–
Gilson et al. (2023)	GPT-3, GPT-3.5, ChatGPT	Question answering	ZS	English	–
Antaki et al. (2023)	ChatGPT	Question answering	ZS	English	–
Lyu et al. (2023a)	ChatGPT, GPT-4	Text simplification	ZS	English	–

Kung et al., 2023; Hulman et al., 2023; Gilson et al., 2023; Antaki et al., 2023), text de-identification (Liu et al., 2023l), dialogue summarization (Giorgi et al., 2023; Tang et al., 2023c; Nair et al., 2023), named entity recognition (Chen et al., 2023a; Gutiérrez et al., 2022), relation extraction (Chen et al., 2023a), text classification (Chen et al., 2023a; Wang et al., 2023n; Moradi et al., 2021; Hernandez et al., 2023), semantic similarity (Chen et al., 2023a; Moradi et al., 2021), text simplification (Yang et al., 2023b; Jeblick et al., 2022; Lyu et al., 2023a), relation classification (Gutiérrez et al., 2022; Moradi et al., 2021), text summarization (Ma et al., 2023b; Shaib et al., 2023), natural language inference (Wu et al., 2023c; Wang et al., 2023n; Moradi et al., 2021; Hernandez et al., 2023), word sense disambiguation (Agrawal et al., 2022), biomedical evidence extraction (Agrawal et al., 2022), coreference resolution (Agrawal et al., 2022), medical status extraction (Agrawal et al., 2022), medical attribute extraction (Agrawal et al., 2022), synonym generation (Carpenter and Altman, 2023), clinical decision support (Rao et al., 2023; Liu et al., 2023i) and diagnostic lists generation (Hirosawa et al., 2023). Most of the research focused on English datasets, except a few focused on other languages like Japanese (Tanaka et al., 2023a; Kasai et al., 2023) and Chinese (Liu et al., 2023n; Xu et al., 2023a; Wang et al., 2023a). Table 13 presents a summary of research works exploring GLLMs for various NLP tasks in the healthcare domain.

Lyu et al. (2023a) investigated the performance of ChatGPT and GPT-4 models in the healthcare domain, specifically the radiology area, by evaluating their ability to simplify the content in radiology reports. Experiment results showed that (i) GPT-4 performs better than ChatGPT. and (ii) optimized prompt with detailed instructions improves the performance for both models by a good margin. Antaki et al. (2023) evaluated the effectiveness of ChatGPT in answering Ophthalmology questions. The test set consists of both easy and moderate-level questions. Experiment results showed that ChatGPT achieves an average accuracy of 49.25%. Specifically, ChatGPT is able to answer the questions with good accuracy in general medicine. However, its performance in specific sub-areas of Ophthalmology is worst. Gilson et al. (2023) evaluated GLLMs like GPT-3, GPT-3.5, and ChatGPT model in answering the medical questions in Step 1 and Step 2 exams of USMLE. Experiment results showed that ChatGPT outperforms the other two models by a good margin. Rao et al. (2023) demonstrated that ChatGPT performs better in the final diagnosis than in the initial diagnosis. This is because ChatGPT has access to more clinical data during the final diagnosis than the initial one.

Carpenter and Altman (2023) demonstrated that GPT-3 can be used for the synonym generation for drugs of abuse. The authors query GPT-3 repeatedly for each drug to generate multiple synonyms, which are later filtered. The generated synonyms are then used to build a lexicon that is helpful for pharmacovigilance on social media platforms. Inspired by the success of the GPT-3 model for text summarization in the general domain, Shaib et al. (2023) explored the GPT-3 model for summarizing biomedical documents. Experiment results revealed that (i) GPT-3 performance is promising in the case of single document summarization and (ii) GPT-3 struggles to summarize the content from multiple biomedical documents. Nair et al. (2023) proposed a novel approach called "MEDSUM-ENT", a multi-stage framework for clinical dialogue summarization. The proposed method leverages the GPT-3 model through multiple intermediate calls to extract medical entities from the conversations. In the final step of summarization, the extracted entities, task instructions and in-context examples help the GPT-3 model to generate high-quality summaries. Based on the evaluation of radiology reports simplified by ChatGPT, Jeblick et al. (2022) reported that ChatGPT-generated simplified radiology reports are not potentially harmful, complete and factually correct. However, further analysis reveals that some simplified reports contain factually incorrect sentences, potentially harmful paragraphs and a lack of essential medical findings.

Hirosawa et al. (2023) investigated the effectiveness of ChatGPT for clinical diagnosis by evaluating its ability to generate accurate diagnosis lists for clinical vignettes with common chief complaints. Experimental results showed that ChatGPT can generate diagnosis lists with good accuracy. However, the accuracy rate of ChatGPT is still less than the accuracy rate of physicians. Wang et al. (2023a) evaluated the performance of the ChatGPT model in answering medical questions in the Chinese language. Here, ChatGPT is prompted with questions in both English and Chinese to avoid language barriers. Experimental results show that the performance of ChatGPT is much lower than the average performance of the medical students. For example, ChatGPT correctly answers 45.8% of questions, while the average answering rate of medical students is 67.9% in 2021.

Some of the research works demonstrated that domain-specific PLMs outperform GLLMs. Hernandez et al. (2023) compared the performance of the GPT-3 model with the performances of general and domain-specific PLMs on three healthcare NLP tasks: natural language inference, question answering and text classification. Experiment results showed that domain-specific PLMs achieve better results even though they are much smaller than GPT-3. Xu et al. (2023a) introduced MedGPTEval, a benchmark to assess LLMs in the healthcare domain. An extensive evaluation showed that domain-specific Chinese LLM outperforms general-purpose models like ChatGPT and ERNINE Bot. Singhal et al. (2023b) introduced MedPaLM2, a healthcare domain-specific LLM obtained by domain-specific finetuning of the PaLM2 (Anil et al., 2023) model. Experiment results showed that MedPaLM2 outperforms few-shot GPT-4 and achieves new state-of-the-art results on the MultiMedQA benchmark. Moradi et al. (2021) investigated the performances of BioBERT and GPT-3 in few-shot settings on five biomedical NLP tasks: text classification, natural language inference, question answering, relation extraction and semantic similarity. The authors observed that BioBERT and GPT-3 models underperform the model fine-tuned using full training data. Moreover, the BioBERT model outperforms GPT-3 in few-shot settings even though the BioBERT model is 514 times smaller than GPT-3.

Some research works showed that GLLMs can outperform domain-specific PLMs. Ma et al. (2023b) proposed ImpressionGPT, a novel approach for summarizing radiology reports using ChatGPT. The proposed method involves dynamic prompt construction and iterative optimization to enhance the performance of ChatGPT further. Evaluation on two standard datasets showed that the proposed framework achieves new SOTA results outperforming fine-tuned models like ChestXrayBERT (Cai et al., 2021). Liu et al. (2023n) introduced CMExam, a dataset with 60k+ multiple-choice medical questions in the Chinese language and evaluated GLLMs like GPT-3.5 and GPT-4 on answer prediction and answer reasoning tasks. The authors observed that GPT-4 achieves the best results for both tasks, outperforming GPT-3.5 and medical domain-specific Chinese LLMs like Huatuo (Antaki et al., 2023) and DoctorGLM (Xiong et al., 2023). Chen et al. (2023a) explored GLLMs like GPT-3.5 and GPT-4 on eight datasets spanning four tasks in zero and few-shot settings. The authors observed that fine-tuned PubMedBERT outperforms both the GLLMs in all the biomedical tasks except question answering. In the case of biomedical question answering, GPT-4 outperforms the fine-tuned PubMedBERT model by a large margin of 17

Giorgi et al. (2023) explored models like Longformer Encoder–Decoder (LED) (Beltagy et al., 2020) based on supervised fine-tuning and GLLMs like GPT-4 based on few-shot ICL for clinical dialogue summarization as a part of MEDIQA-Chat 2023 (Abacha et al., 2023) shared task. Here, the authors used Instructor (Su et al., 2022) to select the most similar examples for few-shot ICL. Experiment results based on automatic metrics like BERTScore and ROUGE demonstrated that GPT-4 not only outperforms the LED model but also achieves first rank in the shared task. For medical text de-identification, Liu et al. (2023l) proposed a novel approach called "DeID-GPT", a two-step approach based on GLLMs. In the first step, HIPAA identifiers are included in the prompt. In the second step, GLLM receives the prompt and the medical record based on which the model generates the de-identified medical record having the personal information masked. The authors observed that GPT-4 outperforms not only ChatGPT but also fine-tuned models based on BERT, RoBERTa and ClinicalBERT.

Table 14
Summary of research works exploring GLLMs for various NLP tasks in the legal domain. Here ZS represents zero-shot, and FS represents few-shot. Here '–' represents there is no comparison between GLLMs and domain-specific PLMs in the paper.

Paper	GLLMs explored	Task(s)	Prompt settings	Language(s)	Outperforms domain-specific models
Yu et al. (2022)	GPT-3	Natural language inference	ZS, FS	English	–
Bommarito and Katz (2022)	GPT-3.5	Question answering	ZS	English	–
Nguyen (2023)	GPT-3	Question answering, text generation	ZS	English	–
Chalkidis (2023)	ChatGPT	Text classification	ZS, FS	English	No
Choi et al. (2023)	ChatGPT	Question answering, text generation	ZS	English	–

Table 15
Summary of research works exploring GLLMs for various NLP tasks in the finance domain. Here ZS represents zero-shot, and FS represents few-shot. Here '–' represents there is no comparison between GLLMs and domain-specific PLMs in the paper.

Paper	GLLMs explored	Task(s)	Prompt settings	Language(s)	Outperforms domain-specific models
Li et al. (2023k)	ChatGPT, GPT-4	News headlines classification, financial sentiment analysis, named entity recognition, question answering	ZS	English	Yes
Fatouros et al. (2023)	ChatGPT	Sentiment analysis	ZS	English	Yes
Leippold (2023)	GPT-3	Sentiment analysis	ZS	English	No
Wiriyathammabhum (2022)	GPT-3.5	Pairwise ranking	FS	Chinese	–
Shah and Chava (2023)	ChatGPT	Sentiment analysis, claim detection, named entity recognition	ZS	English	No
Zhang et al. (2023b)	ChatGPT, GPT-4	Question answering	ZS, FS	Chinese	–
Rajpoot and Parikh (2023)	ChatGPT, GPT-4	Relation extraction	FS	English	–
Lan et al. (2023)	ChatGPT	Sentiment analysis	ZS	Chinese	–
Loukas et al. (2023)	GPT-3.5, GPT-4	Text classification	ZS, FS	English	–

5.2. Legal domain

The recent works explored GLLMs for a variety of legal NLP tasks like natural language inference (Yu et al., 2022), question answering (Bommarito and Katz, 2022; Lan et al., 2023; Choi et al., 2023), text generation (Nguyen, 2023; Choi et al., 2023) and text classification (Chalkidis, 2023). Table 14 presents a summary of research works exploring GLLMs for various NLP tasks in the legal domain. Bommarito and Katz (2022) evaluated the performance of the GPT3.5 model in the legal domain by evaluating its ability to answer bar exam questions. The model answers the questions correctly at a rate of 50%, which is 25% more than the random guess baseline. However, the model performance is almost 18% less than the human performance, and overall model performance is below the passing threshold. Nguyen (2023) presented LawGPT 1.0, the first-ever chatbot model based on GPT-3 for the legal domain. The GPT-3 model is pretrained on mostly generic corpus, so it lacks domain-specific knowledge. To add domain-specific knowledge, LawGPT is developed by fine-tuning the GPT-3 model on the law corpus. Experimental results showed that LawGPT 1.0 performs on par with existing legal assistants.

Chalkidis (2023) investigated how effective ChatGPT is for legal text classification by evaluating the model performance on the LexGLUE (Chalkidis et al., 2022) benchmark, which consists of seven legal text classification datasets. The evaluation is performed in both zero and few-shot settings. Experiment results showed that ChatGPT performs poorly on legal text classification datasets. Choi et al. (2023) demonstrated that the performance of ChatGPT is just above the passing threshold, i.e., equivalent to a C+ grade student. The authors found that advanced prompts like CoT (Wei et al., 2022b) and Ranking prompts performed worse or the same as simple prompts for multiple-choice questions. For essay writing, the authors used carefully crafted simple prompts by including specific instructions at the end of the prompt.

5.3. Finance domain

The recent works explored GLLMs for a variety of finance NLP tasks like text classification (Li et al., 2023k; Loukas et al., 2023), sentiment analysis (Li et al., 2023k; Leippold, 2023; Shah and Chava, 2023; Lan et al., 2023), named entity recognition (Li et al., 2023k; Shah and Chava, 2023), question answering (Li et al., 2023k; Zhang et al., 2023b), pairwise ranking (Wiriyathammabhum, 2022), claim detection (Shah and Chava, 2023) and relation extraction (Rajpoot and Parikh, 2023). Table 15 presents a summary of research works exploring GLLMs for various NLP tasks in the finance domain.

Li et al. (2023k) compared the performances of general LLMs like ChatGPT and GPT-4 in the finance domain with domain-specific models like BloombergGPT (Wu et al., 2023a) and small fine-tuned models like FinBERT (Araci, 2019) and FinQANet (Chen et al., 2021a). The evaluation is done on five different datasets related to four financial NLP tasks: news headlines classification, sentiment analysis, entity extraction, and question answering. The ChatGPT and GPT4 models do well in question-answering task but lag behind in tasks requiring domain-specific knowledge like entity extraction and sentiment analysis. Fatouros et al. (2023) evaluated the effectiveness of ChatGPT for financial sentiment analysis by assessing its performance on the forex-related news headlines dataset. Experiment results showed that ChatGPT outperforms the domain-specific FinBERT (Liu et al., 2021a) model by a large margin of 35% and also exhibits a high correlation with market returns.

Leippold (2023) explored GPT-3 for financial sentiment analysis and to generate adversarial attacks. Experiment results showed that FinBERT outperforms keyword-based approaches and the few-shot GPT-3 model in financial sentiment analysis. To study the robustness of FinBERT-based and keyword-based approaches, the authors explored GPT-3 to generate adversarial attacks. The main advantage of GPT-3 over existing adversarial attack-generating methods is that the model makes more subtle changes to the instances such that they are not noticeable to humans but still can fool the models. Wiriyathammabhum (2022) explored instruction fine-tuned T5 and GPT-3.5 models to evaluate investments-related social media posts in Chinese. The task involves two subtasks, namely pairwise ranking and unsupervised ranking. Experiment results showed that the few-shot prompted GPT-3.5 model outperforms the instruction fine-tuned T5 model and the few-shot prompted GPT-3.5 model with English-translated social media posts.

Table 16
Summary of research works exploring GLLMs for NLP tasks in multilingual settings. Here, ZS represents zero-shot, and FS represents few-shot.

Paper	GLLMs explored	Task(s)	Prompt settings	Language(s)	Domain(s)
Lai et al. (2023a)	ChatGPT	PoS tagging, entity extraction, relation extraction, natural language inference, question answering, text summarization, common sense reasoning	ZS	37 Languages	General
Fang et al. (2023b)	ChatGPT	Grammar error correction	ZS, FS	English, German, Chinese	General
Armengol-Estapé et al. (2022)	GPT-3	Question answering, natural language generation, text summarization	ZS	German, Spanish, Russian, Turkish, Catalan	General
Ahuja et al. (2023)	GPT-3.5, ChatGPT, GPT-4	Natural language inference, paraphrase identification, commonsense reasoning, question answering, parts of speech tagging, sentiment analysis, text summarization	ZS	70 languages	General
Bang et al. (2023)	ChatGPT	Sentiment analysis, language identification, machine translation	ZS	Multiple language including low resource languages like Sudanese, Javanese etc.	General
Kuzman et al. (2023)	ChatGPT	Genre identification	ZS	English, Slovenian	General
Zhang et al. (2023g)	ChatGPT	Question answering, Reasoning	ZS	Six languages including Chinese, German and French	General
Das et al. (2023)	ChatGPT	Hate speech detection	ZS	Eleven languages including Hindi, Arabic and Italian	Social media
Hada et al. (2023)	GPT-4	Three text generation tasks	ZS	Ten languages including Chinese and Japanese.	General
Leong et al. (2023)	ChatGPT, GPT-4	Question answering, sentiment analysis, text summarization, named entity recognition, toxicity detection, machine translation, natural language inference, casual reasoning	ZS, FS	Indonesian, Vietnamese, Thai, Tamil	General, Social Media, News

Shah and Chava (2023) compared the performance of ChatGPT with the performance of fine-tuned PLMs for three different financial NLP tasks: claim detection, sentiment analysis and named entity recognition. The authors observed that fine-tuned models outperform ChatGPT, but ChatGPT performs much better than some open-source LLMs. Zhang et al. (2023b) introduced FinEval, a new benchmark to evaluate the financial domain of knowledge of LLMs in the Chinese language. FinEval includes 4661 multiple-choice questions in Chinese language from four different categories spanning 34 academic subjects. Experiment results showed that GPT-4 achieves around 70% accuracy and outperforms all other LLMs, including ChatGPT and Chinese LLMs.

Rajpoot and Parikh (2023) assessed the effectiveness of ChatGPT and GPT-4 for financial relation extraction in few-shot settings. As the choice of examples is crucial in few-shot ICL, the authors explored learning free and learning-based retriever for example selection. The authors observed that GPT-4 outperforms ChatGPT by a decent margin, and the learning-based retriever performs better than the learning-free retriever.

6. Multilingual performance of GLLMs

Overview. GLLMs are pretrained over large volumes of text data from multiple languages. For example, the corpus used to pretrain the GPT-3 model includes text from around 90 languages, and the percentage of English text is more than 90% (Brown et al., 2020; Ahuja et al., 2023). In the beginning, most of the research focused on assessing the performance of GLLMs on English datasets only. However, it is essential to evaluate these models on datasets from non-English languages, especially low-resource languages, to know how effective GLLMs are for non-English languages, and the insights gained from the comprehensive evaluation help to further improve these models towards non-English languages.

Research works exploring GLLMs in multilingual settings. Recently, some of the research works focused on evaluating GLLMs across various non-English languages. The evaluation is done on various tasks like parts of speech tagging (Lai et al., 2023a; Ahuja et al., 2023), named entity recognition (Lai et al., 2023a; Leong et al., 2023), relation extraction (Lai et al., 2023a), natural language inference (Lai et al., 2023a; Ahuja et al., 2023; Leong et al., 2023), question answering (Lai et al., 2023a; Armengol-Estapé et al., 2022; Ahuja et al., 2023; Zhang et al., 2023g; Leong et al., 2023), text summarization (Lai et al., 2023a; Armengol-Estapé et al., 2022; Ahuja et al., 2023; Leong et al., 2023), commonsense reasoning (Lai et al., 2023a; Ahuja et al., 2023), grammar error correction (Fang et al., 2023b), text generation (Armengol-Estapé et al., 2022; Hada et al., 2023), paraphrase identification (Ahuja et al., 2023), sentiment analysis (Ahuja et al., 2023; Bang et al., 2023; Leong et al., 2023), language identification (Bang et al., 2023), machine translation (Bang et al., 2023; Leong et al., 2023), genre identification (Kuzman et al., 2023), hate speech detection (Das et al., 2023) and toxicity detection (Leong et al., 2023). Most of the research focused on general domain datasets, except a few focused on other domains like social media (Das et al., 2023; Leong et al., 2023) and news (Leong et al., 2023). Table 16 presents a summary of research works exploring GLLMs for NLP tasks in multilingual settings.

Bang et al. (2023) presented an extensive multilingual evaluation of ChatGPT across three tasks: sentiment analysis, language identification and machine translation. When compared to English, the performance of ChatGPT degrades in the case of low-resource languages, particularly in the case of languages with non-Latin scripts. Das et al. (2023) assessed the effectiveness of ChatGPT for emoji-based hate speech detection in multilingual settings. The authors reported that ChatGPT exhibits good performance but tends to misclassify abusive content as hate speech for non-English languages in the case of non-protected groups. Moreover, Armengol-Estapé et al. (2022) reported that the performance of GPT-3 can be improved in the case of low-resource languages with optimized tokenization.

The focus of existing benchmarks like HELM (Bommasani et al., 2023) and BIG-Bench (Srivastava et al., 2023) is on the English language. So, some of the research works focused on introducing new benchmarks to facilitate a systematic and comprehensive evaluation of the multilingual performance of GLLMs (Ahuja et al., 2023; Leong

Table 17
Summary of research works exploring GLLMs for data labelling. Here, '–' represents that the paper does not include a comparison between GLLMs and human annotators.

Paper	GLLMs explored	Task(s)	Prompt settings	Domain(s)	Language(s)	Outperforms human annotators
Gilardi et al. (2023)	ChatGPT	Stance, relevance, frame and topics detection	ZS	Social media, news	English	Yes
He et al. (2023b)	GPT-3.5	Three binary text classification tasks	ZS, FS	General	English	Yes
Törnberg (2023)	GPT-4	Political tweets classification	ZS	Social media	English	Yes
Zhu et al. (2023e)	ChatGPT	Stance detection, sentiment analysis, hate speech detection, bot detection	ZS	Social media	English	No
Li et al. (2023c)	ChatGPT	Detection of hateful, toxic and offensive comments	ZS	Social media	English	No
Gu et al. (2023)	GPT-3.5, GPT-4	Adverse drug reaction extraction	ZS, FS	Healthcare	English	–
Wang et al. (2021b)	GPT-3	Text entailment, topic classification, sentiment analysis, answer type classification, question generation, text generation	ZS	General	English	–
Ding et al. (2022)	GPT-3	Sentiment analysis, relation extraction, named entity recognition	FS	General	English	–
Meoni et al. (2023)	GPT-3.5	Named entity recognition	ZS	Healthcare	English, French, Spanish, Italian, Basque	–
Xu et al. (2023c)	GPT-3.5	Text summarization	ZS, FS	General	English	–
Alizadeh et al. (2023)	ChatGPT	Detection of stance, topics, relevance, general frame and policy frame	ZS, FS	Social media, news	English	Yes
Yang et al. (2023b)	GPT-3	Radiology text simplification	FS	Healthcare	English	–

et al., 2023). For example, Ahuja et al. (2023) presented MEGA, a comprehensive evaluation benchmarking having 16 datasets covering 70 languages. Based on the evaluation of GLLMs like GPT-3.5, ChatGPT and GPT-4, the authors reported that GLLMs perform well in the case of languages with Latin scripts, and the performance is worst in the case of low-resource languages with non-Latin scripts across tasks. One of the possible reasons for this is the quality of tokenization. Similarly, Leong et al. (2023) introduced BHASA, a benchmark to evaluate the performance of LLMs in four Southeast Asian languages. The benchmark consists of 20 datasets covering eight NLP tasks. The authors reported that (i) GPT-4 achieves better results compared to ChatGPT, and (ii) overall, the performance on some of the tasks is promising, with a lot of room for improvement in other tasks.

Some of the existing works demonstrated that using prompts in English improves the performance of GLLMs in the case of non-English languages (Lai et al., 2023a; Kuzman et al., 2023). For example, Lai et al. (2023a) performed a comprehensive evaluation of the multilingual abilities of ChatGPT on seven tasks covering more than 30 languages ranging from high-resource to extremely low-resource languages. The experiment results confirmed the bias of ChatGPT towards the English language, i.e., the performance is better for English compared to other languages and prompts in the English language can enhance the performance for non-English languages. The possible reason for the bias of GLLMs towards the English language is that GLLMs are trained mostly on English text corpus; hence, these models can better understand the prompt if it is in English (Kuzman et al., 2023).

Some of the research works investigated how GLLMs exhibit multilingual capabilities (Zhang et al., 2023g) and how effective GLLM-based evaluators are in scaling up evaluation in multilingual settings (Hada et al., 2023). Zhang et al. (2023g) proposed a novel back translation prompting approach to systematically study how ChatGPT exhibit multilingual capabilities, although these models are largely pretrained on the English text corpus. The authors demonstrated that ChatGPT does translation in multilingual settings. Moreover, the multilingual performance of GLLMs is good only in the case of tasks which can translated. Hada et al. (2023) assessed the effectiveness of GPT-4 as an evaluator for natural language generation tasks in multilingual settings. The authors reported that GPT-4 tends to favour high scores and should be used carefully.

7. Data labelling and data augmentation abilities of GLLMs

7.1. Data labelling

Overview. Large language models, specifically GLLMs, have achieved impressive performances in most of the NLP tasks, highlighting the huge potential of these models. However, large model size, high latency, high inference costs, proprietary access (in the case of GLLMs) and confidentiality concerns (in the case of sensitive domains like medical Meoni et al., 2023) have become bottlenecks for the practical use of these models. Because of these bottlenecks, in environments with constrained resources or confidentiality constraints, PLMs are preferred over GLLMs as these models are much smaller in size and also more efficient compared to GLLMs (Thapa et al., 2023). For example, BERT base and large models contain just 110M and 340M parameters, while the GPT-3 model contains 175B parameters. Moreover, it is reported that GLLMs are trailing the SOTA models, with 4% to 70% lower performance when evaluated across a set of 25 diverse natural language processing tasks (Kocoń et al., 2023).

The performance of fine-tuned PLMs is largely determined by the quality as well as the quantity of labelled data. Human-annotated data is considered the gold standard (Murthy et al., 2019; Van Atteveldt et al., 2021), and we have two strategies for this (Gilardi et al., 2023; Törnberg, 2023). The first one is using trained expert coders like students and research assistants, and the second one is using crowd workers from online platforms like Amazon Mechanical Turk. Although human-labelled data is considered the gold standard, the human annotation process is expensive, laborious and time-consuming. The second

strategy, i.e., using crowd workers, is comparatively less expensive, but there is a growing concern regarding the degrading annotation quality of crowd workers (Chmielewski and Kucker, 2020). Moreover, the annotation quality varies with annotators, and hence it is consistent. To address the challenges associated with the human annotation process, there is a growing interest in the NLP research community to leverage the extraordinary generative abilities of GLLMs to make the data annotation process less expensive, faster and consistent. Similar to the human annotation process, GLLMs are provided with detailed instructions along with some labelled examples to label the data.

Research exploring GLLMs for data labelling. The research community explored GLLMs for data labelling in a variety of NLP tasks like stance detection (Gilardi et al., 2023; Zhu et al., 2023e), political tweets classification (Törnberg, 2023), sentiment analysis (Zhu et al., 2023e; Wang et al., 2021b; Ding et al., 2022), hate speech detection (Zhu et al., 2023e; Li et al., 2023c), bot detection (Zhu et al., 2023e), toxic comments detection (Li et al., 2023c), offensive comments detection (Li et al., 2023c), adverse drug reaction extraction (Gu et al., 2023), text entailment (Wang et al., 2021b), topic classification (Wang et al., 2021b), text generation (Wang et al., 2021b), answer type classification (Wang et al., 2021b), question generation (Wang et al., 2021b), relation extraction (Ding et al., 2022), named entity recognition (Ding et al., 2022; Meoni et al., 2023), text summarization (Xu et al., 2023c), radiology text simplification (Yang et al., 2023b) etc. Most of the research works focused on English datasets, except a few research works focused on other languages like French (Meoni et al., 2023), Spanish (Meoni et al., 2023), Italian (Meoni et al., 2023) and Basque (Meoni et al., 2023). Table 17 presents a summary of research works exploring GLLMs for data labelling.

Gu et al. (2023) labelled sentences from PubMed abstracts using the GPT-3.5 model and then fine-tuned the PubMedBERT model for adverse drug reaction extraction. Experiment results showed that (i) PubMedBERT achieves results comparable to the SOTA model and (ii) PubMedBERT outperforms the GPT-3.5 and GPT-4 models by large margins of 6 and 5 points in F1 score, respectively. Based on the evaluation of multiple NLU and NLG tasks, Wang et al. (2021b) demonstrated that GPT-3 labelled data can result in a 50 to 96% reduction in labelling expenses. Moreover, PLMs fine-tuned on GPT-3 labelled data outperform the few-shot GPT-3 model in both NLU and NLG tasks. Further, the authors proposed an approach based on active learning to make use of both human and GPT-3 labels, which further enhances the performance of the fine-tuned models. Meoni et al. (2023) investigated the effectiveness of GPT-3.5 labelled data and dictionary-based labelled data in fine-tuning PLMs to extract clinical entities in multiple languages like English, Spanish, Basque, Italian and French. The authors reported that (i) the performance of GPT-3.5 labelled data is on par with dictionary-based labelled data, and (ii) combining annotations from both approaches further enhances the results. Xu et al. (2023c) proposed InhertiSumm, a novel approach for training small text summarization models like ZCode++ (He et al., 2022c) using GPT-3.5 generated summaries. The authors showed that the ZCode++ model with just 390M parameters trained using GPT-3.5 generated summaries performs on par with GPT-3.5 in zero and few-shot settings.

Zhu et al. (2023e) investigated how effective ChatGPT is for labelling data for social computing tasks. Based on the evaluation of five datasets spanning over tasks like stance detection, hate speech detection, bot detection and sentiment analysis, the authors reported that ChatGPT achieves an average accuracy of 60.9. Li et al. (2023c) investigated the ability of ChatGPT to label hateful, offensive and toxic comments and compared the performances with MTurk annotations. The authors observed that ChatGPT performance is promising as it is able to label 80% of comments correctly. Moreover, the performance of ChatGPT is more consistent for non-harmful comments than harmful comments.

Some of the research works (Gilardi et al., 2023; He et al., 2023b; Törnberg, 2023; Alizadeh et al., 2023) showed that GLLMs as data annotators can outperform human annotators. Gilardi et al. (2023) investigated the effectiveness of ChatGPT as an annotator in zero-shot settings for four text classification tasks involving tweets and news articles. The authors reported that ChatGPT is more effective than MTurk crowd-workers as (i) ChatGPT achieves 25 points more than crowd-workers in terms of accuracy, (ii) ChatGPT is approximately 30 times cheaper, and (iii) intercoder agreement of ChatGPT is more than crowd-workers. He et al. (2023b) proposed a novel approach called "explain then annotate" to enhance the performance of GLLMs as text data annotators. The proposed approach involves two steps: (i) GLLM generates explanations for the demonstrations and then (ii) annotates the data by leveraging annotation guidelines, demonstrations and explanations through CoT prompting. Evaluation on three binary text classification tasks revealed that GPT-3.5 outperforms crowd-workers on one task and matches the performance of crowd-workers on the other two tasks. Törnberg (2023) demonstrated that zero-shot GPT-4 outperforms human annotators in labelling political English tweets. Further analysis demonstrated that GPT-4 possesses the ability to accurately label tweets that involve logical reasoning from contextual information. Alizadeh et al. (2023) compared the performances of GLLMs like ChatGPT, open-source LLMs like FLAN (Chung et al., 2022) and MTurk annotators in labelling data (tweets and news articles) for five text classification tasks. The authors reported that ChatGPT achieves the best results, outperforming both open-source LLMs and MTurk annotators. One promising observation here is that open-source LLMs outperform MTurk annotators, and the performance is comparable to ChatGPT.

7.2. Data augmentation

Overview. The performance of downstream task-specific models is determined by the quality as well as the quantity of labelled data. Fine-tuning the PLMs on a small amount of labelled data will result in overfitting (Kalyan et al., 2021) and, subsequently, poor performances. However, it is not feasible all the time to label a large number of instances as the annotation process is expensive. So, the research community focused on alternative approaches like data augmentation to increase the size of training sets in a relatively inexpensive way (Shorten and Khoshgoftaar, 2019; Li et al., 2022b; Liu et al., 2020b; Feng et al., 2021; Bayer et al., 2022). The data augmentation approaches focus on generating additional training instances either by making small changes to the existing instances or creating new instances with a distribution similar to the existing instances.

Data augmentation is initially explored in the area of computer vision (Shorten and Khoshgoftaar, 2019) and then explored in natural language processing (Li et al., 2022b; Liu et al., 2020b; Feng et al., 2021; Bayer et al., 2022). When compared to computer vision, text data augmentation is more challenging because of the discrete nature of text. Data augmentation can be done at character, word and sentence levels. Character-level data augmentation approaches involve random deletion, addition, exchange or insertion of characters (Belinkov and Bisk, 2018; Coulombe, 2018). For example, in the case of keyboard augmentation, a random character is replaced with its neighbour based on the QWERTY layout (Belinkov and Bisk, 2018). Similar to character-level data augmentation, word-level data augmentation approaches involve deletion, replacement, exchange or insertion of words at random positions (Wei and Zou, 2019; Wang and Yang, 2015). Sentence-level approaches like back translation and paraphrasing generate augmented instances by rewriting the sentence (Sennrich et al., 2016; Mallikarjuna and Sivanesan, 2022). Overall, the main drawbacks of existing data augmentation approaches are (i) lack of sufficient diversity in the augmented instances and (ii) often struggle to guarantee the accurate labelling of the augmented data (Dai et al., 2023b). To address these drawbacks, the research community focused on leveraging the exceptional generating abilities of GLLMs for data augmentation to ensure sufficient diversity and correct labelling in the augmented data.

Table 18
Summary of research works exploring GLLMs for paraphrasing-based data augmentation.

Paper	GLLMs explored	Task(s)	Prompt settings	Domain(s)	Language(s)
Cegin et al. (2023)	ChatGPT	Intent classification	ZS	General	English
Oh et al. (2023)	ChatGPT	Machine translation	ZS	General	Korean, German
Sharma et al. (2022)	GPT-3	Named entity recognition	ZS	News, social media, general, healthcare	English
Guo et al. (2023b)	ChatGPT, GPT-4	Question answering	ZS	Healthcare	English
Abaskohi et al. (2023)	GPT-3	Text classification	FS	General	English
Sarker et al. (2023)	ChatGPT	Medical event classification, medication identification	ZS	Healthcare	English
Parikh et al. (2023)	GPT-3	Intent classification	ZS	Social media	English
Dai et al. (2023b)	ChatGPT	Text classification	ZS	General, Healthcare	English
Fang et al. (2023a)	ChatGPT	Open intent detection	ZS	General	English

7.2.1. Paraphrasing

Research works exploring GLLMs for paraphrasing-based data augmentation. The research community explored GLLMs for paraphrasing in various NLP tasks like intent classification (Cegin et al., 2023; Parikh et al., 2023; Fang et al., 2023a), machine translation (Oh et al., 2023), named entity recognition (Sharma et al., 2022), question answering (Guo et al., 2023b), medical event classification (Sarker et al., 2023), medication identification (Sarker et al., 2023) etc. GLLM-based paraphrasing is explored in multiple domains like general (Cegin et al., 2023; Oh et al., 2023; Sharma et al., 2022; Abaskohi et al., 2023; Dai et al., 2023b; Fang et al., 2023a), news (Sharma et al., 2022), social media (Sharma et al., 2022; Parikh et al., 2023) and healthcare (Sharma et al., 2022; Guo et al., 2023b; Sarker et al., 2023; Dai et al., 2023b). Table 18 presents a summary of research works exploring GLLMs for paraphrasing-based data augmentation.

Cegin et al. (2023) compared the quality of paraphrases generated by ChatGPT and crowd workers for intent classification. The authors reported that (i) ChatGPT generates more diversified paraphrases compared to crowd-workers and (ii) the robustness of models fine-tuned on ChatGPT is comparable to the models fine-tuned on crowd-workers generated paraphrases. Oh et al. (2023) explored ChatGPT-based data augmentation to generate additional training instances to fine-tune the mBART-50 model (Tang et al., 2020) for machine translation involving Korean–German language pairs. Here, the authors explored three different prompting strategies, out of which the storytelling prompting approach achieves the best results and improves the BLUE score by 0.68. Here, the storytelling prompting approach involves generating a three-sentence story based on the source sentence and then translating each of these sentences into the target language. Abaskohi et al. (2023) proposed a novel approach based on prompt-based tuning and contrastive learning to fine-tune PLMs for text classification. As contrastive learning requires data augmentation, the authors explored models like GPT-3 and OPT-175B (Zhang et al., 2022) for paraphrasing. Experiment results showed that GPT-3 based paraphrasing outperforms existing data augmentation approaches like back translation (Sugiyama and Yoshinaga, 2019) and easy data augmentation (Wei and Zou, 2019).

To overcome the problem of limited training instances for EHR analysis, Sarker et al. (2023) explored ChatGPT to generate additional training instances through paraphrasing. Experiments on medication event classification and medical identification tasks revealed that fine-tuning the PLMs on ChatGPT augmented training set enhances the performance. Dai et al. (2023b) proposed AugGPT, a ChatGPT-based approach to generate additional training instances by paraphrasing existing training instances for few-shot classification. Experiments on general and medical domain text classification datasets revealed that AugGPT outperforms all the existing data augmentation approaches by a good margin. Further analysis showed that AugGPT generates more diversified instances while preserving the original labels.

Paraphrasing-based data augmentation for entity extraction is challenging because of the difficulty in preserving span-level labels. Sharma et al. (2022) explored GPT-3 models, back translation and PEGASUS-based paraphraser for synthetic data generation using paraphrasing. The authors observed that the larger GPT-3 variant with inline annotations achieves the best results for entity extraction across datasets from multiple domains.

7.2.2. Data generation

Research works exploring GLLMs for data generation-based data augmentation. The research community explored GLLMs for data generation-based data augmentation in various NLP tasks like dialogue generation (Wang et al., 2023m), training smaller LLMs (Gunasekar et al., 2023; Eldan and Li, 2023), common sense reasoning (Whitehouse et al., 2023), hate speech detection (Hartvigsen et al., 2022), undesired content detection (Markov et al., 2023), question answering (Guo et al., 2023c; Zhao et al., 2023c), intent classification (Parikh et al., 2023), relation extraction (Xu et al., 2023e; Tang et al., 2023a), instruction tuning (Liu et al., 2023g; Peng et al., 2023c), paraphrase detection (Wahle et al., 2022), tweet intimacy prediction (Michail et al., 2023), named entity recognition (Tang et al., 2023a), machine translation (Yang and Nicolai, 2023) etc. GLLM-based data generation for data augmentation is explored in multiple domains like general (Whitehouse et al., 2023; Parikh et al., 2023; Eldan and Li, 2023; Xu et al., 2023e; Liu et al., 2023g; Peng et al., 2023c; Wahle et al., 2022; Yang and Nicolai, 2023; Zhao et al., 2023c; Xu et al., 2023b), social media (Zhan et al., 2023b; Hartvigsen et al., 2022; Markov et al., 2023; Michail et al., 2023; Yu et al., 2023d), news (Yu et al., 2023d), scientific literature (Xu et al., 2023e; Wahle et al., 2022), healthcare (Wang et al., 2023m; Guo et al., 2023c; Tang et al., 2023a), dialogue (Malkiel et al., 2023), programming (Gunasekar et al., 2023) etc. Table 19 presents a summary of research works exploring GLLMs for data generation-based data augmentation.

Some of the research works explored GLLMs for data generation-based data augmentation in various text classification tasks (Zhan et al., 2023b; Hartvigsen et al., 2022; Markov et al., 2023; Parikh et al., 2023; Michail et al., 2023; Yu et al., 2023d). For example, Hartvigsen et al. (2022) used GPT-3 with demonstration-based prompting to create a large-scale synthetic dataset for the detection of implicit hate speech. Here, the authors explored a variant of constrained beam search to ensure subtle toxicity in the generated examples. Michail et al. (2023) investigated the effectiveness of ChatGPT-generated synthetic data to fine-tune multilingual models for tweet intimacy prediction in the case of languages with no labelled instances. Here, ChatGPT is prompted with instructions and examples from a high-resource language and asked to generate new examples in the target language. Most of the existing research works use simple prompts for data generation, limiting the diversity of the generated synthetic data. To address this,

Table 19
Summary of research works exploring GLLMs for data generation-based data augmentation. Here ZS represents zero-shot and FS represents few-shot.

Paper	GLLMs explored	Task(s)	Prompt settings	Domain(s)	Language(s)
Zhan et al. (2023b)	ChatGPT	Text classification	ZS	Social media	Chinese
Wang et al. (2023m)	ChatGPT	Note2Dialogue generation	ZS	Healthcare	English
Gunasekar et al. (2023)	GPT-3.5	Training Phi-1 LLM	ZS	Programming	English
Whitehouse et al. (2023)	ChatGPT, GPT-4	Cross-lingual common sense reasoning	FS	General	Multiple languages
Hartvigsen et al. (2022)	GPT-3	Hate speech detection	FS	Social media	English
Markov et al. (2023)	GPT-3	Undesired context detection	ZS, FS	Social media	English
Guo et al. (2023c)	ChatGPT, GPT-4	Question answering	ZS	Healthcare	English
Parikh et al. (2023)	GPT-3	Intent classification	ZS	General	English
Eldan and Li (2023)	GPT-3.5, GPT-4	Training smaller LLMs	ZS	General	English
Xu et al. (2023e)	GPT-3.5	Relation extraction	FS	General, scientific literature	English
Liu et al. (2023g)	GPT-4	CoT instruction tuning	FS	General	English
Peng et al. (2023c)	GPT-4	Instruction tuning	ZS	General	English, Chinese
Malkiel et al. (2023)	GPT-3	Call segmentation, topic extraction	ZS	Dialogue	English
Wahle et al. (2022)	GPT-3	Paraphrase detection	ZS	General, scientific literature	English
Michail et al. (2023)	ChatGPT	Tweet intimacy prediction	FS	Social media	Multiple languages
Tang et al. (2023a)	ChatGPT	Named entity recognition, Relation classification	ZS	Healthcare	English
Yu et al. (2023d)	ChatGPT	Topic classification	ZS	News, social media	English
Yang and Nicolai (2023)	ChatGPT	Neural machine translation	ZS	General	Multiple languages
Zhao et al. (2023c)	GPT-3, Codex	Table question answering	ZS	General	English
Xu et al. (2023b)	GPT-4	Text generation evaluation	ZS	General	Multiple languages

Yu et al. (2023d) proposed a novel approach that leverages attributed prompts for data generation to increase the diversity in the generated data. Based on the evaluation on four topic classification datasets, the authors observed that (i) the proposed approach enhances the model performance and (ii) reduces the querying cost of ChatGPT by a large margin.

Some of the research works explored GLLMs for data generation-based data augmentation in various information extraction tasks like relation extraction (Xu et al., 2023e), relation classification (Tang et al., 2023a) and named entity recognition (Tang et al., 2023a). Xu et al. (2023e) evaluated how effective is the GPT-3.5 model for relation classification. To address the data scarcity problem in few-shot settings, the authors used the GPT-3.5 model to generate additional data. The prompt used for data generation consists of instance descriptions along with some example instances. Tang et al. (2023a) used ChatGPT in zero-shot settings to generate synthetic data for tasks like named entity recognition and relation classification in the healthcare domain. The authors showed that the model fine-tuned on this synthetic data outperforms zero-shot ChatGPT by a large margin in both tasks.

Some of the research works explored GLLMs for data generation in LLM development stages, like LLM pretraining (Gunasekar et al., 2023; Eldan and Li, 2023) and instruction tuning (Liu et al., 2023g; Peng et al., 2023c). Gunasekar et al. (2023) trained Phi-1, a code LLM using GPT-3.5 generated synthetic textbook and code data. Here, the training corpus includes 1B tokens of GPT-3.5 generated Python textbook and code data along with 6B tokens of code data from the web. Eldan and Li (2023) explored GLLMs like GPT-3.5 and GPT-4 models to generate TinyStories, a synthetic dataset of stories with only the words understood by typical 3 to 4-year-old kids. The authors demonstrated that the GLLM generated dataset can be used to train smaller LLMs, which can generate coherent and consistent stories with near-perfect grammar. Instruction tuning requires large human-annotated datasets, which are often difficult to obtain. Stanford Alpaca[5] and Vicuna[6] showed the effectiveness of synthetic instruction tuning datasets generated using GPT-3.5 and ChatGPT, respectively. Inspired by the success of these models, Peng et al. (2023c) explored advanced models like GPT-4 to generate instruction-tuning datasets in English and Chinese languages. The experiment results showed that GPT-4 generated instruction tuning datasets further enhance the zero-shot performance of LLaMA models. Liu et al. (2023g) used GPT-4 to generate LogiCoT, a synthetic dataset of CoT rationales. This dataset can be used for instruction tuning the LLMs to enhance their logical reasoning abilities.

8. Detecting GLLM generated text

Overview. GLLMs demonstrated extraordinary human-like capabilities to understand user queries, follow the instructions and then answer the user queries with high-quality content. Apart from responding to user queries, these models can also generate news articles, research papers, code and essays with human-like fluency. With the ability to generate text with human-like fluency, these models are widely adopted in a variety of real-world applications like writing assistants, coding assistants, chatbots, etc. (Mireshghallah et al., 2023). Although there is a lot of excitement about GLLMs and their applications in recent times, there are also growing concerns regarding the potential misuse of these models for illegal activities (Guo et al., 2023d), such as fake

[5] https://crfm.stanford.edu/2023/03/13/alpaca.html.
[6] https://lmsys.org/blog/2023-03-30-vicuna/.

news on social media platforms (Hacker et al., 2023; De Angelis et al., 2023), fake reviews on e-commerce websites (Mitrovi'c et al., 2023), fake research papers (Gao et al., 2023a), academic fraud (Cotton et al., 2023), etc. For example, these models can be easily used by malicious users to create fake news (Hacker et al., 2023; De Angelis et al., 2023) and propagate on social platforms at a large scale to exaggerate or manipulate the facts to get an undue advantage, especially during political campaigns. Similarly, students can use these models to write their assignments or generate code for their projects (Cotton et al., 2023), and GLLM generated fake research papers (Gao et al., 2023a) can have a serious impact on the scientific community as these papers are written without conducting any experiments.

There is a strong need for the development of approaches to detect GLLM generated text, as there are growing concerns regarding the misuse of GLLMs. Such approaches help to distinguish the GLLM generated text from human-generated text and verify the source as well as the authenticity of the information. However, detecting GLLM generated text is more challenging as models like ChatGPT and GPT-4 can generate content with human-like fluency.

Research exploring the detection of GLLM generated text. To avoid misuse and ensure the safe use of these models, the research community focused on developing approaches to identify the GLLM generated text accurately. The recent research works explored the detection of GLLM generated text in multiple domains like scientific literature (Theocharopoulos et al., 2023; Zaitsu and Jin, 2023; Yu et al., 2023a; Yang et al., 2023a), academic (Liu et al., 2023m; Orenstrakh et al., 2023), healthcare (Liao et al., 2023; Zhan et al., 2023a; Yang et al., 2023a), news (Clark et al., 2021), legal (Zhan et al., 2023a; Guo et al., 2023d), social media (Yang et al., 2023a; Mitrovi'c et al., 2023), Finance (Guo et al., 2023d) etc. Most of the research works focused on the English language, while a few research works focused on other languages like Japanese (Zaitsu and Jin, 2023), German (Yang et al., 2023a) and Spanish (Orenstrakh et al., 2023). Table 20 presents a summary of research works exploring the detection of GLLM generated text.

Some of the research works focused on assessing the effectiveness of the existing machine-generated text detection tools to detect GLLM generated text. A number of online tools are available, ranging from simple classifiers based on logistic regression to advanced classifiers based on PLMs to detect ChatGPT-generated text. To assess the effectiveness of these tools, Pegoraro et al. (2023) introduced a dataset having ChatGPT-generated responses for questions from various domains like finance, medicine, etc., and user-generated responses from social media platforms. The comprehensive evaluation showed that the maximum success rate of these tools is less than 50% only, which leaves a lot of room for improvement. Orenstrakh et al. (2023) evaluated the effectiveness of eight popular detectors using three metrics, namely resilience, false positives and accuracy. The authors observed that CopyLeaks, GPTKit and GLTR achieve the best results for the metrics accuracy, false positives and resilience. However, all these detectors struggle with non-English languages and paraphrased LLM-generated text. There is a lack of comprehensive evaluation benchmark to detect machine-generated text as the existing approaches use different models, datasets and settings. To address this, He et al. (2023c) proposed MGTBench, the first machine-generated text detection benchmark. Evaluation on this benchmark showed that, except for the ChatGPT detector (Guo et al., 2023d) and LM detector (Ippolito et al., 2020), the performance of other detectors is not satisfactory. Guo et al. (2023d) introduced the HC3 dataset, having human-authored and ChatGPT-generated responses to questions from multiple domains like legal, healthcare, finance, psychology, etc. The performance of existing detection approaches on the HC3 dataset is just satisfactory, and linguistic analysis showed that human-authored answers are short in length but use a large vocabulary compared to ChatGPT-generated answers.

Some of the research works focused on developing approaches based on trained classifier models to detect GLLM generated text. Theocharopoulos et al. (2023) evaluated the effectiveness of classifiers based on models like logistic regression, support vector machine, LSTM, and BERT to identify GPT-3 generated scientific abstracts. The LSTM-based classifier with word2vec embeddings achieves an accuracy of more than 98% and outperforms other classifiers. Zaitsu and Jin (2023) observed that LLM-generated texts differ significantly from human-written texts in terms of stylometric features. The authors demonstrated that random forest trained with different stylometric features can identify the LLM-generated Japanese text with 100% accuracy. Liu et al. (2023m) reported that fine-tuned RoBERTa model achieves an accuracy of more than 90% on the AruGPT dataset of human-written and GLLM generated argumentative essays. Moreover, linguistic analysis revealed that GLLM generated texts tend to be more complex syntactically, while human-generated texts are lexically more complex. To facilitate the development of a ChatGPT-written abstract detector, Yu et al. (2023a) introduced CHEAT, a large dataset of ChatGPT and human-written abstracts. Based on the evaluation of multiple existing approaches like ZeroGPT, OpenAI detector, ChatGPT-detector-roberta (Guo et al., 2023d) and ChatGPT-qa-detector-roberta (Guo et al., 2023d), the authors reported that performance is away from satisfactory and the human involvement further increases the detection difficulty. Zhan et al. (2023a) treated the detection of LLM generated as a binary classification problem and proposed a novel approach based on fine-tuned RoBERTa model. The authors reported that the proposed approach exhibits good performance and also has the ability to detect the text generated using a detection evasion technique. Mitrovi'c et al. (2023) proposed a novel approach based on DistilBERT (Sanh et al., 2019) and SHAP (Lundberg and Lee, 2017) to detect the machine-generated text and explain the reasoning. The proposed approach achieves an accuracy of 79%, and based on the explanations, the authors observed that ChatGPT-generated text maintains a polite tone, lacks specific details and generally refrains from expressing emotions.

Chen et al. (2023d) introduced OpenGPTText, which includes ChatGPT-generated paraphrased text. The authors reported that fine-tuned classifiers based on models like RoBERTa and T5 can achieve impressive results in detecting ChatGPT-generated text with an accuracy of more than 97%. Yu et al. (2023b) introduced GPT-Pat, a novel approach based on ChatGPT, a Siamese network and binary classifier, to detect machine-generated text effectively. The proposed approach enhances the SOTA accuracy by more than 12% and also exhibits better robustness to attacks like re-translation and text polishing. Yang et al. (2023d) focused on detecting GLLM-polished text, which is more challenging and useful in real-world applications. The proposed approach involves training a classification model to identify the machine-generated text and a polish ratio (regression) model to explain the ChatGPT involvement. A Polish ratio of 0.2 indicates ChatGPT involvement and a value of more than 0.6 represents the text is entirely ChatGPT generated.

Training-based approaches to detect LLM-generated text have limited flexibility, especially when used for new domains (Yang et al., 2023a). To overcome this drawback, some of the research works focused on developing training-free approaches to detect GLLM generated text. Yang et al. (2023a) proposed DNA-GPT, a training-free approach based on divergent n-gram analysis. With the proposed approach, the authors achieved SOTA results on both English and German datasets. Wang et al. (2023g) proposed a novel framework called FLAIR to detect LLM-based bots with a single question in an effective way. The results showed that the proposed approach is effective and a good alternative to existing CAPTCHA-based approaches. Mireshghallah et al. (2023) investigated whether models other than the generator can be used to identify machine-generated text. In general, smaller models serve as more effective universal text detectors. These models exhibit better accuracy in identifying text produced by both small and larger models. For example, OPT-125M achieves better results compared to the GPT-J 6B model in detecting ChatGPT-generated text.

Some of the research works focused on assessing the robustness of machine-generated text detectors towards different attacks. Shi et al.

Table 20
Summary of research works exploring the detection of GLLM generated text.

Paper	Detect	Approach	Satisfactory performance	Training free	Domain(s)	Language(s)
Pegoraro et al. (2023)	ChatGPT generated text	Evaluate multiple online tools	No	–	Multiple domains	English
Theocharopoulos et al. (2023)	GPT-3 generated text	Classifiers based on machine learning models like LR, SVM and deep learning models like LSTM and BERT	Yes	No	Scientific literature	English
Zaitsu and Jin (2023)	ChatGPT and GPT-4 generated text	Classifier based on random forest and stylometric features	Yes	No	Scientific literature	Japanese
Liu et al. (2023m)	GPT-3 and ChatGPT generated text	Classifier based on models like SVM and RoBERTa	Yes	No	Academic	English
Yu et al. (2023a)	ChatGPT generated text	Classifier based on models like RoBERTa	No	No	Scientific literature	English
Liao et al. (2023)	ChatGPT generated text	Classifier based on models like BERT	Yes	No	Healthcare	English
Orenstrakh et al. (2023)	ChatGPT generated text	Evaluate multiple online tools	Yes	–	Academic	English, Spanish
Clark et al. (2021)	GPT-3 generated text	Evaluate human evaluators	No	–	Stories, news, recipes	English
Zhan et al. (2023a)	ChatGPT and GPT-4 generated text	Classifier based on models like BERT and RoBERTa	Yes	No	Law, medical, dialogue, general	English
Yang et al. (2023a)	GPT-3.5, ChatGPT and GPT-4 generated text	Training free divergent N-gram Analysis	Yes	Yes	Healthcare, social media, scientific literature	English, German
Shi et al. (2023)	ChatGPT generated text	Evaluate the robustness of existing detectors	No	–	General	English
Khalil and Er (2023)	ChatGPT generated text	Evaluate existing plagiarism tools	No	–	General	English
He et al. (2023c)	ChatGPT generated text	Propose benchmark and evaluate existing detectors	Yes	–	General	English
Mitrovi'c et al. (2023)	ChatGPT generated text	Propose novel approach based on DistilBERT and SHAP to detect and explain	Yes	No	Social media	English
Guo et al. (2023d)	ChatGPT generated text	Introduce new dataset and evaluate multiple existing detection models	Yes	–	General, finance, healthcare, legal, psychology	English
Wang et al. (2023g)	GPT-3 and ChatGPT-based bots	Propose FLAIR to detect online GPT-3 and ChatGPT-based bots	Yes	Yes	General	English
Chen et al. (2023d)	ChatGPT generated text	Classifiers based on models like RoBERTa and T5	Yes	No	General	English
Mireshghallah et al. (2023)	ChatGPT generated text	Propose a zero-shot approach based on local optimality	Yes	Yes	General	English
Yu et al. (2023b)	ChatGPT generated text	Propose an approach based on Siamese Network and binary classifier	Yes	No	General	English
Yang et al. (2023d)	ChatGPT polished text	Trains classifier and polish ratio models to detect and explain	Yes	No	General	English
Krishna et al. (2023)	GPT-3.5 generated text	Evaluate robustness using paraphrase attacks	No	–	General	English

(2023) evaluated the robustness of existing detectors using attacks like synonym word replacement and writing style modification. The authors implemented both attacks using LLMs. The results showed that the existing detectors are not robust to the attacks, which emphasizes the need for more robust and reliable detectors to detect and avoid the misuse of LLMs. Krishna et al. (2023) showed that existing detectors like OpenAI detector, GPTZero and DetectGPT (Mitchell et al., 2023) are not robust to paraphrase attacks. For example, paraphrase attacks result in a drop of more than 65% accuracy in the case of DetectGPT.

Some of the research works focused on assessing the effectiveness of humans in identifying GLLM generated text. For example, Clark et al. (2021) observed that non-expert evaluators are unable to differentiate GPT-3 generated text from human-authored text in three different domains, namely news, recipes and stories. The reason for this is the evaluators arrived at their decisions based on surface-level features without considering the advanced text generation capabilities of the GPT-3 model.

9. Evaluation of GLLMs

GLLMs with their remarkable performances across a variety of tasks, gained a lot of attention in both industry and academia which eventually led to their use in a lot of real-world applications. GLLMs are double-edged swords i.e., apart from remarkable performances, GLLMs are also associated with a lot of potential risks (Guo et al., 2023a). For example, GLLMs sometimes generate factually incorrect text, biased and harmful text and also tend to leak private data. So, it is highly recommended to have a thorough evaluation of GLLMs to understand their limitations which not only helps the researchers to further improve them but also ensures their safe and reliable use (Guo et al., 2023a; Chang et al., 2023).

In recent times, several benchmarks have been proposed to assess the performance as well as understand the limitations of GLLMs across tasks and domains (Zhuang et al., 2023). A benchmark serves as a standardized method for evaluating a model's ability to generalize

Table 21
Summary of research works exploring GLLMs robustness to out-of-distribution instances, adversarial prompts and adversarial inputs. Here ZS represents zero-shot, and FS represents few-shot.

Paper	GLLMs explored	Task(s)	Prompt settings	Robustness	Domain(s)	Language(s)
Chen et al. (2023g)	GPT-3, GPT-3.5	Nine NLU tasks	ZS, FS	Adversarial input	General	English
Wang et al. (2023b)	GPT-3.5, ChatGPT	Four NLU tasks, machine translation	ZS	Out of distribution	General, medical	English
Zhuo et al. (2023)	Codex	Semantic parsing	ZS, FS	Adversarial input	Programming	English
Zhu et al. (2023d)	ChatGPT	Eight tasks including four NLU tasks	ZS, FS	Adversarial prompt	General	English
Shirafuji et al. (2023)	Codex, InstructGPT, ChatGPT	Code generation	ZS	Adversarial prompt	Programming	English
Zhao et al. (2023c)	GPT-3, Codex	Table question answering	FS	Adversarial input	General	English
Han et al. (2023)	ChatGPT	Fourteen IE tasks	ZS, FS	Adversarial prompt	General	English
Liu et al. (2023f)	ChatGPT, GPT-4	Question answering	ZS, FS	Out-of-distribution	General	English
Liu et al. (2023c)	ChatGPT	Text-to-SQL generation	ZS	Adversarial input	General	English

across different tasks (Kalyan et al., 2021). Typically, it includes a collection of diverse and challenging datasets, an online leaderboard for model comparison and ranking, and a designated metric for assessing overall performance across tasks (Wang et al., 2018). The use of a benchmark is essential to have a consistent evaluation framework, enabling the tracking of progress in the development of LLMs. Without a benchmark, evaluating models lacks a standardized approach and is challenging. Tables 22 and 23 present a summary of various benchmarks which assess the abilities of GLLMs across different tasks and domains.

10. Robustness of GLLMs

Overview. GPT-3 family LLMs achieve impressive performances in zero and few-shot settings in many NLP tasks. In some tasks like text classification (Sun et al., 2023b), relation extraction (Wan et al., 2023), etc. GLLMs without any explicit fine-tuning outperform state-of-the-art fine-tuned models. For example, Sun et al. (2023b) demonstrated that InstructGPT, with the advanced prompting strategy, achieves SOTA results using just 16 examples on four text classification datasets. Similarly, Wan et al. (2023) achieved SOTA results in relation extraction with the GPT-RE framework. However, to increase the reliability of these models in real-world applications, especially in critical domains like medicine, it is essential to systematically study the robustness of these models in various scenarios. Adversarial robustness refers to the model's ability to maintain good performance even in the case of deliberately crafted instances (Goyal et al., 2022; Qiu et al., 2022). These instances are called adversarial instances and are carefully designed by making subtle changes in the original inputs to deceive the model. Out-of-distribution (OOD) instances refer to examples that differ significantly from the data distribution used to train the model (Shen et al., 2021). These instances fall outside the range of the model's training data and present challenges to the model's performance and generalization ability. Some of the recent research works focused on evaluating the robustness of GLLMs to out-of-distribution instances (Wang et al., 2023b; Liu et al., 2023f), adversarial prompts (Zhu et al., 2023d; Shirafuji et al., 2023; Han et al., 2023) and adversarial inputs (Chen et al., 2023g; Zhuo et al., 2023; Zhao et al., 2023c; Liu et al., 2023c) in one or more natural language processing tasks. Table 21 presents a summary of research works assessing GLLMs robustness to out-of-distribution instances, adversarial prompts and adversarial inputs.

Research works exploring GLLMs robustness. Some of the research works evaluated the robustness of GLLMs in specific tasks like semantic parsing (Zhuo et al., 2023), code generation (Shirafuji et al., 2023), table question answering (Zhao et al., 2023c), multi-choice question answering (Liu et al., 2023f) and text-to-SQL generation (Liu et al., 2023c). Zhuo et al. (2023) reported that Codex-based semantic parsers are not robust to adversarial examples, and the robustness can be enhanced using few-shot in-context learning. Shirafuji et al. (2023) studied the robustness of GPT-3 family models like Codex, InstructGPT, and ChatGPT to adversarial prompts in code generation task. The authors observed that InstructGPT and ChatGPT exhibit better robustness compared to Codex. However, there is much room for improvement, indicating that quality code generation requires well-designed prompts. Zhao et al. (2023c) proposed RobuT, a benchmark to systematically study the robustness of LLMs to adversarial inputs in table question answering. The authors reported that GLLMs like GPT-3 and Codex exhibit better robustness than fine-tuned models. Moreover, the authors demonstrated that GLLM generated adversarial inputs can enhance the adversarial robustness of fine-tuned models. Liu et al. (2023f) reported that ChatGPT and GPT-4 perform well in multiple choice question answering but struggle to answer out-of-distribution questions. Liu et al. (2023c) showed that ChatGPT exhibits impressive zero-shot performance in Text-to-SQL generation. Moreover, ChatGPT demonstrates better robustness to adversarial inputs than SOTA models in text-to-SQL generation.

Some of the research works evaluated the GLLM robustness in multiple natural language understanding and generation tasks (Chen et al., 2023g; Wang et al., 2023b; Zhu et al., 2023d; Han et al., 2023). Chen et al. (2023g) assessed the robustness of GPT-3 and GPT-3.5 models on 21 datasets covering nine natural language understanding tasks. Here the authors used adversarial text transformations from TextFlint (Wang et al., 2021a). The authors observed that the models are robust in tasks like machine reading comprehension and exhibit performance degradation of more than 35% in tasks like sentiment analysis and natural language inference. Wang et al. (2023b) evaluated the robustness of GPT-3.5 and ChatGPT models on adversarial and out-of-distribution (OOD) samples on nine datasets covering four NLU tasks and machine translation. The authors observed that ChatGPT exhibits good performances on adversarial and OOD samples, but still, there is much room for improvement.

Zhu et al. (2023d) developed PromptBench, a benchmark with more than 4k adversarial prompts to evaluate the robustness of LLMs to adversarial prompts. The benchmark covers 13 datasets spanning eight tasks, including four NLU tasks. The authors observed that GLLMs are not robust to adversarial prompts. Moreover, word-level attacks are the most effective, which results in a performance drop of more than 30%. Based on the evaluation of ChatGPT on fourteen information extraction sub-tasks, Han et al. (2023) showed that ChatGPT is vulnerable to adversarial prompts, i.e., the performance is greatly affected by including irrelevant context in the prompt.

Table 22
Summary of benchmarks assessing the abilities of GLLMs across various tasks and domains.

Benchmark	Evaluates	Domain	Language(s)	Description
KoLA (Yu et al., 2023c)	World knowledge	General	English	KoLA stands for Knowledge-oriented LLM Assessment benchmark covering 19 tasks and assesses the world knowledge of GLLMs.
SciBench (Wang et al., 2023c)	College-level scientific problem solving	Education	English	SciBench stands for Scientific problem-solving Benchmark and includes two datasets of scientific problems at the college level.
FinEval (Zhang et al., 2023b)	Chinese finance domain knowledge	Finance	Chinese	FinEval includes over a thousand multiple-choice questions covering more than 30 academic subjects from the Finance domain.
LegalBench (Guha et al., 2023)	Legal reasoning	Legal	English	LegalBench is a legal reasoning benchmark created through collaborative efforts, featuring 162 tasks that encompass six distinct categories of legal reasoning.
SciEval (Sun et al., 2023a)	Scientific research ability	Education	English	SciEval benchmark includes both objective and subjective questions from science subjects like biology, physics and chemistry.
LongBench (Bai et al., 2023a)	Long context understanding	Multiple domains	English, Chinese	LongBench consists of 21 datasets spanning 6 task categories, available in both English and Chinese.
LawBench (Fei et al., 2023)	Legal knowledge	Legal	Chinese	LawBench evaluates LLMs in three dimensions namely legal knowledge memorization, understanding and applying. This benchmark covers 20 tasks spanning over 5 task types.
BHASA (Leong et al., 2023)	Language understanding, generation and reasoning	Multiple domains	Indonesian, Thai, Tamil, Vietnamese	BHASA evaluates LLMs in South East Asian languages like Tamil, Thai, Vietnamese and Indonesian. This benchmark includes eight tasks spanning over natural language reasoning, generation and understanding.
L2CEval (Ni et al., 2023)	Language to code generation	Programming	English	L2CEval benchmark systematically evaluates the language-to-code generation capabilities of LLMs across seven different tasks.
XSafety (Wang et al., 2023j)	LLM safety	Multiple domains	Ten languages	XSafety is an LLM safety benchmark which includes fourteen types of frequently encountered safety concerns, spanning ten languages that belong to diverse language families.
TRAM (Wang and Zhao, 2023)	Temporal reasoning	General	English	TRAM, a benchmark for temporal reasoning includes ten datasets that cover a range of events related to temporal aspects like duration, frequency, arithmetic and order.
FELM (Chen et al., 2023i)	Factuality	Multiple domains	English	FELM is a factuality evaluating LLM benchmark focusing diverse domains including math, reasoning and world knowledge.
LAiW (Dai et al., 2023a)	Legal knowledge	Legal	Chinese	LAiW is the first benchmark for Legal LLMs in the Chinese language and it evaluates three levels of legal abilities.
LLMBar (Zeng et al., 2023)	Instruction following ability	General	English	LLMBar is a meta-evaluation benchmark assessing the instruction following ability of LLMs and consists of 419 instances.
BLESS (Kew et al., 2023)	Text simplification ability	Multiple domains	English	BLESS benchmark evaluates the text simplification ability of LLMs and includes instances from three different domains.

11. GLLMs as evaluators

Overview. Natural language processing tasks can be broadly classified into natural language understanding (NLU) and natural language generation (NLG). NLU involves the interpretation of text, while NLG involves generating human-like text. The evaluation of NLU outputs is pretty straightforward, while the evaluation of NLG outputs is challenging because of the diversity and inherent complexity of the text (Chen et al., 2023f). Moreover, the NLG evaluation involves assessing the generated text outputs in multiple dimensions, such as coherence, fluency, naturalness and semantic consistency. Human evaluation and automatic evaluation are two existing approaches for NLG evaluation. The human evaluation depends on competent annotators for an accurate and reliable assessment (Sai et al., 2022).

Human Evaluation vs. Automatic Evaluation. Human evaluation is treated as the gold standard, but it is time-consuming, expensive, difficult to scale, inconsistent, and not reproducible (Chen et al., 2023f; Wang et al., 2023e). To address the issues with human evaluation, automatic evaluation metrics are developed, which fall broadly into two categories: n-gram-based and embedding-based. N-gram-based metrics assess the quality based on the lexical overlap between the generated and reference texts. Some of the commonly used n-gram-based metrics are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005). However, these metrics have a poor correlation with human scores because of their inability to capture semantic meaning (Kocmi et al., 2021). Later, with the evolution of transformers and PLMs, the researchers developed embedding-based metrics like BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019), BARTScore (Yuan et al., 2021), CodeBERTScore (Zhou et al., 2023) etc. These metrics leverage the PLMs and assess the quality based on the semantic similarity between the generated and reference text. The main drawback of the existing automatic evaluation metrics is the requirement for references, which are difficult to obtain, especially in low-resource domains. Moreover, with just a few references, it is not possible to get an accurate and reliable assessment as few references

Table 23
Summary of benchmarks assessing the abilities of GLLMs across various tasks and domains.

Benchmark	Evaluates	Domain	Language(s)	Description
MedEval (He et al., 2023d)	Medical knowledge	Medical	English	MedEval benchmark is designed to evaluate LLMs in medical domain and covers diverse tasks.
XLingEval (Choudhury et al., 2023)	Multilingual medical capabilities	Medical	English, Spanish, Chinese, Hindi	XLingEval is a multilingual benchmark introduced the assess the effectiveness of LLMs in medical domain.
M4LE (Kwan et al., 2023)	Long context understanding	Multiple domains	English, Chinese	M4LE benchmark is designed to evaluate long context understanding of LLMs and includes 36 datasets from twelve domains covering 11 task types.
BizBench (Koncel-Kedziorski et al., 2023)	Quantitative reasoning	Finance	English	BizBench benchmark is introduced to assess the quantitative reasoning abilities of LLMs and includes eight reasoning tasks.
CodeScope (Yan et al., 2023)	Code understanding	Programming	43 programming languages	CodeScope benchmark evaluates code generation and understanding abilities of LLM and covers eight coding tasks from forty-three programming languages.
FollowEval (Jing et al., 2023)	Instruction following	General	English, Chinese	FollowEval benchmark is introduced to evaluate LLMs based on their performance in five essential aspects of instruction following.
FinanceBench (Islam et al., 2023)	Question answering	Finance	English	FinanceBench is designed to assess the capabilities of LLMs in the context of open-book financial question answering (QA) and it includes 10,231 questions related to publicly traded companies, each accompanied by evidence strings and relevant answers.
ARB (Sawada et al., 2023)	Advanced reasoning	Multiple domains	English	ARB is an LLM benchmark that consists of complex reasoning problems spanning various disciplines like mathematics, physics, biology, chemistry, and law.
TimeBench (Chu et al., 2023)	Temporal reasoning	General	English	The TimeBench benchmark is designed to evaluate the temporal reasoning abilities of LLMs and consists of 10 tasks that address a wide range of temporal reasoning phenomena.
TaskEval (Shen et al., 2023c)	Task automation ability	Programming	Python	TaskEval aims to evaluate the proficiency of LLMs in automating tasks through a comprehensive and quantitative evaluation of three different aspects.
PromptBench (Zhu et al., 2023d)	Robustness to adversarial prompts	General	English	PromptBench is introduced to assess the robustness of LLMs to adversarial prompts and this benchmark includes more than 4500 adversarial prompts.

cannot account for all the semantic variations (Chen et al., 2023f). So, there is a strong need for automatic evaluation metrics which are reference-free.

GLLM-based Evaluation. Recently, with the huge success of GLLMs in most of the NLP tasks, the research community focused on developing automatic evaluation metrics based on these models. These models possess the ability of in-context learning, while instruction tuning enables these models to align themselves with human evaluation (Ouyang et al., 2022). These two abilities enable these models to imitate the behaviour of human evaluators, who typically evaluate natural language generation task outputs by understanding instructions and the given examples. The GLLM-based evaluation metrics demonstrate a strong correlation with human scores even in the absence of reference outputs (Liu et al., 2023d; Fu et al., 2023). Table 24 presents a summary of research works exploring GLLM-based evaluation for various natural language generation tasks.

Research works exploring GLLM-based evaluation. The NLP researchers proposed various GLLM-based evaluation frameworks to evaluate the outputs of various NLG tasks like code generation (Zhuo, 2023), text style transfer (Lai et al., 2023b), text summarization (Liu et al., 2023d; Chen et al., 2023f; Luo et al., 2023; Shen et al., 2023a; Fu et al., 2023; Liu et al., 2023b; Gao et al., 2023b; Tang et al., 2023b; Jain et al., 2023; Wang et al., 2023f), dialogue generation (Liu et al., 2023d; Chen et al., 2023f; Fu et al., 2023), machine translation (Kocmi and Federmann, 2023; Lu et al., 2023a; Xu et al., 2023b; Fu et al., 2023; Tang et al., 2023b; Yang et al., 2023f), story generation (Chen et al., 2023f; Wang et al., 2023f), paraphrase generation (Chen et al., 2023f), text-to-image synthesis (Lu et al., 2023b), data-to-text generation (Fu et al., 2023; Wang et al., 2023f), image captioning (Tang et al., 2023b), text generation (Wang et al., 2023e), open-ended question answering (Bai et al., 2023b; Zheng et al., 2023a). Most of the research works proposed evaluation frameworks using direct prompting, while some of the research works introduced evaluation frameworks based on advanced prompting strategies like chain-of-thoughts (Zhuo, 2023; Liu et al., 2023d) and error analysis prompting (Lu et al., 2023a). Some of the proposed evaluation frameworks work with and without Refs. Zhuo (2023), Kocmi and Federmann (2023) and Wang et al. (2023f), while some of them require Refs. Lai et al. (2023b), Lu et al. (2023a), Xu et al. (2023b), Tang et al. (2023b) and Yang et al. (2023f), and some do not require any Refs. Liu et al. (2023d), Chen et al. (2023f), Luo et al. (2023), Shen et al. (2023a), Fu et al. (2023), Liu et al. (2023b), Gao et al. (2023b), Wang et al. (2023e), Jain et al. (2023), Bai et al. (2023b) and Zheng et al. (2023a).

Lai et al. (2023b) investigated how effective ChatGPT is to evaluate text style transfer task along three dimensions: fluency, content and style. The model achieves good correlations with human judgements, and the best results are obtained by using separate prompts for each dimension evaluation. Kocmi and Federmann (2023) proposed GEMBA, a GPT-based metric to assess translation output quality, with references being optional. The authors reported that GPT-3.5 and higher models are only useful for the assessment, and GPT-4 achieves the best results. Based on the evaluation of four natural language generation tasks, paraphrase generation, text summarization, story generation and dialogue response generation, Chen et al. (2023f) showed that explicit score with greedy decoding strategy is the best way to assess NLG outputs using GLLMs like ChatGPT. Luo et al. (2023) evaluated ChatGPT's ability as a factual inconsistency evaluator for text summarization task. Experiment results showed that ChatGPT outperforms existing metrics on most of the datasets.

Table 24
Summary of research works exploring GLLM-based evaluation for natural language generation tasks. Here ZS represents zero-shot, and FS represents few-shot.

Paper	GLLMs explored	Task(s)	Prompt settings	References required	Domain(s)	Language(s)
Zhuo (2023)	ChatGPT	Code generation	ZS	Optional	Programming	Five programming languages
Lai et al. (2023b)	ChatGPT	Text style transfer	ZS	Yes	General	English
Liu et al. (2023d)	ChatGPT, GPT-4	Text summarization, dialogue generation	ZS	No	General	English
Kocmi and Federmann (2023)	GPT, GPT-3.5, ChatGPT, GPT-4	Machine translation	ZS	Optional	General	English, German, Chinese, Russian
Chen et al. (2023f)	GPT-3.5, ChatGPT	Text summarization, dialogue generation, story generation, paraphrase generation	ZS	No	General	English
Lu et al. (2023a)	GPT-3.5, ChatGPT	Machine translation	ZS, FS	Yes	General	English, Chinese, German
Luo et al. (2023)	ChatGPT	Text summarization	ZS	No	General	English
Shen et al. (2023a)	ChatGPT	Text summarization	ZS	No	General	English
Lu et al. (2023b)	GPT-4	Text-to-image synthesis	ZS	N/A	General	English
Xu et al. (2023b)	GPT-4	Machine translation	ZS	Yes	General	English, German, Russian
Fu et al. (2023)	GPT-3, GPT-3.5	Dialogue generation, machine translation, text summarization, data-to-text generation	ZS, FS	No	General	English, Chinese
Liu et al. (2023b)	GPT-3, ChatGPT	Text summarization	ZS	No	General	English
Gao et al. (2023b)	ChatGPT	Text summarization	ZS	No	General	English
Tang et al. (2023b)	GPT-3.5	Machine translation, text summarization, image caption	ZS	Yes	General	English
Wang et al. (2023e)	ChatGPT, GPT-4	Text generation	ZS	No	General	English
Jain et al. (2023)	GPT-3.5	Text summarization	ZS	No	General	English
Wang et al. (2023f)	ChatGPT	Text summarization, story generation, data-to-text generation	ZS	Optional	General	English
Bai et al. (2023b)	GPT-4	Open-ended question answering	ZS	No	General	English
Yang et al. (2023f)	GPT-4	Machine translation	ZS	Yes	General	Multiple languages
Zheng et al. (2023a)	GPT-4	Open-ended question answering	ZS	No	General	English

Shen et al. (2023a) explored how effective ChatGPT can be as a zero-shot evaluator for abstractive summarization systems using different evaluation methods like likert scaling (He et al., 2022b) and head-to-head comparisons (Shen et al., 2022). Extensive analysis showed that likert scaling implemented as a multiple-choice question gives the best and most stable results. Liu et al. (2023b) designed a novel approach which uses BRIO (Liu et al., 2022), a contrastive learning-based method, to train smaller models like BART for text summarization and metrics like GPTScore (Fu et al., 2023) or GPTRank for evaluation. The contrastive learning training method helps the model to effectively utilize the supervision signal offered by the reference LLMs. The evaluation showed that the proposed approach helps the smaller model to outperform LLMs like GPT-3 and ChatGPT.

Gao et al. (2023b) evaluated ChatGPT for text summarization using various human evaluation methods and reported that (i) ChatGPT-based evaluation is both cost-effective and reproducible, unlike human evaluation, (ii) the performance of ChatGPT-based evaluation is highly dependent on the prompt design, and (iii) ChatGPT generated explanations correlates with its scores. Jain et al. (2023) explored the effectiveness of the GPT-3.5 model as a multi-dimensional evaluator of text summarization. The authors reported that using in-context learning, GPT-3.5-based evaluation achieves SOTA performances on factual consistency and relevance dimensions. Based on the evaluation of five datasets covering text summarization, story generation and data-to-text generation, Wang et al. (2023f) reported that ChatGPT as an evaluator (i) exhibits good correlations with human scores, especially in the case of story generation task and (ii) is prompt sensitive. Bai et al. (2023b) introduced a novel evaluation framework called Language-Model-as-an-Examiner to evaluate open-ended questions. In this framework, GLLM acts as a knowledgeable examiner, generates questions using its own knowledge and then does the reference-free evaluation. Yang et al. (2023f) developed the BigTrans model (based on LLaMA -13B model) with a multilingual translation capacity of more than 100 languages. GPT-4 based assessment showed that BigTrans performance is on par with ChatGPT and Google translate. Zheng et al. (2023a) explored GPT-4 as a judge to evaluate open-ended question answering using two newly introduced benchmarks MT-Bench and Chatbot Arena. The experiment results showed that GPT-4 achieves more than 80

Unlike the above-discussed research works, which used direct prompting, some of the works explored advanced prompting to offer better guidance and context for the GLLM evaluator. Zhuo (2023) developed a code generation evaluation framework based on ChatGPT and demonstrated that the proposed framework outperforms CodeBERTScore (Zhou et al., 2023) consistently across multiple programming languages. Moreover, the performance of the evaluation framework can be enhanced using references and zero-shot CoT prompting. Liu et al. (2023d) proposed G-EVAL, a novel framework based on GPT-4 for the assessment of natural language generation tasks. The proposed framework uses CoT prompting and a form-filling paradigm. Here, CoT prompting enhances the performance of G-EVAL by offering more guidance and context. The performance of ChatGPT-based evaluation in segment-level machine translation is poor. To overcome this, Lu et al. (2023a) proposed a novel prompting called Error Analysis (EA) prompting, which combines error analysis (Lu et al., 2022a) and CoT

prompting. The authors showed that with EA prompting, ChatGPT can assess translations at the segment level much better.

Some of the research works explored GLLMs for the evaluation of multi-modal AI tasks (Lu et al., 2023b), fine-tuning open-source LLM evaluators (Xu et al., 2023b), and paraphrasing references to enhance existing metrics based on PLMs (Tang et al., 2023b). For example, Lu et al. (2023b) introduced LLMScore (based on GPT-4), a new metric which can effectively capture both image and object-level compositionality for text-to-image synthesis evaluation. Some of the research works explored these models to fine-tune open-source LLMs so that they can be used as evaluators, which makes the evaluation less expensive. For example, Xu et al. (2023b) introduced InstructScore, a novel and explainable metric based on fine-tuned LLaMA model for text generation evaluation. Here the authors use GPT-4 generated synthetic data to fine-tune the LLaMA model. InstructScore can generate an error diagnostic report having error details along with an explanation. Natural language generation evaluation using few references results in poor correlation with human judgements. To overcome this drawback, Tang et al. (2023b) introduced Para-Ref, which leverages LLMs to increase the number of references by paraphrasing. The evaluation on three NLG tasks, text summarization, machine translation and image caption, showed that the proposed approach enhances the correlation of sixteen automatic evaluation metrics with human judgements by a good margin.

Some of the research works focused on addressing the limitations of using GLLMs as evaluators. For example, Wang et al. (2023e) demonstrated positional bias in GLLM-based evaluation, i.e., the order of candidate responses can significantly influence the results. The authors demonstrated that the two proposed strategies, namely multiple evidence calibration and balanced position calibration, can reduce the bias and enhance the correlation with human judgements.

12. Future research directions

12.1. Enhance robustness of GLLMs

GLLMs achieved promising results across various NLP tasks in zero and few-shot settings across various NLP tasks. In some of the tasks like data labelling (Gilardi et al., 2023; He et al., 2023b; Törnberg, 2023; Alizadeh et al., 2023), text classification (Sun et al., 2023b), relation extraction (Wan et al., 2023), question answering (Yang et al., 2022; Bang et al., 2023), keyphrase generation (Song et al., 2023), etc., these models achieved even SOTA results. However, some of the recent research works exposed the brittleness of these models towards out-of-distribution inputs (Wang et al., 2023b; Liu et al., 2023f), adversarial prompts (Zhu et al., 2023d; Shirafuji et al., 2023; Han et al., 2023) and inputs (Chen et al., 2023g; Zhuo et al., 2023; Zhao et al., 2023c; Liu et al., 2023c) . For example, Liu et al. (2023f) reported that ChatGPT and GPT-4 perform well in multiple choice question answering but struggle to answer out-of-distribution questions. Similarly, Chen et al. (2023g) observed more than 35% performance degradation for GPT-3 and GPT-3.5 models in tasks like sentiment analysis and natural language inference for adversarial inputs. The brittleness towards out-of-distribution and adversarial inputs makes these models unreliable and limits their practical utility, especially in sensitive domains. So, it is necessary for the research community to focus more on this research direction to make GLLMs more robust and enhance their reliability and usage.

12.2. Red teaming

Red teaming involves an assessment to expose undesirable model behaviours like generating harmful text (Bhardwaj and Poria, 2023; Ganguli et al., 2022; Mehrabi et al., 2023; Perez et al., 2022). GLLMs trained over large volumes of text data with a simple next-word prediction objective are surprisingly good at generating text with human-like fluency. However, the other side is that these models sometimes generate harmful text. For example, Bhardwaj and Poria (2023) observed that GLLMs like ChatGPT and GPT-4 generate answers to more than 60% of harmful queries. One of the possible reasons for this undesirable behaviour of GLLMs is that data used for pretraining these models includes toxic, biased and noisy text to some extent (Bhardwaj and Poria, 2023). This unwanted behaviour of generating harmful text raises concerns and limits the scalable deployment of these models for public use. We can expect more research in future to expose such undesirable behaviour in various scenarios and eventually enhance the safety alignment as well as the safe use of GLLMs.

12.3. State-of-the-art results across NLP tasks

In the beginning, GLLMs like GPT-3 achieved impressive performances in zero and few-shot settings across NLP tasks. Advanced GLLMs like ChatGPT and GPT-4 further pushed the results but still lag behind SOTA results achieved by PLMs fine-tuned based on supervised learning. Later, with the evolution of advanced prompting strategies and novel approaches, GLLMs are able to achieve SOTA results in some of the NLP tasks. For example, InstructGPT with CARP prompting strategy using just 16 examples achieves SOTA results on four text classification datasets (Sun et al., 2023b). Similarly, Wan et al. (2023) achieved SOTA results in relation extraction with the novel GPT-RE framework. Yang et al. (2022) proposed a novel approach which uses GPT-3 as an implicit knowledge source and achieves SOTA results in knowledge-based visual question answering. In future, we can expect more focus from the research community to achieve SOTA results using GLLMs in as many NLP tasks as possible, which will be treated as a further push towards artificial general intelligence. Moreover, this eliminates the painful process of labelling large amounts of data and then fine-tuning PLMs separately for each downstream task.

12.4. Robust approaches to detect GLLM generated text

The ability to generate text with human-like fluency resulted in the wide adoption of GLLMs in various real-world applications like writing assistants, coding assistants, and chatbots (Mireshghallah et al., 2023). There is a growing concern regarding the misuse of these models for various illegal activities (Guo et al., 2023d), like fake news on social media platforms (Hacker et al., 2023; De Angelis et al., 2023), fake reviews on e-commerce websites (Mitrovi'c et al., 2023), fake research papers (Gao et al., 2023a), academic fraud (Cotton et al., 2023), etc. The performance of existing approaches like DetectGPT, ZeroGPT, OpenAI detector, ChatGPT-detector-roberta and ChatGPT-qa-detector-roberta is not satisfactory (Pegoraro et al., 2023; Yu et al., 2023a). Moreover, the existing approaches are not robust to various attacks like paraphrasing, synonym word replacement and writing style modification (Shi et al., 2023; Krishna et al., 2023). So, there is a great need for better approaches which can reliably detect GLLM generated text and also robust to various attacks, including paraphrasing. With reliable and robust detection approaches, the misuse of GLLMs for various illegal activities can be reduced to a great extent.

12.5. Reduce inference costs

GLLMs achieve impressive performances across NLP tasks, with SOTA results in some tasks. However, the downside of using GLLMs is the high inference costs (Chen et al., 2023h; Cheng et al., 2023). For example, a small business is required to spend more than $21,000 monthly to use GPT-4 for better customer support.[7] Such high inference costs have become a burden to small and medium-sized companies.

[7] https://neoteric.eu/blog/how-much-does-it-cost-to-use-gpt-models-gpt-3-pricing-explained.

Recently, Chen et al. (2023h) proposed FrugalGPT, a novel framework involving multiple strategies like prompt adaptation and LLM approximation to reduce the inference costs of GLLMs. The inference costs of GLLMs increase with the prompt size as the inference cost is computed based on the number of tokens processed. Prompt adaptation focuses on reducing the size of the prompt by using fewer but effective examples or querying the GLLMs as a batch. LLM approximation uses cache to avoid querying GLLM for similar queries, which eventually reduces overall inference costs. Similarly, Cheng et al. (2023) proposed batch prompting, which involves GLLM inference in batches rather than processing one sample individually. The authors demonstrated that the proposed prompting strategy reduces Codex model inference cost across ten datasets with little or no degradation in the performance. Future research in this direction will result in much better approaches which will further reduce the GLLM inference costs and make GLLM usage more affordable for companies.

12.6. Enhance performance in domain-specific NLP tasks

Inspired by the success of GLLMs in general domain NLP tasks, the research community explored GLLMs for NLP tasks in specific domains like healthcare, legal, finance, etc. However, the performances of GLLMs in domain-specific NLP tasks are not as impressive as those achieved in general domain NLP tasks (Moradi et al., 2021; Hernandez et al., 2023; Chalkidis, 2023; Choi et al., 2023; Li et al., 2023k; Shah and Chava, 2023). For example, Moradi et al. (2021) reported that the BioBERT model outperforms GPT-3 in few-shot settings even though the BioBERT model is 514 times smaller than GPT-3. Chalkidis (2023) evaluated ChatGPT on the LexGLUE benchmark and reported that ChatGPT performs poorly on legal text classification datasets. Analysing domain-specific texts is more challenging because of domain-specific terminology and abbreviations, complex language structures, etc. In domains like healthcare, finance and legal, domain experts use many words and abbreviations that are specific to the domain and not commonly found in general domain texts. There is a lot of scope to improve the performance of GLLMs in domain-specific NLP tasks, which reduces the bottleneck for the widespread adoption of these models in specific domains.

12.7. Handle limited context length

One of the major drawbacks of GLLMs is their limited context length (Li, 2023; Kaddour et al., 2023; Arefeen et al., 2023). The maximum context length of GLLMs lies in the range of 2049 tokens to 32,768 tokens.[8] This limited context length poses a challenge and becomes a bottleneck for GLLMs to handle long documents or maintain long conservations in which the number of tokens falls beyond the maximum context length. Recently, Li (2023) proposed selective context, a novel approach to effectively utilize the limited context length by filtering out the less useful content in the input text. The authors demonstrated the effectiveness of the proposed approach using the ChatGPT model for question-answering and text summarization tasks across datasets having lengthy input instances. Future research in this direction will help in the evolution of more efficient approaches which will effectively utilize the limited context length and eliminate the bottlenecks for the application of GLLMs in tasks that require processing long inputs.

12.8. Ensure fair evaluation of GLLMs

GLLMs achieved impressive performances across NLP tasks and have received much attention recently. However, one concern regarding the evaluation of GLLMs is data contamination, which refers to the presence of test data instances of downstream tasks in the training corpus of GLLMs (Chang et al., 2023; Golchin and Surdeanu, 2023; Aiyappa et al., 2023). The problem of data contamination is more relevant in the case of GLLMs because of their proprietary nature and non-disclosure of training corpus details. Recent research works have reported the problem of data contamination in GLLMs like ChatGPT (Aiyappa et al., 2023) and GPT-4 (Golchin and Surdeanu, 2023). For example, Golchin and Surdeanu (2023) demonstrated that GPT-4 is contaminated with instances from text classification, natural language inference and text summarization datasets like WNLI (Wang et al., 2018), AG News (Zhang et al., 2015) and XSUM (Narayan et al., 2018). Recently, Golchin and Surdeanu (2023) proposed a novel approach to detect data contamination for LLMs. Future research must focus on developing simple and effective approaches to identify data contamination and ensure fair evaluation, enhancing the reliability of impressive performances of GLLMs.

12.9. Reduce hallucinations

Despite the remarkable performances of GLLMs, there is a growing concern regarding their tendency to generate factually incorrect information (Zhang et al., 2023f; Rawte et al., 2023). This tendency to generate text that does not align with existing world knowledge, deviates from the user's input or contradicts the context generated earlier is referred to as hallucination (Zhang et al., 2023f). Hallucination is a serious problem yet to be addressed fully (Dhuliawala et al., 2023), and it reduces the reliability of GLLMs, which becomes a bottleneck for the adoption of GLLMs, especially in sensitive domains like healthcare (Umapathi et al., 2023). Recently, some of the research works focused on evaluating hallucination in GLLMs (Umapathi et al., 2023), assessing the ability of GLLMs to identify hallucinations (Li et al., 2023b) and developing approaches to reduce hallucinations (Peng et al., 2023b). For example, Li et al. (2023b) proposed HaluEval, a novel benchmark to assess the ability of GLLMs to identify hallucinations. Peng et al. (2023b) introduced LLM-AUGMENTER, a novel approach that reduces hallucinations in ChatGPT without impacting the quality of generated responses. Considering the seriousness of the hallucination problem, we can expect more future research to identify and reduce hallucinations in GLLMs, which enhance their reliability and adoption across domains, including sensitive domains like healthcare.

12.10. Enhance the performance of GLLMs for non-english languages

The performance of GLLMs is not impressive in the case of non-English languages, especially in the case of languages with non-Latin scripts (Ahuja et al., 2023; Bang et al., 2023; Lai et al., 2023a; Kuzman et al., 2023). This is because GLLMs are mostly pretrained on English text. For example, more than 90% of text in the pretraining corpus of the GPT-3 model is from the English language (Brown et al., 2020; Ahuja et al., 2023). Some of the possible options to enhance the performance of GLLMs for non-English languages are the use of English prompts (Lai et al., 2023a; Kuzman et al., 2023) and optimized tokenization (Armengol-Estapé et al., 2022). There is a great need for better approaches to greatly enhance the performance of GLLMs for non-English languages, which increase their adoption across the globe and benefit users from non-English communities.

13. Conclusion

In this survey paper, we provide a comprehensive review of GPT-3 family LLMs in multiple dimensions covering more than 350 recent research papers. Here, we present foundation concepts, GPT-3 family LLMs and discuss the performances of these models in various downstream tasks, specific domains and multiple languages. We also discuss data labelling, data augmentation and data generation abilities of GLLMs, the robustness of GLLMs, the effectiveness of GLLMs as evaluators, and finally, conclude with multiple insightful future research directions. Overall, this comprehensive survey paper on GPT-3 family LLMs will serve as a good resource for both academic and industry people to stay updated with the latest research.

[8] https://platform.openai.com/docs/models/overview.

CRediT authorship contribution statement

Katikapalli Subramanyam Kalyan: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The author would like to thank Ajit Rajasekharan for his encouragement and support.

References

Abacha, A.B., Yim, W.-w., Adams, G., Snider, N., Yetisgen-Yildiz, M., 2023. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In: Proceedings of the 5th Clinical Natural Language Processing Workshop. pp. 503–513.

Abaskohi, A., Rothe, S., Yaghoobzadeh, Y., 2023. LM-CPPF: Paraphrasing-guided data augmentation for contrastive prompt-based few-shot fine-tuning. arXiv preprint arXiv:2305.18169.

Adomavicius, G., Tuzhilin, A., 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. 17 (6), 734–749.

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., Sontag, D., 2022. Large language models are few-shot clinical information extractors. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 1998–2022.

Ahmad, W., Chakraborty, S., Ray, B., Chang, K.-W., 2021. Unified pre-training for program understanding and generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2655–2668.

Ahuja, K., Hada, R., Ochieng, M., Jain, P., Diddee, H., Maina, S., Ganu, T., Segal, S., Axmed, M., Bali, K., et al., 2023. Mega: Multilingual evaluation of generative ai. arXiv preprint arXiv:2303.12528.

Aiyappa, R., An, J., Kwak, H., Ahn, Y.-Y., 2023. Can we trust the evaluation on ChatGPT? arXiv preprint arXiv:2303.12767.

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Bermeo, J.D., Korobeynikova, M., Gilardi, F., 2023. Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks. arXiv preprint arXiv:2307.02179.

Amin, M.M., Cambria, E., Schuller, B.W., 2023. Will affective computing emerge from foundation models and general AI? A first evaluation on ChatGPT. IEEE Intell. Syst. 38 (2).

Anand, A., Lyu, L., Idahl, M., Wang, Y., Wallat, J., Zhang, Z., 2022. Explainable information retrieval: A survey. arXiv preprint arXiv:2211.02405.

Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al., 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403.

Antaki, F., Touma, S., Milad, D., El-Khoury, J., Duval, R., 2023. Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. Ophthalmol. Sci. 100324.

Araci, D., 2019. Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

Arefeen, M.A., Debnath, B., Chakradhar, S., 2023. LeanContext: Cost-efficient domain-specific question answering using LLMs. arXiv preprint arXiv:2309.00841.

Armengol-Estapé, J., de Gibert Bonet, O., Melero, M., 2022. On the multilingual capabilities of very large-scale english language models. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 3056–3068.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473.

Bahdanau, D., Cho, K.H., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., et al., 2023a. LongBench: A bilingual, multitask benchmark for long context understanding. arXiv preprint arXiv:2308.14508.

Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., Yu, J., Zeng, K., Xiao, Y., Lyu, H., et al., 2023b. Benchmarking foundation models with language-model-as-an-examiner. arXiv preprint arXiv:2306.04181.

Banerjee, S., Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/Or Summarization. pp. 65–72.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al., 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023.

Barbieri, F., Camacho-Collados, J., Anke, L.E., Neves, L., 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1644–1650.

Bayer, M., Kaufhold, M.-A., Reuter, C., 2022. A survey on data augmentation for text classification. ACM Comput. Surv. 55 (7), 1–39.

Belinkov, Y., Bisk, Y., 2018. Synthetic and natural noise both break neural machine translation. In: International Conference on Learning Representations.

Beltagy, I., Peters, M.E., Cohan, A., 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Bhardwaj, R., Poria, S., 2023. Red-teaming large language models using chain of utterances for safety-alignment. arXiv preprint arXiv:2308.09662.

Bhattacharya, A., Singla, Y.K., Krishnamurthy, B., Shah, R.R., Chen, C., 2023. A video is worth 4096 tokens: Verbalize story videos to understand them in zero shot. arXiv preprint arXiv:2305.09758.

Blitzer, J., Dredze, M., Pereira, F., 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 440–447.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. 5, 135–146.

Bommarito, II, M., Katz, D.M., 2022. GPT takes the bar exam. arXiv preprint arXiv:2212.14402.

Bommasani, R., Liang, P., Lee, T., 2023. Holistic evaluation of language models. Ann. New York Acad. Sci..

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al., 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Cai, X., Liu, S., Han, J., Yang, L., Liu, Z., Liu, T., 2021. Chestxraybert: A pretrained language model for chest radiology report summarization. IEEE Trans. Multimed..

Carpenter, K.A., Altman, R.B., 2023. Using GPT-3 to build a lexicon of drugs of abuse synonyms for social media pharmacovigilance. Biomolecules 13 (2), 387.

Cegin, J., Simko, J., Brusilovsky, P., 2023. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. arXiv preprint arXiv:2305.12947.

Chali, Y., Hasan, S.A., Joty, S.R., 2011. Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels. Inf. Process. Manage. 47 (6), 843–855.

Chalkidis, I., 2023. ChatGPT may pass the bar exam soon, but has a long way to go for the LexGLUE benchmark. arXiv preprint arXiv:2304.12202.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I., 2020. LEGAL-BERT: The muppets straight out of law school. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2898–2904.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D., Aletras, N., 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4310–4330.

Chan, C., Cheng, J., Wang, W., Jiang, Y., Fang, T., Liu, X., Song, Y., 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. arXiv preprint arXiv:2304.14827.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., et al., 2023. A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.-H., Routledge, B.R., et al., 2021a. FinQA: A dataset of numerical reasoning over financial data. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3697–3711.

Chen, Y., Cheng, J., Jiang, H., Liu, L., Zhang, H., Shi, S., Xu, R., 2022. Learning from sibling mentions with scalable graph inference in fine-grained entity typing. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2076–2087.

Chen, Q., Du, J., Hu, Y., Keloth, V.K., Peng, X., Raja, K., Zhang, R., Lu, Z., Xu, H., 2023a. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. arXiv preprint arXiv:2305.16326.

Chen, E., Huang, R., Chen, H.-S., Tseng, Y.-H., Li, L.-Y., 2023b. GPTutor: a ChatGPT-powered programming tool for code explanation. arXiv preprint arXiv:2305.01863.

Chen, H., Jiao, F., Li, X., Qin, C., Ravaut, M., Zhao, R., Xiong, C., Joty, S., 2023c. ChatGPT's one-year anniversary: Are open-source large language models catching up? arXiv preprint arXiv:2311.16989.

Chen, Y., Kang, H., Zhai, V., Li, L., Singh, R., Ramakrishnan, B., 2023d. GPT-sentinel: Distinguishing human and ChatGPT generated content. ArXiv, abs/2305.07969.

Chen, S., Li, Y., Lu, S., Van, H., Aerts, H.J., Savova, G.K., Bitterman, D.S., 2023e. Evaluation of ChatGPT family of models for biomedical reasoning and classification. arXiv preprint arXiv:2304.02496.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al., 2021b. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Chen, Y., Wang, R., Jiang, H., Shi, S., Xu, R., 2023f. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. arXiv preprint arXiv:2304.00723.

Chen, X., Ye, J., Zu, C., Xu, N., Zheng, R., Peng, M., Zhou, J., Gui, T., Zhang, Q., Huang, X., 2023g. How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks. arXiv preprint arXiv:2303.00293.

Chen, L., Zaharia, M., Zou, J., 2023h. FrugalGPT: How to use large language models while reducing cost and improving performance. arXiv preprint arXiv:2305.05176.

Chen, S., Zhao, Y., Zhang, J., Chern, I., Gao, S., Liu, P., He, J., et al., 2023i. Felm: Benchmarking factuality evaluation of large language models. arXiv preprint arXiv:2310.00741.

Cheng, Z., Kasai, J., Yu, T., 2023. Batch prompting: Efficient inference with large language model apis. arXiv preprint arXiv:2301.08721.

Cheshkov, A., Zadorozhny, P., Levichev, R., 2023. Evaluation of ChatGPT model for vulnerability detection. arXiv preprint arXiv:2304.07232.

Chintagunta, B., Katariya, N., Amatriain, X., Kannan, A., 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In: Machine Learning for Healthcare Conference. PMLR, pp. 354–372.

Chiu, K.-L., Collins, A., Alexander, R., 2021. Detecting hate speech with gpt-3. arXiv preprint arXiv:2103.12407.

Chmielewski, M., Kucker, S.C., 2020. An MTurk crisis? Shifts in data quality and the impact on study results. Soc. Psychol. Pers. Sci. 11 (4), 464–473.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, p. 1724.

Choi, J.H., Hickman, K.E., Monahan, A., Schwarcz, D., 2023. Chatgpt goes to law school. Available at SSRN.

Choromanski, K.M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J.Q., Mohiuddin, A., Kaiser, L., et al., 2020. Rethinking attention with performers. In: International Conference on Learning Representations.

Choudhury, D., et al., 2023. Ask me in english instead: Cross-lingual evaluation of large language models for healthcare queries. arXiv preprint arXiv:2310.13132.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

Chu, Z., Chen, J., Chen, Q., Yu, W., Wang, H., Liu, M., Qin, B., 2023. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. arXiv preprint arXiv:2311.17667.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on Deep Learning, December 2014.

Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al., 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., Smith, N.A., 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 7282–7296.

Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D., 2019. ELECTRA: Pre-training text encoders as discriminators rather than generators. In: International Conference on Learning Representations.

Collins, K.M., Wong, C., Feng, J., Wei, M., Tenenbaum, J.B., 2022. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. arXiv preprint arXiv:2205.05718.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451.

Conneau, A., Lample, G., 2019. Cross-lingual language model pretraining. Adv. Neural Inf. Process. Syst. 32.

Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al., 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Cotton, D.R., Cotton, P.A., Shipway, J.R., 2023. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innov. Educ. Teach. Int. 1–12.

Coulombe, C., 2018. Text data augmentation made simple by leveraging nlp cloud apis. arXiv preprint arXiv:1812.04718.

Dai, Y., Feng, D., Huang, J., Jia, H., Xie, Q., Zhang, Y., Han, W., Tian, W., Wang, H., 2023a. LAiW: A Chinese legal large language models benchmark (a technical report). arXiv preprint arXiv:2310.05620.

Dai, A.M., Le, Q.V., 2015. Semi-supervised sequence learning. Adv. Neural Inf. Process. Syst. 28.

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., et al., 2023b. AugGPT: Leveraging ChatGPT for text data augmentation. arXiv preprint arXiv:2302.13007.

Dai, S., Shao, N., Zhao, H., Yu, W., Si, Z., Xu, C., Sun, Z., Zhang, X., Xu, J., 2023c. Uncovering ChatGPT's capabilities in recommender systems. arXiv preprint arXiv:2305.02182.

Das, S.S.S., Katiyar, A., Passonneau, R.J., Zhang, R., 2022. Container: Few-shot named entity recognition via contrastive learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6338–6353.

Das, M., Pandey, S.K., Mukherjee, A., 2023. Evaluating ChatGPT's performance for multilingual and emoji-based hate speech detection. arXiv preprint arXiv:2305.13276.

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G.P., Ferragina, P., Tozzi, A.E., Rizzo, C., 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front. Public Health 11, 1166120.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Derner, E., Batistič, K., Zahálka, J., Babuška, R., 2023. A security risk taxonomy for large language models. arXiv preprint arXiv:2311.11415.

Destefanis, G., Bartolucci, S., Ortu, M., 2023. A preliminary analysis on the code generation capabilities of GPT-3.5 and bard AI models for java functions. arXiv preprint arXiv:2305.09402.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., Weston, J., 2023. Chain-of-verification reduces hallucination in large language models.

Ding, B., Qin, C., Liu, L., Bing, L., Joty, S., Li, B., 2022. Is gpt-3 a good data annotator? arXiv preprint arXiv:2212.10450.

Doddapaneni, S., Ramesh, G., Khapra, M.M., Kunchukuttan, A., Kumar, P., 2021. A primer on pretrained multilingual language models. arXiv preprint arXiv:2107.00676.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Sui, Z., 2022. A survey for in-context learning. arXiv preprint arXiv:2301.00234.

Dong, M., Zeng, X., Koehl, L., Zhang, J., 2020. An interactive knowledge-based recommender system for fashion product design in the big data environment. Inform. Sci. 540, 469–488.

Du, X., Cardie, C., 2020. Event extraction by answering (almost) natural questions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 671–683.

Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., et al., 2022. Glam: Efficient scaling of language models with mixture-of-experts. In: International Conference on Machine Learning. PMLR, pp. 5547–5569.

Eldan, R., Li, Y., 2023. TinyStories: How small can language models be and still speak coherent english? arXiv preprint arXiv:2305.07759.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Çelebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., Joulin, A., 2020. Beyond english-centric multilingual machine translation. arXiv abs/2010.11125.

Fan, Y., Jiang, F., 2023. Uncovering the potential of ChatGPT for discourse analysis in dialogue: An empirical study. arXiv preprint arXiv:2305.08391.

Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y., 2023. Improving CLIP training with language rewrites. arXiv preprint arXiv:2305.20088.

Fang, Y., Li, X., Thomas, S.W., Zhu, X., 2023a. ChatGPT as data augmentation for compositional generalization: A case study in open intent detection. arXiv preprint arXiv:2308.13517.

Fang, T., Yang, S., Lan, K., Wong, D.F., Hu, J., Chao, L.S., Zhang, Y., 2023b. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. arXiv preprint arXiv:2304.01746.

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., Kyriazis, D., 2023. Transforming sentiment analysis in the financial domain with ChatGPT. arXiv preprint arXiv:2308.07935.

Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z., Ge, J., 2023. LawBench: Benchmarking legal knowledge of large language models. arXiv preprint arXiv:2309.16289.

Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E., 2021. A survey of data augmentation approaches for NLP. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 968–988.

Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., et al., 2020. CodeBERT: A pre-trained model for programming and natural languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1536–1547.

Feng, W., Zhu, W., Fu, T.-j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y., 2023. LayoutGPT: Compositional visual planning and generation with large language models. arXiv preprint arXiv:2305.15393.

Fu, J., Ng, S.-K., Jiang, Z., Liu, P., 2023. Gptscore: Evaluate as you desire. arXiv preprint arXiv:2302.04166.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al., 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858.

Gao, C.A., Howard, F.M., Markov, N.S., Dyer, E.C., Ramesh, S., Luo, Y., Pearson, A.T., 2023a. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit. Med. 6 (1), 75.

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., Wan, X., 2023b. Human-like summarization evaluation with chatgpt. arXiv preprint arXiv:2304.02554.

Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., Zhang, J., 2023c. Chat-rec: Towards interactive and explainable llms-augmented recommender system. arXiv preprint arXiv:2303.14524.

Gao, Y., Wang, R., Hou, F., 2023d. How to design translation prompts for ChatGPT: An empirical study. arXiv e-prints, arXiv–2304.

Gao, J., Zhao, H., Yu, C., Xu, R., 2023e. Exploring the feasibility of chatgpt for event extraction. arXiv preprint arXiv:2303.03836.

Geng, M., Wang, S., Dong, D., Wang, H., Li, G., Jin, Z., Mao, X., Liao, X., 2023. An empirical study on using large language models for multi-intent comment generation. arXiv abs/2304.11384.

Gilardi, F., Alizadeh, M., Kubli, M., 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056.

Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A., Chartash, D., et al., 2023. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med. Educ. 9 (1), e45312.

Giorgi, J., Toma, A., Xie, R., Chen, S., An, K., Zheng, G., Wang, B., 2023. WangLab at MEDIQA-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In: Proceedings of the 5th Clinical Natural Language Processing Workshop. pp. 323–334.

Glaese, A., McAleese, N., Trebacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., et al., 2022. Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:2209.14375.

Goertzel, B., 2014. Artificial general intelligence: concept, state of the art, and future prospects. J. Artif. Gener. Intell. 5 (1), 1.

Golchin, S., Surdeanu, M., 2023. Time travel in LLMs: Tracing data contamination in large language models. arXiv preprint arXiv:2308.08493.

González-Gallardo, C.-E., Boros, E., Girdhar, N., Hamdi, A., Moreno, J.G., Doucet, A., 2023. Yes but.. Can ChatGPT identify entities in historical documents? arXiv preprint arXiv:2303.17322.

Goyal, S., Doddapaneni, S., Khapra, M.M., Ravindran, B., 2022. A survey of adversarial defences and robustness in nlp. ACM Comput. Surv..

Gu, W., 2023. Linguistically informed ChatGPT prompts to enhance Japanese–Chinese machine translation: A case study on attributive clauses. arXiv preprint arXiv:2303.15587.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2020. Domain-specific language model pretraining for biomedical natural language processing. arXiv preprint arXiv:2007.15779.

Gu, Y., Zhang, S., Usuyama, N., Woldesenbet, Y., Wong, C., Sanapathi, P., Wei, M., Valluri, N., Strandberg, E., Naumann, T., et al., 2023. Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. arXiv preprint arXiv:2307.06439.

Guha, N., Nyarko, J., Ho, D.E., Re, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., et al., 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In: Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Gui, J., Chen, T., Cao, Q., Sun, Z., Luo, H., Tao, D., 2023. A survey of self-supervised learning from multiple perspectives: Algorithms, theory, applications and future trends. arXiv preprint arXiv:2301.05712.

Gui, L., Wang, B., Huang, Q., Hauptmann, A.G., Bisk, Y., Gao, J., 2022. KAT: A knowledge augmented transformer for vision-and-language. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 956–968.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Giorno, A.D., Gopi, S., Javaheripi, M., Kauffmann, P.C., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H.S., Wang, X., Bubeck, S., Eldan, R., Kalai, A.T., Lee, Y.T., Li, Y.-F., 2023. Textbooks are all you need. ArXiv, abs/2306.11644.

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., Liu, Y., Li, J., Xiong, B., Xiong, D., et al., 2023a. Evaluating large language models: A comprehensive survey. arXiv preprint arXiv:2310.19736.

Guo, D., Ren, S., Lu, S., Feng, Z., Tang, D., Shujie, L., Zhou, L., Duan, N., Svyatkovskiy, A., Fu, S., et al., 2020. GraphCodeBERT: Pre-training code representations with data flow. In: International Conference on Learning Representations.

Guo, Z., Wang, P., Wang, Y., Yu, S., 2023b. Dr. LLaMA: Improving small language models in domain-specific QA via generative data augmentation. arXiv preprint arXiv:2305.07804.

Guo, Z., Wang, P., Wang, Y., Yu, S., 2023c. Dr. LLaMA: Improving small language models on PubMedQA via generative data augmentation. ArXiv, abs/2305.07804.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y., 2023d. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. ArXiv, abs/2301.07597.

Gupta, R., Herzog, I., Park, J.B., Weisberger, J., Firouzbakht, P., Ocon, V., Chao, J., Lee, E.S., Mailey, B.A., 2023. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthetic Surg. J. sjad128.

Gutiérrez, B.J., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., Su, Y., 2022. Thinking about GPT-3 in-context learning for biomedical ie? Think again. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 4497–4512.

Hacker, P., Engel, A., Mauer, M., 2023. Regulating ChatGPT and other large generative AI models. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1112–1123.

Hada, R., Gumma, V., de Wynter, A., Diddee, H., Ahmed, M., Choudhury, M., Bali, K., Sitaram, S., 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation? arXiv preprint arXiv:2309.07462.

Hakimov, S., Schlangen, D., 2023. Images in language space: Exploring the suitability of large language models for vision & language tasks. arXiv preprint arXiv:2305.13782.

Hamidi, A., Roberts, K., 2023. Evaluation of AI chatbots for patient-specific EHR questions. arXiv preprint arXiv:2306.02549.

Han, R., Peng, T., Yang, C., Wang, B., Liu, L., Wan, X., 2023. Is information extraction solved by ChatGPT? An analysis of performance, evaluation criteria, robustness and errors. arXiv preprint arXiv:2305.14450.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al., 2021. Pre-trained models: Past, present and future. AI Open 2, 225–250.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., Kamar, E., 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3309–3326.

He, P., Gao, J., Chen, W., 2022a. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In: The Eleventh International Conference on Learning Representations.

He, J., Kryściński, W., McCann, B., Rajani, N., Xiong, C., 2022b. CTRLsum: Towards generic controllable text summarization. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 5879–5915.

He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S., Wang, X., 2023a. Exploring human-like translation strategy with large language models. arXiv preprint arXiv:2305.04118.

He, X., Lin, Z., Gong, Y., Jin, A., Zhang, H., Lin, C., Jiao, J., Yiu, S.M., Duan, N., Chen, W., et al., 2023b. Annollm: Making large language models to be better crowdsourced annotators. arXiv preprint arXiv:2303.16854.

He, P., Liu, X., Gao, J., Chen, W., 2020. DEBERTA: Decoding-enhanced bert with disentangled attention. In: International Conference on Learning Representations.

He, P., Peng, B., Lu, L., Wang, S., Mei, J., Liu, Y., Xu, R., Awadalla, H.H., Shi, Y., Zhu, C., et al., 2022c. Z-code++: A pre-trained language model optimized for abstractive summarization. arXiv preprint arXiv:2208.09770.

He, X., Shen, X., Chen, Z., Backes, M., Zhang, Y., 2023c. Mgtbench: Benchmarking machine-generated text detection. arXiv preprint arXiv:2303.14822.

He, Z., Wang, Y., Yan, A., Liu, Y., Chang, E.Y., Gentili, A., McAuley, J., Hsu, C.-N., 2023d. MedEval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. arXiv preprint arXiv:2310.14088.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., Awadalla, H.H., 2023. How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210.

Hernandez, E., Mahajan, D., Wulff, J., Smith, M.J., Ziegler, Z., Nadler, D., Szolovits, P., Johnson, A., Alsentzer, E., et al., 2023. Do we still need clinical language models? In: Conference on Health, Inference, and Learning. PMLR, pp. 578–597.

Hirosawa, T., Harada, Y., Yokose, M., Sakamoto, T., Kawamura, R., Shimizu, T., 2023. Diagnostic accuracy of differential-diagnosis lists generated by generative pre-trained transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. Int. J. Environ. Res. Public Health 20 (4), 3378.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al., 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T.T., McGee, L.A., Ashman, J.B., Li, X., Liu, T., Shen, J., et al., 2023a. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front. Oncol. 13, 1219326.

Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T.T., McGee, L.A., Ashman, J.B., Li, X., Liu, T., Shen, J., et al., 2023b. Evaluating large language models on a highly-specialized topic, radiation oncology physics. arXiv preprint arXiv:2304.01938.

Hong, S., Seo, J., Hong, S., Shin, H., Kim, S., 2023. Large language models are frame-level directors for zero-shot text-to-video generation. arXiv preprint arXiv:2305.14330.

Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., Zhao, W.X., 2023a. Large language models are zero-shot rankers for recommender systems. arXiv preprint arXiv:2305.08845.

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., Wang, H., 2023b. Large language models for software engineering: A systematic literature review. arXiv preprint arXiv:2308.10620.

Howard, J., Ruder, S., 2018. Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339.

Hu, Y., Ameer, I., Zuo, X., Peng, X., Zhou, Y., Li, Z., Li, Y., Li, J., Jiang, X., Xu, H., 2023a. Zero-shot clinical entity recognition using chatgpt. arXiv preprint arXiv:2303.16416.

Hu, H., Lu, H., Zhang, H., Lam, W., Zhang, Y., 2023b. Chain-of-symbol prompting elicits planning in large langauge models. arXiv preprint arXiv:2305.10276.

Huang, J., Chang, K.C.-C., 2022. Towards reasoning in large language models: A survey. arXiv preprint arXiv:2212.10403.

Huang, F., Kwak, H., An, J., 2023a. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. arXiv preprint arXiv:2302.07736.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., et al., 2023b. Audiogpt: Understanding and generating speech, music, sound, and talking head. arXiv preprint arXiv:2304.12995.

Huang, X., Ruan, W., Huang, W., Jin, G., Dong, Y., Wu, C., Bensalem, S., Mu, R., Qi, Y., Zhao, X., et al., 2023c. A survey of safety and trustworthiness of large language models through the lens of verification and validation. arXiv preprint arXiv:2305.11391.

Hulman, A., Dollerup, O.L., Mortensen, J.F., Fenech, M., Norman, K., Stoevring, H., Hansen, T.K., 2023. ChatGPT-versus human-generated answers to frequently asked questions about diabetes: a turing test-inspired survey among employees of a Danish diabetes center. medRxiv, pp. 2023-2002.

Hutter, F., Kotthoff, L., Vanschoren, J., 2019. Automated Machine Learning: Methods, Systems, Challenges. Springer Nature.

Huynh, J., Jiao, C., Gupta, P., Mehri, S., Bajaj, P., Chaudhary, V., Eskenazi, M., 2023. Understanding the effectiveness of very large language models on dialog evaluation. arXiv preprint arXiv:2301.12004.

Ippolito, D., Duckworth, D., Callison-Burch, C., Eck, D., 2020. Automatic detection of generated text is easiest when humans are fooled. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1808–1822.

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., Vidgen, B., 2023. FinanceBench: A new benchmark for financial question answering. arXiv preprint arXiv:2311.11944.

Iyer, S., Lin, X.V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P.S., et al., 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. arXiv preprint arXiv:2212.12017.

Jain, S., Keshava, V., Sathyendra, S.M., Fernandes, P., Liu, P., Neubig, G., Zhou, C., 2023. Multi-dimensional evaluation of text summarization with in-context learning. arXiv preprint arXiv:2306.01200.

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A.T., Topalis, J., Weber, T., Wesp, P., Sabel, B., Ricke, J., et al., 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. arXiv preprint arXiv:2212.14882.

Jiao, W., Wang, W., Huang, J., Wang, X., Tu, Z., 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q., 2020. TinyBERT: Distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4163–4174.

Jing, Y., Jin, R., Hu, J., Qiu, H., Wang, X., Wang, P., Xiong, D., 2023. FollowEval: A multi-dimensional benchmark for assessing the instruction-following capability of large language models. arXiv preprint arXiv:2311.09829.

Joshi, I., Budhiraja, R., Dev, H., Kadia, J., Ataullah, M.O., Mitra, S., Kumar, D., Akolekar, H.D., 2023. ChatGPT–a blessing or a curse for undergraduate computer science students and instructors? arXiv preprint arXiv:2304.14993.

Just, R., Jalali, D., Ernst, M.D., 2014. Defects4J: A database of existing faults to enable controlled testing studies for java programs. In: Proceedings of the 2014 International Symposium on Software Testing and Analysis. pp. 437–440.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., McHardy, R., 2023. Challenges and applications of large language models. arXiv preprint arXiv:2307.10169.

Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N., Bhattacharyya, A., Khapra, M.M., Kumar, P., 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4948–4961.

Kalakonda, S.S., Maheshwari, S., Sarvadevabhatla, R.K., 2022. Action-GPT: Leveraging large-scale language models for improved and generalized zero shot action generation. arXiv preprint arXiv:2211.15603.

Kalchbrenner, N., Grefenstette, E., Blunsom, P., 2014. A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 655–665.

Kalyan, K.S., Rajasekharan, A., Sangeetha, S., 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. arXiv preprint arXiv:2108.05542.

Kalyan, K.S., Rajasekharan, A., Sangeetha, S., 2022. AMMU: a survey of transformer-based biomedical pretrained language models. J. Biomed. Inform. 126, 103982.

Kalyan, K.S., Sangeetha, S., 2020a. Medical concept normalization in user-generated texts by learning target concept embeddings. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis. pp. 18–23.

Kalyan, K.S., Sangeetha, S., 2020b. Target concept guided medical concept normalization in noisy user-generated texts. In: Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. pp. 64–73.

raj Kanakarajan, K., Kundumani, B., Sankarasubbu, M., 2021. Bioelectra: pretrained biomedical text encoder using discriminators. In: Proceedings of the 20th Workshop on Biomedical Language Processing. pp. 143–154.

Kang, S., Chen, B., Yoo, S., Lou, J.-G., 2023a. Explainable automated debugging via large language model-driven scientific debugging. arXiv preprint arXiv:2304.02195.

Kang, W.-C., Ni, J., Mehta, N., Sathiamoorthy, M., Hong, L., Chi, E., Cheng, D.Z., 2023b. Do LLMs understand user preferences? Evaluating LLMs on user rating prediction. arXiv preprint arXiv:2305.06474.

Karpinska, M., Iyyer, M., 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. arXiv preprint arXiv:2304.03245.

Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y., Radev, D., 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. arXiv preprint arXiv:2303.18027.

Kashefi, A., Mukerji, T., 2023. ChatGPT for programming numerical methods. arXiv abs/2303.12093.

Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., Shardlow, M., 2023. BLESS: Benchmarking large language models on sentence simplification. arXiv preprint arXiv:2310.15773.

Khalil, M., Er, E., 2023. Will ChatGPT get you caught? Rethinking of plagiarism detection. arXiv preprint arXiv:2302.04335.

Khan, J.Y., Uddin, G., 2022. Automatic code documentation generation using gpt-3. In: Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering. pp. 1–6.

Kim, Y., 2014. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

Kocmi, T., Federmann, C., 2023. Large language models are state-of-the-art evaluators of translation quality. arXiv preprint arXiv:2302.14520.

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., Menezes, A., 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In: Proceedings of the Sixth Conference on Machine Translation. pp. 478–494.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., et al., 2023. Chatgpt: Jack of all trades, master of none. arXiv preprint arXiv:2302.10724.

Koncel-Kedziorski, R., Krumdick, M., Lai, V., Reddy, V., Lovering, C., Tanner, C., 2023. BizBench: A quantitative reasoning benchmark for business and finance. arXiv preprint arXiv:2311.06602.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., Iyyer, M., 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. arXiv preprint arXiv:2303.13408.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

Kulkarni, M., Mahata, D., Arora, R., Bhowmik, R., 2022. Learning rich representation of keyphrases from text. In: Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, Seattle, United States, pp. 891–906. http://dx.doi.org/10.18653/v1/2022.findings-naacl.67, URL https://aclanthology.org/2022.findings-naacl.67.

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., et al., 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS Digit. Health 2 (2), e0000198.

Kuzman, T., Ljubešić, N., Mozetič, I., 2023. Chatgpt: Beginning of an end of manual annotation? Use case of automatic genre identification. arXiv preprint arXiv:2303.03953.

Kwan, W.-C., Zeng, X., Wang, Y., Sun, Y., Li, L., Shang, L., Liu, Q., Wong, K.-F., 2023. M4LE: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. arXiv preprint arXiv:2310.19240.

Lai, V.D., Ngo, N.T., Veyseh, A.P.B., Man, H., Dernoncourt, F., Bui, T., Nguyen, T.H., 2023a. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613.

Lai, H., Toral, A., Nissim, M., 2023b. Multidimensional evaluation for text style transfer using ChatGPT. arXiv preprint arXiv:2304.13462.

Lamichhane, B., 2023. Evaluation of chatgpt for nlp-based mental health applications. arXiv preprint arXiv:2303.15727.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2019. ALBERT: A lite BERT for self-supervised learning of language representations. In: International Conference on Learning Representations.

Lan, Y., Wu, Y., Xu, W., Feng, W., Zhang, Y., 2023. Chinese fine-grained financial sentiment analysis with large language models. arXiv preprint arXiv:2306.14096.
Larson, S., Leach, K., 2022. A survey of intent classification and slot-filling datasets for task-oriented dialog. arXiv preprint arXiv:2207.13211.
Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), 1234–1240.
Leinonen, J., Denny, P., MacNeil, S., Sarsa, S., Bernstein, S., Kim, J., Tran, A., Hellas, A., 2023. Comparing code explanations created by students and large language models. arXiv preprint arXiv:2304.03938.
Leippold, M., 2023. Sentiment spin: Attacking financial sentiment with GPT-3. Finance Res. Lett. 103957.
Leivaditi, S., Rossi, J., Kanoulas, E., 2020. A benchmark for lease contract review. arXiv preprint arXiv:2010.10386.
Leong, W.Q., Ngui, J.G., Susanto, Y., Rengarajan, H., Sarveswaran, K., Tjhi, W.C., 2023. BHASA: A holistic southeast Asian linguistic and cultural evaluation suite for large language models. arXiv preprint arXiv:2309.06085.
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880.
Li, Y., 2023. Unlocking context constraints of LLMs: Enhancing context efficiency of LLMs with self-information-based content filtering. arXiv preprint arXiv:2304.12102.
Li, R., Allal, L.B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., Li, J., Chim, J., et al., 2023a. StarCoder: may the source be with you! arXiv preprint arXiv:2305.06161.
Li, J., Cheng, X., Zhao, W.X., Nie, J.-Y., Wen, J.-R., 2023b. HaluEval: A large-scale hallucination evaluation benchmark for large language models. arXiv e-prints, arXiv–2305.
Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al., 2022a. Competition-level code generation with alphacode. Science 378 (6624), 1092–1097.
Li, L., Fan, L., Atreja, S., Hemphill, L., 2023c. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. arXiv preprint arXiv:2304.10619.
Li, B., Fang, G., Yang, Y., Wang, Q., Ye, W., Zhao, W., Zhang, S., 2023d. Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. arXiv preprint arXiv:2304.11633.
Li, B., Hou, Y., Che, W., 2022b. Data augmentation approaches in natural language processing: A survey. Ai Open 3, 71–90.
Li, X., Li, Z., Luo, X., Xie, H., Lee, X., Zhao, Y., Wang, F.L., Li, Q., 2023e. Recurrent attention networks for long-text modeling. arXiv preprint arXiv:2306.06843.
Li, J., Li, H., Pan, Z., Pan, G., 2023f. Prompt ChatGPT in MNER: Improved multimodal named entity recognition method based on auxiliary refining knowledge from ChatGPT. arXiv preprint arXiv:2305.12212.
Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S., He, L., 2022c. A survey on text classification: From traditional to deep learning. ACM Trans. Intell. Syst. Technol. 13 (2), 1–41.
Li, P., Sun, T., Tang, Q., Yan, H., Wu, Y., Huang, X., Qiu, X., 2023g. CodeIE: Large code generation models are better few-shot information extractors. arXiv preprint arXiv:2305.05711.
Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J., 2023h. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890.
Li, X.-Y., Xue, J.-T., Xie, Z., Li, M., 2023i. Think outside the code: Brainstorming boosts large language models in code generation. arXiv preprint arXiv:2305.10679.
Li, X., Yao, Y., Jiang, X., Fang, X., Meng, X., Fan, S., Han, P., Li, J., Du, L., Qin, B., et al., 2023j. FLM-101b: An open LLM and how to train it with 100 K budget. arXiv preprint arXiv:2309.03852.
Li, X., Zhu, X., Ma, Z., Liu, X., Shah, S., 2023k. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? An examination on several typical tasks. arXiv preprint arXiv:2305.05862.
Li, T.-O., Zong, W., Wang, Y., Tian, H., Wang, Y., Cheung, S.-C., 2023l. Finding failure-inducing test cases with ChatGPT. arXiv preprint arXiv:2304.11686.
Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., Liu, T., et al., 2023. Differentiate chatgpt-generated and human-written medical texts. arXiv preprint arXiv:2304.11567.
Lieber, O., Sharir, O., Lenz, B., Shoham, Y., 2021. Jurassic-1: Technical Details and Evaluation. White Paper. AI21 Labs.
Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81.
Lin, S., Hilton, J., Evans, O., 2022a. TruthfulQA: Measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3214–3252.
Lin, D., Koppel, J., Chen, A., Solar-Lezama, A., 2017. QuixBugs: A multi-lingual program repair benchmark set based on the quixey challenge. In: Proceedings Companion of the 2017 ACM SIGPLAN International Conference on Systems, Programming, Languages, and Applications: Software for Humanity. pp. 55–56.
Lin, T., Wang, Y., Liu, X., Qiu, X., 2022b. A survey of transformers. AI Open.
Lin, Y., Xie, Y., Chen, D., Xu, Y., Zhu, C., Yuan, L., 2022c. Revive: Regional visual representation matters in knowledge-based visual question answering. arXiv preprint arXiv:2206.01201.
Liu, C., Bao, X., Zhang, H., Zhang, N., Hu, H., Zhang, X., Yan, M., 2023a. Improving ChatGPT prompt for code generation. arXiv preprint arXiv:2305.08360.
Liu, Y., Fabbri, A.R., Liu, P., Radev, D., Cohan, A., 2023b. On learning to summarize with large language models as references. arXiv preprint arXiv:2305.14239.
Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L., 2020a. Multilingual denoising pre-training for neural machine translation. Trans. Assoc. Comput. Linguist. 8, 726–742.
Liu, A., Hu, X., Wen, L., Yu, P.S., 2023c. A comprehensive evaluation of ChatGPT's zero-shot Text-to-SQL capability. arXiv preprint arXiv:2303.13547.
Liu, Z., Huang, D., Huang, K., Li, Z., Zhao, J., 2021a. Finbert: A pre-trained financial language representation model for financial text mining. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 4513–4519.
Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C., 2023d. Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634.
Liu, J., Liu, C., Lv, R., Zhou, K., Zhang, Y., 2023e. Is chatgpt a good recommender? a preliminary study. arXiv preprint arXiv:2304.10149.
Liu, Y., Liu, P., Radev, D., Neubig, G., 2022. BRIO: Bringing order to abstractive summarization. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2890–2903.
Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., Zhang, Y., 2023f. Evaluating the logical reasoning ability of chatgpt and gpt-4. arXiv preprint arXiv:2304.03439.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N., 2021. Self-alignment pretraining for biomedical entity representations. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4228–4238.
Liu, H., Teng, Z., Cui, L., Zhang, C., Zhou, Q., Zhang, Y., 2023g. LogiCoT: Logical chain-of-thought instruction-tuning data collection with GPT-4. arXiv preprint arXiv:2305.12147.
Liu, X., Wang, J., Sun, J., Yuan, X., Dong, G., Di, P., Wang, W., Wang, D., 2023h. Prompting frameworks for large language models: A survey. arXiv preprint arXiv:2311.12785.
Liu, P., Wang, X., Xiang, C., Meng, W., 2020b. A survey of text data augmentation. In: 2020 International Conference on Computer Communication and Network Security (CCNS). IEEE, pp. 191–195.
Liu, S., Wright, A.P., Patterson, B.L., Wanderer, J.P., Turer, R.W., Nelson, S.D., McCoy, A.B., Sittig, D.F., Wright, A., 2023i. Assessing the value of ChatGPT for clinical decision support optimization. medRxiv, pp. 2023-2002.
Liu, J., Xia, C.S., Wang, Y., Zhang, L., 2023j. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. arXiv preprint arXiv:2305.01210.
Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Cheng, R.G.H., Klochkov, Y., Taufiq, M.F., Li, H., 2023k. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. arXiv preprint arXiv:2308.05374.
Liu, Z., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Liu, W., Shen, D., Li, Q., et al., 2023l. Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:2303.11032.
Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., 2021c. Self-supervised learning: Generative or contrastive. IEEE Trans. Knowl. Data Eng. 35 (1), 857–876.
Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y., Hu, H., 2023m. ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. arXiv preprint arXiv:2304.07666.
Liu, J., Zhou, P., Hua, Y., Chong, D., Tian, Z., Liu, A., Wang, H., You, C., Guo, Z., Zhu, L., et al., 2023n. Benchmarking large language models on cmexam–a comprehensive Chinese medical exam dataset. arXiv preprint arXiv:2306.03030.
Lopez-Lira, A., Tang, Y., 2023. Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619.
Loukas, L., Stogiannidis, I., Malakasiotis, P., Vassos, S., 2023. Breaking the bank with ChatGPT: Few-shot text classification for finance. arXiv preprint arXiv:2308.14634.
Lu, Q., Ding, L., Xie, L., Zhang, K., Wong, D.F., Tao, D., 2022a. Toward human-like evaluation for natural language generation with error analysis. arXiv preprint arXiv:2212.10179.
Lu, S., Guo, D., Ren, S., Huang, J., Svyatkovskiy, A., Blanco, A., Clement, C., Drain, D., Jiang, D., Tang, D., et al., 2021. CodeXGLUE: A machine learning benchmark dataset for code understanding and generation. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1).
Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., Sun, L., Wu, H., 2022b. Unified structure generation for universal information extraction. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5755–5772.
Lu, Q., Qiu, B., Ding, L., Xie, L., Tao, D., 2023a. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. arXiv preprint arXiv:2303.13809.

Lu, Y., Yang, X., Li, X., Wang, X.E., Wang, W.Y., 2023b. LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation. arXiv preprint arXiv:2305.11116.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Luo, Z., Xie, Q., Ananiadou, S., 2023. ChatGPT as a factual inconsistency evaluator for text summarization.

Luong, M.-T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1412–1421.

Lyu, Q., Tan, J., Zapadka, M.E., Ponnatapura, J., Niu, C., Myers, K.J., Wang, G., Whitlow, C.T., 2023a. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis. Comput. Ind. Biomed. Art 6 (1), 9.

Lyu, C., Xu, J., Wang, L., 2023b. New trends in machine translation using large language models: Case examples with chatgpt. arXiv preprint arXiv:2305.01181.

Ma, Y., Cao, Y., Hong, Y., Sun, A., 2023a. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! arXiv preprint arXiv:2303.08559.

Ma, Y., Wang, Z., Cao, Y., Li, M., Chen, M., Wang, K., Shao, J., 2022. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6759–6774.

Ma, C., Wu, Z., Wang, J., Xu, S., Wei, Y., Liu, Z., Guo, L., Cai, X., Zhang, S., Zhang, T., et al., 2023b. ImpressionGPT: an iterative optimizing framework for radiology report summarization with chatGPT. arXiv preprint arXiv:2304.08448.

Mahowald, K., Ivanova, A.A., Blank, I.A., Kanwisher, N., Tenenbaum, J.B., Fedorenko, E., 2023. Dissociating language and thought in large language models: a cognitive perspective. arXiv preprint arXiv:2301.06627.

Malkiel, I., Alon, U., Yehuda, Y., Keren, S., Barkan, O., Ronen, R., Koenigstein, N., 2023. GPT-calls: Enhancing call segmentation and tagging by generating synthetic conversations via large language models. arXiv preprint arXiv:2306.07941.

Mallikarjuna, C., Sivanesan, S., 2022. Question classification using limited labelled data. Inf. Process. Manage. 59 (6), 103094.

Markov, T., Zhang, C., Agarwal, S., Nekoul, F.E., Lee, T., Adler, S., Jiang, A., Weng, L., 2023. A holistic approach to undesired content detection in the real world. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37. pp. 15009–15018, no. 12.

Martínez-Cruz, R., López-López, A.J., Portela, J., 2023. ChatGPT vs state-of-the-art models: A benchmarking study in keyphrase generation task. arXiv preprint arXiv:2304.14177.

Mehrabi, N., Goyal, P., Dupuy, C., Hu, Q., Ghosh, S., Zemel, R., Chang, K.-W., Galstyan, A., Gupta, R., 2023. FLIRT: Feedback loop in-context red teaming. arXiv preprint arXiv:2308.04265.

Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M.D., Zou, Y., Wang, W., 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. arXiv preprint arXiv:2303.17395.

Meng, R., Yuan, X., Wang, T., Zhao, S., Trischler, A., He, D., 2021. An empirical study on neural keyphrase generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4985–5007.

Meoni, S., De la Clergerie, E., Ryffel, T., 2023. Large language models as instructors: A study on multilingual clinical entity extraction. In: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. pp. 178–190.

Michail, A., Konstantinou, S., Clematide, S., 2023. UZH_CLyp at SemEval-2023 task 9: Head-first fine-tuning and ChatGPT data generation for cross-lingual learning in tweet intimacy prediction. arXiv preprint arXiv:2303.01194.

Michalopoulos, G., Wang, Y., Kaka, H., Chen, H., Wong, A., 2020. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. arXiv preprint arXiv:2010.10391.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mireshghallah, F., Mattern, J., Gao, S., Shokri, R., Berg-Kirkpatrick, T., 2023. Smaller language models are better black-box machine-generated text detectors. ArXiv, abs/2305.09859.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C., 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. arXiv preprint arXiv:2301.11305.

Mitrovi'c, S., Andreoletti, D., Ayoub, O., 2023. ChatGPT or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text. ArXiv, abs/2301.13852.

Moradi, M., Blagec, K., Haberl, F., Samwald, M., 2021. Gpt-3 models are poor few-shot learners in the biomedical domain. arXiv preprint arXiv:2109.02555.

Moslem, Y., Haque, R., Way, A., 2023. Adaptive machine translation with large language models. arXiv preprint arXiv:2301.13294.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T.L., Bari, M.S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al., 2022. Crosslingual generalization through multitask finetuning. arXiv preprint arXiv:2211.01786.

Murthy, J.S., Siddesh, G., Srinivasa, K., 2019. TwitSenti: a real-time Twitter sentiment analysis and visualization framework. J. Inf. Knowl. Manag. 18 (02), 1950013.

Mysore, S., McCallum, A., Zamani, H., 2023. Large language model augmented narrative driven recommendations. arXiv preprint arXiv:2306.02250.

Nair, V., Schumacher, E., Kannan, A., 2023. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. arXiv preprint arXiv:2305.05982.

Narayan, S., Cohen, S.B., Lapata, M., 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1797–1807.

Nascimento, N., Alencar, P., Cowan, D., 2023. Comparing software developers with ChatGPT: An empirical investigation. arXiv preprint arXiv:2305.11837.

Nguyen, H.-T., 2023. A brief report on LawGPT 1.0: A virtual legal assistant based on GPT-3. arXiv preprint arXiv:2302.05729.

Nguyen, D.Q., Vu, T., Nguyen, A.T., 2020. BERTweet: A pre-trained language model for english tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14.

Ni, A., Yin, P., Zhao, Y., Riddell, M., Feng, T., Shen, R., Yin, S., Liu, Y., Yavuz, S., Xiong, C., et al., 2023. L2CEval: Evaluating language-to-code generation capabilities of large language models. arXiv preprint arXiv:2309.17446.

Nijkamp, E., Hayashi, H., Xiong, C., Savarese, S., Zhou, Y., 2023. Codegen2: Lessons for training llms on programming and natural languages. arXiv preprint arXiv:2305.02309.

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., Xiong, C., 2022. CodeGen: An open large language model for code with multi-turn program synthesis. In: The Eleventh International Conference on Learning Representations.

Nogueira, R., Jiang, Z., Pradeep, R., Lin, J., 2020. Document ranking with a pretrained sequence-to-sequence model. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 708–718.

Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E., 2023a. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.

Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E., 2023b. Capabilities of GPT-4 on medical challenge problems. arXiv abs/2303.13375.

Nunes, D., Primi, R., Pires, R., Lotufo, R., Nogueira, R., 2023. Evaluating GPT-3.5 and GPT-4 models on Brazilian university admission exams. arXiv preprint arXiv:2303.17003.

Oh, S., Jung, W., et al., 2023. Data augmentation for neural machine translation using generative language model. arXiv preprint arXiv:2307.16833.

Olmo, A., Sreedharan, S., Kambhampati, S., 2021. GPT3-to-plan: Extracting plans from text using GPT-3. arXiv preprint arXiv:2106.07131.

OpenAI, 2023. GPT-4 technical report. arXiv:2303.08774.

Orenstrakh, M.S., Karnalim, O., Suarez, C.A., Liut, M., 2023. Detecting LLM-generated text in computing education: A comparative study for ChatGPT cases. arXiv preprint arXiv:2307.07411.

Otter, D.W., Medina, J.R., Kalita, J.K., 2020. A survey of the usages of deep learning for natural language processing. IEEE Trans. Neural Netw. Learn. Syst. 32 (2), 604–624.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 35, 27730–27744.

Pagliardini, M., Gupta, P., Jaggi, M., 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 528–540.

Pan, W., Chen, Q., Xu, X., Che, W., Qin, L., 2023. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. arXiv preprint arXiv:2304.04256.

Pan, S.J., Yang, Q., 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318.

Parikh, S., Vohra, Q., Tumbade, P., Tiwari, M., 2023. Exploring zero and few-shot techniques for intent classification. arXiv preprint arXiv:2305.07157.

Pegoraro, A., Kumari, K., Fereidooni, H., Sadeghi, A.-R., 2023. To ChatGPT, or not to ChatGPT: That is the question!. arXiv preprint arXiv:2304.01487.

Peng, Y., 2022. A survey on modern recommendation system based on big data. arXiv preprint arXiv:2206.02631.

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., Tao, D., 2023a. Towards making the most of chatgpt for machine translation. arXiv preprint arXiv:2303.13780.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al., 2023b. Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv preprint arXiv:2302.12813.

Peng, B., Li, C., He, P., Galley, M., Gao, J., 2023c. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543.

Pereira, J., Fidalgo, R., Lotufo, R., Nogueira, R., 2023. Visconde: Multi-document QA with GPT-3 and neural reranking. In: European Conference on Information Retrieval. Springer, pp. 534–543.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., Irving, G., 2022. Red teaming language models with language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3419–3448.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237. http://dx.doi.org/10.18653/v1/N18-1202, URL https://aclanthology.org/N18-1202.

Phan, L., Tran, H., Le, D., Nguyen, H., Annibal, J., Peltekian, A., Ye, Y., 2021. CoTexT: Multi-task learning with code-text transformer. In: Proceedings of the 1st Workshop on Natural Language Processing for Programming (NLP4Prog 2021). pp. 40–47.

Phung, T., Padurean, V.-A., Cambronero, J.P., Gulwani, S., Kohn, T., Majumdar, R., Singla, A.K., Soares, G., 2023. Generative AI for programming education: Benchmarking ChatGPT, GPT-4, and human tutors. arXiv abs/2306.17156.

Poldrack, R.A., Lu, T., Beguš, G., 2023. AI-assisted coding: Experiments with GPT-4. arXiv preprint arXiv:2304.13187.

Prenner, J.A., Robbes, R., 2021. Automatic program repair with openai's codex: Evaluating QuixBugs. arXiv preprint arXiv:2111.03922.

Prodan, G.P., Pelican, E., 2022. Prompt scoring system for dialogue summarization using GPT-3. ACM Trans. Audio Speech Lang. Process. 1–9.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D., 2023. Is ChatGPT a general-purpose natural language processing task solver? arXiv preprint arXiv:2302.06476.

Qiu, S., Liu, Q., Zhou, S., Huang, W., 2022. Adversarial attack and defense technologies in natural language processing: A survey. Neurocomputing 492, 278–307.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X., 2020. Pre-trained models for natural language processing: A survey. Sci. China Technol. Sci. 63 (10), 1872–1897.

Radford, A., Jozefowicz, R., Sutskever, I., 2017. Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9.

Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al., 2021. Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (1), 5485–5551.

Rajpoot, P.K., Parikh, A., 2023. GPT-FinRE: In-context learning for financial relation extraction using large language models. arXiv preprint arXiv:2306.17519.

Ranjit, M., Ganapathy, G., Manuel, R., Ganu, T., 2023. Retrieval augmented chest X-Ray report generation using openai gpt models. arXiv preprint arXiv:2305.03660.

Rao, A.S., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A.K., Landman, A., Dryer, K., Succi, M.D., 2023. Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv, pp. 2023-2002.

Raunak, V., Menezes, A., Post, M., Awadallah, H.H., 2023a. Do GPTs produce less literal translations? arXiv preprint arXiv:2305.16806.

Raunak, V., Sharaf, A., Awadallah, H.H., Menezes, A., 2023b. Leveraging GPT-4 for automatic translation post-editing. arXiv preprint arXiv:2305.14878.

Rawte, V., Sheth, A., Das, A., 2023. A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922.

Rehana, H., Çam, N.B., Basmaci, M., He, Y., Özgür, A., Hur, J., 2023. Evaluation of GPT and BERT-based models on identifying protein-protein interactions in biomedical text. arXiv preprint arXiv:2303.17728.

Rezaimehr, F., Dadkhah, C., 2021. A survey of attack detection approaches in collaborative filtering recommender systems. Artif. Intell. Rev. 54, 2011–2066.

Robinson, J., Wingate, D., 2022. Leveraging large language models for multiple choice question answering. In: The Eleventh International Conference on Learning Representations.

Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Remez, T., Rapin, J., et al., 2023. Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950.

Sai, A.B., Mohankumar, A.K., Khapra, M.M., 2022. A survey of evaluation metrics used for NLG systems. ACM Comput. Surv. 55 (2), 1–39.

Salehinejad, H., Sankar, S., Barfett, J., Colak, E., Valaee, S., 2017. Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078.

Samaan, J.S., Yeo, Y.H., Rajeev, N., Hawley, L., Abel, S., Ng, W.H., Srinivasan, N., Park, J., Burch, M., Watson, R., et al., 2023. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes. Surg. 1–7.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Sarker, S., Qian, L., Dong, X., 2023. Medical data augmentation via ChatGPT: A case study on medication identification and medication event classification. arXiv preprint arXiv:2306.07297.

Savelka, J., Agarwal, A., Bogart, C., Sakr, M., 2023. Large language models (gpt) struggle to answer multiple-choice questions about code. arXiv preprint arXiv:2303.08033.

Sawada, T., Paleka, D., Havrilla, A., Tadepalli, P., Vidas, P., Kranias, A., Nay, J., Gupta, K., Komatsuzaki, A., 2023. ARB: Advanced reasoning benchmark for large language models. In: The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23.

Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al., 2022. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

Schaeffer, R., Miranda, B., Koyejo, S., 2023. Are emergent abilities of large language models a mirage? arXiv preprint arXiv:2304.15004.

Sengupta, N., Sahu, S.K., Jia, B., Katipomu, S., Li, H., Koto, F., Afzal, O.M., Kamboj, S., Pandit, O., Pal, R., et al., 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. arXiv preprint arXiv:2308.16149.

Sennrich, R., Haddow, B., Birch, A., 2016. Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 86–96.

Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J., 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. Dial. Discourse 9 (1), 1–49.

Shah, A., Chava, S., 2023. Zero is not hero yet: Benchmarking zero-shot performance of LLMs for financial tasks. arXiv preprint arXiv:2305.16633.

Shaib, C., Li, M.L., Joseph, S., Marshall, I.J., Li, J.J., Wallace, B.C., 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). arXiv preprint arXiv:2305.06299.

Shao, Z., Yu, Z., Wang, M., Yu, J., 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14974–14983.

Sharma, S., Joshi, A., Mukhija, N., Zhao, Y., Bhathena, H., Singh, P., Santhanam, S., Biswas, P., 2022. Systematic review of effect of data augmentation using paraphrasing on Named entity recognition. In: NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research.

Shen, C., Cheng, L., Bing, L., You, Y., Si, L., 2022. SentBS: Sentence-level beam search for controllable summarization. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 10256–10265.

Shen, C., Cheng, L., You, Y., Bing, L., 2023a. Are large language models good evaluators for abstractive summarization? arXiv preprint arXiv:2305.13091.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., Cui, P., 2021. Towards out-of-distribution generalization: A survey. arXiv preprint arXiv:2108.13624.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y., 2023b. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580.

Shen, Y., Song, K., Tan, X., Zhang, W., Ren, K., Yuan, S., Lu, W., Li, D., Zhuang, Y., 2023c. TaskBench: Benchmarking large language models for task automation.

Shi, Z., Wang, Y., Yin, F., Chen, X., Chang, K.-W., Hsieh, C.-J., 2023. Red teaming language model detectors with language models. arXiv preprint arXiv:2305.19713.

Shirafuji, A., Watanobe, Y., Ito, T., Morishita, M., Nakamura, Y., Oda, Y., Suzuki, J., 2023. Exploring the robustness of large language models for solving programming problems. arXiv preprint arXiv:2306.14583.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 1–48.

Siddiq, M.L., Santos, J.C.S., Tanvir, R.H., Ulfat, N., Rifat, F.A., Lopes, V.C., 2023. Exploring the effectiveness of large language models in generating unit tests. arXiv abs/2305.00418.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015). Computational and Biological Learning Society.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al., 2023a. Large language models encode clinical knowledge. Nature 1–9.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al., 2023b. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al., 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990.

Soltan, S., Ananthakrishnan, S., FitzGerald, J., Gupta, R., Hamza, W., Khan, H., Peris, C., Rawls, S., Rosenbaum, A., Rumshisky, A., et al., 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. arXiv preprint arXiv:2208.01448.

Song, M., Jiang, H., Shi, S., Yao, S., Lu, S., Feng, Y., Liu, H., Jing, L., 2023. Is ChatGPT a good keyphrase generator? A preliminary study. arXiv preprint arXiv:2303.13001.

Srivastava, P., Ganu, T., Guha, S., 2022. Towards zero-shot and few-shot table question answering using GPT-3. arXiv preprint arXiv:2210.17284.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al., 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Trans. Mach. Learn. Res..

Stahlberg, F., 2020. Neural machine translation: A review. J. Artificial Intelligence Res. 69, 343–418.

Stammbach, D., Antoniak, M., Ash, E., 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In: Proceedings of the 4th Workshop of Narrative Understanding (WNU2022). pp. 47–56.

Su, H., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N.A., Zettlemoyer, L., Yu, T., et al., 2022. One embedder, any task: Instruction-finetuned text embeddings. arXiv preprint arXiv:2212.09741.

Sugiyama, A., Yoshinaga, N., 2019. Data augmentation using back-translation for context-aware neural machine translation. In: Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019). pp. 35–44.

Sun, L., Han, Y., Zhao, Z., Ma, D., Shen, Z., Chen, B., Chen, L., Yu, K., 2023a. SciEval: A multi-level large language model evaluation benchmark for scientific research. arXiv preprint arXiv:2308.13149.

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G., 2023b. Text classification via large language models. arXiv preprint arXiv:2305.08377.

Sun, W., Yan, L., Ma, X., Ren, P., Yin, D., Ren, Z., 2023c. Is ChatGPT good at search? Investigating large language models as re-ranking agent. arXiv preprint arXiv:2304.09542.

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., Zhou, D., 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2158–2170.

Sundar, A., Heck, L., 2022. Multimodal conversational AI: A survey of datasets and approaches. In: Proceedings of the 4th Workshop on NLP for Conversational AI. pp. 131–147.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Adv. Neural Inf. Process. Syst. 27.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Tan, Y., Min, D., Li, Y., Li, W., Hu, N., Chen, Y., Qi, G., 2023. Evaluation of ChatGPT as a question answering system for answering complex questions. arXiv preprint arXiv:2303.07992.

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., Liu, Y., 2020. Neural machine translation: A review of methods, resources, and tools. AI Open 1, 5–21.

Tanaka, Y., Nakata, T., Aiga, K., Etani, T., Muramatsu, R., Katagiri, S., Kawai, H., Higashino, F., Enomoto, M., Noda, M., Kometani, M., Takamura, M., Yoneda, T., Kakizaki, H., Nomura, A., 2023a. Performance of generative pretrained transformer on the national medical licensing examination in Japan. medRxiv.

Tanaka, Y., Nakata, T., Aiga, K., Etani, T., Muramatsu, R., Katagiri, S., Kawai, H., Higashino, F., Enomoto, M., Noda, M., et al., 2023b. Performance of generative pretrained transformer on the national medical licensing examination in Japan. medRxiv, pp. 2023-2004.

Tang, R., Han, X., Jiang, X., Hu, X., 2023a. Does synthetic data generation of llms help clinical text mining? arXiv preprint arXiv:2303.04360.

Tang, T., Lu, H., Jiang, Y.E., Huang, H., Zhang, D., Zhao, W.X., Wei, F., 2023b. Not all metrics are guilty: Improving NLG evaluation with LLM paraphrasing. arXiv preprint arXiv:2305.15067.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., Fan, A., 2020. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401.

Tang, X., Tran, A., Tan, J., Gerstein, M., 2023c. GersteinLab at MEDIQA-chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. arXiv preprint arXiv:2305.05001.

Tay, Y., Dehghani, M., Bahri, D., Metzler, D., 2022. Efficient transformers: A survey. arXiv:2009.06732.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., Stojnic, R., 2022. Galactica: A large language model for science. arXiv preprint arXiv:2211.09085.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I., 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Thapa, S., Naseem, U., Nasim, M., 2023. From humans to machines: can ChatGPT-like LLMs effectively replace human annotators in NLP tasks. In: Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media.

Theocharopoulos, P.C., Anagnostou, P., Tsoukala, A., Georgakopoulos, S.V., Tasoulis, S.K., Plagianakos, V.P., 2023. Detection of fake generated scientific abstracts. arXiv preprint arXiv:2304.06148.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al., 2022. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239.

Tian, H., Lu, W., Li, T.O., Tang, X., Cheung, S.-C., Klein, J., Bissyandé, T.F., 2023. Is ChatGPT the ultimate programming assistant–how far is it? arXiv preprint arXiv:2304.11938.

Torfi, A., Shirvani, R.A., Keneshloo, Y., Tavaf, N., Fox, E.A., 2020. Natural language processing advancements by deep learning: A survey. arXiv preprint arXiv:2003.01200.

Törnberg, P., 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. arXiv preprint arXiv:2304.06588.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023a. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Umapathi, L.K., Pal, A., Sankarasubbu, M., 2023. Med-halt: Medical domain hallucination test for large language models. arXiv preprint arXiv:2307.15343.

Valmeekam, K., Olmo, A., Sreedharan, S., Kambhampati, S., 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In: NeurIPS 2022 Foundation Models for Decision Making Workshop.

Van Atteveldt, W., Van der Velden, M.A., Boukes, M., 2021. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. Commun. Methods Meas. 15 (2), 121–140.

Van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109 (2), 373–440.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wadhwa, S., Amir, S., Wallace, B.C., 2023. Revisiting relation extraction in the era of large language models. arXiv preprint arXiv:2305.05003.

Wahle, J.P., Ruas, T., Kirstein, F., Gipp, B., 2022. How large language models are transforming machine-paraphrase plagiarism. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 952–963.

Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., Kurohashi, S., 2023. Gpt-re: In-context learning for relation extraction using large language models. arXiv preprint arXiv:2305.02105.

Wang, X., Gong, Z., Wang, G., Jia, J., Xu, Y., Zhao, J., Fan, Q., Wu, S., Hu, W., Li, X., 2023a. Chatgpt performs on the chinese national medical licensing examination.

Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., Yang, L., Huang, H., Ye, W., Geng, X., et al., 2023b. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. arXiv preprint arXiv:2302.12095.

Wang, X., Hu, Z., Lu, P., Zhu, Y., Zhang, J., Subramaniam, S., Loomba, A., Zhang, S., Sun, Y., Wang, W., 2023c. SCIBENCH: Evaluating college-level scientific problem-solving abilities of large language models. In: The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23.

Wang, Y., Le, H., Gotmare, A.D., Bui, N.D., Li, J., Hoi, S.C., 2023d. Codet5+: Open code large language models for code understanding and generation. arXiv preprint arXiv:2305.07922.

Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., Sui, Z., 2023e. Large language models are not fair evaluators. arXiv preprint arXiv:2305.17926.

Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H., 2020a. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768.

Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., Qu, J., Zhou, J., 2023f. Is chatgpt a good nlg evaluator? a preliminary study. arXiv preprint arXiv:2303.04048.

Wang, L., Lim, E.-P., 2023. Zero-shot next-item recommendation using large pretrained language models. arXiv preprint arXiv:2304.03153.

Wang, X., Liu, Q., Gui, T., Zhang, Q., Zou, Y., Zhou, X., Ye, J., Zhang, Y., Zheng, R., Pang, Z., et al., 2021a. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. pp. 347–355.

Wang, S., Liu, Y., Xu, Y., Zhu, C., Zeng, M., 2021b. Want to reduce labeling cost? GPT-3 can help. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 4195–4205.

Wang, H., Luo, X., Wang, W., Yan, X., 2023g. Bot or human? Detecting ChatGPT imposters with a single question. ArXiv, abs/2305.06424.

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., Tu, Z., 2023h. Document-level machine translation with large language models. arXiv preprint arXiv:2304.02210.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A.S., Arunkumar, A., Stap, D., et al., 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 5085–5109.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R., 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: International Conference on Learning Representations.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G., 2023i. Gpt-ner: Named entity recognition via large language models. arXiv preprint arXiv:2304.10428.

Wang, S., Sun, Y., Xiang, Y., Wu, Z., Ding, S., Gong, W., Feng, S., Shang, J., Zhao, Y., Pang, C., et al., 2021c. Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2112.12731.

Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-t., Jiao, W., Lyu, M.R., 2023j. All languages matter: On the multilingual safety of large language models. arXiv preprint arXiv:2310.00905.

Wang, Y., Wang, W., Joty, S., Hoi, S.C., 2021d. CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 8696–8708.

Wang, H., Wang, R., Mi, F., Wang, Z., Xu, R., Wong, K.-F., 2023k. Chain-of-thought prompting for responding to in-depth dialogue questions with LLM. arXiv preprint arXiv:2305.11792.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M., 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. arXiv preprint arXiv:2002.10957.

Wang, Z., Xie, Q., Ding, Z., Feng, Y., Xia, R., 2023l. Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv:2304.04339.

Wang, W.Y., Yang, D., 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2557–2563.

Wang, J., Yao, Z., Mitra, A., Osebe, S., Yang, Z., Yu, H., 2023m. UMASS_BioNLP at MEDIQA-Chat 2023: Can LLMs generate high-quality synthetic note-oriented doctor-patient conversations? arXiv preprint arXiv:2306.16931.

Wang, Y., Zhao, Y., 2023. TRAM: Benchmarking temporal reasoning for large language models. arXiv preprint arXiv:2310.00835.

Wang, Y., Zhao, Y., Petzold, L., 2023n. Are large language models ready for healthcare? A comparative study on clinical language understanding. arXiv preprint arXiv:2304.05368.

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., Liu, Q., 2023o. Aligning large language models with human: A survey. arXiv preprint arXiv:2307.12966.

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., et al., 2023. Zero-shot information extraction via chatting with chatgpt. arXiv preprint arXiv:2302.10205.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al., 2022a. Emergent abilities of large language models. Trans. Mach. Learn. Res..

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al., 2022b. Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. 35, 24824–24837.

Wei, J., Zou, K., 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 6382–6388.

Weng, Y., Li, B., Xia, F., Zhu, M., Sun, B., He, S., Liu, K., Zhao, J., 2023. Large language models need holistically thought in medical conversational QA. arXiv preprint arXiv:2305.05410.

Whitehouse, C., Choudhury, M., Aji, A.F., 2023. LLM-powered data augmentation for enhanced crosslingual performance. ArXiv, abs/2305.14288.

Wiriyathammabhum, P., 2022. PromptShots at the FinNLP-2022 ERAI task: Pairwise comparison and unsupervised ranking. In: Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP). pp. 104–110.

Wu, S., He, Y., 2019. Enriching pre-trained language model with entity information for relation classification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 2361–2364.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G., 2023a. Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N., 2023b. Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671.

Wu, Z., Zhang, L., Cao, C., Yu, X., Dai, H., Ma, C., Liu, Z., Zhao, L., Li, G., Liu, W., et al., 2023c. Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology NLI task. arXiv preprint arXiv:2304.09138.

Xia, C.S., Zhang, L., 2023. Keep the conversation going: Fixing 162 out of 337 bugs for 0.42 each using ChatGPT. arXiv preprint arXiv:2304.00385.

Xie, Y., Gao, J., Zhou, P., Ye, Q., Hua, Y., Kim, J., Wu, F., Kim, S., 2023a. Rethinking multi-interest learning for candidate matching in recommender systems. arXiv preprint arXiv:2302.14532.

Xie, Y., Yu, C., Zhu, T., Bai, J., Gong, Z., Soh, H., 2023b. Translating natural language to planning goals with large-language models. arXiv preprint arXiv:2302.05128.

Xiong, H., Wang, S., Zhu, Y., Zhao, Z., Liu, Y., Wang, Q., Shen, D., 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097.

Xu, J., Lu, L., Yang, S., Liang, B., Peng, X., Pang, J., Ding, J., Shi, X., Yang, L., Song, H., et al., 2023a. MedGPTEval: A dataset and benchmark to evaluate responses of large language models in medicine. arXiv preprint arXiv:2305.07340.

Xu, W., Wang, D., Pan, L., Song, Z., Freitag, M., Wang, W.Y., Li, L., 2023b. Instructscore: Towards explainable text generation evaluation with automatic feedback. arXiv preprint arXiv:2305.14282.

Xu, Y., Xu, R., Iter, D., Liu, Y., Wang, S., Zhu, C., Zeng, M., 2023c. InheritSumm: A general, versatile and compact summarizer by distilling from GPT. arXiv preprint arXiv:2305.13083.

Xu, P., Zhu, X., Clifton, D.A., 2023d. Multimodal learning with transformers: A survey. IEEE Trans. Pattern Anal. Mach. Intell..

Xu, X., Zhu, Y., Wang, X., Zhang, N., 2023e. How to unleash the power of large language models for few-shot relation extraction? arXiv preprint arXiv:2305.01555.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C., 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 483–498.

Yan, W., Liu, H., Wang, Y., Li, Y., Chen, Q., Wang, W., Lin, T., Zhao, W., Zhu, L., Deng, S., et al., 2023. CodeScope: An execution-based multilingual multitask multidimensional benchmark for evaluating LLMs on code understanding and generation. arXiv preprint arXiv:2311.08588.

Yang, X., Cheng, W., Petzold, L., Wang, W.Y., Chen, H., 2023a. DNA-GPT: Divergent N-gram analysis for training-free detection of GPT-generated text. arXiv preprint arXiv:2305.17359.

Yang, Z., Cherian, S., Vucetic, S., 2023b. Data augmentation for radiology report simplification. In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1877–1887.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. 32.

Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L., 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. pp. 3081–3089.

Yang, K., Ji, S., Zhang, T., Xie, Q., Ananiadou, S., 2023c. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. arXiv preprint arXiv:2304.03347.

Yang, L., Jiang, F., Li, H., 2023d. Is chatgpt involved in texts? Measure the polish ratio to detect ChatGPT-generated text. ArXiv, abs/2307.11380.

Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L., 2023e. Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381.

Yang, W., Li, C., Zhang, J., Zong, C., 2023f. BigTrans: Augmenting large language models with multilingual translation capability over 100 languages. arXiv preprint arXiv:2305.18098.

Yang, H., Liu, X.-Y., Wang, C.D., 2023g. FinGPT: Open-source financial large language models. arXiv preprint arXiv:2306.06031.

Yang, W., Nicolai, G., 2023. Neural machine translation data generation and augmentation using ChatGPT. arXiv preprint arXiv:2307.05779.

Yang, Y., Uy, M.C.S., Huang, A., 2020a. Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.

Yang, S., Wang, Y., Chu, X., 2020b. A survey of deep learning techniques for neural machine translation. arXiv abs/2002.07526.

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., et al., 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. arXiv preprint arXiv:2303.10420.

Ye, D., Lin, Y., Li, P., Sun, M., 2022. Packed levitated marker for entity and relation extraction. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4904–4917.

Yetiştiren, B., Özsoy, I., Ayerdem, M., Tüzün, E., 2023. Evaluating the code quality of AI-assisted code generation tools: An empirical study on GitHub copilot, amazon CodeWhisperer, and ChatGPT. arXiv preprint arXiv:2304.10778.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E., 2023. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549.

Young, T., Hazarika, D., Poria, S., Cambria, E., 2018. Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. 13 (3), 55–75.

Yu, P., Chen, J., Feng, X., Xia, Z., 2023a. CHEAT: A large-scale dataset for detecting ChatGPT-writtEn AbsTracts. arXiv preprint arXiv:2304.12008.

Yu, X., Qi, Y., Chen, K., Chen, G., Yang, X., Zhu, P., Zhang, W., Yu, N.H., 2023b. GPT paternity test: GPT generated text detection with GPT genetic inheritance. ArXiv, abs/2305.12519.

Yu, F., Quartey, L., Schilder, F., 2022. Legal prompting: Teaching a language model to think like a lawyer. arXiv preprint arXiv:2212.01326.

Yu, J., Wang, X., Tu, S., Cao, S., Zhang-Li, D., Lv, X., Peng, H., Yao, Z., Zhang, X., Li, H., et al., 2023c. KoLA: Carefully benchmarking world knowledge of large language models. arXiv preprint arXiv:2306.09296.

Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A., Krishna, R., Shen, J., Zhang, C., 2023d. Large language model as attributed training data generator: A tale of diversity and bias. arXiv preprint arXiv:2306.15895.

Yuan, Z., Lou, Y., Liu, M., Ding, S., Wang, K., Chen, Y., Peng, X., 2023a. No more manual tests? Evaluating and improving ChatGPT for unit test generation. arXiv abs/2305.04207.

Yuan, W., Neubig, G., Liu, P., 2021. Bartscore: Evaluating generated text as text generation. Adv. Neural Inf. Process. Syst. 34, 27263–27277.

Yuan, X., Wang, T., Meng, R., Thaker, K., Brusilovsky, P., He, D., Trischler, A., 2020. One size does not fit all: Generating and evaluating variable number of keyphrases. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7961–7975.

Yuan, C., Xie, Q., Ananiadou, S., 2023b. Zero-shot temporal relation extraction with chatgpt. arXiv preprint arXiv:2304.05454.

Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al., 2020. Big bird: Transformers for longer sequences. In: NeurIPS.

Zaib, M., Zhang, W.E., Sheng, Q.Z., Mahmood, A., Zhang, Y., 2022. Conversational question answering: A survey. Knowl. Inf. Syst. 64 (12), 3151–3195.

Zaitsu, W., Jin, M., 2023. Distinguishing ChatGPT (-3.5,-4)-generated and human-written papers through Japanese stylometric analysis. arXiv preprint arXiv:2304.05534.

Zan, D., Chen, B., Yang, D., Lin, Z., Kim, M., Guan, B., Wang, Y., Chen, W., Lou, J.-G., 2022. CERT: Continual pre-training on sketches for library-oriented code generation. arXiv preprint arXiv:2206.06888.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al., 2022. GLM-130b: An open bilingual pre-trained model. In: The Eleventh International Conference on Learning Representations.

Zeng, Z., Yu, J., Gao, T., Meng, Y., Goyal, T., Chen, D., 2023. Evaluating large language models at evaluating instruction following. arXiv preprint arXiv:2310.07641.

Zhan, H., He, X., Xu, Q., Wu, Y., Stenetorp, P., 2023a. G3Detector: General GPT-generated text detector. arXiv preprint arXiv:2305.12680.

Zhan, H., Li, Z., Wang, Y., Luo, L., Feng, T., Kang, X., Hua, Y., Qu, L., Soon, L.-K., Sharma, S., et al., 2023b. SocialDial: A benchmark for socially-aware dialogue systems. arXiv preprint arXiv:2304.12026.

Zhang, J., Bao, K., Zhang, Y., Wang, W., Feng, F., He, X., 2023a. Is chatgpt fair for recommendation? Evaluating fairness in large language model recommendation. arXiv preprint arXiv:2305.07609.

Zhang, L., Cai, W., Liu, Z., Yang, Z., Dai, W., Liao, Y., Qin, Q., Li, Y., Liu, X., Liu, Z., et al., 2023b. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. arXiv preprint arXiv:2308.09975.

Zhang, B., Fu, X., Ding, D., Huang, H., Li, Y., Jing, L., 2023c. Investigating chain-of-thought with ChatGPT for stance detection on social media. arXiv preprint arXiv:2304.03087.

Zhang, S., Gong, C., Wu, L., Liu, X., Zhou, M., 2023d. AutoML-GPT: Automatic machine learning with GPT. arXiv preprint arXiv:2305.02499.

Zhang, K., Gutiérrez, B.J., Su, Y., 2023e. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. arXiv preprint arXiv:2305.11159.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2019. BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al., 2023f. Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

Zhang, X., Li, S., Hauer, B., Shi, N., Kondrak, G., 2023g. Don't trust GPT when your question is not in english. arXiv preprint arXiv:2305.16339.

Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., Qiu, X., 2023h. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. arXiv preprint arXiv:2305.11000.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al., 2022. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

Zhang, B., Soh, H., 2023. Large language models as zero-shot human models for human-robot interaction. arXiv preprint arXiv:2303.03548.

Zhang, Y., Yang, Q., 2021. A survey on multi-task learning. IEEE Trans. Knowl. Data Eng. 34 (12), 5586–5609.

Zhang, Z., Yao, Y., Zhang, A., Tang, X., Ma, X., He, Z., Wang, Y., Gerstein, M., Wang, R., Liu, G., et al., 2023i. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. arXiv preprint arXiv:2311.11797.

Zhang, L., Zhang, Y., Ren, K., Li, D., Yang, Y., 2023j. MLCopilot: Unleashing the power of large language models in solving machine learning tasks. arXiv preprint arXiv:2304.14979.

Zhang, T., Zhang, Y., Vineet, V., Joshi, N., Wang, X., 2023k. Controllable text-to-image generation with GPT-4. arXiv preprint arXiv:2305.18583.

Zhang, X., Zhao, J., LeCun, Y., 2015. Character-level convolutional networks for text classification. Adv. Neural Inf. Process. Syst. 28.

Zhang, J., Zhao, Y., Saleh, M., Liu, P., 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: International Conference on Machine Learning. PMLR, pp. 11328–11339.

Zhao, Z., Guo, L., Yue, T., Chen, S., Shao, S., Zhu, X., Yuan, Z., Liu, J., 2023a. ChatBridge: Bridging modalities with large language model as a language catalyst. arXiv preprint arXiv:2305.16103.

Zhao, K., Jin, X., Bai, L., Guo, J., Cheng, X., 2022a. Knowledge-enhanced self-supervised prototypical network for few-shot event detection. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 6266–6275.

Zhao, W.X., Liu, J., Ren, R., Wen, J.-R., 2022b. Dense text retrieval based on pretrained language models: A survey. arXiv preprint arXiv:2211.14876.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C.M., Eger, S., 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 563–578.

Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S., 2021. Calibrate before use: Improving few-shot performance of language models. In: International Conference on Machine Learning. PMLR, pp. 12697–12706.

Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., Qin, B., 2023b. Is ChatGPT equipped with emotional dialogue capabilities? arXiv preprint arXiv:2304.09582.

Zhao, Y., Zhao, C., Nan, L., Qi, Z., Zhang, W., Tang, X., Mi, B., Radev, D., 2023c. RobuT: A systematic study of table QA robustness against human-annotated adversarial perturbations. arXiv preprint arXiv:2306.14321.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al., 2023d. A survey of large language models. arXiv preprint arXiv:2303.18223.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al., 2023a. Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv preprint arXiv:2306.05685.

Zheng, S., Huang, J., Chang, K.C.-C., 2023b. Why does ChatGPT fall short in answering questions faithfully? arXiv preprint arXiv:2304.10513.

Zheng, M., Su, X., You, S., Wang, F., Qian, C., Xu, C., Albanie, S., 2023c. Can GPT-4 perform neural architecture search? arXiv preprint arXiv:2304.10970.

Zhiyuli, A., Chen, Y., Zhang, X., Liang, X., 2023. BookGPT: A general framework for book recommendation empowered by large language model. arXiv preprint arXiv:2305.15673.

Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D., 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arXiv preprint arXiv:2302.10198.

Zhou, S., Alon, U., Agarwal, S., Neubig, G., 2023. Codebertscore: Evaluating code generation with pretrained models of code. arXiv preprint arXiv:2302.05527.

Zhu, X., Li, J., Liu, Y., Ma, C., Wang, W., 2023a. A survey on model compression for large language models. arXiv preprint arXiv:2308.07633.

Zhu, W., Liu, H., Dong, Q., Xu, J., Kong, L., Chen, J., Li, L., Huang, S., 2023b. Multilingual machine translation with large language models: Empirical results and analysis. arXiv preprint arXiv:2304.04675.

Zhu, F., Wang, Y., Chen, C., Zhou, J., Li, L., Liu, G., 2021. Cross-domain recommendation: challenges, progress, and prospects. arXiv preprint arXiv:2103.01696.

Zhu, W., Wang, X., Lu, Y., Fu, T.-J., Wang, X.E., Eckstein, M., Wang, W.Y., 2023c. Collaborative generative AI: Integrating GPT-k for efficient editing in text-to-image generation. arXiv preprint arXiv:2305.11317.

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Gong, N.Z., Zhang, Y., et al., 2023d. PromptBench: Towards evaluating the robustness of large language models on adversarial prompts. arXiv preprint arXiv:2306.04528.

Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., Tyson, G., 2023e. Can chatgpt reproduce human-generated labels? a study of social computing tasks. arXiv preprint arXiv:2304.10145.

Zhuang, Z., Chen, Q., Ma, L., Li, M., Han, Y., Qian, Y., Bai, H., Feng, Z., Zhang, W., Liu, T., 2023. Through the lens of core competency: Survey on evaluation of large language models. arXiv preprint arXiv:2308.07902.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020. A comprehensive survey on transfer learning. Proc. IEEE 109 (1), 43–76.

Zhuo, T.Y., 2023. Large language models are state-of-the-art evaluators of code generation. arXiv preprint arXiv:2304.14317.

Zhuo, T.Y., Li, Z., Huang, Y., Li, Y.-F., Wang, W., Haffari, G., Shiri, F., 2023. On robustness of prompt-based semantic parsing with large pre-trained language model: An empirical study on codex. arXiv preprint arXiv:2301.12868.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D., 2023a. Can large language models transform computational social science? arXiv preprint arXiv:2305.03514.

Ziems, N., Yu, W., Zhang, Z., Jiang, M., 2023b. Large language models are built-in autoregressive search engines. arXiv preprint arXiv:2305.09612.