Journal of Intelligent Learning Systems and Applications, 2024, 16, 448-480
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Pedagogical Alignment of Large Language Models (LLM) for Personalized Learning: A Survey, Trends and Challenges

Mahefa Abel Razafinirina*, **William Germain Dimbisoa**, **Thomas Mahatody**

School of Computer Science, University of Fianarantsoa, Fianarantsoa, Madagascar
Email: *abelmahefa@gmail.com

How to cite this paper: Razafinirina, M.A., Dimbisoa, W.G. and Mahatody, T. (2024) Pedagogical Alignment of Large Language Models (LLM) for Personalized Learning: A Survey, Trends and Challenges. *Journal of Intelligent Learning Systems and Applications*, **16**, 448-480.
https://doi.org/10.4236/jilsa.2024.164023

Received: August 11, 2024
Accepted: November 26, 2024
Published: November 29, 2024

Abstract

This survey paper investigates how personalized learning offered by Large Language Models (LLMs) could transform educational experiences. We explore Knowledge Editing Techniques (KME), which guarantee that LLMs maintain current knowledge and are essential for providing accurate and up-to-date information. The datasets analyzed in this article are intended to evaluate LLM performance on educational tasks, such as error correction and question answering. We acknowledge the limitations of LLMs while highlighting their fundamental educational capabilities in writing, math, programming, and reasoning. We also explore two promising system architectures: a Mixture-of-Experts (MoE) framework and a unified LLM approach, for LLM-based education. The MoE approach makes use of specialized LLMs under the direction of a central controller for various subjects. We also discuss the use of LLMs for individualized feedback and their possibility in content creation, including the creation of videos, quizzes, and plans. In our final section, we discuss the difficulties and potential solutions for incorporating LLMs into educational systems, highlighting the importance of factual accuracy, reducing bias, and fostering critical thinking abilities. The purpose of this survey is to show the promise of LLMs as well as the issues that still need to be resolved in order to facilitate their responsible and successful integration into the educational ecosystem.

Keywords

Chain of Thought, Education, IA, LLM, Machine Learning, NLP, Personalized Learning, Prompt Optimization, Video Generation

1. Introduction and Generality

The advent of Large Language Models (LLMs) is set to revolutionize the educational landscape by introducing new paradigms in personalized learning and content generation. These advanced AI models, leveraging cutting-edge techniques in Artificial Intelligence (AI) and Natural Language Processing (NLP), have demonstrated significant potential to enhance educational practices by tailoring learning experiences to individual needs, optimizing content delivery, and enabling new forms of interaction between learners and educational material [1] [2].

LLMs, trained on vast datasets comprising both text and code, have shown exceptional proficiency across a wide array of language processing tasks, such as text generation, translation, and question answering [3]-[6]. This extensive capability positions them as powerful tools capable of addressing diverse educational needs, from creating personalized learning environments to supporting sophisticated intelligent tutoring systems [7]-[9]. These models not only generate adaptive learning materials but also provide real-time feedback, allowing for the personalization of educational content based on individual learning styles and teaching strategies [10]-[14]. Furthermore, their integration with pedagogical frameworks such as the Pedagogical Chain-of-Thought (PedCoT) enhances their ability to support reasoning and instructional capabilities in educational settings.

One of the most promising applications of LLMs in education is their ability to support personalized learning by adapting to the unique learning styles and needs of individual students. This adaptability allows LLMs to significantly enhance the effectiveness of educational interventions. They provide immediate, contextually relevant feedback, thereby allowing educators to deliver more targeted instruction [15] [16]. Moreover, the integration of personalized video learning and prompt optimization techniques further augments the capacity of LLMs to create tailored educational content and experiences [17] [18].

LLMs are built upon foundational advancements in AI, particularly the development of transformer architectures that enable complex language tasks with remarkable accuracy [19]. The introduction of techniques like tokenization, parsing, and semantic analysis has further refined the ability of these models to process and generate human-like text, which is crucial for their application in educational settings [20] [21]. These advancements make LLMs highly adaptable, allowing them to cater to diverse educational requirements across various subjects and grade levels.

This paper provides a comprehensive survey of the current state of LLMs in education, focusing on several key areas. We begin with an exploration of Knowledge Editing Techniques (KME), essential for keeping LLMs up-to-date with the latest information. Next, we analyze the datasets utilized for evaluating LLM performance in educational tasks, such as question answering and error correction [22]. The core capabilities of LLMs, including their application in mathematics, writing, programming, and reasoning, are critically examined [23]. In addition, we explore advanced techniques such as the Mixture-of-Experts (MoE) framework and the unified LLM approach, which offer novel solutions for personalized learning [24]. The introduction of the Pedagogical Chain-of-Thought (PedCoT) framework is high-lighted as a key innovation in improving reasoning and instructional capabilities of LLMs [25]. Furthermore, this survey delves into the integration of personalized learning systems with video generation and prompt optimization, emphasizing the transformative potential of these technologies in educational contexts [26].

However, several challenges have emerged regarding the application of LLMs in education. Studies such as [1] highlight issues like bias, factual inaccuracies, and the lack of interpretability, which can undermine trust in these tools. There are also concerns that relying too heavily on LLM-generated answers may hinder students' critical thinking. Similarly, [16] emphasizes the difficulty of creating personalized learning experiences without extensive personal data and the high computational costs, further complicating large-scale adoption. Addressing these challenges is essential for the successful integration of LLMs into personalized learning environments.

By critically examining the current landscape and future directions, this paper aims to pave the way for the responsible and effective integration of LLMs into the educational ecosystem, ensuring that they serve as powerful tools for enhancing learning outcomes while addressing the associated challenges [27].

Figure 1 provides an overview of the representative works in this domain, categorized based on their technical and pedagogical categories.

2. LLM Models

This section delves into the intricacies of LLM inference, the challenges posed by computational demands, and the strategies for fine-tuning these models to enhance their educational utility.

2.1. Theoretical Background

The theoretical foundation of Large Language Models (LLM models) is rooted in deep learning, specifically the transformer architecture introduced by [19]. LLM models, such as GPT-4[1] and BERT[2] [141], are highly parameterized neural networks designed to process and generate natural language text. These models leverage self-attention mechanisms, which allow them to capture long-range dependencies within text sequences efficiently, unlike earlier models such as RNNs[3] or LSTMs[4]. The scale of LLM models enables them to learn intricate patterns in language by training on vast amounts of data. This process, known as unsupervised pretraining, equips LLM models with general language understanding capabilities, which can then be fine-tuned on specific tasks through supervised

[1]https://openai.com/index/gpt-4/.

[2]https://huggingface.co/docs/transformers/en/model_doc/bert.

[3]Recurrent Neural Network.

[4]Long Short-Term Memory.

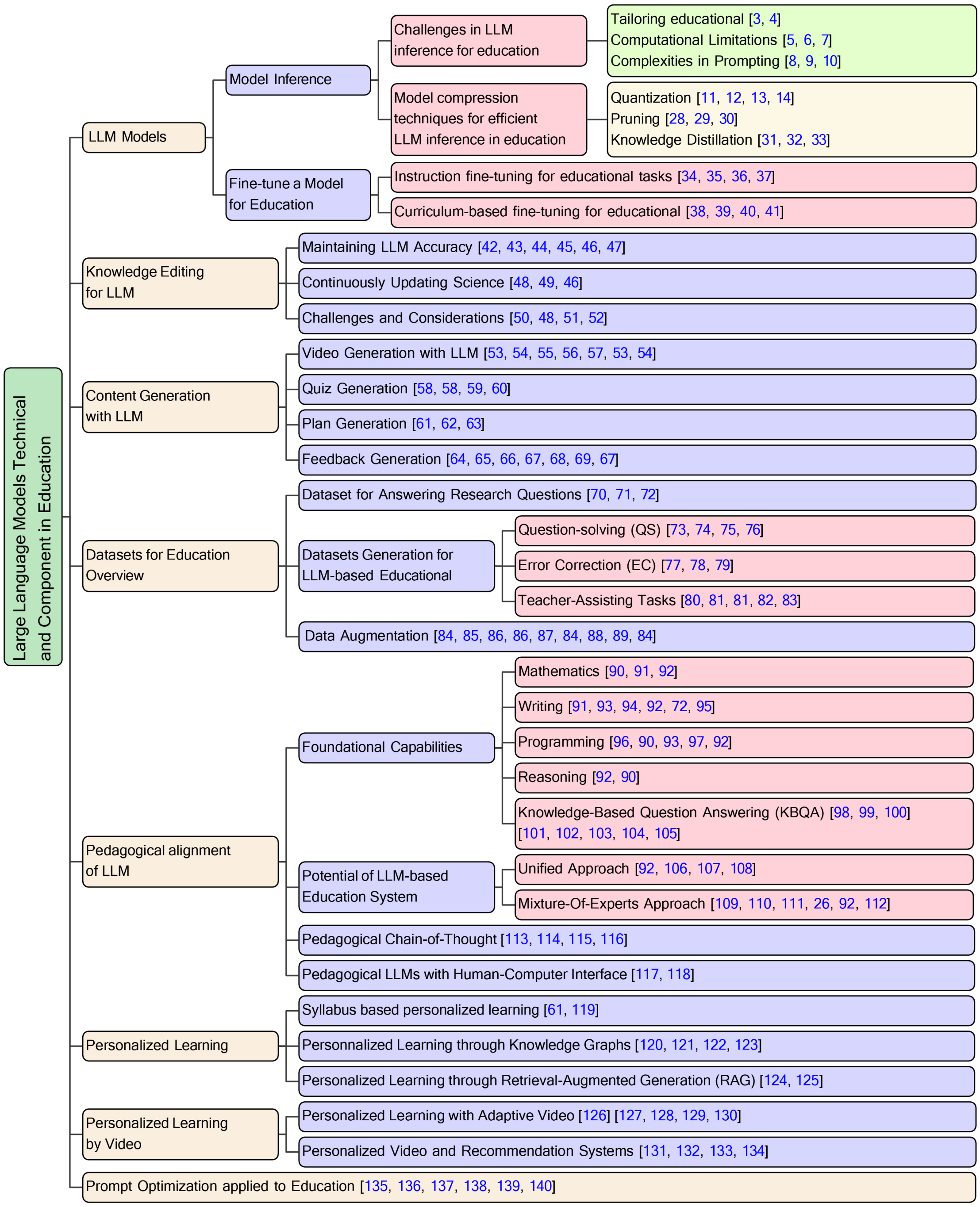

Figure 1. Representative works of LLMs for education alignment.

learning. Additionally, theoretical advancements in model optimization, such as gradient descent and distributed training, have made it possible to scale these models effectively.

2.2. Model Inference

The process of inference, which is fundamental to their ability to derive meaning from complex data such as student interactions with learning materials, is a crucial aspect of this understanding. By examining LLM inference, a framework was developed for evaluating current research on their application in educational settings, as outlined in [142]. This framework will elucidate the various inference methods utilized and their impact on educational outcomes. Moreover, a comprehensive examination of LLM inference will elucidate the challenges and prospects inherent in integrating these models into educational settings [143]. By elucidating these aspects, we can establish a foundation for future research and pedagogical practices that fully leverage the capabilities of LLMs to enhance the learning experience [144].

The broader adoption of LLMs in educational settings is constrained by their substantial computational and memory demands during the inference process [6]. This section explores the challenges, recent developments, and practical considerations associated with LLM inference within educational contexts, as outlined by [145].

2.2.1. Challenges in LLM Inference for Education

Realizing the potential of the LLM requires overcoming notable obstacles associated with computational constraints and the intricacies of controlling these models [3] [4].

One significant obstacle is the extensive computational resources that LLMs require for inference [6]. These models, often involving billions of parameters, often overwhelm traditional computing systems. This issue is particularly evident in educational contexts where high-performance computing accessibility may be restricted. For instance, expansive models such as LaMDA-137B[5] or LLaMa-70B[6] can surpass the Video RAM (VRAM) capacity of standard computers [5], impeding their utilization in educational institutions and home learning settings. Moreover, the sheer volume of parameters can notably decelerate response times, disrupting the learning process. Addressing the optimization of inference in such resource-limited environments requires further research and development endeavors [7].

Prompting LLMs poses an additional layer of complexity beyond computational constraints [8]. The construction of a prompt serves as a guiding principle for the model's generation, although its unpredictable nature can introduce challenges during inference. Anticipating the length and intricacy of an LLM's output in advance is a challenging task [9]. This variability results in challenges in effectively managing memory and computational resources. If the response exceeds expectations, models may face memory constraints, or conversely, may underutilize resources with a shorter response. Furthermore, LLMs exhibit high sensitivity to even subtle prompt variations, emphasizing the need for careful prompt design to align with the desired learning objectives and prevent unintended results [10].

[5]LaMDA-137B is a large-scale language model developed by Google, specialized for dialog applications, consisting of 137 billion parameters.
[6]LLaMA-70B is a large language model developed by Meta, consisting of 70 billion parameters.

2.2.2. Model Compression Techniques for Efficient LLM Inference in Education

LLMs provide a set of tools for individualized learning via inference[7], which involves generating responses or predictions based on input data [146]. Nevertheless, the considerable size of their parameters may present notable obstacles for educational purposes, especially within settings with limited resources. Recent studies have delved into a variety of model compression methods tailored specifically for LLMs to tackle these challenges, with the goal of improving inference efficiency while maintaining performance standards [147].

Quantization: Quantization is a technique that reduces the precision of the model's parameters, typically from 32-bit floating-point numbers to lower precision formats like 8-bit integers [11]. This significantly reduces memory footprint and improves inference speed, making LLMs more amenable to deployment in resource-constrained educational settings. Studies have shown that quantization techniques can achieve significant compression ratios while maintaining acceptable levels of accuracy in LLM inference tasks [12].

Pruning: The objective of pruning is to identify and remove any redundant or superfluous parameters within the LLM framework. This procedure results in a reduction in model size and computational complexity, consequently leading to quicker inference times [28]. The methodologies for pruning commonly entail the assessment of the contribution of individual parameters to the overall performance of the model, followed by the selective elimination of those with minimal impact. Recent studies have explored a range of pruning approaches tailored to LLMs, considering their unique architectural and training characteristics [29].

Knowledge Distillation: Knowledge distillation is a methodology in which a smaller, more streamlined model, known as the student model, acquires knowledge from a larger, already trained model, which is referred to as the teacher model [31]. Throughout the process of distillation, the student model undergoes training not solely based on the original training data but also on the "soft" results (probability distributions) produced by the teacher model [32]. This approach enables the student model to assimilate the expertise and competencies of the larger model in a more concise manner, thereby facilitating efficient inference for educational purposes [33].

2.3. Fine-Tune a Model for Education

The broad applicability of LLM hinders their performance in particular domains such as education. The method of fine-tuning, which exploits pre-existing LLMs and customizes them for particular tasks, presents a strategy for educational purposes [148].

2.3.1. Instruction Fine-Tuning for Educational Tasks

Traditional fine-tuning is dependent on extensive datasets containing labeled

[7]Inference in the context of LLM refers to the process by which the model generates predictions or responses based on input data, leveraging the knowledge encoded within its parameters.

instances. However, within the realm of education, the acquisition of high-caliber labeled data can prove to be both costly and time-intensive. An alternative approach known as instruction fine-tuning has emerged, which involves the utilization of natural language instructions to direct the Language Model (LLM) towards the desired results [34]. This particular technique empowers educators to capitalize on their expertise by crafting precise instructions for the LLM, thereby diminishing the need for pre-labeled data. To illustrate, an educator could furnish an instruction such as "Compose a concise historical overview of the French Revolution comprising 200 words" to fine-tune an LLM for tasks related to summarizing historical texts [35].

2.3.2. Curriculum-Based Fine-Tuning for Educational Progression

Educational content is constructed upon existing knowledge. Conventional fine-tuning often views educational assignments as separate units. The method of curriculum-based fine-tuning tackles this issue by integrating a learning progression into the fine-tuning procedure [38]. In this context, the Language Model (LLM) is exposed to a series of increasingly intricate assignments that mirror a student's educational path. This strategy aims to enhance the LLM's capacity not only to respond to queries but also to showcase a comprehension of fundamental principles and the capability to apply knowledge across diverse assignments [39].

3. Knowledge Editing for LLM

The efficacy of models is significantly contingent upon the quality and comprehensiveness of the training data. Given the perpetual generation of new information globally, LLMs run the risk of becoming obsolete, thereby yielding outputs that are erroneous or deceptive. Knowledge-based Model Editing (KME) presents a promising remedy to mitigate this challenge [22] [48] [149].

KME techniques aim to update pre-trained LLMs efficiently and precisely with new knowledge. This helps LLMs to enhance their performance in education-related applications like question answering and tutoring systems [22].

Here's a closer look at KME[8] in the context of LLMs for education:

Maintaining LLM Accuracy in Education: Educational content evolves with new discoveries, facts, and perspectives. KME ensures LLMs in education access the latest information. Crucial for tasks like question answering, where students rely on the LLMs to provide accurate and up-to-date answers [42].

Continuously Updating Science LLMs: The integration of the Knowledge Management Engine (KME) with a Learning Management System (LMS) in a science classroom offers the potential for continuous updating of the LMS with the latest scientific discoveries. The objective of this process is to provide students with access to the most up-to-date information. To illustrate, in the event of a new

[8]Knowledge Model Editing (KME) is a comprehensive approach that aims to precisely modify pre-trained language models to incorporate new knowledge while preserving existing information, addressing specific needs such as reducing biases, correcting errors, or updating factual knowledge.

species being identified, the KME can be employed to incorporate this information into the LLM, thus enabling it to respond accurately to student queries regarding the newly discovered species [48].

Challenges and Considerations: Although KME shows promise for educational LLMs, it is important to consider the challenges that may arise. It is of paramount importance to achieve a balance between editing specific knowledge (locality[9]) and maintaining overall model performance (generality). This is a crucial point that has been highlighted by [48]. Furthermore, it is imperative to guarantee the resilience of the edits against the influence of misinformation and bias in educational contexts, as elucidated by [50].

4. Content Generation with LLM

This section delves into various applications of LLMs in content generation, highlighting their impact and challenges across different domains, including video generation, quiz creation, plan development, and feedback provision.

4.1. Video Generation with LLM

The reviewed works examine the capability of LLMs in the context of video generation, situating them within the broader domain of generative AI (GAI) techniques for video creation, as outlined by [55] [56]. It is worth noting that the paper identifies two significant challenges: maintaining temporal consistency, which entails ensuring smooth and realistic transitions between video frames, and the high computational demands required for processing and generating video content using LLMs [53] [54]. To address these challenges, the paper references various strategies, including temporal attention layers and specialized training datasets [57].

4.2. Quiz Generation

In their study, [58] developed an AI-based quiz generation system using GPT-4 and the Math-Vista dataset[10]. This system was designed to enhance personalized learning through the use of adaptive quizzes. [59] employs reinforcement learning with the FLAN-T5 model to enhance the accuracy of questions. In a recent study, [60] explored the integration of LLMs and knowledge graphs[11] in cybersecurity education, with the aim of enhancing the accuracy and engagement of the educational content.

4.3. Plan Generation

[61] presents a framework for the generation and evaluation of teaching plans

[9]Locality refers to the precise update of specific knowledge within the model, while generality involves preserving the model's overall performance across a broad range of tasks.

[10]MathVista is a benchmark designed to evaluate the mathematical reasoning abilities of LLMs in visual contexts, consisting of 6,141 examples derived from diverse multimodal datasets.

[11]Knowledge graphs are structured representations of knowledge that connect information across different domains, enhancing the context and accuracy of educational content generated by LLMs.

utilising GPT-4. It identifies strengths in the setting of objectives and the organisation of activities, while also proposing improvements in teacher training and personalised instructional design. In their discussion of LangChain's LLM-powered chatbot, [62] highlight how this technology can enhance engagement, comprehension, and accessibility. This is achieved by analysing multimedia syllabus content[12], which allows for the delivery of personalised responses and a reduction in study time. In their study, [63] investigate the use of T5[13] and GPT-3.5[14] for the generation of student plans for adaptive scaffolding in game-based learning. Their findings demonstrate an effective alignment of LLM capabilities with pedagogical goals, which enhances self-regulated learning.

4.4. Feedback Generation

A review of the literature on LLM-based feedback generation reveals several key trends and challenges. The study by [64] explores the impact of diverse prompting strategies for LLMs on the quality of essay scoring and feedback generation. The study finds that, while LLMs enhance feedback quality, the impact of integrated scoring on feedback is minimal. As evidenced by [65], there is a pressing need for evidence-based approaches to enhance the quality of LLM feedback in educational contexts. This can be achieved by integrating intelligent tutoring systems and learning sciences. As demonstrated by [66], distinguishing between directional and non-directional feedback[15] is crucial in understanding the impact of LLM-based feedback on performance. As evidenced by the findings of [67], there are both advantages and limitations to the use of LLM-generated feedback in the context of programming courses. The optimal approach, it suggests, is to combine this with automated test-based feedback (ATF)[16] in order to achieve comprehensive results. In their research, [68] discuss OpineBot, a conversational LLM that enhances student engagement and feedback quality through interactive feedback processes. Both [67] and [69] examine the role of LLMs, such as ChatGPT[17], in providing feedback for concurrent programming. They identify significant limitations in the accuracy of error detection, underscoring the need for further refinement and integration with existing systems to ensure the reliability of feedback [18].

5. Datasets for Education Overview

This section provides an overview of how datasets can be leveraged in education,

[12]Multimedia content, including text, video, and interactive elements, supports personalized learning by catering to different learning styles and improving engagement, which can reduce the time needed to master content.

[13]https://huggingface.co/docs/transformers/model_doc/t5.

[14]https://platform.openai.com/docs/models/gpt-3-5-turbo.

[15]Directional feedback provides specific guidance on how to improve, while non-directional feedback is more general and evaluative, influencing how effectively students can apply corrections.

[16]ATF involves the use of predefined tests to automatically assess and provide feedback on a student's work, complementing LLM-generated feedback by ensuring accuracy and thoroughness.

[17]https://www.openai.com/chatgpt.

[18]Accurate error detection is essential in educational feedback as it directly impacts a student's ability to correct mistakes and improve learning outcomes.

covering their application in research, the development of LLM-based educational tools, and data augmentation techniques.

5.1. Dataset for Answering Research Questions

The KIWI dataset, introduced by [70], addresses this issue by focusing on knowledge-intensive writing tasks, such as revising long-form answers to research questions with expert-issued instructions [71]. The instructions include directives for information-seeking, stylistic modifications, and precise edits. The extant literature indicates that current LLMs, including GPT-4, are unable to perform these tasks adequately, particularly with regard to integrating new information and following precise edits. This evidence serves to underscore the challenges these models face in maintaining coherence [72]. The KIWI dataset offers invaluable insights for the development of LLMs that can effectively support educational applications.

5.2. Datasets Generation for LLM-Based Educational

The reviewed article elucidates the contemporary landscape of datasets and benchmarks utilized to evaluate the performance of LLM in educational contexts, as referenced in [150]. It underscores the extensive range of educational LLM applications, which encompass student data, learning resources, and educational game data [151]. Nevertheless, the emphasis remains on text-rich tasks where LLMs demonstrate particular proficiency, as evidenced by the findings of [152].

In their study, authors identify several publicly available datasets and benchmarks designed for evaluating LLMs in specific educational tasks [92]. These datasets primarily target the following areas:

- **Question-solving (QS):** This is a pervasive task for both education and NLP, and a plethora of datasets exist for the purpose of evaluating a system's capacity to transform a narrative description into a mathematical expression (e.g., word problems) [73]. Some datasets incorporate supplementary complexities, such as images, tables, and scientific textbook passages, in conjunction with textual descriptions [74].
- **Error correction (EC):** In order to facilitate error correction, it is essential to ensure that large language models are trained on diverse datasets. In the context of foreign language training, the incorporation of datasets containing grammatical and spelling errors proves beneficial for LLMs, as it facilitates their ability to identify and rectify mistakes, thereby aiding language learners [77]. In the field of computer science, the inclusion of erroneous code in training datasets enables LLMs to gain an understanding of fundamental coding principles, thereby facilitating the detection and recommendation of fixes for such bugs [78]. These capabilities render LLMs a valuable asset for programmers striving to enhance code quality [79].
- **Teacher-Assisting Tasks:** Researchers are creating specialized training datasets for teacher-assisting tasks[19] [80]. One area is question generation (QG), where

[19]Teacher-assisting tasks, like question generation and automatic grading, benefit significantly from tailored datasets.

datasets evaluate an LLM's ability to create educational questions based on a learning context [81]. For instance, an LLM could generate multiple-choice questions after a lecture, allowing teachers to focus on other instructional aspects [82]. Another area is automatic grading (AG), with datasets assessing LLMs' effectiveness in grading assignments like essays [83]. While not replacing human evaluation, LLMs could handle initial grading, enabling teachers to provide more personalized feedback [81].

5.3. Data Augmentation

[84] presents a new data augmentation method for few-shot named entity recognition (NER) using LLMs to address limited labeled data [85]. Traditional few-shot NER depends on manually curated datasets, which are costly and time-consuming [86]. LLM-DA[20] generates high-quality synthetic data to augment existing datasets [87]. It prompts LLMs to create text with specific named entities[21] based on user instructions [89], expanding training data for NER models [84] [88]. The authors' evaluation on benchmark NER datasets[22] shows significant improvements in few-shot NER performance over baseline models, indicating LLM-DA's promise for enhancing NER models with limited labeled data.

6. Pedagogical Alignment of LLM

To understand the pedagogical alignment of LLMs, we will explore their foundational capabilities, examine the potential of LLM-based education systems, and discuss the pedagogical chain-of-thought for detecting reasoning mistakes.

6.1. Foundational Capabilities

Building an LLM-based educational system hinges on the development of several core capabilities. Here, we explore these foundational functionalities, drawing insights from the work of [153].

6.1.1. Mathematics

While LLMs are capable of performing **basic calculations** with a reasonable degree of accuracy, their performance deteriorates as the complexity of the problems increases [90]. Complex **mathematical reasoning**, such as solving problems typically encountered in college-level courses or proving theorems, remains a significant challenge for these systems, as evidenced by recent studies [104].

An area of significant potential for advancement lies in multi-modal integration, which entails enabling LLMs to process problems that combine text and visuals, such as those encountered in geometry [91]. However, this area faces

[20]LLM-DA is a data augmentation technique using LLMs' rewriting capabilities and extensive knowledge.

[21]Named entities refer to specific entities such as people, organizations, locations, etc., which are crucial for various NLP tasks, including information extraction and question answering.

[22]The benchmark NER datasets used for evaluation include widely recognized datasets like CoNLL-2003 and OntoNotes 5.0.

challenges related to the sheer amount of data required to train LLMs effectively for such tasks [92].

6.1.2. Writing

The LLM's capacity to **summarize text** can prove invaluable for students grappling with the challenge of distilling information into a concise form, as evidenced by the findings of [91]. However, for educational purposes, it is imperative to develop enhanced evaluation metrics that extend beyond the mere measurement of factual accuracy [93]. The capture of elements such as the maintenance of pivotal concepts, clarity, and relevance to the learning objective is vital to guarantee that these summaries genuinely facilitate comprehension [92].

By identifying and suggesting corrections, LLM can provide valuable feedback to students engaged in the process of developing their writing skills [72]. However, it is essential to recognize that LLM-based correction tools may occasionally result in overcorrection or lack of precision [95]. Consequently, integrating such tools should be undertaken with caution, emphasizing human oversight and fostering the development of students' critical review skills [94].

6.1.3. Programming

LLM's capabilities are currently limited [93] [96]. Training them to **write code** effectively often requires extensive datasets, and they can struggle with complex algorithms [90]. However, LLMs show promise in **refining existing code**. They can identify and suggest improvements [97], but further research is needed to ensure these suggestions are interpretable by human programmers and don't compromise code efficiency [92].

6.1.4. Reasoning

LLMs demonstrate good ability in **problem-solving**, particularly when aided by well-designed prompts and their vast pre-trained knowledge, but currently face limitations in handling implicit reasoning and complex scenarios. This lack of **ability to explain** their thought processes or provide **clear guidance** can hinder their effectiveness in certain educational settings. For LLMs to truly excel as educational tools, further research is needed to bridge the gap between their impressive capabilities and the need for transparent and comprehensive reasoning, especially when tackling intricate problems [92].

6.1.5. Knowledge-Based Question Answering (KBQA)

One significant concern with LLM is their susceptibility to generating inaccurate or misleading information, sometimes referred to as "hallucinations"[23]. Improving **answer accuracy** and implementing methods for real-world **information verification** are crucial steps towards ensuring LLMs provide learners with trustworthy information (as discussed in [100]).

[23]The study of hallucinations in LLMs originated from a focus on natural language generation tasks, with early attention drawn by researchers like [99].

When it comes to answering open-ended questions (KBQA), two promising approaches are emerging: information retrieval from web sources (Open-domain) [101] and integration with **domain-specific** knowledge bases [102]. However, both approaches necessitate careful consideration to avoid perpetuating misinformation (as explored in [98]).

While these models hold potential for educational applications, addressing the limitations mentioned above is critical for their successful implementation in classrooms, as emphasized in [103] [105].

6.2. Potential of LLM-Based Education System

LLMs can revolutionize online education by understanding a wide range of student questions [103], similar to human teachers. They aim to provide support across different subjects and skill levels. [92] proposes two approaches for creating LLM-based education systems:

6.2.1. Unified Approach

This straightforward approach involves training a single, comprehensive LLM to handle questions from various subjects. Students can directly interact with this LLM, asking questions just as they would a human teacher. Research suggests promise for LLMs in some educational tasks, such as improving teaching strategies [92]. However, challenges remain in areas requiring deeper understanding, like grading student work or creating new problems [92].

6.2.2. Mixture-of-Experts Approach

The Mixture of Experts (MoE) framework addresses the limitations of single-purpose LLMs by using multiple specialized models for different subjects such as math, science, and history [109] [110]. An LLM controller coordinates student interactions with these experts, ensuring relevant responses by reformatting requests and aggregating outputs [111]. This approach simplifies training and leverages LLM strengths while mitigating their weaknesses, despite challenges in communication between the controller and expert models [26]. The MoE framework promises effective LLM-powered educational assistants tailored to diverse learning needs [92] [112].

6.3. Pedagogical Chain-of-Thought

The Pedagogical Chain-of-Thought (PedCoT) framework is highlighted across several studies as a crucial approach to enhancing reasoning and instructional capabilities of LLMs in various educational contexts. [113] discusses how PedCoT, combined with educational principles like Bloom's Cognitive Model, significantly improves the detection and correction of mathematical reasoning mistakes by LLMs. Other studies, such as [115], explore its application in automated grading systems, where structured reasoning processes are integrated to enhance the accuracy of student assessments in Earth Science. Similarly, [114] introduces the Chain of Thought with Landmarks (CoTL) to improve navigation instruction generation, further aligning with the PedCoT framework by embedding structured, step-by-step reasoning.

Additionally, [154] examines neuron activation in LLMs to understand the effectiveness of CoT prompting in arithmetic reasoning, offering insights into the underlying mechanisms that support the PedCoT approach. This is complemented by [155], which applies CoT reasoning to manage complex dialogues in sales scenarios, demonstrating its broader applicability.

[156] introduces a hierarchical graphical model to explain how LLMs generate coherent chains of thought during reasoning tasks, emphasizing the role of context and ambiguity in successful CoT generation, thus providing a theoretical foundation for educational applications. The AuRoRA platform presented by [157] and the structured CoT approach discussed by [158] both aim to refine and enhance the reasoning capabilities of LLMs in educational settings. Finally,

[159] provides a comprehensive survey of CoT reasoning techniques, categorizing them and emphasizing their importance in advancing personalized learning experiences through the PedCoT framework, while [116] highlights the usefulness of CoT reasoning to enhance transparency and accuracy in LLMs within an educational context, building user trust by justifying AI-driven decisions and ensuring alignment with educational goals.

6.4. Pedagogical LLMs with Human-Computer Interface

Integrating Human-Computer Interface (HCI) with LLMs in education seeks to create more interactive and personalized learning experiences. [117] discusses how LLMs can enhance pedagogical tools by providing adaptive learning environments tailored to individual student needs, emphasizing the importance of aligning educational content with LLM capabilities to improve learning outcomes. Additionally, Dimbisoa *et al.* [118] focuses on developing platform-independent metamodels for UI components, ensuring reusability and adaptability across various educational platforms. This combination of HCI design principles and LLM capabilities has the potential to create sophisticated educational tools that are both user-centric and pedagogically effective.

7. Personalized Learning

7.1. Syllabus and Plan Based Personalized Learning

[61] explores LLMs' capabilities in creating high school math teaching plans, excelling in setting learning objectives and organizing instructional content, though needing improvements in cultural context and interdisciplinary assessments. Similarly, [119] introduces a Personalized Learning System (PLS) that leverages LLMs and web technologies to generate tailored educational content such as summaries, quizzes, and answer keys, adapting to individual learning styles and providing real-time feedback. Despite occasional inaccuracies in generated content, these systems demonstrate significant promise in enhancing personalized education, optimizing exam preparation, and fostering individualized academic success.

7.2. Personnalized Learning through Knowledge Graphs

[120] highlights the importance of providing clear and accurate explanations for personalized learning recommendations. Using Knowledge Graphs (KGs) to provide factual context for LLM prompts, this approach reduces errors and increases the relevance of explanations, enhancing student engagement and understanding. Similarly, [121] explores personalized learning through KGs and LLMs, emphasizing components like LLM-generated flashcards and Dynamic Competence Maps (DCMs) to tailor content to individual learners, creating a cost-effective and adaptive learning experience. [122] discusses integrating personalized learning within intelligent tutoring systems, using LLMs to assess students' cognitive and affective states and learning styles, delivering customized instructional strategies to enhance engagement and effectiveness. Lastly, [123] focuses on integrating Generative AI, including LLMs and diffusion models, in educational platforms to overcome language barriers and create tailored educational content, addressing ethical concerns and biases to ensure fairness and accuracy in personalized learning.

7.3. Personalized Learning through Retrieval-Augmented Generation (RAG)

[124] discusses personalized learning within the context of improving response generation in language models. The ERAGent framework introduces a Personalized LLM Reader module that tailors responses based on user profiles, which are dynamically updated by the Experiential Learner module, learning from historical interactions to refine the AI's understanding of individual preferences. This approach ensures responses are accurate and aligned with user needs, enhancing overall user experience and model efficiency. Similarly, [125] explores advancements in personalized learning through the automated creation of multiple-choice questions (MCQs). MCQGen leverages a LLM combined with retrieval-augmented generation and advanced prompt engineering techniques to generate relevant and challenging MCQs tailored to individual learning paces and comprehension levels, providing a customized learning experience.

8. Personalized Learning by Video

Personalized Learning with Adaptive Video: Adaptive video learning leverages AI to create interactive and personalized learning experiences for children, enhancing engagement and effectiveness through tailored feedback and tasks [126]. Optimal learning can be achieved at playback speeds of 1.25× and 1.5×, as supported by cognitive load theory[24], which varies by student ability and major [127]. Recommendation methods using collaborative filtering algorithms improve the accuracy and efficiency of video recommendations by analyzing learner preferences,

[24]Cognitive load theory suggests that varying playback speeds can optimize the processing of information by balancing the cognitive demands placed on learners.

making learning more tailored [128]. Additionally, FedABR, a personalized federated learning[25] approach for adaptive video streaming, enhances personalized learning by training a global model that adapts to various network conditions without compromising user privacy, maximizing user Quality of Experience (QoE) through customized bitrate selection [129]. Finally, adaptive video technology customizes content delivery based on individual needs, significantly enhancing engagement and learning outcomes [130].

Personalized Video and Recommendation Systems: A meta-learning framework enhances Quality of Experience (QoE) in personalized 360-degree video streaming by using a metabased LSTM for accurate viewport prediction and meta-based reinforcement learning for bitrate selection, quickly adapting to user preferences [131]. For personalized video recommendations, a system using the DBSCAN[26] clustering algorithm constructs user profiles from attributes and behavior data, effectively clustering users to recommend relevant educational videos, thus improving the accuracy and relevance of recommendations [132]. Educational video games designed with adaptive learning scenarios show how personalized puzzle games enhance game-based learning by adjusting content and difficulty based on student performance, supporting dynamic personalization and adaptation [133]. Additionally, personalized learning for adaptive video generation is advanced through a memory-augmented GAN, which creates high-quality talking face videos with individualized head poses, enhancing realism with attention mechanisms and memory networks for identity feature retrieval [134].

9. Prompt Optimization Applied to Education

The literature highlights the integration of LLMs like ChatGPT in various educational settings, emphasizing prompt engineering for personalized learning. [135] introduces CourseGPT-zh, which constructs high-quality question-answer pairs by mining textbook knowledge and optimizing prompts through LLM-as-Judge[27], enhancing response quality and alignment with user needs. [136] examines generative AI tools in computing education, revealing both benefits and concerns from interviews with students and instructors. [137] and [138] explore LLMs in academic and medical education, respectively, focusing on pedagogical alignment, ethical use, academic integrity, and data privacy. They highlight the need for effective prompt design to maximize educational benefits. [160] discusses prompt engineering in generating educational questions, emphasizing AI-teacher collaboration and the importance of few-shot learning. [139] investigates LLMs in

[25]Federated learning allows models to be trained across multiple decentralized devices without sharing data.

[26]"Density-Based Spatial Clustering of Applications with Noise" is a clustering algorithm that groups users based on similarities in their behavior, allowing for more personalized and accurate recommendations by identifying patterns in the data.

[27]The "LLM-as-Judge" mentioned in the original article refers to the use of a large language model as an evaluator to assess the quality of responses generated by other models, focusing on alignment with human preferences and factual accuracy.

computer programming education, emphasizing systematic categorization of prompts and the continuous refinement to enhance learning outcomes.

[140] underscores the integration of prompt engineering into medical education, highlighting trends, challenges, and the potential for personalized, interactive learning experiences.

10. Challenges, Trends & Future Directions

This section focuses on the critical challenges, emerging trends, and possible future directions for the integration of Large Language Models (LLMs) in educational contexts. It examines the technical and pedagogical barriers to their widespread adoption, evaluates current strategies and innovations aimed at overcoming these barriers, and outlines areas where further research is essential to unlock the full potential of LLMs in transforming personalized learning experiences.

LLM Models Despite its widespread adoption in education, it is hampered by major technical challenges, such as high computational resource demands during inference, which limits their deployment in resource-constrained educational environments. Compression techniques such as quantization, pruning, and knowledge distillation are being explored to enhance the efficiency of LLM inference without compromising performance . Additionally, personalized fine-tuning using natural instruction-based and curriculum-based approaches shows promising potential for improving the relevance and effectiveness of LLMs in specific educational contexts.

Future research should focus on resource optimization, improving prompt robustness, and long-term evaluation of the impact of LLMs on learning outcomes.

Knowledge Editing for Large Language Models One promising trend is Knowledge-based Model Editing (KME), which focuses on updating pre-trained LLMs with new information, ensuring they remain accurate and relevant for educational purposes. This technique allows LLMs to stay current on various topics, improving their performance in downstream educational applications. However, the widespread adoption of LLMs in education is challenged by the need for continuous updates and the computational resources required for such tasks. The practical implementation of KME and other techniques is complex and resource-intensive, despite their potential for maintaining LLM accuracy.

However, the balance between locality (specific knowledge updates) and generality (over-all model performance) remains a significant challenge. Additionally, ensuring that updates are robust against misinformation and bias is crucial, especially in educational settings. Gaps in the current literature include the need for more efficient update mechanisms and strategies to mitigate biases and misinformation. Future research should focus on optimizing these update processes and exploring long-term impacts on learning outcomes.

Content Generation with LLM The analysis of recent studies on LLMs in educational content generation identifies trends in video generation, quiz creation, plan development, and feedback. A consistent challenge is maintaining high-quality outputs, such as temporal consistency in videos, accuracy in quizzes, and alignment of teaching plans with pedagogical goals. Advanced methods like reinforcement learning and knowledge graph integration are proposed to enhance LLM effectiveness. The literature also emphasizes adaptive learning, where personalized content generation improves engagement and outcomes.

Despite these advancements, with some providing robust frameworks, while others show issues in scalability and reliability, particularly in feedback generation. Reliance on specific datasets and high computational demands suggest the need for sustainable solutions. The literature also lacks focus on the long-term impact of LLMs on educational outcomes, indicating future research should prioritize longitudinal studies and integration with existing systems to validate efficacy. Optimizing LLMs for diverse learning environments and ensuring ethical deployment remain crucial areas for future investigation.

Datasets for Education Overview The synthesis of research on datasets for educational applications of Large Language Models (LLMs) highlights trends in developing datasets that capture the complexities of human instruction. Studies like those on the KIWI dataset focus on refining LLMs for knowledge-intensive tasks, showing current models' struggles with integrating new information and coherence. There is also a growing use of specialized datasets for question-solving and teacher-assisting tasks, essential for advancing LLM capabilities in education. However, these datasets often remain narrow, limiting broader applicability in diverse educational scenarios.

Some datasets, such as those for question generation and error correction, enhance LLM performance, while others face limitations in scalability and real-world applicability. Many studies lack evaluations across diverse contexts, highlighting the need for more inclusive datasets. Unresolved questions about the long-term effectiveness and ethical implications of LLM-based tools underscore the need for future research to focus on diverse datasets and longitudinal studies to validate LLM impacts in education.

Pedagogical alignment of LLM The review of studies on the pedagogical alignment highlights key trends, particularly in foundational skills like mathematical reasoning, writing, programming, and KBQA. LLMs show promise in basic tasks but struggle with complex mathematical reasoning and writing error correction. Challenges include the integration of multimodal data and the development of robust evaluation metrics for educational relevance.

LLMs' potential in programming is more evident in refining existing code than in generating complex algorithms.

Critical evaluations reveal significant gaps. Some studies propose innovative approaches, like the Mixture-of-Experts (MoE) framework, addressing LLM shortcomings, but issues of scalability and real-world applicability persist, particularly in handling implicit reasoning. The problem of LLM "hallucinations" remains critical, requiring further research for more reliable educational tools.

Unresolved questions, such as the long-term effectiveness of LLMs in diverse learning environments and ethical considerations, underscore the need for more comprehensive studies.

Personalized Learning The synthesis of research on personalized learning through LLMs highlights significant trends, especially in integrating LLMs with personalized learning systems, knowledge graphs, and retrieval-augmented generation (RAG) techniques. Studies like [61] and [119] showcase the potential of LLMs in creating customized educational content, such as teaching plans and quizzes, tailored to individual learning styles. The use of knowledge graphs, explored by [120] and [121], further enhances personalized learning by providing factual context and dynamic maps that improve engagement and understanding.

However, critical evaluation reveals both strengths and weaknesses. While integrating LLMs with knowledge graphs and RAG frameworks offers a tailored learning experience challenges like inaccuracies in content generation and biases persist. The literature also highlights gaps in addressing ethical implications and potential biases, as discussed by [123]. Future research should focus on improving LLM precision, cultural adaptability, and developing robust frameworks to assess the effectiveness of these personalized learning systems while integrating ethical safeguards.

Personalized Learning by Video The examination of personalized learning through adaptive video technology highlight how AI and collaborative filtering algorithms optimize content delivery based on learner preferences, while FedABR adapts models to network conditions, preserving privacy.

However, the effectiveness of these technologies varies across learning contexts, and predicting user preferences remains challenging. Techniques like memory-augmented GANs require further refinement to address scalability and ethical issues [134]. Future research should focus on improving accessibility and effectiveness across diverse educational environments.

Prompt Optimization applied to Education Integrating prompt optimization techniques into educational applications of Large Language Models (LLMs) shows improvement for personalized learning. Studies like [135] demonstrate how CourseGPT-zh utilizes optimized prompts to generate question-answer pairs by leveraging textbook knowledge, thereby improving the relevance of responses. [160] emphasizes the importance of AI-teacher collaboration in question generation and the role of few-shot learning. Research such as [137] and [140] highlights the need for academic integrity and data privacy in prompt engineering.

However, gaps remain. Although promising, the effectiveness of prompts varies by subject and educational context. Studies like [136] and [139] raise concerns about the reliability of AIgenerated content, particularly in programming education, requiring continuous refinement. Ethical challenges, such as data privacy and potential biases, also demand further research to develop robust frameworks for applying LLMs in personalized learning.

Table 1 provides a synthesis of the main challenges associated with the use of large language models (LLMs) in an educational context, along with proposed solutions and future potential to address these limitations.

Table 1. Challenges and solutions for LLMs in an educational context.

Challenge	Description	Future Potential	Proposed Approach
Multimodal LLMs	Difficulty in integrating visual and auditory data for a complete educational experience.	Improve student engagement through richer multimodal interactions, including using multimodal integration techniques such as Pedagogical Chain-of-Thought (PedCoT) to enhance understanding and educational interaction [115].	Develop techniques to integrate textual, visual, and auditory data into LLMs, as suggested by research on PedCoT and improving coherence across different modalities [91].
Multilingual LLMs	Issues of interpretation and bias across multiple languages.	Increase global accessibility to personalized educational resources by reducing linguistic biases through the use of diversified multilingual corpora [105].	Train LLMs on multilingual corpora and use Knowledge Graphs to contextualize and improve translation accuracy, as proposed to enhance the quality of responses [121].
Prompt Optimization	Prompts may lack robustness and predictability, limiting the effectiveness of responses.	Enhance the effectiveness of LLMs to personalize educational responses through prompt optimization techniques that ensure more relevant answers [135].	Use techniques like LLM-as-Judge to optimize prompts based on pedagogical objectives, as shown in studies that utilize this technique to improve response quality [160].
Cultural Context and Adaptability	Limitation of LLMs in contextualizing content for diverse cultural environments.	Create learning experiences adapted to each cultural context by reducing errors related to cultural differences [121].	Develop approaches based on Knowledge Graphs to contextualize responses and provide appropriate explanations, allowing LLMs to adjust responses to cultural specifics [120].
Ethics and Bias	Risk of bias in LLM responses and lack of transparency, affecting user trust.	Ensure ethical adoption of LLMs in education by using methods aimed at reducing biases and ensuring accuracy [42].	Introduce Knowledge Editing techniques to keep LLMs up-to-date with correct and relevant information, and limit biases, as discussed in approaches for continuous model updates [48].

11. Conclusions

Large language models are positioned to markedly transform educational experiences through the facilitation of personalized learning. This review has explored various aspects of LLMs in the educational context, including Knowledge Editing Techniques (KME), which ensure that models remain up-to-date with evolving information, and the crucial importance of educational datasets in the development of these models. Furthermore, we have examined the core capabilities of LLMs, including mathematics, writing, programming, and reasoning. In doing so, we have identified both the strengths and limitations of these models, with a particular focus on their capacity for transparent reasoning.

The prospective impact of LLMs on the future of education is significant. Promising system architectures, such as the unified LLM approach and the Mixture-of-

Experts (MoE) framework, offer novel avenues for personalized learning and on-demand support. Nevertheless, in order to fully actualize the advantages of LLMs in the field of education, it is essential to confront and overcome significant obstacles, including the assurance of factual precision, the diminution of biases, and the encouragement of critical thinking abilities in conjunction with LLM-facilitated learning.

The capacity of LLMs to generate content, including videos, quizzes, and lesson plans, presents a novel opportunity to enhance educational content and delivery. Furthermore, personalized learning can be enhanced through methodologies such as syllabus-based learning, knowledge graphs, and retrieval-augmented generation, all of which are tailored to meet the distinctive requirements of individual students.

In contemplating the future, it is essential to envisage a scenario in which LLMs function as invaluable collaborators in the field of education, enhancing the work of human educators and enabling students to reach their full potential. By encouraging the responsible development and implementation of LLMs, and addressing the ethical considerations that arise, they can play a transformative role in creating a more engaging, effective, and accessible educational environment for all.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., *et al*. (2023) ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, **103**, Article ID: 102274. https://doi.org/10.1016/j.lindif.2023.102274

[2] Eloundou, T., Manning, S., Mishkin, P. and Rock, D. (2023) GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv: 2303.10130.

[3] Yuzbashyan, N., Banar, N., Markov, I. and Daelemans, W. (2023) An Exploration of Zero-Shot Natural Language Inference-Based Hate Speech Detection. In: Chakravarthi, B.R., Bharathi, B., Griffith, J., Bali, K. and Buitelaar, P., Eds., *Proceedings of the Third Workshop on Language Technology for Equality*, *Diversity and Inclusion*, INCOMA Ltd., 1-9. https://aclanthology.org/2023.ltedi-1.1

[4] Goldzycher, J., Preisig, M., Amrhein, C. and Schneider, G. (2023) Evaluating the Effectiveness of Natural Language Inference for Hate Speech Detection in Languages with Limited Labeled Data. *The* 7 *th Workshop on Online Abuse and Harms* (*WOAH*), Toronto, 13 July 2023, 187-201. https://doi.org/10.18653/v1/2023.woah-1.19

[5] Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A. (2024) A Comprehensive Overview of Large Language Models. arXiv: 2307.06435.

[6] Abzianidze, L., Zwarts, J. and Winter, Y. (2023) SpaceNLI: Evaluating the Consistency of Predicting Inferences in Space. *Proceedings of the* 4 *th Natural Logic Meets Machine Learning Workshop*, Nancy, June 2023, 12-24. https://aclanthology.org/2023.naloma-1.2

[7] Han, X., Zeng, G., Zhao, W., Liu, Z., Zhang, Z., Zhou, J., *et al*. (2022) BMInF: An Efficient Toolkit for Big Model Inference and Tuning. *Proceedings of the* 60*th Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations*, 22-27 May 2022, Dublin, 224-230. https://doi.org/10.18653/v1/2022.acl-demo.22

[8] Liu, S., Wen, T., Pattamatta, A.S.L.S. and Srolovitz, D.J. (2024) A Prompt-Engineered Large Language Model, Deep Learning Workflow for Materials Classification. *Materials Today*. https://doi.org/10.1016/j.mattod.2024.08.028

[9] Winata, G., Xie, L., Radhakrishnan, K., Gao, Y. and Preotiuc-Pietro, D. (2023) Efficient Zero-Shot Cross-Lingual Inference via Retrieval. *Proceedings of the* 13*th International Joint Conference on Natural Language Processing and the* 3*rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (*Volume* 2: *Short Papers*), Nusa Dua, November 2023, 93-104. https://doi.org/10.18653/v1/2023.ijcnlp-short.11

[10] Conceição, S.I.R., F. Sousa, D., Silvestre, P. and Couto, F.M. (2023) LasigeBioTM at SemEval-2023 Task 7: Improving Natural Language Inference Baseline Systems with Domain Ontologies. *Proceedings of the the* 17*th International Workshop on Semantic Evaluation* (*SemEval*-2023), Toronto, 13-14 July 2023, 10-15. https://doi.org/10.18653/v1/2023.semeval-1.2

[11] Jin, R.R., Du, J.C., Huang, W.W., Liu, W., Luan, J., Wang, B. and Xiong, D.Y. (2024) A Comprehensive Evaluation of Quantization Strategies for Large Language Models. arXiv: 2402.16775.

[12] Chavan, A., Magazine, R., Kushwaha, S., Debbah, M. and Gupta, D. (2024) Faster and Lighter Llms: A Survey on Current Challenges and Way Forward. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, Jeju, 3-9 August 2024, 7980-7988. https://doi.org/10.24963/ijcai.2024/883

[13] Li, L., Jiang, B., Wang, P., Ren, K., Yan, H. and Qiu, X. (2023) Watermarking LLMs with Weight Quantization. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 3368-3378. https://doi.org/10.18653/v1/2023.findings-emnlp.220

[14] Gong, Z., Liu, J., Wang, Q., Yang, Y., Wang, J., Wu, W., *et al*. (2023) PreQuant: A Task-Agnostic Quantization Approach for Pre-Trained Language Models. *Findings of the Association for Computational Linguistics*: *ACL* 2023, Toronto, 9-14 July 2023, 8065-8079. https://doi.org/10.18653/v1/2023.findings-acl.511

[15] Smith, A., Hachen, S., Schleifer, R., Bhugra, D., Buadze, A. and Liebrenz, M. (2023) Old Dog, New Tricks? Exploring the Potential Functionalities of ChatGPT in Supporting Educational Methods in Social Psychiatry. *International Journal of Social Psychiatry*, **69**, 1882-1889. https://doi.org/10.1177/00207640231178451

[16] Kolagar, Z. and Zarcone, A. (2024) HumSum: A Personalized Lecture Summarization Tool for Humanities Students Using LLMs. *Proceedings of the 1st Workshop on Personalization of Generative AI Systems* (*PERSONALIZE* 2024), St. Julians, March 2024, 36-70. https://aclanthology.org/2024.personalize-1.4

[17] Jawahar, G., Mukherjee, S., Liu, X., Kim, Y.J., Abdul-Mageed, M., Lakshmanan, V.S., L., *et al*. (2023) AutoMoE: Heterogeneous Mixture-Of-Experts with Adaptive Computation for Efficient Neural Machine Translation. *Findings of the Association for Computational Linguistics*: *ACL* 2023, Toronto, 9-14 July 2023, 9116-9132. https://doi.org/10.18653/v1/2023.findings-acl.580

[18] Yen, A. and Hsu, W. (2023) Three Questions Concerning the Use of Large Language Models to Facilitate Mathematics Learning. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 3055-3069.

https://doi.org/10.18653/v1/2023.findings-emnlp.201

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. arXiv: 1706.03762.

[20] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., *et al.* (2020) Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, **33**, 1877-1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[21] Wang, P., Zhang, N., Tian, B., Xi, Z., Yao, Y., Xu, Z., *et al.* (2024) EasyEdit: An Easy-To-Use Knowledge Editing Framework for Large Language Models. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 3: *System Demonstrations*), Bangkok, 11-16 August 2024, 82-93. https://doi.org/10.18653/v1/2024.acl-demos.9

[22] Zhong, Z., Wu, Z., Manning, C., Potts, C. and Chen, D. (2023) MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 15686-15702. https://doi.org/10.18653/v1/2023.emnlp-main.971

[23] Chan, C., Jiayang, C., Wang, W.Q., Jiang, Y.X., Fang, T., Liu, X. and Song, Y.Q. (2024) Exploring the Potential of ChatGPT on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations. *Findings of the Association for Computational Linguistics*: *EACL* 2024, St. Julian's, 17-22 March 2024, 684-721. https://aclanthology.org/2024.findings-eacl.47

[24] Das, S.S.S., Zhang, H., Shi, P., Yin, W. and Zhang, R. (2023) Unified Low-Resource Sequence Labeling by Sample-Aware Dynamic Sparse Finetuning. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 6998-7010. https://doi.org/10.18653/v1/2023.emnlp-main.433

[25] Bowen, C., Sætre, R. and Miyao, Y. (2024) A Comprehensive Evaluation of Inductive Reasoning Capabilities and Problem Solving in Large Language Models. *Findings of the Association for Computational Linguistics*: *EACL* 2024, St. Julian's, 17-22 March 2024, 323-339. https://aclanthology.org/2024.findings-eacl.22

[26] Li, J., Su, Q., Yang, Y., Jiang, Y., Wang, C. and Xu, H. (2023) Adaptive Gating in Mixture-Of-Experts Based Language Models. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 3577-3587. https://doi.org/10.18653/v1/2023.emnlp-main.217

[27] Moon, H., Lee, J., Eo, S., Park, C., Seo, J. and Lim, H. (2024) Generative Interpretation: Toward Human-Like Evaluation for Educational Question-Answer Pair Generation. *Findings of the Association for Computational Linguistics*: *EACL* 2024, St. Julian's, 17-22 March 2024, 2185-2196. https://aclanthology.org/2024.findings-eacl.145

[28] Campos, D., Marques, A., Kurtz, M. and Xiang Zhai, C. (2023) oBERTa: Improving Sparse Transfer Learning via Improved Initialization, Distillation, and Pruning Regimes. *Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing* (*SustaiNLP*), Toronto, 13 July 2023, 39-58. https://doi.org/10.18653/v1/2023.sustainlp-1.3

[29] Men, X., Xu, M.Y., Zhang, Q.Y., Wang, B.N., Lin, H.Y., Lu, Y.J., Han, X.P. and Chen, W.P. (2024) ShortGPT: Layers in Large Language Models Are More Redundant than You Expect. arXiv: 2403.03853.

[30] Azeemi, A., Qazi, I. and Raza, A. (2023) Data Pruning for Efficient Model Pruning in

Neural Machine Translation. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 1127-1149. https://doi.org/10.18653/v1/2023.findings-emnlp.18

[31] Lewis, A. and White, M. (2023) Mitigating Harms of LLMs via Knowledge Distillation for a Virtual Museum Tour Guide. *Proceedings of the 1st Workshop on Taming Large Language Models*: *Controllability in the Era of Interactive Assistants*, Prague, 12 September 2023, 31-45. https://aclanthology.org/2023.tllm-1.4

[32] West, P., Bras, R., Sorensen, T., Lin, B., Jiang, L., Lu, X., *et al*. (2023) NovaCOMET: Open Commonsense Foundation Models with Symbolic Knowledge Distillation. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 1127-1148. https://doi.org/10.18653/v1/2023.findings-emnlp.80

[33] Hubert, R., Sokolov, A. and Riezler, S. (2023) Improving End-to-End Speech Translation by Imitation-Based Knowledge Distillation with Synthetic Transcripts. *Proceedings of the* 20*th International Conference on Spoken Language Translation* (*IWSLT* 2023), Toronto, July 2023, 89-101. https://doi.org/10.18653/v1/2023.iwslt-1.4

[34] Faysse, M., Viaud, G., Hudelot, C. and Colombo, P. (2023) Revisiting Instruction Fine-Tuned Model Evaluation to Guide Industrial Applications. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 9033-9048. https://doi.org/10.18653/v1/2023.emnlp-main.559

[35] Zhou, W., Tahmasebi, N. and Dubossarsky, H. (2023) The Finer They Get: Combining Fine-Tuned Models for Better Semantic Change Detection. *Proceedings of the* 24*th Nordic Conference on Computational Linguistics* (*NoDaLiDa*), Tórshavn, 22-24 May 2023, 518-528. https://aclanthology.org/2023.nodalida-1.52

[36] Qi, Z., Tan, X., Shi, S., Qu, C., Xu, Y. and Qi, Y. (2023) PILLOW: Enhancing Efficient Instruction Fine-Tuning via Prompt Matching. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*: *Industry Track*, Singapore, 6-10 December 2023, 471-482. https://doi.org/10.18653/v1/2023.emnlp-industry.45

[37] Anuranjana, K. (2023) DiscoFlan: Instruction Fine-Tuning and Refined Text Generation for Discourse Relation Label Classification. *Proceedings of the* 3*rd Shared Task on Discourse Relation Parsing and Treebanking* (*DISRPT* 2023), Toronto, 14 July 2023, 22-28. https://doi.org/10.18653/v1/2023.disrpt-1.2

[38] Arriola, J.M., Iruskieta, M., Arrieta, E. and Alkorta, J. (2023) Towards Automatic Essay Scoring of Basque Language Texts from a Rule-Based Approach Based on Curriculum-Aware Systems. *Proceedings of the NoDaLiDa* 2023 *Workshop on Constraint Grammar—Methods*, *Tools and Applications*, Tórshavn, 22 May 2023, 20-28. https://aclanthology.org/2023.nodalida-cgmta.4.

[39] Ranaldi, L., Pucci, G. and Massimo Zanzotto, F. (2023) Modeling Easiness for Training Transformers with Curriculum Learning. *Proceedings of the* 14*th International Conference on Recent Advances in Natural Language Processing*, Varna, 4-6 September 2023, 937-948. https://doi.org/10.26615/978-954-452-092-2_101

[40] Vakil, N. and Amiri, H. (2023) Complexity-Guided Curriculum Learning for Text Graphs. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 2610-2626. https://doi.org/10.18653/v1/2023.findings-emnlp.172

[41] Zhou, J., Zeng, Z., Gong, H. and Bhat, S. (2023) Non-Compositional Expression Generation Based on Curriculum Learning and Continual Learning. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December

2023, 4320-4335. https://doi.org/10.18653/v1/2023.findings-emnlp.286

[42] Zhang, X., Ju, T.J., Liang, H.J., Fu, Y. and Zhang, Q. (2024) LLMs Instruct LLMs: An Extraction and Editing Method. arXiv: 2403.15736v1.

[43] Heitmann, M. (2020) More than a Feeling: Benchmarks for Sentiment Analysis Accuracy. Elsevier.

[44] Vansh, R., Rank, D., Dasgupta, S. and Chakraborty, T. (2023) Accuracy Is Not Enough: Evaluating Personalization in Summarizers. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 2582-2595. https://doi.org/10.18653/v1/2023.findings-emnlp.169

[45] Schmidtova, P. (2023) Semantic Accuracy in Natural Language Generation: A Thesis Proposal. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics* (*Volume* 4: *Student Research Workshop*), Toronto, 9-14 July 2023, 352-361. https://doi.org/10.18653/v1/2023.acl-srw.48

[46] Lee, K., Han, W., Hwang, S., Lee, H., Park, J. and Lee, S. (2022) Plug-and-Play Adaptation for Continuously-Updated QA. *Findings of the Association for Computational Linguistics*: *ACL* 2022, Dublin, 22-27 May 2022, 438-447. https://doi.org/10.18653/v1/2022.findings-acl.37

[47] Lan, W., Qiu, S., He, H. and Xu, W. (2017) A Continuously Growing Dataset of Sentential Paraphrases. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 9-11 September 2017, 1224-1234. https://doi.org/10.18653/v1/d17-1126

[48] Wang, S., Zhu, Y.C., Liu, H.C., Zheng, Z.Y., Chen, C. and Li, J.D. (2023) Knowledge Editing for Large Language Models: A Survey. arXiv: 2310.16218.

[49] Bittermann, A. and Rieger, J. (2022) Finding Scientific Topics in Continuously Growing Text Corpora. *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, 12-17 October 2022, 7-18. https://aclanthology.org/2022.sdp-1.2

[50] Zhang, N., Tian, B., Cheng, S., Liang, X., Hu, Y., Xue, K., *et al*. (2024) Instructedit: Instruction-Based Knowledge Editing for Large Language Models. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, Jeju, 3-9 August 2024, 6633-6641. https://doi.org/10.24963/ijcai.2024/733

[51] Onoe, Y., Zhang, M., Padmanabhan, S., Durrett, G. and Choi, E. (2023) Can Lms Learn New Entities from Descriptions? Challenges in Propagating Injected Knowledge. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Toronto, 9-14 July 2023, 5469-5485. https://doi.org/10.18653/v1/2023.acl-long.300

[52] Pandya, H.A. and Bhatt, B.S. (2021) Question Answering Survey: Directions, Challenges, Datasets, Evaluation Matrices. arXiv: 2112.03572.

[53] Li, C.X., Huang, D., Lu, Z.Y., Xiao, Y., Pei, Q.Q. and Bai, L. (2024) A Survey on Long Video Generation: Challenges, Methods, and Prospects. arXiv: 2403.16407.

[54] Yang, D.S., Hu, L.H., Tian, Y., Li, Z.H., Kelly, C., Yang, B., Yang, C. and Zou, Y. (2024) WorldGPT: A Sora-Inspired Video AI Agent as Rich World Models from Text and Image Inputs. arXiv: 2403.07944.

[55] Zhou, P., Wang, L., Liu, Z., Hao, Y.B., Hui, P., Tarkoma, S. and Kangasharju, J. (2024) A Survey on Generative AI and LLM for Video Generation, Understanding, and Streaming. arXiv: 2404.16038.

[56] Renella, N. and Eger, M. (2023) Towards Automated Video Game Commentary

Using Generative AI. *CEUR Workshop Proceedings*: *AIIDE Workshop on Experimental Artificial Intelligence in Games*, 8 October 2023, Utah, 341-350. https://ceur-ws.org/Vol-3626/paper7.pdf

[57] Bhagwatkar, R., Bachu, S., Fitter, K., Kulkarni, A. and Chiddarwar, S. (2020) A Review of Video Generation Approaches. 2020 *International Conference on Power*, *Instrumentation*, *Control and Computing* (*PICC*), Thrissur, 17-19 December 2020, 1-5. https://doi.org/10.1109/picc51425.2020.9362485

[58] Sreekanth, D. and Dehbozorgi, N. (2023) Enhancing Engineering Education through LLM-Driven Adaptive Quiz Generation. https://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1399&context=cday

[59] Lamsiyah, S., El Mahdaouy, A., Nourbakhsh, A. and Schommer, C. (2024) Fine-Tuning a Large Language Model with Reinforcement Learning for Educational Question Generation. In: Olney, A.M., Chounta, IA., Liu, Z., Santos, O.C. and Bittencourt, I.I., Eds., *Artificial Intelligence in Education*, Springer Nature Switzerland, 424-438. https://doi.org/10.1007/978-3-031-64302-6_30

[60] Agrawal, G., Pal, K., Deng, Y., Liu, H. and Chen, Y. (2024) CyberQ: Generating Questions and Answers for Cybersecurity Education Using Knowledge Graph-Augmented LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 23164-23172. https://doi.org/10.1609/aaai.v38i21.30362

[61] Hu, B., Zheng, L., Zhu, J., Ding, L., Wang, Y. and Gu, X. (2024) Teaching Plan Generation and Evaluation with GPT-4: Unleashing the Potential of LLM in Instructional Design. *IEEE Transactions on Learning Technologies*, **17**, 1471-1485. https://doi.org/10.1109/tlt.2024.3384765

[62] Annie Micheal, A., Prasanth, A., Aswin, T.S., *et al*. (2024) Advancing Educational Accessibility: The LangChain LLM Chatbot's Impact on Multimedia Syllabus-Based Learning. https://doi.org/10.21203/rs.3.rs-4399670/v1

[63] Goslen, A., Kim, Y.J., Rowe, J. and Lester, J. (2024) LLM-Based Student Plan Generation for Adaptive Scaffolding in Game-Based Learning Environments. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00421-1

[64] Stahl, M., Biermann, L., Nehring, A. and Wachsmuth, H. (2024) Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. arXiv: 2404.15845. https://arxiv.org/abs/2404.15845

[65] Stamper, J., Xiao, R. and Hou, X. (2024) Enhancing LLM-Based Feedback: Insights from Intelligent Tutoring Systems and the Learning Sciences. *Communications in Computer and Information Science*, **2150**, 32-43. https://doi.org/10.1007/978-3-031-64315-6_3

[66] Nie, A., Cheng, C.A., Kolobov, A. and Swaminathan, A. (2024) The Importance of Directional Feedback for LLM-Based Optimizers. arXiv: 2405.16434. https://arxiv.org/abs/2405.16434

[67] Gabbay, H. and Cohen, A. (2024) Combining LLM-Generated and Test-Based Feedback in a MOOC for Programming. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, Atlanta, 18-20 July 2024, 177-187. https://doi.org/10.1145/3657604.3662040

[68] Tanwar, H., Shrivastva, K., Singh, R. and Kumar, D. (2024) OpineBot: Class Feedback Reimagined Using a Conversational LLM. arXiv: 2401.15589. https://arxiv.org/abs/2401.15589

[69] Estévez-Ayres, I., Callejo, P., Hombrados-Herrera, M.Á., Alario-Hoyos, C. and

Delgado Kloos, C. (2024) Evaluation of LLM Tools for Feedback Generation in a Course on Concurrent Programming. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00406-0

[70] Liu, Y., Cao, J.H., Liu, C.Y., Ding, K. and Jin, L.W. (2024) Datasets for Large Language Models: A Comprehensive Survey. https://doi.org/10.21203/rs.3.rs-3996137/v1

[71] Xu, F.Y., Lo, K.L., Soldaini, L.C., Kuehl, B., Choi, E. and Wadden, D. (2024) KIWI: A Dataset of Knowledge-Intensive Writing Instructions for Answering Research Questions. arXiv: 2403.03866.

[72] Wang, S., Xu, T.L., Li, H., Zhang, C.L., Liang, J., Tang, J.L., Yu, P.S. and Wen, Q.S. (2024) Large Language Models for Education: A Survey and Outlook. arXiv: 2403.18105.

[73] Lu, R., Tang, Z., Hu, G., Liu, D. and Li, J. (2023) NetEase.AI at Semeval-2023 Task 2: Enhancing Complex Named Entities Recognition in Noisy Scenarios via Text Error Correction and External Knowledge. *Proceedings of the* 17*th International Workshop on Semantic Evaluation* (*SemEval*-2023), Toronto, 13-14 July 2023, 987-904. https://doi.org/10.18653/v1/2023.semeval-1.124

[74] Li, Q., Yang, X.Y., *et al*. (2024) From Beginner to Expert: Modeling Medical Knowledge into General LLMs. arXiv: 2312.01040.

[75] Chen, W., Verga, P., de Jong, M., Wieting, J. and Cohen, W.W. (2023) Augmenting Pre-Trained Language Models with QA-Memory for Open-Domain Question Answering. *Proceedings of the* 17*th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, 2-6 May 2023, 1597-1610. https://doi.org/10.18653/v1/2023.eacl-main.117

[76] Talmor, A., Herzig, J., Lourie, N. and Berant, J. (2018) CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. arXiv: 1811.00937.

[77] Marini, T. and Brant-Ribeiro, T. (2024) Comparative Analysis of Intentional Grammatical Error Correction Techniques on Twitter/X. *Proceedings of the* 16*th International Conference on Computational Processing of Portuguese*, Santiago de Compostela, 14-15 March 2024, 527-531. https://aclanthology.org/2024.propor-1.55

[78] Luhtaru, A., Korotkova, E. and Fishel, M. (2024) No Error Left Behind: Multilingual Grammatical Error Correction with Pretrained Translation Models. *Proceedings of the* 18*th Conference of the European Chapter of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), St. Julian's, 17-22 March 2024, 1209-1222. https://aclanthology.org/2024.eacl-long.73

[79] Ponce, A.D.H., Jadie, J.S.A., Espiritu, P.E.A. and Cheng, C. (2023) Balarila: Deep Learning for Semantic Grammar Error Correction in Low-Resource Settings. *Proceedings of the First Workshop in South East Asian Language Processing*, Nusa Dua, November 2023, 21-29. https://doi.org/10.18653/v1/2023.sealp-1.3

[80] Veerubhotla, A.S., Poddar, L., Yin, J., Szarvas, G. and Eswaran, S. (2023) Few Shot Rationale Generation Using Self-Training with Dual Teachers. *Findings of the Association for Computational Linguistics*: *ACL* 2023, Toronto, 9-14 July 2023, 4825-4838. https://doi.org/10.18653/v1/2023.findings-acl.297

[81] Ho, N., Schmid, L. and Yun, S. (2023) Large Language Models Are Reasoning Teachers. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Toronto, 9-14 July 2023, 14852–14882. https://doi.org/10.18653/v1/2023.acl-long.830

[82] Warstadt, A., Mueller, A., *et al*. (2023) Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning. Association for Computational Linguistics. https://aclanthology.org/2023.conll-babylm

[83] Huang, P.W. (2022) Domain Specific Augmentations as Low Cost Teachers for Large Students. *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, November 2022, 84-90. https://aclanthology.org/2022.wiesp-1.10

[84] Chen, Q.L., Liu, T. and Guo, J. (2024) LLM-Da: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition. arXiv: 2402.14568v1.

[85] Zheng, H., Zhong, Q., Ding, L., Tian, Z., Niu, X., Wang, C., *et al*. (2023) Self-Evolution Learning for Mixup: Enhance Data Augmentation on Few-Shot Text Classification Tasks. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 8964-8974. https://doi.org/10.18653/v1/2023.emnlp-main.555

[86] Li, Z., Haroutunian, L., Tumuluri, R., Cohen, P. and Haf, R. (2024) Improving Cross-domain Low-Resource Text Generation through LLM Post-Editing: A Programmer-Interpreter Approach. *Findings of the Association for Computational Linguistics*: *EACL* 2024, St. Julian's, 17-22 March 2024, 347-354. https://aclanthology.org/2024.findings-eacl.24

[87] Zhu, Y., Si, J., Zhao, Y., Zhu, H., Zhou, D. and He, Y. (2023) EXPLAIN, EDIT, GENERATE: Rationale-Sensitive Counterfactual Data Augmentation for Multi-Hop Fact Verification. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 13377-13392. https://doi.org/10.18653/v1/2023.emnlp-main.826

[88] Krzymiński, M. (2023) Take the Most Out of Text Data Augmentation Strategies for Intent Clustering and Induction Based on DSTC 11 track 2. *Proceedings of the* 19*th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems*, Prague, 11-12 September 2023, 47-48. https://aclanthology.org/2023.yrrsds-1.17

[89] Lai, V., Nguyen, C., Ngo, N., Nguyen, T., Dernoncourt, F., Rossi, R., *et al*. (2023) Okapi: Instruction-Tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Singapore, 6-10 December 2023, 318-327. https://doi.org/10.18653/v1/2023.emnlp-demo.28

[90] Bansal, R., Samanta, B., Dalmia, S., Gupta, N., Vash-ishth, S., *et al*. (2024) LLM Augmented LLMs: Expanding Capabilities through Composition. arXiv: 2401.02412.

[91] Cheng, Y.H., Zhang, C.Y., *et al*. (2024) Exploring Large Language Model Based Intelligent Agents: Definitions, Methods, and Prospects. arXiv: 2401.03428.

[92] Li, Q.Y., Fu, L.Y., *et al*. (2024) Adapting Large Language Models for Education: Foundational Capabilities, Potentials, and Challenges. arXiv: 2401.08664.

[93] Sun, Z.H., Lyu, C., Li, B.L., Wan, Y., Zhang, H.Y., Li, G. and Jin, Z. (2024) Enhancing Code Generation Performance of Smaller Models by Distilling the Reasoning Ability of LLMs. arXiv: 2403.13271.

[94] Zheng, C., Sun, K., Wu, H., Xi, C.G. and Zhou, X. (2024) Balancing Enhancement, Harmlessness, and General Capabilities: Enhancing Conversational LLMs with Direct RLHF. arXiv: 2403.02513.

[95] Hu, S.J., Zhou, L., *et al*. (2024) WavLLM: Towards Robust and Adaptive Speech Large Language Model. arXiv: 2404.00656.

[96] Lee, C., Xia, C.S., Huang, J., Zhu, Z., Zhang, L. and Lyu, M.R. (2024) A Unified Debugging Approach via LLM-Based Multi-Agent Synergy. arXiv: 2404.17153.

[97] Guo, S.Y., Deng, C., Wen, Y., Chen, H.C., Chang, Y. and Wang, J. (2024) DS-Agent:

Automated Data Science by Empowering Large Language Models with Case-Based Reasoning. arXiv: 2402.17453.

[98] Ma, K., Cheng, H., Zhang, Y., Liu, X., Nyberg, E. and Gao, J. (2023) Chain-of-Skills: A Configurable Model for Open-Domain Question Answering. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Toronto, 9-14 July 2023, 1599-1618. https://doi.org/10.18653/v1/2023.acl-long.89

[99] Huang, L., *et al*. (2023) A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv: 2311.05232. https://arxiv.org/pdf/2311.05232

[100] O'Neill, L., Anantharama, N., Borgohain, S. and Angus, S.D. (2023) Models Teaching Models: Improving Model Accuracy with Slingshot Learning. *Proceedings of the* 17*th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, 2-6 May 2023, 3233-3247. https://doi.org/10.18653/v1/2023.eacl-main.236

[101] Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., *et al*. (2021) Recipes for Building an Open-Domain Chatbot. *Proceedings of the* 16*th Conference of the European Chapter of the Association for Computational Linguistics*: *Main Volume*, 19-23 April 2021, 300-325. https://doi.org/10.18653/v1/2021.eacl-main.24

[102] Lassner, D., Brandl, S., Baillot, A. and Nakajima, S. (2023) Domain-Specific Word Embeddings with Structure Prediction. *Transactions of the Association for Computational Linguistics*, **11**, 320-335. https://doi.org/10.1162/tacl_a_00538

[103] Arefeen, M.A., Debnath, B. and Chakradhar, S. (2024) LeanContext: Cost-Efficient Domain-Specific Question Answering Using LLMs. *Natural Language Processing Journal*, **7**, Article ID: 100065. https://doi.org/10.1016/j.nlp.2024.100065

[104] Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R. and Yin, W.P. (2024) Large Language Models for Mathematical Reasoning: Progresses and Challenges. *Proceedings of the* 18*th Conference of the European Chapter of the Association for Computational Linguistics*: *Student Research Workshop*, St. Julian's, Malta, 21-22 March 2024, 225-237. https://aclanthology.org/2024.eacl-srw.17

[105] Handa, K., Clapper, M., Boyle, J., Wang, R., Yang, D., Yeager, D., *et al*. (2023) "Mistakes Help Us Grow": Facilitating and Evaluating Growth Mindset Supportive Language in Classrooms. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 8877-8897. https://doi.org/10.18653/v1/2023.emnlp-main.549

[106] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. arXiv: 1705.07874.

[107] Yu, W., Zhu, C., Zhang, Z., Wang, S., Zhang, Z., Fang, Y., *et al*. (2022) Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. *Proceedings of the* 2022 *Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, 7-11 December 2022, 4364–4377. https://doi.org/10.18653/v1/2022.emnlp-main.294

[108] Sultana, A., Chowdhury, N.K. and Chy, A.N. (2022) CSECU-DSG@SMM4H'22: Transformer Based Unified Approach for Classification of Changes in Medication Treatments in Tweets and WebMD Reviews. *Proceedings of the Seventh Workshop on Social Media Mining for Health Applications*, *Workshop & Shared Task*, Gyeongju, 12-17 October 2022, 118-122. https://aclanthology.org/2022.smm4h-1.33

[109] Si, C., Shi, W., Zhao, C., Zettlemoyer, L. and Boyd-Graber, J. (2023) Getting More

Out of Mixture of Language Model Reasoning Experts. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 8234–8249. https://doi.org/10.18653/v1/2023.findings-emnlp.552

[110] Shen, S., Yao, Z., Li, C., Darrell, T., Keutzer, K. and He, Y. (2023) Scaling Vision-Language Models with Sparse Mixture of Experts. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 11329-11344. https://doi.org/10.18653/v1/2023.findings-emnlp.758

[111] Li, R., Murray, G. and Carenini, G. (2023) Mixture-of-Linguistic-Experts Adapters for Improving and Interpreting Pre-Trained Language Models. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 9456-9469. https://doi.org/10.18653/v1/2023.findings-emnlp.634

[112] Artetxe, M., Bhosale, S., Goyal, N., Mihaylov, T., Ott, M., Shleifer, S., *et al*. (2022) Efficient Large Scale Language Modeling with Mixtures of Experts. *Proceedings of the* 2022 *Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, 7-11 December 2022, 11699-11732. https://doi.org/10.18653/v1/2022.emnlp-main.804

[113] Jiang, Z., Peng, H., Feng, S., Li, F. and Li, D. (2024) LLMs Can Find Mathematical Reasoning Mistakes by Pedagogical Chain-of-thought. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, Jeju, 3-9 August 2024, 3439-3447. https://doi.org/10.24963/ijcai.2024/381

[114] Kong, X.H., Chen, J.Y., Wang, W.G., Su, H., Hu, X.L., Yang, Y. and Liu, S. (2024) Controllable Navigation Instruction Generation with Chain of Thought Prompting. arXiv: 2407.07433.https://arxiv.org/abs/2407.07433

[115] Cohn, C., Hutchins, N., Le, T. and Biswas, G. (2024) A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 23182-23190. https://doi.org/10.1609/aaai.v38i21.30364

[116] Parker, M.J., Anderson, C., Stone, C., *et al*. (2024) A Large Language Model Approach to Educational Survey Feedback Analysis. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00414-0

[117] Pozdniakov, S., Brazil, J., Abdi, S., Bakharia, A., Sadiq, S., Gašević, D., *et al*. (2024) Large Language Models Meet User Interfaces: The Case of Provisioning Feedback. *Computers and Education*: *Artificial Intelligence*, **7**, Article ID: 100289. https://doi.org/10.1016/j.caeai.2024.100289

[118] Dimbisoa, W.G., Mahatody, T. and Razafimandimby, J.P. (2018) Creating a Metamodel of UI Components in Form of Model Independent of the Platform. *International Journal of Conceptions on Computing and Information Technology*, **6**, 48-52. http://wairco.org/IJCCIT/November2018Paper12.pdf

[119] Logaprakash, M., Manjunath, N., Rubanraaj, K. and Srinivas, V. (2024) Personalised Learning System Using LLM. *International Journal of Creative Research Thoughts* (*IJCRT*), **12**, c24-c26. https://www.ijcrt.org/papers/IJCRT2405220.pdf

[120] Abu-Rasheed, H., Weber, C. and Fathi, M. (2024) Knowledge Graphs as Context Sources for LLM-Based Explanations of Learning Recommendations. arXiv: 2403.03008.

[121] Fahl, W. (2024) GraphWiseLearn: Personalized Learning through Semantified TEL, Leveraging QA-Enhanced LLM-Generated Content. https://2024.eswc-conferences.org/wp-content/uploads/2024/05/77770405.pdf

[122] Park, M., Kim, S., Lee, S., Kwon, S. and Kim, K. (2024) Empowering Personalized Learning through a Conversation-Based Tutoring System with Student Modeling.

Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Honolulu, 11-16 May 2024, 1-10. https://doi.org/10.1145/3613905.3651122

[123] Shoeibi, N. (2023) Cross-Lingual Transfer in Generative AI-Based Educational Platforms for Equitable and Personalized Learning. *Learning Analytics Summer Institute* (*LASI*), Madrid, June 29-30 2023, 524-540. https://ceur-ws.org/Vol-3542/paper8.pdf

[124] Shi, Y.X., Zi, X., Shi, Z.J., Zhang, H.M., Wu, Q. and Xu, M. (2024) ERAGent: Enhancing Retrieval-Augmented Language Models with Improved Accuracy, Efficiency, and Personalization. arXiv: 2405.06683.

[125] Hang, C.N., Wei Tan, C. and Yu, P. (2024) MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning. *IEEE Access*, **12**, 102261-102273. https://doi.org/10.1109/access.2024.3420709

[126] Teresa, L.A., Sunil, N.M., Andrews, S.R., Thengumpallil, T.T., Thomas, S. and V A, B. (2023) Enhancing Children's Learning Experience: Interactive and Personalized Video Learning with AI Technology. 2023 *IEEE International Conference on Recent Advances in Systems Science and Engineering* (*RASSE*), Kerala, 8-11 November 2023, 1-5. https://doi.org/10.1109/rasse60029.2023.10363506

[127] Mo, C., Wang, C., Dai, J. and Jin, P. (2022) Video Playback Speed Influence on Learning Effect from the Perspective of Personalized Adaptive Learning: A Study Based on Cognitive Load Theory. *Frontiers in Psychology*, **13**, Article 839982. https://doi.org/10.3389/fpsyg.2022.839982

[128] Cui, Y. and Hu, Y. (2024) Personalized Recommendation Method for the Video Teaching Resources of Folk Sports Shehuo Based on Mobile Learning. In: Wang, B., Hu, Z., Jiang, X. and Zhang, Y.D., Eds., *Multimedia Technology and Enhanced Learning*, Springer Nature Switzerland, 254-267. https://doi.org/10.1007/978-3-031-50574-4_18

[129] Xu, Y., Li, X., Yang, Y., Lin, Z., Wang, L. and Li, W. (2023) FedABR: A Personalized Federated Reinforcement Learning Approach for Adaptive Video Streaming. 2023 *IFIP Networking Conference* (*IFIP Networking*), Barcelona, 12-15 June 2023, 1-9. https://doi.org/10.23919/ifipnetworking57963.2023.10186404

[130] Gorban, A.N., Mirkes, E.M. and Zinovyev, A.Y. (2023) Exploring the Impact of Adaptive Video on Personalized Learning Experiences. *Proceedings of the Workshop on the Influence of Adaptive Video Learning*, Plovdiv, 13-14 October 2022, 9-16. https://ceur-ws.org/Vol-3372/paper01.pdf

[131] Lu, Y., Zhu, Y. and Wang, Z. (2022) Personalized 360-Degree Video Streaming. *Proceedings of the* 30*th ACM International Conference on Multimedia*, Lisboa, 10-14 October 2022, 3143-3151. https://doi.org/10.1145/3503161.3548047

[132] Liu, X.D. and Xue, X.W. (2023) Research on Learning Video Recommendation System Based on DBSCAN Clustering Algorithm. *International Conference on Algorithms, High Performance Computing, and Artificial Intelligence* (*AHPCAI* 2023), Yinchuan, 18-19 August 2023, 129-137.

[133] Bontchev, B., Antonova, A. and Dankov, Y. (2020) Educational Video Game Design Using Personalized Learning Scenarios. In: Gervasi, O., *et al*., Eds., *Computational Science and Its Applications—ICCSA* 2020, Springer, 829-845.

[134] Yi, R., Ye, Z.P., Zhang, J.Y., Bao, H.J. and Liu, Y.J. (2020) Audio-Driven Talking Face Video Generation with Learning-Based Personalized Head Pose. arXiv: 2002.10137. https://arxiv.org/abs/2002.10137

[135] Qu, Z.Y., Yin, L., Yu, Z.T., Wang, W.B. and Zhang, X. (2024) CourseGPT-ZH: An Educational Large Language Model Based on Knowledge Distillation Incorporating Prompt Optimization. arXiv: 2405.04781.

[136] Zastudil, C., Rogalska, M., Kapp, C., Vaughn, J. and MacNeil, S. (2023) Generative AI in Computing Education: Perspectives of Students and Instructors. 2023 *IEEE Frontiers in Education Conference* (*FIE*), College Station, 18-21 October 2023, 1-9. https://doi.org/10.1109/fie58773.2023.10343467

[137] Wang, H., Dang, A., Wu, Z. and Mac, S. (2024) Generative AI in Higher Education: Seeing ChatGPT through Universities' Policies, Resources, and Guidelines. *Computers and Education: Artificial Intelligence*, **7**, Article ID: 100326. https://doi.org/10.1016/j.caeai.2024.100326

[138] Heston, T. and Khun, C. (2023) Prompt Engineering in Medical Education. *International Medical Education*, **2**, 198-205. https://doi.org/10.3390/ime2030019

[139] Wang, T.Y., Zhou, N.J. and Chen, Z.X. (2024) Enhancing Computer Programming Education with LLMs: A Study on Effective Prompt Engineering for Python Code Generation. arXiv: 2407.05437. https://arxiv.org/abs/2407.05437

[140] Taylor Gonzalez, D.J., Djulbegovic, M.B. and Bair, H. (2024) We Need to Add Prompt Engineering Education to Optimize Generative Artificial Intelligence in Medicine. *Academic Medicine*, **99**, 1050-1051.

[141] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) Bert: Pretraining of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805.

[142] Zhang, H.J., Xu, Y.M. andPerez-Beltrachini, L. (2024) Fine-Grained Natural Language Inference Based Faithfulness Evaluation for Diverse Summarisation Tasks. *Proceedings of the* 18*th Conference of the European Chapter of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), St. Julian's, 17-22 March 2024, 1701-1722. https://aclanthology.org/2024.eacl-long.102.

[143] Akoju, S.A., Vacareanu, R., Blanco, E., Riaz, H. and Surdeanu, M. (2023) Synthetic Dataset for Evaluating Complex Compositional Knowledge for Natural Language Inference. *Proceedings of the* 1*st Workshop on Natural Language Reasoning and Structured Explanations* (*NLRSE*), Toronto, 13 June 2023, 157-168. https://doi.org/10.18653/v1/2023.nlrse-1.12

[144] Tian, R., Zhao, Z., Liu, W., Liu, H., Mao, W., Zhao, Z., *et al*. (2023) SAMP: A Model Inference Toolkit of Post-Training Quantization for Text Processing via Self-Adaptive Mixed-precision. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*: *Industry Track*, Singapore, 6-10 December 2023, 123-130. https://doi.org/10.18653/v1/2023.emnlp-industry.13

[145] Austin, E., Zaïane, O.R. and Largeron, C. (2022) Community Topic: Topic Model Inference by Consecutive Word Community Discovery. *Proceedings of the* 29*th International Conference on Computational Linguistics*, Gyeongju, 12-17 October 2022, 971-983. https://aclanthology.org/2022.coling-1.81

[146] Pletenev, S., Chekalina, V., Moskovskiy, D., Seleznev, M., Zagoruyko, S. and Panchenko, A. (2023) A Computational Study of Matrix Decomposition Methods for Compression of Pre-Trained Transformers. *Proceedings of the* 37*th Pacific Asia Conference on Language*, *Information and Computation*, Hong Kong, 2-5 December 2023, 723-742. https://aclanthology.org/2023.paclic-1.73

[147] Volosincu, M., Lupu, C., Trandabat, D. and Gifu, D. (2023) FII SMART at Semeval 2023 Task7: Multi-Evidence Natural Language Inference for Clinical Trial Data. *Proceedings of the* 17*th International Workshop on Semantic Evaluation* (*SemEval*-2023), Toronto, 13-14 July 2023, 212-220. https://doi.org/10.18653/v1/2023.semeval-1.30

[148] Kotitsas, S., Kounoudis, P., Koutli, E. and Papageorgiou, H. (2024) Leveraging Fine-tuned Large Language Models with LoRA for Effective Claim, Claimer, and Claim

Object Detection. *Proceedings of the* 18*th Conference of the European Chapter of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), St. Julian's, 17-22 March 2024, 2540-2554. https://aclanthology.org/2024.eacl-long.156

[149] Power, R. and Scott, D. (1998) WYSIWYM: Knowledge Editing with Natural Language Feedback. Association for Computational Linguistics. https://aclanthology.org/W98-1437

[150] Yehudai, A., Carmeli, B., Mass, Y., Arviv, O., Mills, N., Toledo, A., Shnarch, E. and Choshen, L. (2024) Genie: Achieving Human Parity in Content-Grounded Datasets Generation. arXiv: 2401.14367.

[151] Nayak, N., Nan, Y., Trost, A. and Bach, S. (2024) Learning to Generate Instruction Tuning Datasets for Zero-Shot Task Adaptation. *Findings of the Association for Computational Linguistics ACL* 2024, Bangkok, 11-16 August 2024, 12585-12611. https://doi.org/10.18653/v1/2024.findings-acl.748

[152] Xu, X.H., Li, M., Tao, C.Y., Shen, T., Cheng, R., Li, J.Y., Xu, C., Tao, D.C. and Zhou, T.Y. (2024) A Survey on Knowledge Distillation of Large Language Models. arXiv: 2402.13116.

[153] Li, Q.Y., Fu, L.Y., Zhang, W.M., Chen, X.Y., Yu, J.W., Xia, W., Zhang, W.N., Tang, R.M. and Yu, Y. (2023) Adapting Large Language Models for Education: Foundational Capabilities, Potentials, and Challenges. arXiv: 2401.08664.

[154] Rai, D. and Yao, Z.Y. (2024) An Investigation of Neuron Activation as a Unified Lens to Explain Chain-of-Thought Eliciting Arithmetic Reasoning of LLMs. arXiv: 2406.12288. https://arxiv.org/abs/2406.12288

[155] Chang, W. and Chen, Y. (2024) Injecting Salesperson's Dialogue Strategies in Large Language Models with Chain-Of-Thought Reasoning. *Findings of the Association for Computational Linguistics ACL* 2024, Bangkok, 11-16 August 2024, 3798-3812. https://doi.org/10.18653/v1/2024.findings-acl.228

[156] Tutunov, R., Grosnit, A., Ziomek, J., Wang, J. and Bou-Ammar, H. (2024) Why Can Large Language Models Generate Correct Chain-of-Thoughts? arXiv: 2310.13571. https://arxiv.org/abs/2310.13571

[157] Zou, A., Zhang, Z.S. and Zhao, H. (2024) AuRoRA: A One-for-All Platform for Augmented Reasoning and Refining with Task-Adaptive Chain-of-Thought Prompting. *Proceedings of the* 2024 *Joint International Conference on Computational Linguistics*, *Language Resources and Evaluation* (*LREC-COLING* 2024), Torino, 20-25 May 2024, 1801-1807. https://aclanthology.org/2024.lrec-main.160

[158] Sultan, A., Ganhotra, J. and Astudillo, R.F. (2024) Structured Chain-of-Thought Prompting for Few-Shot Generation of Content-Grounded QA Conversations. arXiv: 2402.11770. https://arxiv.org/abs/2402.11770

[159] Chu, Z., Chen, J.C., *et al*. (2024) Navigate through Enigmatic Labyrinth a Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. arXiv: 2309.15402. https://arxiv.org/abs/2309.15402

[160] Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., *et al*. (2023) Few-shot Is Enough: Exploring ChatGPT Prompt Engineering Method for Automatic Question Generation in English Education. *Education and Information Technologies*, **29**, 11483-11515. https://doi.org/10.1007/s10639-023-12249-8