Issue 4/2024
pp. 156–176
ISSN 0038 0474

Matej Urbančič

Artificial intelligence in education: comparing the responses of different large language models

Abstract: Currently, there are several advanced large language models (LLMs) freely available for testing, and their use is steadily increasing. This paper aims to compare the results produced by some selected models for use in an educational setting. A qualitative research design was employed to identify the structure of the outputs and to analyse the information and key ideas related to questions about the purpose of education. The findings raise concerns about the reliability and relevance of the results, as they are neither equally informative nor consistent across different LLMs, with variations occurring even when the same questions are repeatedly tested. At present, there is no consensus on the optimal approach to integrating AI into education, nor on the potential impact of AI on learning, teaching, work and society. While it appears that the risks associated with AI can be managed, training in the use of LLMs is crucial at present, as these models will significantly impact various educational domains.

Key words: education, large language models, digitalisation, teaching and learning, purpose of education.

UDC: 37.091.64

Scientific article

Matej Urbančič, PhD, docent, Institute of the Republic of Slovenia for Vocational Education and Training, Kajuhova 32U, SI-1000 Ljubljana, Slovenija; e-naslov: matej.urbancic@cpi.si

Introduction

In 2020, OpenAI (OpenAI 2022, Roumeliotis and Tselikas 2023) introduced a *large language model* (LLM) capable of generating comprehensible and fluent text after analysing billions of pages of articles, reports, books, websites and other materials the developers deemed useful for testing. The texts produced were so sophisticated that many individuals had *difficulty grasping their origin*. This newly developed system was the basis for ChatGPT, a tool designed specifically for *conversational tasks*. This development demonstrated that a machine could perform natural language processing, generating human-like text with context and coherence. However, this is not the ultimate goal of machine learning. The theory of machine learning (Jordan and Mitchell 2015, Bellomarini et al. 2018) seeks to determine how computers can improve their performance and capabilities through *experience*. This development relies on sophisticated algorithms and the growing availability of online data.

At present, there are over 37,000 variants of language models (LLM Explorer 2024), each built with specific features. From the user's perspective, the most important characteristic for text production is the *context length*[1]. It significantly affects the performance, accuracy and relevance of the output. Combined with the challenge of *difficulty grasping the origin*, this creates a strong impression of usability, easily enhancing personal productivity (Ju and Stewart 2024). A longer context length generally enables higher-quality output, while shorter context lengths result in faster performance. With longer context lengths, users can gradually

1 In theory, the *prompt* (command line input) opens the *context*, a memory space within the machine that holds data. Once, the prompt has been sent to the model, it analyses the input and predicts the most appropriate next words as a response. This process, known as *completion*, is generated by considering the context of all prior prompts and completions held in memory. Completion is therefore the sum of entered prompts and generated responses (A comparison of LLMs 2024).

build coherent outputs, much like in a conversation[2]. Artificial intelligence, therefore, learns to respond to users by considering the sequence of questions posed and the relevance of the answers, as assessed by the users. Larger models also enable the scanning of emails, documents, multimedia and personal preferences to enhance efficiency, aligning responses more closely with users' expectations. As such, conversational technology can neither be understood as neutral nor as objective. It can be shaped by specific personal preferences, purposes, values and interests, which influence its development and use (Zheng 2024). Users who cannot grasp the process may base their conversations on unfounded assumptions or rely too heavily on their own reasoning.

Large language models

An LLM must be capable of interpreting human language and forming speech patterns that make computer-generated text appear more *genuinely human* (Wei et al. 2021, Manning 2022). In addition to *statistical* models that calculate the probability of word order, neural models use neural networks to perform complex tasks like *natural language processing* (Khurana 2023), enabling machines to actually comprehend *natural human speech*. Computers, supported by linguistics, not only derive meaning but also capture context, mood and intent, as humans do. Considering that language models are used to generate *meaningful* and *appropriate language for the user*, learning models could be particularly problematic as *virtual assistants* (Alsafari et al. 2024, Zhang et al. 2024), because users expect accuracy, relevance and credibility in the outputs. To meet that demand, outputs should be consistent across different *assistants* or, at least, provide sufficient information for drawing similar conclusions. While this may seem straightforward in mathematics, it is often not (Salpute 2024). It can be significantly more complicated in non-science and technology fields, such as history, politics and education. Furthermore, written texts produced by LLMs constitute an immense repository of data gathered from a multitude of sources, including objective, evidence-based, ethical articles as well as general texts from written media and the web. These texts may include emotionally charged responses that arise in intense public discussions. Humans might be overwhelmed by the unexpected twists and turns in public discourse, while AI does not share the same perception of unpredictability and bias (Hovy and Prabhumoye 2021). The output of LLMs should be moder-

2 The *context length* refers to the maximum number of elements (tokens) a language model can consider simultaneously while processing a *prompt* or generating a response. These elements can be words or parts of words, depending on the language model's tokenization method. A larger context allows the model to maintain coherence over longer passages of text, leading to more contextually relevant and accurate responses. With a longer context, the model can understand and respond to more complex questions or tasks that require knowledge of previous interactions or detailed prompts. As of now, models like GPT-4 have context windows of up to 128K tokens, with some versions capable of handling 1M tokens (LLM Leaderboard 2024, Groq 2024).

ated[3] to ensure it is informative and constructive, as users actively engage with AI (Brandtzaeg et al. 2022). Large-scale language technologies are increasingly used in various forms of communication with humans across different contexts as *conversational agents* (Kasirzadeh and Gabriel 2023). While AI systems may appear seamless in conversations, they still exhibit significant flaws. In a study on emergency remote teaching (Tülübaş et al. 2023), researchers conducted an AI-supported research process to evaluate its potential to generate *accurate*, *clear*, *concise* and *unbiased* information–essential elements of rigorous scientific work. The study concluded that while ChatGPT has value, its use should not go *unchecked*. Similar conclusions have been drawn in various disciplines (Scaringi and Loche 2023). Iskander (2023) used ChatGPT as an interviewee to analyse the impact of AI on higher education and academic publishing. In the interview, the model *acknowledged* the risk of *diminishing critical thinking*, particularly when it is over-relied upon as the sole source. It also cannot be a substitute for human creativity and intellect, because it lacks *originality in generated outputs*. Analysis showed (Uludag 2023) that there is a need to develop methods to test the creativity of language models and assess their potential for generating *novel and valuable* responses, as they currently rely heavily on pre-existing content. There are also frequently emerging questions of bias (Liu 2022, Ferrara 2023, Hajikhani and Cole 2024). Do the responses maintain the same standards of adequacy and relevance across different topics, cultures, values and levels of complexity? Or are these standards adjusted to the specific topic for which the *agent* is being fine-tuned (Parthasarathy et al. 2024)? It is important to recognise that outputs may rely on outdated sources and may not be *open* to explore various possibilities, especially when dealing with complex issues, such as those found in *human-like* debates.

As LLM systems are already active assistants that help personalise and evaluate students' work (Chen et al. 2024), educators require structured training and guidance to understand how these systems are used and to assess the usability, accuracy and relevance of the provided models. It is vital to understand the capabilities of this technology and to continually refine one's ability to use it effectively (Jeon and Lee 2023, Albadarin et al. 2024, Liao et al. 2024). There are many LLMs, not all of which are equally accessible. Free models may have limitations or inherent biases, but their features and capabilities are continually evolving and improving (Schur and Groenjes 2023).

3 *Level of moderation* refers to the extent to which a language model is regulated to filter out inappropriate, harmful, or sensitive content. This can include profanity, hate speech, misinformation and any content that violates community guidelines. Moderation helps ensure that interactions with the model are safe and appropriate, protecting users from harmful content. Effective moderation also contributes to the overall quality of responses. The level of moderation can vary based on the context of the interaction, with some models allowing for more lenient responses in specific scenarios, such as in educational settings, while maintaining stricter guidelines in public or general contexts.

Purpose of the Research

The purpose of this research was to explore and compare the outputs of several freely available LLMs by posing basic questions about education, with a focus on the usefulness of their responses for students' informed conclusions (Kumar et al. 2023, Agarwal et al. 2024). The chosen underlying model is *ChatGPT 4o*, which is often the first choice in academic settings (Meyer et al. 2023). It is frequently recognised as the LLM that provides the most comprehensive responses to questions. Initially, we examined the AI outputs on the importance of education by asking three similar questions: (1) 'Why educate?', (2) 'What is the purpose of education?' and (3) 'What is education for?' This inquiry aimed to capture the *nuances in responses* and to reveal various *viewpoints of prioritisation*. Based on the AI's responses, follow-up questions were posed to further refine and clarify the answers. The next three questions addressed critical considerations regarding the *direction of education*. They explored the issues over whether education should (4) prioritise developing specialised expertise or fostering a well-rounded general knowledge base, (5) identify the single most promising factor shaping the future of education and raise concern about the (6) single most harmful factor threatening education. Together, these questions shape AI's outputs on the strategic *priorities of educational systems*, the challenges they face and the opportunities that can be leveraged to foster a more effective and equitable learning environment.

A question regarding the sources of the answers and whether they could be linked to a specific author or book was raised; however, the responses provided no useful information, merely suggesting methods for finding texts *that represent foundational works in educational philosophy*. Consequently, this question was excluded from the analysis.

Research questions

- How do the responses of different LLMs vary when asked fundamental questions about the purpose and importance of education?
- What differing perspectives do LLMs offer regarding the direction of education?
- What factors do LLMs identify as the most promising for shaping the future of education or the most harmful to the educational landscape, and how do these factors differ across models?

Description of the analysed LLMs

Currently, several advanced LLMs (LLM Explorer 2024, LLM Leaderboard 2024) have emerged as prominent tools in the field of artificial intelligence. Some models, such as *GPT-4o by OpenAI* and *Gemini by Google*, support multimodal capabilities, allowing them to process both text and images. Others, like *Claude by*

Anthropic, *Mistral by Mistral* and *Ernie 4.0 by Baidu*, rely on enhanced conversational abilities and reasoning, while models like *Llama and OPT by Meta* focus on research capabilities. This may, however, change in future versions. Additionally, other models, though less widely known, are readily available without requiring special software or registration. Examples include *GPT-4o mini*, *Claude 3 Haiku*, *Llama 3.1 70B* and *Mixtral 8x7B*, all of which are accessible through platforms like DuckDuckGo's Duck.ai. These LLMs serve general-purpose applications with varying levels of built-in moderation:

– *ChatGPT 4o*, *GPT-4o mini*, *Claude 3 Haiku* and *Gemini 1.5 Flash* feature high levels of built-in moderation, employing robust filtering systems to maintain safe and appropriate interactions.
– *Llama 3.1 70B* offers a medium level of moderation, balancing flexibility and usability.
– *Mixtral 8x7B* has low built-in moderation, allowing greater freedom in interactions, making it suitable for contexts with less restrictive content management.
– *Copilot free* and *Copilot Office 365* operate with medium levels of moderation, managing various types of data and documents within a single cloud space.

These models offer a range of moderation capabilities, enabling selection of a suitable supervision level based on particular needs and application contexts.

Methodology

Research Design

The research design for this study is qualitative, focusing on the comparison of textual outputs from a selected set of LLMs. Additionally, contextual and content analysis, along with the length of the outputs, is used to compare the extent of the answers and their consistency, particularly for similar questions.

Data Collection and Data Analysis

Obtaining the output was straightforward in the analysed LLMs, as their user interfaces are organised into conversation-like segments resembling chat interactions. Prompts were sent sequentially into each LLM to enable a structured analysis and provide context. After the third question, the process was restarted with the second question: (2b) 'What is the purpose of education?' Outputs were copied into tables and analysed as interview answers. The outputs produced by the models were categorised and examined according to several criteria, including output structure, length, relevance and context. This comparative analysis sought to highlight the different patterns in how various LLMs form answers to the same questions.

Basic Limitations

When analysing LLMs, various ethical considerations and limitations must be taken into account. Models are trained on vast datasets that may contain inherent biases, which can affect their outputs. The sources of these outputs cannot always be confirmed and may reflect inherent biases. Additionally, the output can vary significantly based on context and conversation history. Responses to the same questions may differ in style and coherence, potentially leading to nonsensical answers.

Since updates to LLMs are not publicly documented, analyses reflecting a specific version may quickly become outdated. The interaction between users and LLMs can also influence outputs, making it challenging to isolate model behaviour from user input and context.

Results

Word count and depth of the output

The first notable difference among the collected answers was the variation in the lengths of the outputs. Table 1 displays the character counts for outputs from different LLMs in response to questions about education. It reveals significant differences among the evaluated set of LLMs in their responses to the queries.

	ChatGPT 4o	GPT-4o mini	Claude 3 Haiku	Llama 3.1 70B	Mixtral 8x7B	Copilot free	Copilot Office 365	Gemini 1.5 flash
(1) Why educate?	**2426**	1426	1510	2025	1000	*485*	650	1214
(2a) What is the purpose of education?	**2284**	1331	1390	1415	1906	*350*	889	1729
(3) What is education for?	**2212**	1456	1379	1415	583	*380*	866	1661

(2b) What is the purpose of education?	**3341**	1248	1390	1681	2277	*349*	1298	1532
(4) Should the focus of education be on developing specialised expertise or a well-rounded general knowledge?	**5526**	2513	2210	2232	2443	*548*	1591	2606
(5) What is the single most promising factor for the future of education?	**4029**	1734	2445	1508	2658	*380*	883	1558
(6) What is the single most harmful factor threatening education?	**3801**	2052	2200	1765	2718	*348*	1030	1403

Table 1: Character counts (with spaces) for outputs from different LLMs. Bold values indicate the highest character counts, while italicised values represent the lowest counts.

ChatGPT 4o consistently leads in both character count and output depth. Longer answers can provide more detailed insights and greater context, but they also occupy more *conversation space* (Liu et al. 2024). The variability among the other models suggests that the choice of LLM may heavily depend on whether one requires depth or brevity.

In general, the output is structured into three sections: the *Introduction* as a brief explanation, *Key sections* with bullet points, setting out the ideas of the question and *the Conclusion*. This structure organizes the information, making it easier to understand. All but *Copilot Free* follow this basic structure. The introduction sets the context and outlines expectations for the subsequent text or serves as a brief opening to begin the list. Each bullet point in the list represents a distinct idea or argument. The conclusion summarises the main points discussed. Copilot Office 365 also provides examples of additional questions to help users refine their research ideas. However, this feature is not consistently used across all types of questions. In some cases, key sections are structured with ordered lists, which also indicate priority[4]. Interestingly, *ChatGPT 4o, Gemini* and both *Copilots* begin their *introductions* differently each time, using varied words and synonyms. In contrast, other models tend to use a consistent format, such as '*The question of whether …'* and other similar ones. This may not affect the content but gives an impression of false diversity and, through word choice, a sense of elevated rele-

4 It is important to understand that questions may follow different structures. The structure of questions can vary based on the complexity of the topic or information requested, user preferences, context and purpose.

vance, linguistic diversity and professionalism. The rich language range, however, enhances the perceived quality of the response (Takase et. al 2024). The wording of the introduction is notable, as some cases include a summary introducing the text, while others simply begin the list. Characteristics are presented in Table 2.

Element	**ChatGPT 4o**	**GPT-4o mini**	**Claude 3 Haiku**	**Llama 3.1 70B**	**Mixtral 8x7B**	**Copilot free**	**Copilot Office 365**	**Gemini 1.5 flash**
Includes introduction?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Introduction length (characters)	217	80	96	136	126	485	144	121
Introduction outlines the content of a list?	Yes	No	No	Yes	No	Yes	No	No
Has ordered or unordered lists?	Ordered	List	List	List	Ordered	No list	List	Ordered
(1) Why educate? (bullets)	8	7	7	8	5	0	5	4
(2a) What is the purpose of education? (bullets)	6	7	6	6	9	0	6	5
(2b) What is the purpose of education? (bullets)	9	7	6	7	9	0	6	5
(3) What is education for? (bullets)	6	7	6	6	0	0	6	7
Has a conclusion?	Yes	Yes	Yes	Yes	Yes	No	No	Yes
Conclusion length (characters)	302	126	245	160	152	0	0	222
Outputs additional info?	No	No	No	No	No	No	Questions	No

Table 2: Characteristics for outputs from different LLMs on question 2b, What is the purpose of education?

Aims of education

The first three questions (1, 2a, 3) were used to explore the importance of education, capture nuances in the responses and reveal specific viewpoints. Despite their differences, the outlined main ideas define some fundamental characteristics of education.

The *purpose of education* revolves around fundamental categories, particularly the goals that should guide the process. Various aims have been proposed, including the acquisition of knowledge and skills, personal development and the cultivation of character traits. These traits promote qualities such as curiosity, creativity, rationality, critical thinking and moral tendencies to think, feel and

ChatGPT 4o	GPT-4o mini	Claude 3 Haiku	Llama 3.1 70B	Mixtral 8x7B	Copilot Free	Copilot Office 365	Gemini 1.5 Flash
9 key topics	**7 key topics**	**6 key topics**	**7 key topics**	**9 key topics**	**1 key topics**	**6 key topics**	**5 key topics**
Personal Development	Knowledge Acquisition	Knowledge acquisition	Personal growth and development	Acquisition of knowledge	Critical thinking	Knowledge Acquisition	Knowledge and Skill Acquisition
Skill and Knowledge Acquisition	Personal Development	Critical thinking development	Preparation for career and workforce	Intellectual development	Social awareness	Personal Development	Personal Development
Civic Responsibility and Social Awareness	Socialisation	Personal growth and fulfilment	Socialisation and community building	Personal growth		Socialisation	Socialisation and Citizenship
Economic Empowerment and Workforce Preparation	Economic Opportunity	Societal advancement	Critical thinking and problem solving	Career preparation		Economic Benefits	Economic Development
Promoting Equity and Reducing Inequality	Civic Engagement	Preparation for the workforce	Preservation and transmission of knowledge	Lifelong learning		Civic Engagement	Cultural Preservation and Enrichment
Cultural Transmission and Preservation	Cultural Transmission	Fostering lifelong learning	Empowerment and social mobility	Social mobility		Innovation and Progress	
Innovation and Progress	Innovation and Progress		Personal fulfilment and enjoyment	Civic engagement			
Emotional and Social Development				Cultural preservation and diversity			
Inspiring Purpose and Meaning				Health and well-being			

Table 3: Key sections reflecting basic educational aims in response to the question: What is the purpose of education? Categories: □ Personal, Knowledge, Societal, Economic and Cultural

act ethically (Table 3). This framework aligns with ideas presented in Wikipedia's 'Aims and ideologies', 2024. Scholars differ on whether education should prioritise – *personal development*, questioning authority and dispelling false beliefs and illusory ideas – or cultivating individuals *into productive members of society* (Dewey 1922, Randall 1997, Biesta 2015, Selwyn 2019). Biesta argues that education serves three functions: *qualification*, which encompasses the knowledge and skills needed for activity in social spheres; *socialisation*, that defines culture and traditions that identify an individual as a member of a society; and *subjectification*, which empowers individuals to think and act independently. However, when compared to the outputs produced by the LLMs, this division appears too complex and intertwines aspects that LLMs treat as separate. Since LLMs do not cite their outputs, the widely recognised *Robinson's model of educational aims* (Robinson 2022) has been employed as a background framework. Buzzwords like the 8Cs (curiosity, creativity, criticism, communication, collaboration, compassion, composure, citizenship), seem to align more closely with the outputs of LLMs than with Biesta's proposal. Within the Robinson's model, a new category, *Knowledge*, has been identified separately from the *Personal* category, as it is central to the content of the outputs. These categories form a comprehensive structure of the outputs and are colour coded in Table 3.

In most LLMs, all five categories appear in response to this question (2b). However, *Copilot* Free which defines only critical thinking and social awareness due to the brevity of the response, and *Claude 3 Haiku*, *Llama 3.1 70B* and *Copilot Office 365*, which explicitly do not mention *cultural preservation*. Despite these significant differences, the outputs can be considered comprehensive overall.

Among more spiritual concepts, the notion of *inspiring purpose* and *meaning* supports the development of a sense of direction, guiding individuals toward meaningful goals that align with their values and interests. Overall, these outputs illustrate the many roles of education in shaping well-rounded individuals. No LLM argued against trusting authority.

Table 4 highlights the ideas that significantly diverge from the common elements identified across LLM outputs.

Diverged points	**(Number of the question) Large language model**
Improving Health and Well-being	(1,2) ChatGPT 4o, (1) Llama 3.1 70B, (1) Gemini 1.5 Flash, (3) Mixtral 8x7B
Environmental Awareness and Action	(1) ChatGPT 4o, (1) Mixtral 8x7B, (3) Mixtral 8x7B
Equity and Inclusion	(2) GPT-4o mini, (1) Mixtral 8x7B, (3) Claude 3 Haiku, (3) Gemini 1.5 Flash
Sustainable Development	(1) ChatGPT 4o, (2) Mixtral 8x7B
Digital	(2) ChatGPT 4o
Entrepreneurship	/
Wrong focus	(1) Mixtral 8x7B: *Educating users about privacy and security.*

Table 4: Outputs of LLMs that significantly diverge from the common outputs. The number in brackets indicates the questions: (1) Why educate?, (2) What is the purpose of education?, (3) What is education for?

Particularly noteworthy is *Mixtral 8x7B*'s misplaced focus on education about privacy and security as, presented an *essential part of creating a responsible and trustworthy digital service*. This output stands out as significantly different from all other responses across all questions, appearing out of place in its context. Given that no prior questions posed in the environment addressed this topic, the content of the response is intriguing. It can be assumed that such questions are more common in contexts related to digital services, which LLMs are designed to address. As a result, the response is directed toward a specific area rather than the broader concept of education.

Digital literacy, entrepreneurship and sustainable development are key competences and essential skills necessary for personal fulfilment, employability, social inclusion and active citizenship, as supported by European institutions (Collective council EC 2018). Sustainable development is mentioned twice, digital literacy once, and entrepreneurship not at all.

Comparing two outputs of the same question

For the purpose of education (questions 2a and 2b), bullet points form the primary structure of the outputs, creating coherent ideas. However, the responses are inconsistent between the two attempts. When the question was repeated, the response differed significantly, altering the core idea of the answer. The most significant change occurred with *ChatGPT 4o*, while *Claude 3 Haiku* and *Gemini 1.5 Flash* produced a perfect match on both occasions. The comparison is presented in Table 4.

	Same argument in a bullet point in the first and the second output	Missing in the second output	Extra in the second output
ChatGPT 4o 6 / 9 bullets	Personal Development Skill and Knowledge Acquisition (this are two separate arguments in the second output)	Character and Citizenship Economic and Social Mobility Adaptability and Lifelong Learning	Civic Responsibility and Social Awareness Economic Empowerment and Workforce Preparation Promoting Equity and Reducing Inequality Cultural Transmission and Preservation Innovation and Progress Emotional and Social Development Inspiring Purpose and Meaning
GPT-4o mini 7 / 7 bullets	Knowledge Acquisition Personal Development Socialisation Economic Opportunity Civic Engagement Innovation and Progress	Equity and Inclusion	Cultural Transmission
Llama 3.1 70B 6 / 7 bullets	Personal growth and development Preparation for career and professional life Socialisation and community building Critical thinking and problem solving Empowerment and social mobility	Cultural transmission and preservation	Preservation and transmission of knowledge Personal fulfilment and enjoyment
Mixtral 8x7B 9 / 9 bullets	Acquisition of Knowledge Personal Development Career Preparation Social Mobility Lifelong Learning Cultural Preservation and Transmission Health and Well-being	Citizenship Sustainable Development	Intellectual development Civic engagement
Copilot Office 365 6 / 6 bullets	Personal Growth Socialisation Economic Empowerment Civic Responsibility Innovation and Progress	Intellectual Development	Knowledge Acquisition

Table 4: Differences among the answers. Copilot is excluded because it does not have a properly structured response. Claude 3 Haiku and Gemini 1.5 Flash produced a perfect match. Synonyms are considered identical in this comparison.

The variation in the general content of points across responses denotes incoherence in the LLM outputs. Based on prior use of the tool – where users often save their conversations (Mayer 2023) – it appears that the generated list is

not fully representative. It may exclude elements that are fundamental to understanding the problem, regardless of the underlying source. Although users can extract individual keywords from the context based on the descriptions, doing so requires critical reading. The differences may reflect a momentary fluctuation in the tool's processing rather than a deliberate approach. Therefore, answers to important questions cannot be relied upon without broader knowledge of the topic.

Developing specialised expertise or a well-rounded general knowledge is a simple question

The responses to question (4), 'Should education prioritise developing specialised expertise or foster a well-rounded general knowledge', are highly consistent across models. Although the models highlight different points characteristic of specialised or generalised knowledge, all but the *Mixtral 8x7B* model propose an ideal *balanced approach*. The emphasis lies in *early general education, followed by later specialisation* and *cross-disciplinary learning in specialist fields*. This approach *incorporates general skills within specialised knowledge*. Striking a balance between the two is crucial, as this approach provides students with a comprehensive education. It fosters *flexibility and adaptability*, allowing them to explore different disciplines.

The *Mixtral 8x7B* takes a slightly different approach by emphasising individual circumstances and goals. Educational systems should offer a range of options to ensure that students are well-prepared for their chosen paths, starting with a solid general foundation *before allowing* for specialisation. The final decision should be based on a careful consideration of individual goals, career aspirations and societal needs. In all cases, the LLMs provided arguments for and against focusing exclusively on one approach. As expected, a *contra option* was included even though the question did not explicitly seek a comparison.

Most promising and most harmful factors

'What is the single most promising or most harmful factor?' (questions 5 and 6) also produces diverse outputs.

LLM	Single most promising factor	Short overview	Single most harmful factor	Short overview
ChatGPT 4o	**Personalised learning**	Personalised learning holds immense potential to transform education by focusing on each student's unique journey.	**Inequality in access to education and resources**	Whether due to socioeconomic status, geographic location, or biases, educational inequality undermines the promise of education as a pathway to opportunity.
GPT-4o mini	**Personalised learning**	Personalised learning holds great promise for transforming education by making it more responsive to the unique needs of students.	**Standardised testing**	Standardised testing can lead to a rigid, one-size-fits-all approach that fails to address the diverse needs of students and stifles innovation in teaching and learning.
Claude 3 Haiku	**Fostering of critical thinking and problem-solving skills**	With the development of critical thinking and problem-solving skills, education can prepare students to thrive in an increasingly complex, rapidly evolving world.	**Suppression of objective, knowledge and critical thinking**	When education becomes overly ideological, politicised, or beholden to narrow interests, it risks undermining the fundamental purpose of learning.
Llama 3.1 70B	**Critical thinking**	Critical thinking is the ability to analyse information, evaluate evidence and form informed opinions.	**Dogmatic thinking**	By recognising and addressing dogmatic thinking, educators can promote a more open-minded, critical and creative approach to learning.
Mixtral 8x7B	**The use of technology to support teaching and learning**	By leveraging technology in a thoughtful and strategic way, we can help to create more personalised, engaging and effective learning experiences for all students.	**Overemphasis on standardised testing**	An overemphasis on these measures can lead to a narrow and superficial approach to learning, increased stress and anxiety and inequity and bias.

Copilot free	**Adaptability**	It's about cultivating a mindset that's curious, flexible and resilient. Focusing on adaptability, education can continuously meet the needs of society and the individual.	**Complacency**	When education stops evolving and adapting, it risks becoming irrelevant. Outdated methods and content can stifle creativity and critical thinking.
Copilot Office 365	**Personalised learnings**	Personalised learning tailors educational experiences to meet the individual needs, strengths and interests of each student.	**Inequity**	Addressing educational inequality is crucial for creating a more just and prosperous society.
Gemini 1.5 flash	**Personalised learning**	It has the potential to revolutionise education by creating more equitable, effective and engaging learning experiences for all students.	**Inequity**	By working to eliminate inequality in education, we can create a brighter future for all.

Table 5: The most promising and the most harmful factors threatening education

The predominant responses are *Personalised Learning* and *Critical Thinking*. In this context, problem-solving skills and adaptability are also closely tied to personalised approaches to education. Conversely, the most harmful factors identified are *Inequality and Inequity*. An intriguing response is the emphasis on the harm caused by *Standardised testing*, which is linked to the suppression of objective knowledge and critical thinking, potentially resulting in *dogmatic thinking* (*Llama 3.1 70B*). The output is presented in Table 5.

Conclusion and implications

The *purpose of education* is not a simple concept to grasp. Although the purpose of this comparative study was not to find a definitive answer, but rather to compare the content produced by LLMs on this topic, the results indicate that the answer is not easily obtained. Instead, it requires serious study and deeper analysis. The results indicate that the answers to the questions share a similar focus. It is important to *acquire knowledge and skills* for personal empowerment, independence and the cultivation of critical thinking. Education supports *personal growth and fulfilment*, both of which are essential in the contemporary world. Civic engagement promotes *social cohesion and promotion* by fostering a sense of community and encouraging active participation. Problem solving, innovation and progress are vital for *economic opportunities, mobility* and workforce preparation. *Lifelong learning* facilitates these goals while fostering *cultural and interpersonal understanding and preservation*, providing meaning and a sense of

belonging to all individuals. LLMs, on the other hand, encompass diverse perspectives, ranging from humanistic individual growth to global technological and economic competition. In a sense, this raises the question of who determines the value and prioritisation of different educational purposes (Tenam-Zemach and Flynn 2011). The problem lies in the source of the content. The AI-generated content already found its place in Wikipedia (Brooks et al. 2024, Ashkinaze et al. 2024). This suggests that AI-generated content may eventually replace Wikipedia as the primary source for students seeking a general overview of a topic (Fessakis and Zoumpatianou 2013). According to some scholars (Thomas 2023), this represents a plausible future. The key difference is that Wikipedia cites its sources and employs moderators to oversee them. In contrast, LLMs cannot conclusively reference their sources, as the underlying texts are too numerous and varied.

It is essential to use LLMs with caution. As demonstrated, various LLMs produce significantly different outputs, which must be carefully assessed to understand their proposed content and account for the models' limitations. Since AI tools cannot be held accountable for the accuracy and integrity of their content (Stokel-Walker 2022, 2023), the primary responsibility in education falls on educators to prepare students to work with these tools. Since the use of LLMs in education is viewed as a *transformative technology*, it is also important to recognise its contradictory nature. AI can be highly efficient for users who critically evaluate its responses and refine their questions with subsequent prompts. However, if users settle for the first output they receive, they risk obtaining incorrect, inappropriate, or incomplete answers. Iskander (2023) demonstrated that optimising queries enhances a model's ability to generate clearer and more concise responses. However, using LLMs to discover genuinely novel solutions remains unreliable. Frequent questions might also result in prioritised answers, which could be further influenced by the fine-tuning process. Since the authorship of LLM outputs is unclear and will likely remain so, the iterative process of using generated answers as new inputs undermines the potential for genuine novelty. While a lack of originality might suggest similar answers, this is not always the case. Differences can occur with each prompt and may vary based on the user's perspective, which influences how the tool is used. The analysis shows that all responses reflect a commonly shared understanding of the purpose of education. However, individual nuances—such as emotional development, joyfulness and seeking purpose—add unique perspectives shaped by the underlying structure of the LLMs. These nuances may also align with user preferences. LLM outputs also place a strong emphasis on economic empowerment, particularly workforce preparation. This highlights how the promotion of education's purpose evolves with advancements in technology.

This tool is now a reality and will inevitably be used; therefore, it is crucial to ensure its proper and effective use. Without strategies to integrate teachers' oversight into learning activities involving AI tools (Albadarin et al. 2024), education risks significant shortcomings and unfulfilled expectations. The potential of AI is also evident in its ability to adapt and refine algorithms. This means its usefulness depends on the user's knowledge and their ability to design, refine and enhance the algorithms used to query sources.

AI will continue to play an increasingly significant role in teaching and learning. It will become ever more sophisticated, and produce more accurate information and faster prompts (Toczauer 2024). Advancements in AI will eventually enable common sense reasoning in computers (Chowdhary and Chowdhary 2020). However, this progress will not occur without deliberate effort. Evaluation of usage, queries and responses should be treated as an essential discipline assisting the development of LLMs (Chang 2024). Currently, there is no consensus on the extent to which AI should be integrated into education or its potential effects. Researchers suggest (Kasneci et al. 2023) that despite the challenges, the associated risks are manageable and should be addressed to ensure trustworthy and equitable access to LLMs for education and research (Liao et al. 2024). Towards this goal, the mitigation strategies proposed in this commentary could serve as a starting point.

LLMs will inevitably affect learning, teaching and work. Efficiency-oriented modernity makes their use virtually irresistible.

References

Agarwal, V., Thureja, N., Garg, M. K., Dharmavaram, S. and Kumar, D. (2024). "Which LLM should I use?": Evaluating LLMs for tasks performed by Undergraduate Computer Science Students in India. a*rXiv preprint* arXiv:2402.01687.

Alsafari, B., Atwell, E., Walker, A. and Callaghan, M. (2024). Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants. *Natural Language Processing Journal*, 8, 100101.

Ashkinaze, J., Guan, R., Kurek, L., Adar, E., Budak, C. and Gilbert, E. (2024). Seeing like an AI: How LLMs apply (and misapply) Wikipedia neutrality norms. *arXiv preprint* arXiv:2407.04183.

Bellomarini, L., Fayzrakhmanov, R. R., Gottlob, G., Kravchenko, A., Laurenza, E., Nenov, Y. and Wu, L. (2018). Data science with Vadalog: Bridging machine learning and reasoning. In *Model and Data Engineering: 8th International Conference, MEDI 2018, Marrakesh, Morocco, October 24-26, 2018, Proceedings 8*. Springer International Publishing. pp. 3–21.

Biesta, G. (2015). What is education for? On good education, teacher judgement, and educational professionalism. *European Journal of education*, 50, issue 1, pp. 75–87.

Brandtzaeg, P. B., Skjuve, M. and Følstad, A. (2022). My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research*, 48, issue 3, pp. 404–429.

Brooks, C., Eggert, S. and Peskoff, D. (2024). The Rise of AI-Generated Content in Wikipedia. *arXiv preprint* arXiv:2410.08044.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K. and Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15, issue 3, pp. 1–45.

Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G. and Chen, E. (2024). When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27, issue 4, pp. 42–54.

Chowdhary, K. and Chowdhary, K. R. (2020). Natural language processing. In. *Fundamentals of artificial intelligence*, Springer, New Delhi. pp. 603–649.

Collective Council, E. U. (2018). *Recommendation On Key Competences For Lifelong Learning*. European Commission, Brussels, Belgium, 18.

Dewey, J. (1910). *How we think*. D C Heath. https://doi.org/10.1037/10903-000

Education: Aims and ideologies. (2024, November 22). *In Wikipedia*. Retrieved from: https://en.wikipedia.org/wiki/Education#Aims_and_ideologies (accessed on 15 November 2024).

Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint* arXiv:2304.03738.

Fessakis, G. and Zoumpatianou, M. (2013). Wikipedia uses in learning design: A literature review. *Themes in Science and Technology Education*, 5, issue 1–2, 97–106.

Groq (2024). *The Crucial Role of Context Length in Large Language Models for Business Applications*. Retrieved from: https://groq.com/the-crucial-role-of-context-length-in-large-language-models-for-business-applications/ (accessed on 15 March 2024).

Hajikhani, A. and Cole, C. (2024). A critical review of large language models: Sensitivity, bias, and the path toward specialized AI. *arXiv*, arXiv:2307.15425 pp. 1–22.

Hovy, D. and Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and linguistics compass*, 15, issue 8, e12432.

Introducing ChatGPT (2022). Retrieved from: https://openai.com/index/chatgpt/ (accessed on 15 November 2024).

Iskender, A. (2023). Holy or unholy? Interview with open AI's ChatGPT. *European Journal of Tourism Research*, *34*, pp. 3414–3414.

Jeon, J. and Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 28, issue 12, pp. 15873–15892.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349, issue 6245, pp. 255–260.

Ju, B. and Stewart, J. B. (2024). Empowering Users with ChatGPT and Similar Large Language Models (LLMs): Everyday Information Needs, Uses, and Gratification. *Proceedings of the Association for Information Science and Technology*, 61, issue 1, pp. 172–182.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F. and Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, 102274.

Khurana, D., Koli, A., Khatter, K. and Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82, issue 3, 3713–3744.

Kumar, H., Musabirov, I., Reza, M., Shi, J., Wang, X., Williams, J. J. and Liut, M. (2023). Impact of guidance and interaction strategies for LLM use on Learner Performance and perception. *arXiv preprint* arXiv:2310.13712.

Liao, Z., Antoniak, M., Cheong, I., Cheng, E. Y. Y., Lee, A. H., Lo, K. and Zhang, A. X. (2024). LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. *arXiv preprint* arXiv:2411.05025.

Liu, R., Jia, C., Wei, J., Xu, G. and Vosoughi, S. (2022). Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304, 103654.

Liu, Y., Li, D., Wang, K., Xiong, Z., Shi, F., Wang, J. and Hang, B. (2024). Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs. *Information Processing & Management*, 61, issue 5, 103809.

LLM Leaderboard – Comparison of GPT-4o, Llama 3, Mistral, Gemini and over 30 models (November 2024). Retrieved from: https://artificialanalysis.ai/leaderboards/models (accessed 24. 11. 2024)

Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, 151, issue 2, 127–138.

Meyer, J. G., Urbanowicz, R. J., Martin, P. C., O'Connor, K., Li, R., Peng, P. C. and Moore, J. H. (2023). ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, 16, issue 1, pp. 20–30.

Parthasarathy, V. B., Zafar, A., Khan, A. and Shahid, A. (2024). The ultimate guide to fine-tuning LLMs from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint* arXiv:2408.13296.

Randall V. Bass (1997) The Purpose of Education, *The Educational Forum*, 61, issue 2, 128–132.

Roumeliotis, K. I. and Tselikas, N. D. (2023). ChatGPT and open-AI models: A preliminary review. *Future Internet*, 15, issue 6, pp. 192.

Satpute, A., Gießing, N., Greiner-Petter, A., Schubotz, M., Teschke, O., Aizawa, A. and Gipp, B. (2024, July). Can LLMs master math? Investigating Large Language Models On Math Stack Exchange. *In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2316–2320.

Scaringi, G. and Loche, M. (2023). An interview with ChatGPT: Discussing artificial intelligence in teaching, research, and practice. *Computer Science and Engineering*. Preprint.

Schur, A. and Groenjes, S. (2023, July). Comparative Analysis for Open-Source Large Language Models. In *International Conference on Human-Computer Interaction* Cham: Springer Nature Switzerland. pp. 48–54.

Selwyn, D. (2019). *The purpose of education*. Counterpoints, 529, pp. 81–100.

Solulab – A Detailed Comparison of Large Language Models (November 2024). Dostopno na: https://www.solulab.com/comparison-of-all-llm/ (accessed on 24. November 2024).

Stokel-Walker, C. (2022, December 9). AI Bot ChatGPT writes Smart Essays-should professors worry? *Nature News*. Retrieved from: https://www.nature.com/articles/d41586-022-04397-7 (accessed on 24. November 2024).

Stokel-Walker, C. (2023, January 18). CHATGPT listed as author on research papers: Many scientists disapprove. *Nature News*. Retrieved from: https://www.nature.com/articles/d41586-023-00107-z (accessed on 24. November 2024).

Takase, S., Ri, R., Kiyono, S. and Kato, T. (2024). Large Vocabulary Size Improves Large Language Models. *arXiv preprint* arXiv:2406.16508.

Thomas, P. A. (2023). Wikipedia and large language models: perfect pairing or perfect storm?. *Library Hi Tech News*, 40, issue 10, 6–8.

Tülübaş, T., Demirkol, M., Ozdemir, T. Y., Polat, H., Karakose, T. and Yirci, R. (2023). An interview with ChatGPT on emergency remote teaching: A comparative analysis based on human–AI collaboration. *Educational Process: International Journal*, 12, issue 2, 93–110.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B. and Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint* arXiv:2109.01652.

Zhang, Z., Zhang-Li, D., Yu, J., Gong, L., Zhou, J., Liu, Z. and Li, J. (2024). Simulating classroom education with llm-empowered agents. *arXiv preprint* arXiv:2406.19226.

Zheng, W., Yang, A., Lin, N. and Zhou, D. (2024, August). From Bias to Fairness: The Role of Domain-Specific Knowledge and Efficient Fine-Tuning. *In International Conference on Intelligent Computing*. Singapore: Springer Nature Singapore. pp. 354–365.

Matej URBANČIČ (Center RS za poklicno izobraževanje Slovenija)

UMETNA INTELIGENCA V IZOBRAŽEVANJU: PRIMERJAVA ODZIVOV RAZLIČNIH VELIKIH JEZIKOVNIH MODELOV

Povzetek: V prispevku je predstavljena primerjava odzivov, ki jih vrnejo različni prosto dostopni veliki jezikovni modeli (LLM). Za ugotavljanje strukture odzivov in analize informacij in ključnih zamisli o predlaganih vprašanjih o namenu izobraževanja, je bila uporabljena kvalitativna raziskovalna zasnova. Ugotovitve vzbujajo pomisleke glede zanesljivosti in ustreznosti rezultatov, saj ti niso enako informativni in konsistentni pri različnih LLM, razlike pa se pojavijo celo pri večkratnem preizkušanju istih vprašanj. Trenutno ni soglasja o optimalnem pristopu k vključevanju umetne inteligence v izobraževanje niti o morebitnem vplivu umetne inteligence na učenje, poučevanje, delo in družbo. Čeprav se zdi, da je tveganja, povezana z UI, mogoče obvladovati, je trenutno ključnega pomena usposabljanje za uporabo teh modelov, saj bodo ti modeli pomembno vplivali na številna področja izobraževanja.

Ključne besede: izobraževanje, veliki jezikovni modeli, digitalizacija, poučevanje in učenje, namen izobraževanja

Elektronski naslov: matej.urbancic@cpi.si