JKAU: Comp. IT. Sci., **Vol. 13** No. 1, pp: 1 – 31 (2024 A.D.)
DOI: 10.4197/Comp.13-1.1

Task-Oriented Authoring Tool Using ChatGPT to Create Educational Textbooks

Miada Ahmeddeb Almasre[1] and Alanoud Talal Subahi[2]

[1]*Faculty of Computers and Information Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia*
[2]*Faculty of Computing and Information Technology, Department of Information Technology, King Abdulaziz University, Rabigh, Saudi Arabia*
malmasre@kau.edu.sa,
asubahi@kau.edu.sa

Abstract. the accelerated development of technology has led to the emergence of cutting-edge smart tools, such as artificial intelligence (AI) chatbots and machine learning algorithms, which possess substantial potential for improving learning and education. Conventional content creation tools frequently lack these sophisticated features, rendering the incorporation of AI, including OpenAI's ChatGPT, an appealing area to investigate. This study aims to assess the effectiveness, cognitive load, usability, and potential challenges of a task-oriented authoring tool integrated with ChatGPT for producing personalized educational content. Design considerations using the SDLC Waterfall Model and prompt engineering were discussed. The research involved a total of 25 participants: experts (n=5) and novices (n=20), who utilized the authoring tool to generate academic content. A 5-likert questionnaire that consisted of 41 items was designed to investigate the users' agreement about the tool's effectiveness, cognitive load, usability, and AI-associated challenges, with mean comparison and t-tests being used for analysis. The primary findings revealed overall positive impressions among users, particularly concerning the tool's efficiency and cognitive load management. Nevertheless, small differences in usability perceptions arose between experts and novices. These findings provide valuable insights for refining and augmenting AI-integrated authoring tools to better accommodate varying user requirements in the educational domain.
Keywords—ChatGPT, Authoring tool, E-learning, AI task-oriented, personalized educational content.

I. INTRODUCTION

the constant demand for up-to-date information exacerbate these challenges. Consequently, stakeholders are constantly searching for innovative ways to optimize resources and manage expenses,

The utilization of artificial intelligence (AI) driven solutions in the realm of education has the potential to revolutionize content creation and management processes, making them more efficient, targeted, and learner-centric. By incorporating AI technologies such as natural language processing, machine learning, and advanced analytics, educational stakeholders cannot only expedite the development of tailored learning materials but also gain insights into individual learners' needs, progress, and potential areas for improvement [1]. Such insights can enable educators to better adapt their teaching strategies, implement targeted interventions, and foster a more inclusive learning environment that addresses the diverse needs of all learners, including those from underrepre- sented or disadvantaged groups [2]. Considering the current state of affairs, educational institu- tions, educators, and libraries continuously face the challenge of providing quality educational content while adhering to tight bud- gets and resource

constraints. Inefficiencies in content creation, high costs associated with acquiring or developing materials, and
thereby enhancing overall educational processes.

Traditionally created content may not always meet the needs of diverse learners or address current educational standards and trends. However, the significant concern for stakeholders in the education sector is in time-consuming of creating high-quality, accessible, and engaging educational content. Current traditional content creation methods may not involve significant time invest- ment, impacting both educator productivity and the speed at which new resources are made available. Although inclusive education aims to ensure equal opportunities for all learners, creating tailored content that caters to a wide range of individual requirements is often resource intensive [3]. Thus, the diverse needs of learners with different abilities, linguistic backgrounds, and learning pref- erences are a central challenge in the field of education. Therefore, there is an inherent need for approaches that increase efficiency and expedite content development without compromising quality and adherence to curriculum standards. This necessitates the con-tinuous exploration of new methods and innovative solutions to
facilitate the development of accessible, adaptable, and differen- tiated content that promotes greater inclusively in education and enhances learning outcomes for all students [4].

The emergence of AI-driven solutions presents a promising opportunity to address some of the persistent challenges faced by educational institutions, educators, and libraries, particularly in the area of content creation. Within this context, the advent of large language models, particularly ChatGPT, serves as a promising ad-vancement in the integration of AI within educational systems [5]. ChatGPT is a cutting-edge AI language model developed by OpenAI, which has generated significant interest in the fields of natural language processing and machine learning. Trained on vast amounts of textual data and utilizing deep learning algorithms, ChatGPT demonstrates a remarkable ability to comprehend con- text, grasp language nuances, and generate coherent, relevant, and engaging textual output. Originally designed for conversational AI applications, its versatile capabilities extend to various domains, including content generation, sentiment analysis, and summariza- tion. As this language model continues to evolve, its potential for transforming industries such as education is becoming increasingly apparent.

Therefore, to investigate the potential of AI solutions in ad- dressing the previously discussed research problem, this paper aims to:

• Develop an innovative AI task-oriented authoring tool that integrates OpenAI ChatGPT to assist subject matter experts (SMEs) in efficiently producing high-quality educational con- tent.

• Evaluate the effectiveness, cognitive load, and usability of the developed authoring tool when used to create educational materials in the context of identifiable challenges.

To achieve the objectives of this study, the researchers aim to design, develop, and evaluate an AI-driven authoring tool integrat- ing chat-based language models like ChatGPT to cater to SMEs in creating educational content. The assumption is that adopt- ing the waterfall system development life cycle model (SDLC) ensures a systematic and well-structured approach. This is partic- ularly useful in designing AI solutions for the educational sector, as it promotes increased efficacy, usability, and reliability of the AI-driven authoring tool, consequently contributing to enhanced learning experiences and greater adoption by SMEs and educa- tional institutions. By following the SDLC process, the study will identify development requirements and constraints, incorporate a

chat-based language model, design user-friendly interfaces, ad- dress key educational content creation challenges, and assess the tool's potential in fostering capacity building and professional de- velopment. The culmination of this research will be an evaluation framework assessing the AI-driven authoring tool's performance and effectiveness, providing insights for future research and devel-opment in the field of AI applications in education.

As we will be evaluating the performance of the AI task- oriented authoring tool considering the perspectives of users, we will be investigating three main hypotheses:

1. There is no significant difference ($p > 0.05$) in the effective- ness of the task-oriented authoring tool between expert and novice users.
2. There is no significant difference ($p > 0.05$) in the cognitive load associated with using a task-oriented authoring tool be- tween expert and novice educators.
3. There is no significant difference ($p > 0.05$) in the usability of the task-oriented authoring tool between expert and novice users.

The remainder of this paper is structured as follows: Section 2 provides an overview of the related work, focusing on the use of ChatGPT for educational purposes and the process of engineering the prompt. In Section (3), we present different phases used in our methodology to develop the proposed task-oriented authoring tool, followed by an evaluation stage where we measure the ef-fectiveness, cognitive load, and usability of the tool in Section 4. Finally, we discuss our findings and provide recommendations for our future research in Section 5.

II. LITERATURE REVIEW

This section will discuss recent studies that focus on three main points. Firstly, we will explore studies related to AI Generative models. Secondly, we will focus on studies that used ChatGPT for content creation in academia. Finally, we will discuss studies that employed the ChatGPT Prompt for designing and evaluation purposes.

A AI Generative models

Multimodal AI progress has provided individuals with potent means of generating text and images through text. Recent research has demonstrated that text-to-image creations can depict diverse topics and artistic techniques [1]. Such advancements have facili- tated addressing various users' needs, especially in the field of ed- ucation and skill enhancement, as these innovations assisted in the development of technical, creative, and motivational approaches [3].

DALLE 2, for example, is a smaller version of OpenAI's Large Language Model (LLM) that uses NLP and diffusion techniques to generate various styles of art from a text prompt. The 2021 DALLE 2 release has demonstrated better results compared to the earlier version of the same model especially with regards to the breadth and quality of generated art. A similar innovative release, DALLE 3 has been introduced in 2023 which ultimately became the driving force in various applications which rely on effective visual design like gaming and entertainment [6].

With the same objective in mind, in late November 2022, Ope- nAI released ChatGPT (the GPT 3.5 model), which has captured the attention of various industries including business, healthcare, entertainment, and education [2, 7]. ChatGPT's ability to effi-ciently perform complex tasks within the field of education has

caused mixed feelings among educators, as it appears to challenge existing educational practices [4]. ChatGPT's foundation lies in a Generative Pre-trained Transformer (GPT), which utilizes an ex- tensive amount of publicly available digital content data NLP to create human-like text in multiple languages. Its writing capa- bilities range from a paragraph to a full research article and can convincingly (or almost convincingly) cover a wide variety of

top- ics [8]. In less than a week after its release, ChatGPT has already amassed over one million subscribers due to its potential to revolu- tionize various professions. Developed using OpenAI [9, 10], this chatbot can converse like a person, and users can interact with it by inputting prompts based on OpenAI’s language model. As studies have shown, the limitless potential of ChatGPT in the field of ed- ucation has yet to be fully explored. We believe that ChatGPT can be utilized by academics to create innovative learning resources.

We will divide these studies into two phases: the first phase examines the studies that used ChatGPT in developing new tools for libraries and content creation in academia, while the second phase examines the studies that evaluate the effectiveness of ChatGPT in education through prompt engineering and evaluation.

B ChatGPT and content creation in academia

Currently, NLP studies have started to address novel applica- tions and methodologies due to the recent developments in the field of Large Language Models (LLMs). Such models were very instrumental in the development of artificial intelligence applica- tions that emulate human behavioral and linguistic styles with near distinguished accuracy. This led to their recent constant integra- tion in tasks like text summarization, translation, and automatic content generation [11].

For instance, the researchers in [12] examined the possible ap- plications of ChatGPT in educational contexts and proposed that it might be used as a replacement to search engines by learners and educators. Such technologies allow students accessibility to aggregated learning content and facilitate educators’ content cre- ations. Nonetheless, issues related to unethical usage cases might lead to worry about academic integrity. The researchers address this concern by recommendation that educators encourage the use of ChatGPT, for example, as a supplementary tool or resource for knowledge not as a replacement for actual practice and learning on the part of the student.

Researchers in [9], assumed that potentially ChatGPT can cre- ate a revolution in the fields of education and libraries, which can be perceived both positively and negatively. To investigate their assumption, they interviewed ChatGPT, which yielded positive re- sponses related to its capability to enhance libraries search services, help develop automated reference services, generate pipelines for cataloging and metadata assignment to library references. The negative responses where very much related to ethical and privacy issues when deploying ChatGPT application as part of the library service offering.

In a recent literature compilation by the authors of [8], the po- tential benefits of ChatGPT for enhancing teaching and learning

were explored. These benefits include personalized and interac- tive learning, formative assessment practices, and more. However, limitations such as the generation of misinformation, training data biases leading to existing biases, and privacy issues were also high- lighted. The study offers recommendations for maximizing the use of ChatGPT in teaching and learning. Policymakers, researchers, educators, and technology experts are encouraged to collaborate and discuss safe and constructive ways to utilize these advanced generative AI tools to support student learning and improve teach- ing practices.

The objective of the research conducted in [13] was to eval- uate the satisfaction level of users who utilized chatbot applica- tions within library settings with a special focus on comparing traditional library chatbots and ChatGPT. Findings of the research demonstrated that participants preferred using the chatbot, but they commented on issues related to privacy and capability of handling complex operations. To investigate this further, the re- searchers compared traditional library chatbots to ChatGPT con- sidering prompts used to reference and generate articles. The study

offered librarians a view of the advantages and disadvantages of using such tools, especially with regards to insights which chat- bots might offer them if successfully integrated into their reference services.

The study in [14] discusses the potential of deploying Chat- GPT applications to replace the traditional knowledge-based type of chatbot often used in libraries and information centers (LICs). The researchers suggested that using ChatGPT can enhance the process of information retrieval for users, ultimately improving the quality of library services. They addressed, nonetheless, some limitations of implementing ChatGPT like training data bias and outdated information as it might affect the accuracy and relevance of generated responses. Potentially, a ChatGPT application in li- brary context might be effective in transforming the library service provision, but this study asserts the importance of considering the challenges of implementation as well.

The researchers in [15] conducted a survey study to investigate the perceptions of library and information science professionals of ChatGPT. Basically, the researchers implemented a content anal- ysis methodology of the comments and content shared on social media by those professionals. They have considered the poten- tial as well of ChatGPT from the perspective of academics who worked on enhancing their writing in terms of language usage and structure.

Similarly, the authors of [16] presented a qualitatively designed case study to investigate educational uses of ChatGPT in education through three stages. Basically, the study studied three aspects of ChatGPT use. Firstly, they discovered via social media interactions that there is a positive view of ChatGPT's future potential in educa- tion, Secondly, they examined in depth stage examined ChatGPT's impact on education considering the quality, usefulness, personal- ity, emotion, and ethics related to generated responses and content. Thirdly, they conducted ten educational scenarios to study user ex- perience. These experiments demonstrated academia's concerns over academic integrity, plagiarism, privacy issues. The study recommended conducting further research to ensure safe and re- sponsible applications of ChatGPT in education.

[17] " A Framework for Applying Generative AI in Educa- tion," is a detailed analysis of how ChatGPT's applications can be used in educational settings. This paper introduces the 'IDEE' framework, which emphasizes the integration of generative AI, like ChatGPT, in education by focusing on outcomes, automation levels, ethical considerations, and effectiveness evaluation. The research highlights ChatGPT's potential in personalizing learning and providing efficient feedback. However, it also acknowledges the challenges that come with it, such as unproven effectiveness, data quality concerns, and ethical issues. This work emphasizes the need for further research and policy development to incorporate ChatGPT effectively in educational contexts.

The study [18] involved 143 students from 7 online college- level chemistry courses who generated short answer questions re- lated to their current learning content. The study assessed the quality of these questions using both human and automatic meth- ods, including GPT-3. It was found that 32% of the questions were of high quality and could be used directly in the course, and 23% assessed higher cognitive processes according to Bloom's Tax- onomy. The study recommends combining expert and automatic evaluation methods for better results.

C ChatGPT Prompt Designing and Evaluation

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template

measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Designing prompts, which is sometimes referred to prompt engineering, for book authoring using ChatGPT requires a careful balance between specificity and flexibility. The primary goal is to elicit relevant and targeted content while allowing the tool to generate well-structured and coherent text.

The authors of [19]have introduced a comprehensive frame- work for documenting and implementing prompt patterns for large language models (LLMs) such as ChatGPT. These patterns are de- signed to offer practical solutions to common issues faced by users while interacting with LLMs for various tasks. Similar to software patterns, prompt patterns are classified into different types, and Prompt Improvement is one of them. This category specifically emphasizes enhancing the quality of LLM conversations, both in terms of input and output. The authors have also provided ex- amples of how prompt patterns can be combined to create more effective prompts.

The framework proposed by the authors of [20] provides a quantitative evaluation method for interactive LLMs like Chat- GPT using publicly available data sets. They studied the language comprehension ability of ChatGPT across three different languages

from various language categories in NusaX, English, Indonesian, and Japanese. The authors conducted a comprehensive technical evaluation of ChatGPt, using 23 data sets that cover eight common NLP application tasks. They found that ChatGPT is best suited for open-domain dialogue tasks, but they also explored how its emergent abilities and interactivity could potentially be useful for task-oriented dialogue.

III. METHODOLOGY

To achieve the objectives of this study, the researchers designed a task-oriented tool that utilizes the text generation capabilities of ChatGPT to assist authors in creating academic content, specifi- cally textbooks or teaching textual packages that can be used in educational contexts. The following subsection will explain the stages used to develop such a tool.

A Development of the task-oriented authoring tool

The creation and design of a new task-oriented textbook au- thoring tool involves several stages, with the development of an intuitive user interface being a crucial step. Waterfall Software Development Life Cycle (SDLC) is a linear model for software development that follows a sequential process from requirements gathering to maintenance. This approach is best suited for projects with well-defined and predictable requirements. Waterfall model has distinct phases [21], as follows:

A.1 Requirement Gathering and Analysis phase

This is the first phase in the waterfall SDLC model, which presents findings of the requirement gathering and analysis phase for a ChatGPT-based educational content generation tool. Insights were obtained from key stakeholders involving educators, content creators, and subject matter experts. Interviews with educators provided specific requirements on tool features and need assess- ment. A use case was developed, focusing on an educator teaching an Arabic course in need of content creation. Functional require- ments include content generation, multilingual support, content export options, and visual and multimedia content creation. Non- functional requirements emphasize effectiveness, cognitive load reduction, and usability. The tool should also provide guidance on content organization, layout, and hierarchical structure for bet- ter learner engagement. Prioritization was determined based on the expected impact,

placing content generation as a high priority. This process provides a solid foundation for further development using the Waterfall SDLC, ultimately ensuring the tool meets users' diverse needs.

During the requirement gathering and analysis phase, inter- views were conducted with experts in the field of authoring to assess their needs and challenges in the process.

• The first requirement identified was the need for specific, ded- icated authoring tools, particularly for the Arabic language.

Currently, authors rely on basic text editors, open-source im- agery, and collaborations with designers, but these methods do not fulfill all their needs. The development of tailored au- thoring tools, complete with translation and linguistic proof- reading functions, would be highly beneficial.

• The second requirement addressed time management chal- lenges that authors face during their projects. The lengthy process of content development and organization may lead to project delays or abandonment. Solutions to assist authors in staying focused, organized, and efficient throughout the project's duration are also a critical need.

• The third crucial requirement gathered from the interviews is the importance of considering the goal of the content, tar- get audience, and main themes while starting an authoring project. This insight highlights the importance of incorporat- ing these aspects into the proposed authoring tool, allowing authors to keep their focus aligned with their initial intentions.

Based on the results of the interviews, the provision of a user- friendly tool through a website with minimal technical require- ments emerges as a significant need for authors. The tool should facilitate strategic interaction with GPT elements while consider- ing authoring terminology and the constraints of dialogue window lengths. To optimize prompt engineering, the collaboration of specialist authors should be integrated into the Requirement Gath- ering and Analysis phase, ensuring the effective use of essential phrases and terms in authoring-related tasks. The gathered re- quirements from this interview were categorized into functional and non-functional [22] requirements are:

• Content generation: The tool must be able to generate educa- tional content in Arabic/English using ChatGPT technology, encompassing a wide range of materials and resources.

• Customization: the designed tool should offer access to tem- plates that can be customized as per the requirements of dif- ferent types of educational content

• Integration: The authoring tool must be designed in a such a way that allows for its integration with standard instructional methodologies, learning management systems (LMSs), and learning analytics systems.

• Collaboration: The tool should include features that allow for real-time interactivity and collaboration between instruc- tors and instructional designers to streamline the process of content creation.

While the non-functional requirements are:

• Usability: Form a user's perspective the tool must be per- ceived as user-friendly and easy to navigate, emphasizing that it should be so to users coming from various backgrounds of technical expertise.

• Accessibility: To ensure all users being able to access the tool, it should be working and accessible via multiple devices and platforms.

• Performance: Without affecting the quality of generated con- tent, the tool must have a high performance in terms of how fast it generates content.

A.2 System Design Phase

According to the waterfall SDLC model, the second stage of system design includes developing a blueprint of the system's main structure, as well as creating relevant flowcharts, and required sequence diagrams.

The tool's structure consists of three layers: the user, i.e., the author, the task-oriented authoring tool interface, and the ChatGPT API. Figure 1 shows the flowchart depicting the sequence of steps an author follows to use the task-oriented authoring tool. First, the tool must be provided with the primary inputs, which are the book's title and the target group's level to read such a book. After that, the tool will formulate the prompt based on the previous inputs, which will be explained in detail in the programming phase 3.

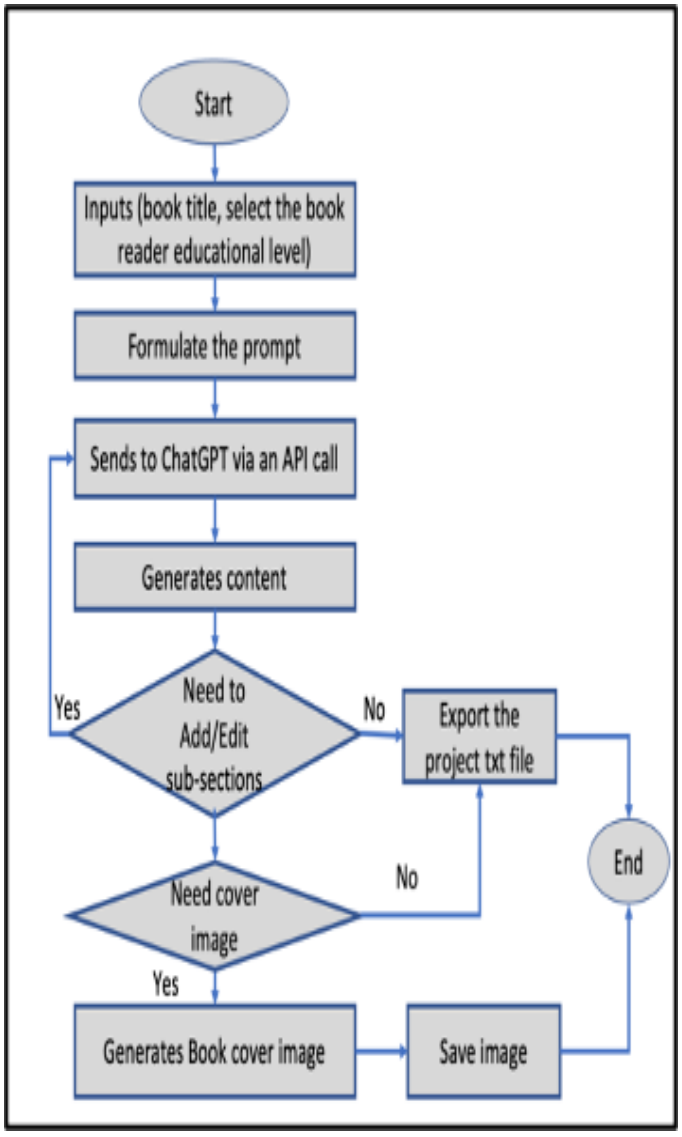


Fig. 1. Flowchart detailed step-by-step process for using the task- oriented authoring tool

Sequence Diagrams are interaction diagrams that illustrate how operations are executed. They depict the communication between the user (author) and ChatGPT and/or DALEE models during a collaboration. Figure 2 illustrates the sequence diagram of cre- ating educational content, while Figure 3 illustrates the sequence diagram for image generation and steps on

how users can save images on their device for later use in the final book cover design.

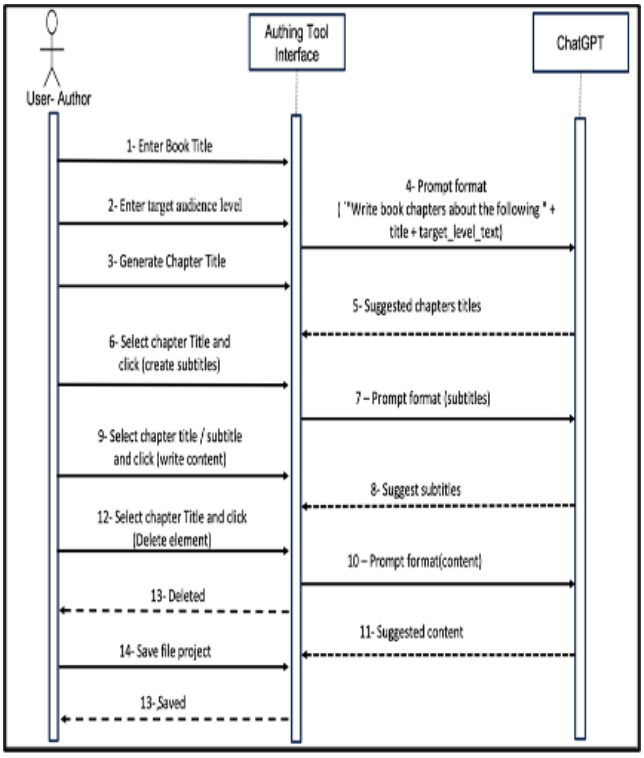


Fig. 2. Content generating sequence diagram.

Moreover, this phase contains the detailed designs of the in- terfaces to ensure that the different components can communicate effectively. Figure 4 shows the main page of the tool, which dis- plays many sections. First of all, the author needs to select their preferred authoring language, i.e., Arabic or English. Second of all, the author should read carefully the instructions on how to ef- fectively utilize the interface for producing high-quality authored texts. Third of all, the author must provide two entries (steps No.3 and No.4 as shown in Figure 4 in order to proceed with content authoring.

The first input is providing a suggested title for the book. The second input is to select from the drop-down list the appropriate ed- ucational level which corresponds with the intended audience (e.g., primary school stages, upper school stages, university stages). By clicking on the ”Generate Chapter Titles” button (step No.5 in Fig- ure 4, a prompt is formatted and compiled to communicate with ChatGPT to generate several chapters’ titles. The prompt format- ting process is explained in detail in the subsequent paragraph. Once chapter titles have been generated, users can edit and gen- erate more text cumulatively, using interface details outlined in Appendix section6.

Furthermore, the tool allows the authors to add or edit sub- sections within the book’s chapters to provide a smooth and effi- cient user experience while creating Arabic textbooks with the help of ChatGPT. In steps No.6 and No.7 Figure 4, users can generate

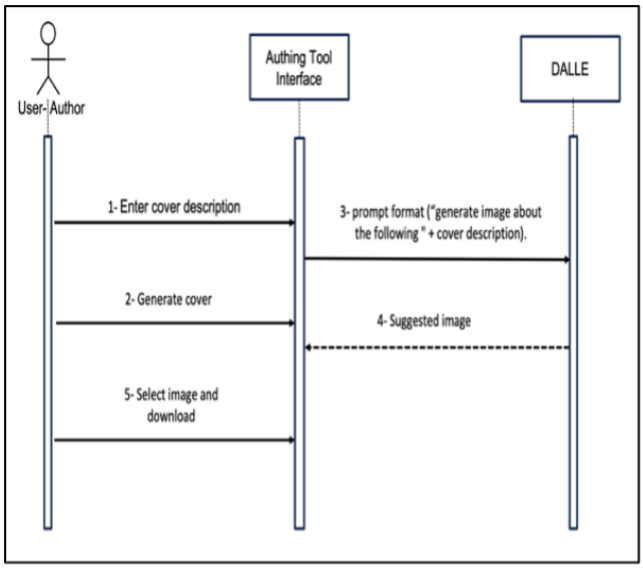


Fig. 3. Cover image generating sequence diagram

a suggested image for their educational content cover by entering text expressing their desired image and then clicking on "Generate Cover." This request is sent to the smart DALLE model, which will generate an image from a text sentence. Finally, in step No.8, users can save project files as text divided by chapter titles and contents authored within each chapter, as shown in Figure 5.

To briefly recap the previous steps, the authoring tool sends the formulated prompt to ChatGPT via an API call, and the user can interact with generated content from ChatGPT and generate more content. ChatGPT processes the prompt and returns the generated content, which is streamed into the authoring tool. Users can add or edit sub-sections within the book's chapters, further refining the content and structure of the book, then export the book in a text format.

A.3 Implementation and Coding Phase

The Implementation and Coding phase is the third phase in the waterfall SDLC model; it involves the actual coding of the software. This phase takes the design created from the previous phase and turns it into a functional component, i.e., the integration of ChatGPT and the formulated prompt.

Integrating ChatGPT: The first component involved integrat- ing ChatGPT (Gpt-3 5-turbo model) as the primary content gen- eration engine for the authoring tool. The GPT series of lan- guage models have been iteratively improved, culminating in the advanced GPT-3.5-turbo model. This version retains the rich ca- pabilities of GPT-3 while offering significant cost-efficiency im- provements in terms of tokens, making it a fitting choice for var- ious applications. GPT-3.5-turbo is a massive AI model that has been trained on an extensive collection of internet text data. This training process allows the model to gain a deep understanding of language structures, context, grammar, and even world facts

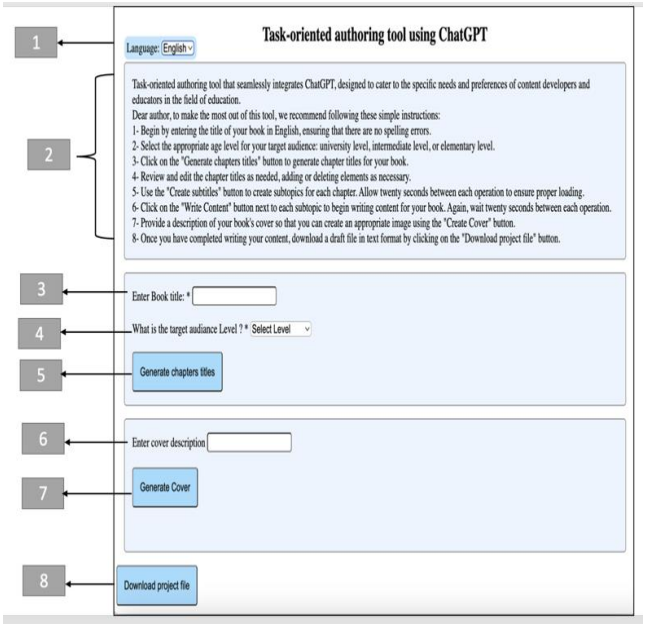
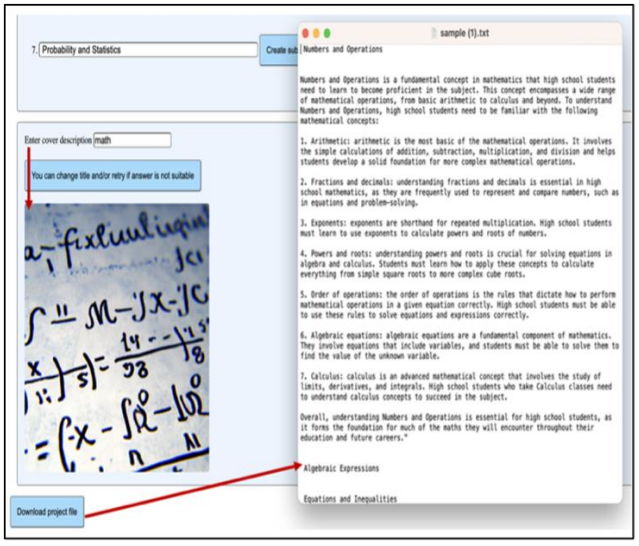


Fig. 4. The main page of the Authoring tool



Fig, 5. Prompt to generate cover image and save the file in text format

[22]. The model learns to perform a wide array of tasks solely based on the textual information it was trained on and can operate without any additional fine-tuning for specific tasks. The model is equipped with state-of-the-art language processing capabilities, such as understanding grammar, context, tone, and style. This makes it particularly suitable for generating text that matches the user's requirements and mimics different writing styles. GPT-3.5- turbo models have a token limit of 4096 tokens for both input and output during an API call. If the input text surpasses this limit, you will need to truncate or reduce the content. The general steps in subnet a prompt in the proposed tool is:

1) A list of messages is initialized with system, assistant, and user prompts: a system message is crafted to provide the AI assistant's role and high-level guidance for generating book

content. An optional assistant message is included in the example to describe the AI's readiness to generate content according to user requests. A user message is formulated with specific instructions to generate Arabic chapter titles for the selected book title and target age group.

2) A call is made to the GPT-3.5-turbo API using the 'ope- nai.ChatCompletion.create()' method, which takes the list of messages as input.

3) The API returns a response that includes the generated chapter titles in Arabic.

4) The response is parsed to extract and print the generated
chapter titles.

Key capabilities offered by GPT-3.5-turbo include:

• Language understanding: GPT-3.5-turbo demonstrates a re- markable ability to comprehend and process natural language. It can interpret user input and generate creative, contextually appropriate output that closely aligns with the user's intent.

• Generalization: The model exhibits a strong aptitude for gen- eralization, which means it can intelligently infer concepts and apply them across a range of contexts. This characteristic comes in handy when generating book content that caters to different audience types and age groups.

• Task versatility: Owing to its language understanding and generalization capabilities, GPT-3.5-turbo can manage tasks beyond content generation. Additional tasks it can perform include text summarization, translation, sentiment analysis, question-answering, and code generation.

Formulate the Prompt The second component focused on programming efficient prompts for the ChatGPT API using the guidelines mentioned previously. The ability of AI-based mod- els like ChatGPT to generate high-quality content significantly depends on creating an efficient and precise prompt. A compre- hensive approach to prompt design must incorporate language, the subject of the book, the target audience's age, user messages, and overall prompt formation to produce contextually accurate and age- appropriate content. A more detailed discussion of these crucial aspects of prompt design follows:

• Selecting an appropriate language through the 'lang' parame- ter, such as "ar" for Arabic or "en" for English, ensures that the generated content is suitable for the desired linguistic back- ground. This approach makes the textbook content inclusive and accessible to readers from diverse language communities.

• Including the 'book title' parameter enables the AI model to concentrate on specific topics or themes related to the book's subject matter, ensuring that the generated content aligns with the intended educational objectives.

• To produce age-appropriate content, the 'target-level' param- eter must account for various educational stages, such as "target-college" for college students, "target-high-school" for high school students, and "target-primary" for primary school students. Aligning the content with the cognitive and learning capabilities of the target age group is essential for effective learning material.

In this way, the prompt formation and integration of user mes- sages further personalize and customize the content to meet spe- cific user requirements. By considering users' queries, sugges- tions, or feedback within the prompt, the AI model can generate even more focused and refined content that caters to the user's unique needs and preferences. Constructing a prompt that seam-lessly combines language, title, and target-level parameters as user messages ensures the generation of tailored content that matches the desired context and audience. For example, a comprehen- sive prompt may follow this format: '"Write book chapters about the following " + title + target-level-text.' By using this forma- tion, the AI model can produce

subject titles, chapter names, and content adhering to the specified language, subject, age group, and user input. Implementing these key considerations in prompt design allows AI-based models like ChatGPT to generate meaningful, level-appropriate, and contextually precise content for edu- cational textbooks spanning an array of subjects and languages while taking user messages into account for more customized output. We provide the complete code of this tool available on GitHub, enabling any researcher to utilize it by following this link: https://github.com/Malmasre/AuthoringChatGPT

A.4 Testing and Maintenance phase

The testing phase of the waterfall model focuses on verifying the functionality of each requirement for the task-oriented tool uti- lizing ChatGPT technology. Overall, the testing phase seeks to establish that all these functional requirements are met satisfac- torily before proceeding to the final implementation of the tool. During this phase, it will be crucial to ensure that content gen- eration adequately covers a wide array of materials and resources while maintaining quality and accuracy. The customization feature should be thoroughly examined to confirm that the templates can effectively adapt to various educational content types.

B Evaluation the Task-oriented authoring tool

The study aims to evaluate the Task-oriented authoring tool through examining four facets of the tool's usage: effectiveness, cognitive load, usability, and challenges which users face while generating various educational content types, text and images. This aim is realized through designing and deploying a survey that gathered the responses of 25 participants coming from different academic backgrounds, like public schools' teachers, university faculty, and postgraduate students. This sample of users were included in the study considering their expertise in the field of education and to allow for various types of academic content to be generated and evaluated. The sample has 4 male participants and 21 females, who have intermediate (6 users) to advanced (19 users) computer skills. They demonstrated a considerably high level of computer proficiency (76%). Furthermore, the majority of users had experience in creating teaching content in Arabic, and 17 participants had previously authored educational textbooks. Appendix A contains four sections of the evaluation survey that pertain to the research objectives and utilize parametric statistical tests to evaluate the data. The four sections are:

1) Effectiveness: A 5-point Likert scale survey was used to assess participants' experience with the authoring process and their satisfaction with the generated products (i.e., textbooks). We conducted a t-test and Fisher's Combined Probability Test to analyze the data for significant differences in effectiveness.
2) Cognitive Load: A 5-point Likert scale survey was employed to measure the cognitive load experienced by users during the authoring process, and the statements in the survey followed the NASA-TLX, which is a tool used to estimate workload from one or more operators while they perform a task or immediately afterward [23], [24].
3) Usability: In this section, participants rated various aspects of the tool, such as its ease of use and learnability, using a 7-point ranking scale survey. The BOT Usability Scale, BUS-15 [25], then the data were analyzed using a t-test.
4) Challenges: A 5-point Likert-scale survey was used to evalu- ate the challenges encountered by users when using the proposed tool. We calculated the means of the responses to provide an indication of the most recurrent issues faced by users.

For each evaluated aspect, we conducted the respective statistical tests, such as the t-test. Once the t-statistic is calculated, it can be compared to the critical value from the t-

distribution table or a t-distribution probability function to obtain the p-value. A p-value that is smaller than 0.05 will indicate that we can reject the null hypothesis, as well as the presence of a significant difference when considering the means of the two groups. Implementing these tests provided us with insights about the performance of the Task-Oriented Authoring Tool in terms of its effectiveness, cognitive load, usability, as well as the challenges which users faced.

IV. RESULTS AND DISCUSSION

This section presents our findings with regards to testing the tool's effectiveness, cognitive load levels, usability, and challenges experienced by the users.

A Tool Effectiveness

Basically, our research targets the evaluation of the effec- tiveness of an AI-task-oriented tool enhanced with ChatGPT, considering both process effectiveness (the actual generation process) and the final product effectiveness (generated content). Our investigated the hypothesis is: "There is no significant difference (0.05) in the effectiveness of the task-oriented authoring tool between expert and novice users." The agreement levels between expert and novice users were assessed, and possible statistical differences between these groups were investigated using a t-test. A Likert scale was used to rate the level of agreement or disagreement for each challenge, with the following ranges: strong disagreement (1-1.8), disagreement (1.9-2.6), neutral (2.7-3.4), agreement (3.4-4.2), and strong agreement (4.2-5). Initially, we examined the normality of the distribution of the responses related to effectiveness assessment. For the categories related to effectiveness, the "Effectiveness (Process)" displays non-normal distribution for both expert and novice groups according to the Shapiro-Wilk test, with significance values of .021 and .003, respectively. In the "Effectiveness (Product)" category, both expert and novice groups exhibit normal distribution, with significance values of .770 and .138 in the Shapiro-Wilk test.

Table I demonstrates the results related to process effective- ness, which includes aspects such as writing process facilitation, organization, and guidance provided, both expert and novice users' overall mean scores fell within the "Agree" range on the Likert scale. Expert users' mean scores ranged from 3.8 to 4.2, while novice users' mean scores ranged from 3.4 to 4.15. These mean scores indicate that both user groups were satisfied with the tool's features in aiding the writing process, as it provided a smooth user experience and supported various authoring functionalities.

In terms of product effectiveness Table II, which pertains to the quality of the authored content, factors such as content reliability, comprehension, organization, and natural language were assessed. The overall mean scores for both user groups once again fell within the "Agree" range, with expert users' mean scores ranging from 2.6 to 4.0 and novice users' mean scores ranging from 2.35 to 3.95. These scores highlight that users were satisfied with the final output generated by the AI-task-oriented tool, suggesting that the content created with the help of the tool was generally perceived as well-structured, comprehensible, and reliable. In addition, the p-values for all comparisons between expert and novice users were higher than the 0.05 threshold. The absence of significant differences between the two groups indicates that both sets of users had consistent experiences with the AI-enhanced authoring tool in terms of process and product effectiveness, suggesting that the tool is adaptable and beneficial for a wide range of users.

Figure 6 presents the results in a boxplot format, illustrating the distribution and central tendency of process and product effectiveness scores for both expert and novice users.

Table I. T-Test Results (Effectiveness – Process)

Item	t-test	p-value	Overall Mean	Likert Agree- ment
The tool allows writing in the language of your choice.	0.280	0.782	4.04	Agree
The tool facilitates the writing process (such as selecting a book title, creating chapters).	-0.221	0.827	4.12	Agree
The tool provides options for the author (creating chapter titles, creating topics, writing content, deleting, etc.).	-0.413	0.684	4.04	Agree
The tool encourages organization during the writing process.	-0.135	0.894	3.88	Agree
The interface and writing options of the tool increase focus by reducing distractions.	0.355	0.726	3.60	Agree
The tool provides guidance for the author (contextual help instructions and tips).	0.524	0.605	3.48	Agree
The tool provides an option to design a book cover.	-0.489	0.630	3.52	Agree
The tool facilitates downloading a draft of the written content.	-0.568	0.576	3.76	Agree

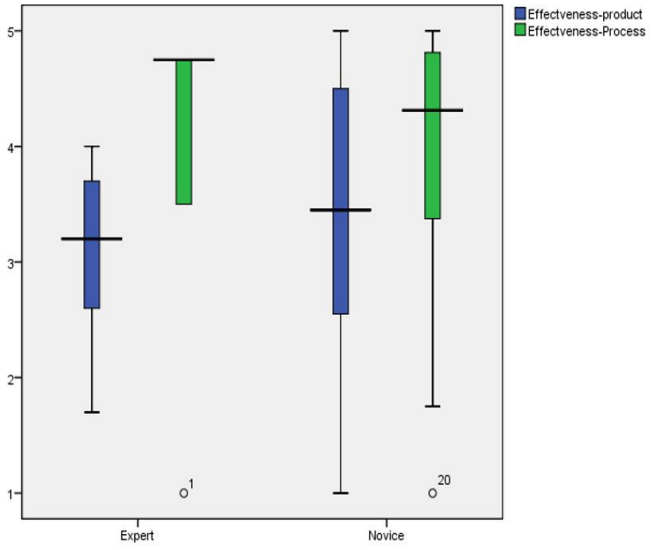
The boxplot displays the median score as a horizontal line within the box, which represents the interquartile range (IQR), encompassing the 25th percentile (Q1) and the 75th percentile (Q3). The whiskers extend from the box, denoting the variability outside the IQR, and the potential outliers are depicted as individual points beyond the whiskers.

For process effectiveness, expert users illustrated a mean score of 3.75 and a slightly lower median of 4.75, indicating data skewness and variability in the responses. In contrast, novice users exhibited a mean of 3.82, a median of 4.31, and a more consistent range of 4.00, suggesting overall satisfaction with the tool's assistance in facilitating the writing process.

Regarding product effectiveness, expert users demonstrated a mean score of 3.04, a median of 3.20, and a less diverse range of 2.30, reflecting more homogenous satisfaction with authored content. Similarly, novice users recorded a mean score of 3.40, a median of 3.45, and a range of 4.00, implying general contentment with the quality of the authored content.

Table II. T-Test Results (Effectiveness – Product)

Item	t-test	p-value	Overall Mean	Likert Agree- ment
The length of the written content is ap- propriate.	-1.446	.162	3.4	Agree
The written content is reliable.	-1.384	.18	2.96	Neutral
The written content is comprehensive.	1.517	.143	3	Neutral
The written content is organized.	0.192	.849	3.48	Agree
References are docu- mented in the written content.	0.463	.648	2.28	Disagree
Each chapter's writ- ten content relates to its topic.	0.88	.388	3.52	Agree
Reading the written content is easy.	-0.196	.846	3.92	Agree
Understanding the written content is easy.	-0.064	.949	3.84	Agree
The written content is natural language (mimics human- written content).	-0.063	0.95	3.64	Agree
The created cover re- lates to the entered description.	-0.991	.332	3.24	Neutral



Fig, 6. Boxplot of tool effectiveness results

In conclusion, both expert and novice users appreciated the features offered by the AI-task-oriented tool driven by ChatGPT for educational content creation, thus proving the validity of our initial hypothesis.

B Tool Cognitive Load

In this study, we aimed as well to investigate the

perceived cog- nitive load of an AI-task-oriented tool driven by ChatGPT, designed to help educators generate academic content, such as textbooks. Our hypothesis in this context is "there is no significant difference (0.05) in the cognitive load associated with using a task-oriented authoring tool between expert and novice educators." The levels of agreement between expert and novice users were evaluated, and potential statistical differences between these groups were ex- plored utilizing a t-test. A test of normality has been conducted on the users' responses which reveals that normality is observed for both expert and novice groups with values of .344 and .357 in the Shapiro-Wilk test.

Table III presents the results, which are based on a Likert scale, with lower values signifying stronger levels of agreement and higher values indicating stronger disagreement. As we are dealing with a reversed scale, the Likert ranges for agreement can be pre- sented as follows: (Strong Agreement= 1-1.8), (Agreement= 1.9-2.6), (Neutral= 2.7-3.4), (Disagreement= 3.4-4.2), (Strong Dis-
agreement = 4.2-5).

Overall, participants found that the task of using the AI-based, task-oriented authoring tool is mentally demanding, with a mean score of 2.6, placing the users' experiences in the neutral range. This suggests that both expert and novice users neither strongly agreed nor disagreed with the mental demand of the task. When looking at how physically demanding the task was, the overall mean score was 1.92, showing a disagreement. This indicates that users found the task not too physically demanding to complete. The pace of the task had an overall mean of 3.76, placing it in the agreement range. This suggests that users felt the task was somewhat hurried or rushed. Regarding the level of success in ac- complishing the task, the overall mean score was 3.68, placing it in the neutral category. This signifies those users had varying levels of success while using the authoring tool. Similarly, the overall mean for work needed to accomplish the desired performance level was 3.28, also in the neutral range, meaning that neither group had a clear agreement or disagreement on their effort level. Lastly, the overall mean score for users feeling insecure, discouraged, ir- ritated, stressed, and annoyed was 2.76. This puts the result in the neutral range, suggesting that participants had mixed experiences regarding their emotional state during the task.

In the reported results for cognitive load experienced by expert and novice users while utilizing the AI-task-oriented tool driven by ChatGPT, the p-value offers insights into the statistical signif- icance of the differences observed between the two groups. It is important to note that a low p-value (typically less than 0.05) indicates a statistically significant difference. However, based on the results, the p-values for all comparisons appear to be higher than the 0.05 threshold. This suggests that the differences ob-
served between expert and novice users concerning cognitive load are not statistically significant. In other words, the experiences of both groups seem consistent with each other when utilizing the AI-enhanced authoring tool based on large language models such as ChatGPT.

Table III. T-Test Results Cognitive Load

Item	t-test	p-value	Overall Mean	Likert Agree- ment
How mentally de- manding was the task.	0.547	0.590	2.6	Agree
How physically demanding was the task.	0.795	0.435	1.92	Agree
How hurried or rushed was the pace of the task.	0.321	0.751	3.76	Disagree
How successful were you in accomplishing what you were asked to do?	-1.469	0.155	3.68	Disagree
How hard did you have to work to ac- complish your level of performance?	-0.105	0.917	3.28	Neutral
How insecure, dis- couraged, irritated, stressed, and annoyed were you?	-0.685	0.500	2.76	Neutral

Figure 7 demonstrates that the cognitive load experiences of expert (N=5) and novice (N=20) users while utilizing a task- oriented authoring tool with ChatGPT were assessed. Experts had a mean score of 2.9, whereas novices had a slightly higher mean of 3.025, indicating a relatively neutral cognitive load for both groups. Though the median and range values were similar, the dispersion of cognitive load scores was higher among experts (Std. Deviation = 1.59252) compared to novices (Std. Deviation
= 1.19975). Overall, these results suggest that the cognitive load experiences of both expert and novice users were generally neu- tral, with minor differences in variability and precision of mean estimates owing to differing sample sizes.

Users exhibited variations in their agreement levels regarding the cognitive load experienced while using the task-oriented author- ing tool with ChatGPT. The pace of the task was perceived as hurried, but participants disagreed about its physical demand. Ad- ditionally, users showed diverse levels of agreement concerning their success, effort, and emotions throughout the task, reflecting the differences in individual experiences. However, and with re- gards to our hypothesis, it is proved that there are no significant differences between the two types of users.

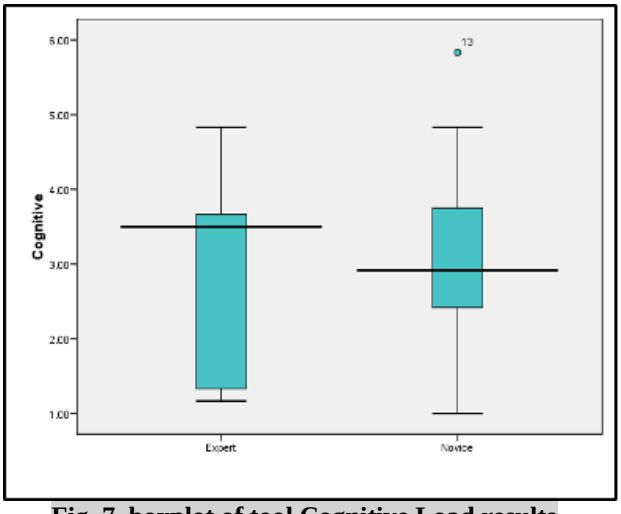


Fig. 7. boxplot of tool Cognitive Load results

C Tool Usability

In this study, we investigated the usability of a task-oriented authoring tool enhanced with ChatGPT based on several questions. Basically, our hypothesis assumes that "there is no significant difference $p > 0.05$ in the usability of the task-oriented authoring tool between expert and novice users". A Likert scale was used to rate the level of agreement or disagreement for each challenge, with the following ranges: strong disagreement (1-1.8), disagreement (1.9-2.6), neutral (2.7-3.4), agreement (3.4-4.2), and strong agreement (4.2-5). The analysis focuses on various domains, including learnability, efficiency, satisfaction, and errors.

Also, our primary objective was to determine if there are statistically significant differences between Expert and Novice users regarding their experience with usability while using the tool. We conducted a normality test of usability responses distribution for experts and novices across the domains of Learnability, Efficiency, Satisfaction, and Error, which show relatively similar distributions among both groups. While there are slight differences in the means for these categories, they do not exhibit any extreme deviations. Therefore, based on the available data, it can be cautiously concluded that the distribution of responses is approximately normal for both experts and novices in the usability categories assessed (Table IV).

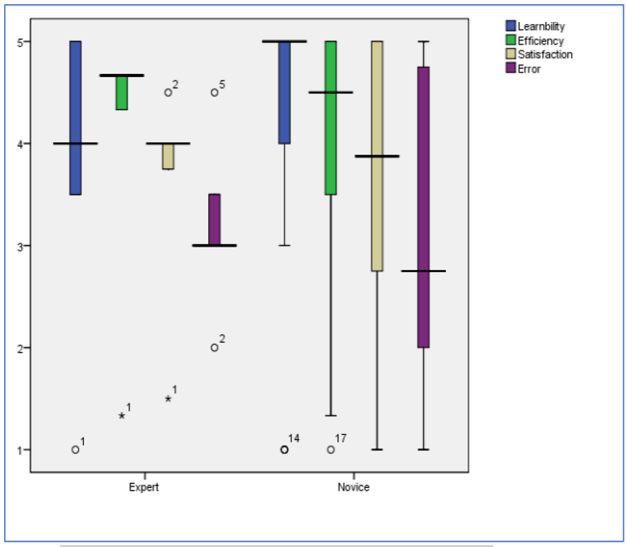
Regarding learnability, participants agreed that the chatbot function was easily detectable, with an overall mean score of 4.12. They also agreed that it was easy to find the chatbot, as reflected by the overall mean score of 3.96. In terms of efficiency, users generally found communication with the chatbot clear and easy to understand, as evidenced by the overall mean scores of 4.16 and 3.84, respectively. Additionally, users agreed that the chatbot was able to keep track of the context, with an overall mean score of

3.84. When assessing satisfaction, participants agreed that the chatbot understood their needs, achieved their goals, and provided an appropriate amount of information, with overall mean scores of 4.04 and 3.76, respectively. However, opinions were neutral regarding the chatbot's response accuracy, as indicated by the overall mean score of 3.28.

Lastly, in the domain of errors, participants showed neutral opinions on whether the chatbot informed them about potential privacy issues and the waiting time for responses, with overall mean scores of 3.04 in both cases. These findings suggest that users encountered a mix of positive and neutral experiences in these aspects of the chatbot's usability.

Based on the results, the p-values for all comparisons appear to be higher than the 0.05 threshold. This suggests that the differences observed between expert and novice users, concerning challenges and usability domains, are not statistically significant. In other words, the experiences of both groups seem consistent with each other when addressing challenges or utilizing the AI-enhanced authoring tool based on large language models such as ChatGPT.

Overall, the boxplot (Figure 8) representation of the usabil- ity domains reveals that novice users reported slightly higher mean scores in the learnability, efficiency, and satisfaction domains than expert users. However, expert users reported a marginally higher mean score in the errors domain. This suggests



Fig, 8. boxplot of tool Usability results

that the usability aspects of the AI-enhanced authoring tool, based on large language models such as ChatGPT, appear to be relatively close between the two user groups. Future improvements in the tool should prioritize addressing the identified challenges and enhancing user experiences in these domains to accommodate the specific needs of both expert and novice users.

Overall, the results prove our assumption and highlight the importance of addressing various usability domains to enhance user experience and promote the adoption of AI-enhanced authoring tools based on large language models such as ChatGPT.

Table IV. T-Test Results of Usability

Item	Domain	t-test	p-value	Overall Mean	Likert Agree- ment
The chatbot function was easily detectable	Learnability Q1	-0.198	.844	4.12	Agree
It was easy to find the chatbot	Learnability Q2	-0.922	0.366	3.96	Agree
Communicating with the chatbot was clear	Efficiency Q1	-0.263	0.795	4.16	Agree
The chatbot was able to keep track of the context	Efficiency Q2	0.619	0.542	3.84	Agree
The chatbot's responses were easy to understand	Efficiency Q3	-0.429	0.672	3.84	Agree
I find that the chatbot under- stands what I want and helps me achieve my goal	Satisfaction-Q1	0.280	0.782	4.04	Agree
The chatbot gives me the appropri- ate amount of information	Satisfaction-Q2	0.066	0.948	3.76	Agree
The chatbot only gives me the in- formation I need	Satisfaction Q3	-0.653	0.521	3.76	Agree

I feel like the chatbot's responses were accurate	Satisfaction Q4	-0.852	0.403	3.28	Neutral
I believe the chatbot informs me of any possible pri- vacy issues	Errors Q1	-0.070	0.945	3.04	Natural
My waiting time for a response from the chatbot was short	Errors-Q1	0.566	0.577	3.04	Natural

D Usage Challenges

The experiment has investigated the challenges faced by both expert and novice users when utilizing a task-oriented authoring tool enhanced with ChatGPT (Table V). The basic assumption and hypothesis in this context is that "there is no significant difference

(p ¿ 0.05) in the evaluation of challenges associated with using the task-oriented authoring tool between expert and novice users." A Likert scale was used to rate the level of agreement or disagreement for each challenge, with the following ranges: strong disagreement (1-1.8), disagreement (1.9-2.6), neutral (2.7-3.4), agreement (3.4- 4.2), and strong agreement (4.2-5). Initially, the normality of data distribution has been investigated. The expert group's responses show non-normal distribution in the Shapiro-Wilk test (.038), while the novice group presents non-normal distribution in both tests with values of .100 and .006. The analysis of the results focuses on the overall mean column and Likert Level for each challenge experi- enced by the users. The users' responses demonstrate that they have several concerns related to the use of AI tools in the author- ing of educational content. The experiment aimed to evaluate the challenges experienced by expert and novice users when using a task-oriented authoring tool enhanced with ChatGPT. By analyzing the overall mean column, the results provide insights into the users' agreement on various issues related to the AI tool. The challenges identified include violations of intellectual property rights (overall mean: 3.9600), academic integrity issues (overall mean: 4.12), lack of originality (overall mean: 4.00), challenges in accountabil- ity (overall mean: 3.72), and limited creativity and personalization (overall mean: 3.8). These challenges were acknowledged by both user groups, illustrating areas of improvement for AI tools. The overall mean values indicate that there is general agreement on the existence of these challenges. The findings emphasize the need to address these significant concerns to enhance the overall user experience of AI-enhanced authoring tools. The understanding of these challenges, as evidenced by the overall mean values, can be instrumental for developers and researchers working on improv- ing AI technologies, ensuring that future iterations can overcome the identified limitations and cater to the diverse needs of users in the expert and novice domains. Upon examining the p-values reported for the challenge items, all appear to be greater than the

0.05 threshold. This implies that the observed differences between expert and novice users regarding the challenges faced are not statistically significant. The findings suggest that both expert and novice users showed a similar level of agreement on the challenges they encountered with the AI-enhanced authoring tool based on large language models such as ChatGPT.

The boxplot (Figure 9) representation of the study comparing ex- pert and novice users of a

task-oriented authoring tool enhanced with ChatGPT revealed that both user groups faced similar chal- lenges. Although the expert users showed a higher standard error of the mean and a wider range of challenges, the overall level of challenges experienced was comparable between the two groups. This insight suggests that future development of AI tools should address the diverse challenges faced by both expert and novice users to improve the user experience for a broader audience.

Table V. T-Test Results Usage Challenges

Item	t-test	p-value	Overall Mean	Likert Agree- ment
Violation of intellec- tual property rights	0.075	0.941	3.9600	Agree
Academic integrity challenges	0.159	0.875	4.12	Agree
lack originality.	0.000	1.000	4	Agree
Challenges account-ability	0.555	0.584	3.72	Agree
Lack creativity and personalization.	0.000	1.000	3.8	Agree
Lack originality.	0.000	1.000	4.2	Agree

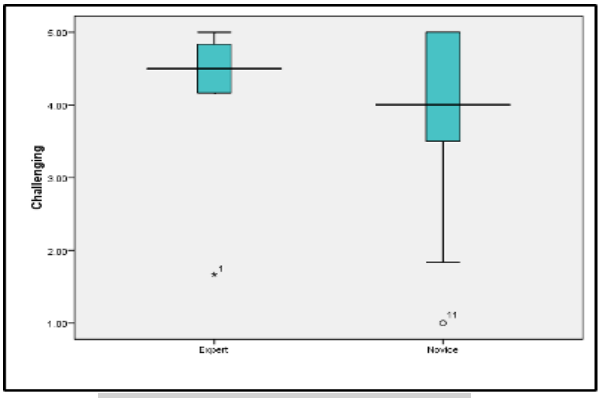


Fig. 9. Boxplot Usage Challenges

To summarize the critical results of our study:

- The study found that there were no significant differences in effectiveness, cognitive load, usability, and challenge evalua- tion between expert and novice users.
- Both expert and novice users reported similar levels of satis- faction with the process and product effectiveness, indicating that the tool successfully aids the authoring process and pro- duces reliable content. This was surprising given the expec- tations of variance.
- The cognitive load assessments also showed that the tool's mental demands are manageable for all users.
- Usability factors such as learnability, efficiency, and satis- faction were consistently rated high by all users. However, the evaluation of challenges such as intellectual property and originality concerns did not differ significantly between the groups.
- The outcomes suggest that the AI tool is adaptable and pro- vides a consistent user experience, highlighting its poten- tial for broad applicability. However, specific user concerns should be addressed in future developments.

V. CONCLUSION AND RECOMMENDATION

The study aimed to examine various aspects of using a ChatGPT-based task-oriented authoring tool for generating aca- demic content, including effectiveness, cognitive load, usability, and challenges through the

perspective of SDLC. By incorporat- ing evaluations during the SDLC, developers can refine the tool's features based on users' feedback, boosting performance and user satisfaction. The results from this study reveal that large language models like ChatGPT can play a crucial role in the educational sector. They assist in streamlining the writing process by provid- ing suggestions and organizational support, which can save time and enable educators to focus on creative tasks, improved peda- gogy, and direct student interactions. Hence, these models show great potential in enhancing learning experiences in educational contexts. However, to foster mainstream adoption and trust in AI- enhanced authoring tools, it is essential to address critical chal- lenges such as ensuring academic integrity, encouraging original content, and maintaining accountability during content creation. This study highlights the effectiveness of a ChatGPT-powered AI- task-oriented tool in assisting educational content creation for both expert and novice users. The positive feedback from participants showcases the tool's potential to streamline the writing process, reduce manual content generation, ensure high-quality material, and promote collaboration between content creators and educa- tors. By addressing changing requirements and diverse learning needs, AI-enhanced authoring tools can greatly impact the educa- tional arena. The cognitive load experienced by expert and novice users while using the tool, along with their respective agreement levels, indicates that both groups generally found the tool satis- factory and effective. It is important to consider the variation in users' cognitive loads and prior experiences when developing a tool like the one presented in this research. This will guarantee that such AI-enhanced tools address the needs of multiple experts' levels and offer custom experiences to users as per their needs, ultimately improving these tools' overall success, satisfaction, and effectiveness. In addition, investigating usability focusing on var- ious domains like the tool's learnability, efficiency, satisfaction, and error detection capability, helps pinpoint important insights that can effectively lead to the adoption of AI applications in ed- ucational environments. In our experiment, we noted that expert and novice users' responses indicate their comparable evaluation of the authoring tool which indicates its ability to adapt to the needs of various types of users with different levels of expertise. Thus, we believe that usability is a decisive factor in creating ef- fective, accessible, and engaging authoring solutions that integrate AI. The study also underscores the importance of tackling chal- lenges related to intellectual property rights, academic integrity, originality, and accountability when developing AI-enhanced au- thoring tools. Furthermore, fostering creativity and personaliza- tion while using these tools is vital for user trust and confidence in the technology. Addressing these concerns will not only help in mainstream adoption but also ensure that AI-generated content complies with ethical and legal norms, which establishes increased trust and acceptance for AI-based solutions in various industries. To enhance the usability and acceptance of the ChatGPT-integrated authoring tool, several recommendations were proposed by users like: providing downloadable files in readable formats to address unreadable symbols and improve accessibility; enhancing filter- ing options to accommodate various study fields and education levels, resulting in more accurate and relevant results; allowing customization of chapter arrangements and integrating references and illustrations for richer content presentation; implementing an efficient content writing process to streamline content generation; and improving support for non-English languages, such as Arabic, especially for elements like book covers. By effectively

integrat- ing insights gained from the study of effectiveness, cognitive load, usability, and challenges into the SDLC, developers can create a robust and reliable AI-enhanced authoring tool that meets users' diverse needs in the educational context. The continuous improve- ment and optimization of the ChatGPT-integrated authoring tool, driven by the SDLC framework, will contribute significantly to the advancement of AI-based tools in education and other pro- fessional settings. Future research could delve deeper into user feedback and analysis, investigating specific areas where improve- ments may be needed. By gathering more qualitative data, such as user interviews, researchers can obtain a comprehensive view of the interaction between users and the tool, driving the devel- opment of potential enhancements that increase its effectiveness and usability. Additionally, future research could explore features designed to cater to diverse user groups, as well as adaptive tech- nologies that promote personalized experience - all contributing to the continuous improvement and optimization of AI-based tools in the education sector. With recent advancement in ChatGPT, specifically, the feature which allows users to create their own ver- sion of GPTs (released November 2023) research about creating automated authoring tools can gain further momentum. This fea- ture can allow for training this ChatGPT large model on custom domain-specific data. So, our proposed tool can be further be cus- tomized by domain knowledge so that content authors can benefit from this in targeting specific learning groups and topics.

VI. APPENDIX

A Survey on evaluating the Task-oriented authoring tool

The following tables (VI, VII, VIII, IX)show all survey ques- tions that used in this paper.

Table VI. Effectiveness Domain Questions

Question	Category	Number
The tool allows writing in the language of your choice.	Process	Q1
The tool facilitates the writing process (such as selecting a book title, creating chapters).	Process	Q2
The tool provides options for the author (creating chapter titles, creating topics, writing content, deleting, etc.).	Process	Q3
The tool encourages organization dur- ing the writing process.	Process	Q4
The interface and writing options of the tool increase focus by reducing distrac- tions.	Process	Q5
The tool provides guidance for the au- thor (contextual help instructions and tips).	Process	Q6
The tool provides an option to design a book cover.	Process	Q7
The tool facilitates downloading a draft of the written content.	Process	Q8
The length of the written content is ap- propriate.	Product	Q9
The written content is reliable.	Product	Q10
The written content is comprehensive.	Product	Q11
The written content is organized.	Product	Q12
References are documented in the writ- ten content.	Product	Q13
Each chapter's written content relates to its topic.	Product	Q14
Reading the written content is easy.	Product	Q15
Understanding the written content is easy.	Product	Q16
The written content is natural language (mimics human-written content).	Product	Q17
The created cover relates to the entered description.	Product	Q18

Table VII. Usability Domain Questions

Question	Category	Number
The chatbot function was easily de-tectable	Learnability	Q1
It was easy to find the chatbot	Learnability	Q2
Communicating with the chatbot was clear	Efficiency	Q3
The chatbot was able to keep track of the context	Efficiency	Q4
The chatbot's responses were easy to understand	Satisfaction	Q5
I find that the chatbot understands what I want and helps me achieve my goal	Satisfaction	Q6
The chatbot gives me the appropriate amount of information	Satisfaction	Q7
The chatbot only gives me the informa- tion I need	Satisfaction	Q8
I feel like the chatbot's responses were accurate	Errors	Q9
I believe the chatbot informs me of any possible privacy issues	Errors	Q10
My waiting time for a response from the chatbot was short	Efficiency	Q11

Table VIII. Cognitive Load Questions

Question	Number
How mentally demanding was the task.	Q1
How physically demanding was the task.	Q2
How hurried or rushed was the pace of the task	Q3
How successful were you in accomplishing what you were asked to do?	Q4
How hard did you have to work to accomplish your level of performance?	Q5
How insecure, discouraged, irritated, stressed, and annoyed were you?	Q6

Table IX. Challenges Questions

Question	Number
Violation of intellectual property rights	Q1
Academic integrity challenges	Q2
Challenges accountability	Q3
Lack creativity and personalization.	Q4
Lack originality.	Q5
Requires human editing and revision.	Q6

B Task-oriented authoring tool Screens

In this appendix, the process and order of navigating the tool are outlined. It begins with selecting a language and entering book details such as the title and target age group for the written content. The tool also offers an option to generate a recommended cover image and save the content. Figure 10

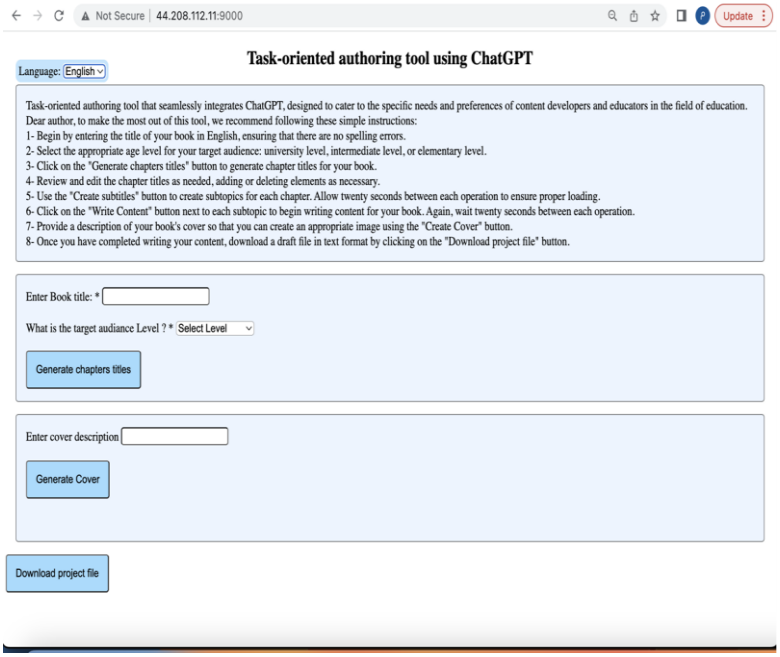


Figure 10. Author Select the English Language

Figure 11 shows the generated titles for mathematical funda- mental title Figure 5, Figure 3, and Figure 4 show the more details about the interface coding which allowed users to generate chapter titles one by one, with the capability to rewrite them if required. The tool can assist authors in creating cover images that are relevant to their text. This is achieved by inputting the text and then using a prompt to activate the DALEE model, which provides image suggestions, then the author can download the generated content in a text format as one file.

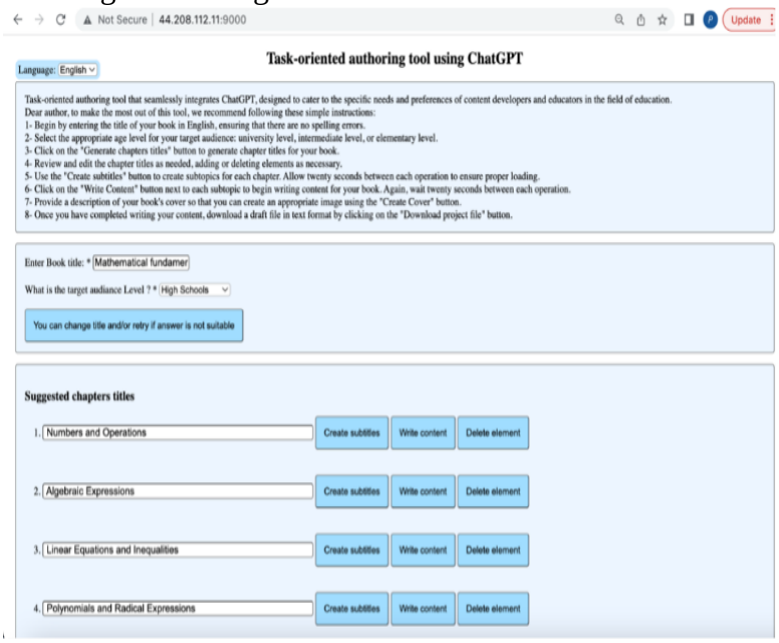


Figure 11. Generated Titles for Math Book

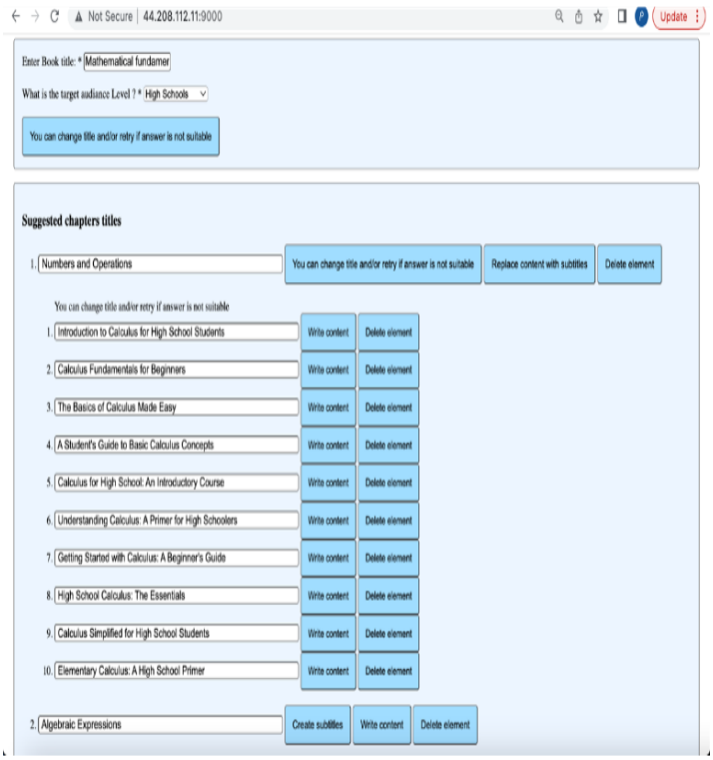


Figure 12. Author Generate Subtitles for First Chapter

Figure 13. The Tool Reply with Generated Content

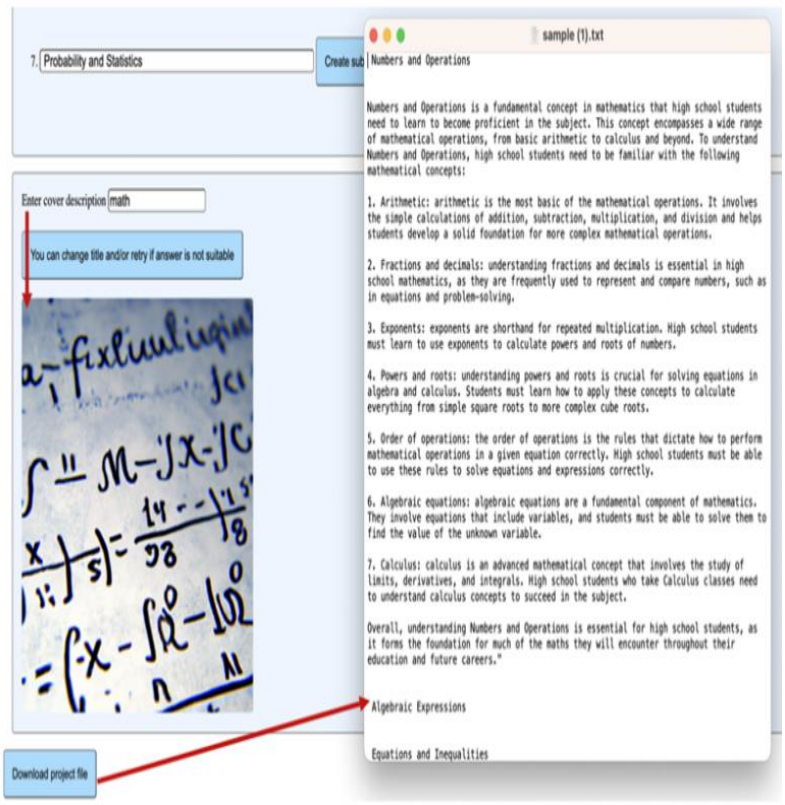


Figure 14. Author Request to Generate a Cover Image and Then Save the Generated Content in a Text Fil

ACKNOWLEDGMENT

Both authors are working and sponsored by King Abdulaziz University in Saudi Arabia. We would like to thank the Deanship of Scientific Research (DSR), King Abdulaziz University, Saudi Arabia, Jeddah. For facilitating the activities of this research.

REFERENCES

[1] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100033, 2021.

[2] A. Haleem, M. Javaid, and R. P. Singh, "An era of chatgpt as a significant futuristic support tool: A study on features, abil- ities, and challenges," *BenchCouncil transactions on bench- marks, standards and evaluations*, vol. 2, no. 4, p. 100089, 2022.

[3] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100033, 2021.

[4] R. Peres, M. Schreier, D. Schweidel, and A. Sorescu, "On chatgpt and beyond: How generative artificial intelligence may affect research, teaching, and practice," *International Journal of Research in Marketing*, 2023.

[5] D. Mhlanga, "Open ai in education, the responsible and eth- ical use of chatgpt towards lifelong learning," *Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023)*, 2023.

[6] V. Liu, H. Qiao, and L. Chilton, "Opal: Multimodal image generation for news illustration," in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–17, 2022.

[7] P. Kowalczyk, M. Ro¨der, and F. Thiese, "Nudging creativity in digital marketing with generative artificial intelligence: Opportunities

and limitations," 2023.
[8] D. Baidoo-Anu and L. Owusu Ansah, "Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learn- ing," *Available at SSRN 4337484*, 2023.
[9] B. D. Lund and T. Wang, "Chatting about chatgpt: how may ai and gpt impact academia and libraries?," *Library Hi Tech News*, vol. 40, no. 3, pp. 26–29, 2023.
[10] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos,
L. De Leon, C. Elepan˜o, M. Madriaga, R. Aggabao, G. Diaz- Candido, J. Maningo, *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *PLoS digital health*, vol. 2, no. 2,
p. e0000198, 2023.
[11] M. Abdullah, A. Madain, and Y. Jararweh, "Chatgpt: Fun- damentals, applications and social impacts," in *2022 Ninth International Conference on Social Networks Analysis, Man- agement and Security (SNAMS)*, pp. 1–8, IEEE, 2022.
[12] M. A. AlAfnan, S. Dishari, M. Jovic, and K. Lomidze, "Chat- gpt as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses," *Journal of Artificial Intelligence and Technology*, vol. 3, no. 2, pp. 60–68, 2023.
[13] X. Chen, "Chatgpt and its possible impact on library refer- ence services," *Internet Reference Services Quarterly*, pp. 1– 9, 2023.
[14] S. Panda and N. Kaur, "Exploring the viability of chatgpt as an alternative to traditional chatbot systems in library and information centers," *Library Hi Tech News*, vol. 40, no. 3,
pp. 22–25, 2023.
[15] A. Subaveerapandiyan, A. Vinoth, and N. Tiwary, "Netizens, academicians, and information professionals' opinions about ai with special reference to chatgpt," 2023.
[16] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, no. 1,
p. 15, 2023.
[17] S. Moore, H. A. Nguyen, N. Bier, T. Domadia, and J. Stam- per, "Assessing the quality of student-generated short answer questions using gpt-3," in *European conference on technol- ogy enhanced learning*, pp. 243–257, Springer, 2022.
[18] J. Su and W. Yang, "Unlocking the power of chatgpt: A framework for applying generative ai in education," *ECNU Review of Education*, p. 20965311231168423, 2023.
[19] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert,
A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.
[20] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie,
H. Lovenia, Z. Ji, T. Yu, W. Chung, *et al.*, "A mul- titask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *arXiv preprint arXiv:2302.04023*, 2023.
[21] P. Ragunath, S. Velmourougan, P. Davachelvan, S. Kay- alvizhi, and R. Ravimohan, "Evolving a new model (sdlc model-2010) for software development life cycle (sdlc)," *In- ternational Journal of Computer Science and Network Secu- rity*, vol. 10, no. 1, pp. 112–119, 2010.
[22] L. W. D. Quan, "Openai launches new gpt 3.5 turbo and whisper ai models, 10x cheaper with better re- sults."
https://www.gizmochina.com/2023/03/01/ openai-new-gpt-3-5-turbo-whisper-ai-models/, 2023. [Accessed 10-JUn-2023].
[23] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the*

human factors and ergonomics society annual meeting, vol. 50, pp. 904–908, Sage publications Sage CA: Los Angeles, CA, 2006.

[24] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical re- search," in *Advances in psychology*, vol. 52, pp. 139–183, Elsevier, 1988.

[25] S. Borsci, A. Malizia, M. Schmettow, F. Van Der Velde,
G. Tariverdiyeva, D. Balaji, and A. Chamberlain, "The chat- bot usability scale: the design and pilot of a usability scale for interaction with ai-based conversational agents," *Personal and Ubiquitous Computing*, vol. 26, pp. 95–119, 2022.

أداة التأليف لإنتاج محتوى أكاديمي باستخدام ChatGPT

ميادة المصري[١]، العنود سبحي[٢]

[١] قسم تقنية المعلومات، كلية الحاسبات وتقنية المعلومات

جامعة الملك عبد العزيز، جدة، المملكة العربية السعودية

[٢] قسم تقنية المعلومات، كلية الحاسبات وتقنية المعلومات

جامعة الملك عبد العزيز، جدة، المملكة العربية السعودية

malmasre@kau.edu.sa , asubahi@kau.edu.sa

مستخلص. لقد أدى التطور المتسارع للتكنولوجيا إلى ظهور أدوات ذكية متطورة مثل الروبوتات الدردشة الذكية وخوارزميات التعلم الآلي، والتي تمتلك إمكانات كبيرة لتحسين التعليم والتعلم. وباعتبار أن غالباً ما تفتقر أدوات إنشاء المحتوى التقليدية إلى هذه الميزات المتطورة، مما يجعل دمج الذكاء الاصطناعي، بما في ذلك ChatGPT، مجالاً واسعاً للبحث.

تهدف هذه الدراسة إلى تقييم فعالية أداة التأليف الموجهة للمهام والمتكاملة مع ChatGPT لإنتاج محتوى تعليمي شخصي، وتحميلها الإدراكي، وسهولة استخدامها، والتحديات المحتملة. شمل البحث مجموع ٢٥ مشاركاً في استخدام الأداة وهم: عدد ٥ من الخبراء وعدد ٢٠ من المبتدئين، وجميعهم استخدموا أداة التأليف لإنتاج محتوى أكاديمي. تم تصميم استبيان ليكرت المكون من ٤١ عنصراً لاستطلاع آراء المستخدمين حول فعالية الأداة، وتحميلها الإدراكي، وسهولة استخدامها، والتحديات المرتبطة بالذكاء الاصطناعي، مع استخدام المقارنة المتوسطة واختبارات للتحليل.

كشفت النتائج الرئيسية عن انطباعات إيجابية بشكل عام بين المستخدمين، خاصةً فيما يتعلق بكفاءة الأداة وإدارة التحميل الإدراكي. ومع ذلك، ظهرت اختلافات صغيرة في تصورات سهولة الاستخدام بين الخبراء والمبتدئين. توفر هذه النتائج رؤى قيمة لتحسين وتعزيز أدوات التأليف المدمجة بالذكاء الاصطناعي لتلبية احتياجات المستخدمين المتنوعة بشكل أفضل في المجال التعليمي.

الكلمات المفتاحيةـ أدوات التأليف، نموذج توليد النصوص الشات جي بي تي، التعلم الالكتروني، تخصيص التعليم