applied sciences

MDPI

Review

Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications

Jakub Swacha * **and Michał Gracel**

Department of Information Technology in Management, University of Szczecin, 71-004 Szczecin, Poland; michal.gracel@usz.edu.pl
* Correspondence: jakub.swacha@usz.edu.pl

Abstract: Retrieval-Augmented Generation (RAG) overcomes the main barrier for the adoption of LLM-based chatbots in education: hallucinations. The uncomplicated architecture of RAG chatbots makes it relatively easy to implement chatbots that serve specific purposes and thus are capable of addressing various needs in the educational domain. With five years having passed since the introduction of RAG, the time has come to check the progress attained in its adoption in education. This paper identifies 47 papers dedicated to RAG chatbots' uses for various kinds of educational purposes, which are analyzed in terms of their character, the target of the support provided by the chatbots, the thematic scope of the knowledge accessible via the chatbots, the underlying large language model, and the character of their evaluation.

Keywords: RAG chatbot; educational chatbot; large language models in education; LLMs for education; AI teaching support; AI learning support

check for updates

Academic Editors: Jenny Pange and Zoi Nikiforidou

Received: 11 March 2025
Revised: 5 April 2025
Accepted: 9 April 2025
Published: 11 April 2025

Citation: Swacha, J.; Gracel, M. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Appl. Sci.* **2025**, *15*, 4234. https://doi.org/10.3390/app15084234

1. Introduction

While many education researchers believe that dialogue plays a significant role in learning [1] (p. 1), in practice, many students are deprived of the opportunity to talk to their teachers (e.g., when the latter have to teach too many students to be able to find sufficient time to speak with each of them) or fellow students (e.g., in the case of asynchronous online learning when there is no one else online at the given moment). A revolutionary solution to this issue materialized with the introduction of chatbots based on large language models (LLMs), such as OpenAI's ChatGPT, capable of instantly generating highly realistic and convincing conversational responses [2]. An important advantage is that talking to chatbots is free from the apprehension that the students face in exposing their lack of knowledge or understanding to teachers or peers, fostering a more supportive learning environment, as they introduce neither social distance nor a power relationship [3].

Despite their numerous virtues, LLMs also have some flaws; in particular, it is difficult to update or extend their knowledge and they are prone to hallucinating unintended text [4]. These form a serious problem in the use of LLMs for educational purposes, especially the hallucinations, i.e., generating *content that is nonsensical or unfaithful to the provided source content* [5].

A capable solution addressing these flaws comes in the form of Retrieval-Augmented Generation (RAG) [4]. RAG efficiently enhances the performance of LLMs, not by optimizing their training but by integrating information retrieval techniques [6]. In short, RAG uses the input sequence x to retrieve text documents z and use them as additional context when generating the target sequence y [4]. Although RAG has its limitations, such as a possible failure to retrieve documents most relevant to the query, a possible failure of the generative module to effectively incorporate the retrieved information into its response, and the computational overhead because of both a retrieval and a generation step being performed for each query [7], nonetheless, it has established itself as a practical means for developing numerous LLM-based applications serving various purposes [8]. It is especially useful in areas where model reliability is of primary importance, such as healthcare [9]. Although education can also be considered such an area, so far there has been no survey dedicated to unveiling the progress made in applying RAG there. This research gap requires attention for a number of reasons, including the expected continued quick development of the field of RAG applications in the educational domain, which calls for defining the state of the progress attained so far; the revolutionary character of chatbots featuring RAG, making them notably distinct from chatbots developed in the pre-RAG era and covered in the educational chatbot surveys published so far [10–12]; and the unique character of RAG chatbots developed for the educational domain (in particular, their specific range of targets of support), different from the specific traits of RAG chatbots developed for other domains, such as healthcare where, e.g., patient safety is of primary concern [9].

This paper aims to address this gap by investigating the development of the primary kind of RAG applications, that is, RAG-based chatbots in the educational domain. Based on works published so far, not only in peer-reviewed venues, but also as preprints and theses available online, this survey attempts to answer the following research questions:

RQ1: How many and what kinds of works have been published on the topic so far?

RQ2: What are the purposes served by the RAG-based chatbots in the educational domain?

RQ3: What are the themes covered by the RAG-based chatbots in the educational domain?

RQ4: Which of the available large language models are used by the RAG-based chatbots in the educational domain?

RQ5: What are the criteria according to which the RAG-based educational chatbots were evaluated?

This paper is structured as follows. Right after this introduction, Section 2 presents the used data sources, search procedure, parameters, and outcomes. In the respective subsections of Section 3, the results relevant to all five research questions stated above are provided. Section 4 discusses the obtained results and forms recommendations, whereas the final section concludes the paper, also presenting the study limitations and suggesting directions for the future work.

2. Materials and Methods

The search for the relevant publications has been performed using three bibliographic data access platforms: Scopus, Web of Science, and Google Scholar. The first two platforms feature a reliable search procedure based on multiple user-defined criteria that can be combined in various ways and a large database of quality publication venues. Their inclusion is motivated by the assumption that, if there is a high-quality work on the chosen topic, it can highly probably be found by an exact search in either of these two sources. The third (Google Scholar) uses a black-box search procedure based on relevancy (so an exact match with the search phrase is not necessary to qualify a paper for the results) and includes various web sources, regardless of their quality. Its inclusion is motivated by the novelty of the covered topic (in the context of the long publication times of many publication venues) and strives to identify relevant works that were not captured by the query on the first two platforms because of, e.g., not being indexed (e.g., preprints, Master's theses, or local journals and conference proceedings) or not matching the search criteria (e.g., unusual wording used by their authors or their keywords being defined using completely different schemes).

In all three analyzed sources, we considered only publications written in English for two practical reasons: English is currently the lingua franca of scientific writing in which most papers are published and our command of other languages is not sufficient to reliably survey all non-English papers.

The Scopus search term has been defined to target all original publications in English from 2022 onward that featured the terms *education**, *chatbot*, and *Retrieval-Augmented Generation* or *RAG* in their titles, abstracts, or keywords specified by their respective authors:

```
TITLE-ABS-KEY ( "education*" ) AND TITLE-ABS-KEY ( "chatbot" )
AND ( TITLE-ABS-KEY ( "RAG" ) OR
TITLE-ABS-KEY ( "Retrieval Augmented Generation" ) )
AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE , "cp" ) )
AND PUBYEAR > 2021 AND ( LIMIT-TO ( LANGUAGE , "English" ) )
```

A similar approach was used to define the search term for Web of Science:

```
(((((TI=(education*)) OR AK=(education*))) OR (AB=(education*)))
AND ((((TI=(chatbot)) OR AK=(chatbot))) OR (AB=(chatbot))))
AND (((((TI=("Retrieval Augmented Generation"))
OR AK=("Retrieval Augmented Generation")))
OR (AB=("Retrieval Augmented Generation"))) OR ((((TI=(RAG)) OR (AK=(RAG)))
OR (AB=RAG)))) AND (PY>2021)
```

As for querying Google Scholar, the Publish or Perish software version 8.17 has been used with the following search terms: *education* and *rag chatbot*, excluding citations and patents.

Both bibliographic data providers were queried on 17 February 2025. Figure 1 shows the flow of the data collection and qualification process. The Scopus query yielded 24 results and Web of Science provided only 8 results, whereas 83 results were obtained from Google Scholar. While seven out of eight Web of Science results were also present in Scopus, only four works found in Scholar were duplicated in the results from Scopus, which indicates the mutual complementarity of Google Scholar with the classic bibliographic data sources, confirming the rationale for including the former data source in the query. After screening the titles and abstracts, only 1 of the 24 works found in Scopus but as many as 59 of the 83 works found in Scholar were assessed as unrelated to the area of education. Furthermore, nine works found in Scholar and one in Web of Science were excluded for being written in a language other than English.

A snowballing procedure was then performed on the resulting set of 34 identified papers, whose references provided 13 additional relevant papers, resulting in a total of 47 research papers that were qualified for further analysis.

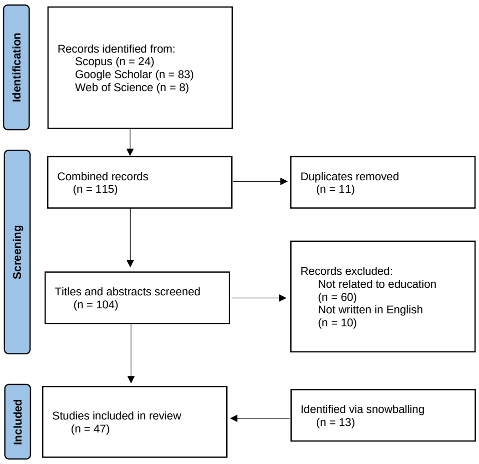

Figure 1. Data collection and qualification.

3. Results

3.1. Overview of Identified Studies

The analyzed set consists of 23 conference proceeding papers (including those published in book series and journals devoted solely to disseminating conference proceedings), 13 refereed journal articles, 9 preprints published on the web by their authors, and 2 Master's theses published in the universities' online repositories. The majority of identified works (37) were published in 2024, 7 in 2023, and 3 were from 2025. The papers included in the analyzed set were arranged according to their target of support (see Section 3.2) and are listed in Tables 1–4. Apart from data identifying the respective works, the table columns provide detailed information that is synthesized in the form of figures presented in subsequent subsections.

Table 1. List of works in the analyzed dataset. Target of support: Learning.

Paper	Title	Thematic Scope	LLM
[13]	Beyond Chatbots: Enhancing Luxembourgish Language Learning Through Multi-agent Systems and Large Language Model	Language	GPT-4o
[14]	ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: A comparative pilot study	Health sciences	GPT-3.5
[15]	Design, architecture and safety evaluation of an AI chatbot for an educational approach to health promotion in chronic medical conditions	Health sciences	GPT-3.5-turbo-0125
[16]	Embodied AI-Guided Interactive Digital Teachers for Education	Any	LLaMA-3-8B
[17]	Enhancing Programming Education with Open-Source Generative AI Chatbots	Computer science	Mistral-7B-Openorca
[18]	Harnessing GenAI for Higher Education: A Study of a Retrieval Augmented Generation Chatbot's Impact on Human Learning	Computer science	GPT4 changed to Claude 3-Sonnet
[19]	Interactive Video Virtual Assistant Framework with Retrieval Augmented Generation for E-Learning	Any	LLaMA-2-7B
[20]	Into the unknown unknowns: Engaged human learning through participation in language model agent conversations	Any	GPT-4o-2024-05-13
[21]	An LLM-Driven Chatbot in Higher Education for Databases and Information Systems	Computer science	GPT-4
[22]	Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference	Math	GPT-3.5-turbo-0613
[23]	Scalable Mentoring Support with a Large Language Model Chatbot	Any	GPT-3.5-turbo
[24]	Transforming Healthcare Education: Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation	Health sciences	GPT-3.5, GPT-4

Table 2. List of works in the analyzed dataset. Target of support: Organizational matters.

Paper	Title	LLM
[25]	AI-driven Chatbot Implementation for Enhancing Customer Service in Higher Education: A Case Study from Universitas Negeri Semarang	LLaMA-3
[26]	Closed Domain Question-Answering Techniques in an Institutional Chatbot	LLaMA-2
[27]	Development and Evaluation of a University Chatbot Using Deep Learning: A RAG-Based Approach	Unspecified
[28]	Facilitating university admission using a chatbot based on large language models with retrieval-augmented generation	GPT-3.5-turbo, Text-davinci-003
[29]	From Questions to Insightful Answers: Building an Informed Chatbot for University Resources	GPT-3.5-turbo

Table 3. List of works in the analyzed dataset. Target of support: Various.

Paper	Title	Target of Support	Thematic Scope	LLM
[30]	AI Assistants in Teaching and Learning: Compliant with Data Protection Regulations and Free from Hallucinations!	Learning + Question generation	Any	GPT-4
[31]	AI-TA: Towards an Intelligent Question-Answer Teaching Assistant using Open-Source LLMs	Learning + Organizational matters	Computer science	GPT-4, LLaMA-2-7B, LLaMA-2-13B, LLaMA-2-70B
[32]	Ask NAEP: A Generative AI Assistant for Querying Assessment Information	Access to information on assessment	Education	GPT-3.5 changed to GPT-4o
[33]	Developing a RAG system for automatic question generation: A case study in the Tanzanian education sector	Question generation	Any	GPT-3.5-turbo-0125, GPT-4-turbo-2024-04-09, LLaMA-3-70b-8192
[34]	The end of multiple choice tests: using AI to enhance assessment	Learning analytics support	Any	GPT (Unspecified)
[35]	Enhancing International Graduate Student Experience through AI-Driven Support Systems: A LLM and RAG-Based Approach	Organizational matters, Local culture	Local-culture and institutional-domain knowledge	GPT-3.5
[36]	Gaita: A RAG System for Personalized Computer Science Education	Course selection	Institutional-domain knowledge	GPT-4o Mini
[37]	An Intelligent Retrieval Augmented Generation Chatbot for Contextually-Aware Conversations to Guide High School Students	Course selection	Institutional-domain knowledge	Unspecified
[38]	RAMO: Retrieval-Augmented Generation for Enhancing MOOCs Recommendations	Course selection	Institutional-domain knowledge	GPT-3.5-turbo, GPT-4, LLaMA-2, LLaMA-3
[39]	VizChat: Enhancing Learning Analytics Dashboards with Contextualised Explanations Using Multimodal Generative AI Chatbots	Learning analytics support	Education	GPT-4V

Table 4. List of works in the analyzed dataset. Target of support: Access to source knowledge.

Paper	Title	Thematic Scope	LLM
[40]	Bio-Eng-LMM AI Assist chatbot: A Comprehensive Tool for Research and Education	Various	GPT-4
[41]	CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering	Law	Mistral-7B
[42]	ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge	Health sciences	LLaMA-7B
[43]	ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases	Law	Ziya-LLaMA-13B
[44]	ChatPapers: An AI Chatbot for Interacting with Academic Research	Computer science	Unspecified
[45]	Chatbot to chat with medical books using Retrieval-Augmented Generation Model	Health sciences	GPT-4
[46]	Chatbots in Academia: A Retrieval-Augmented Generation Approach for Improved Efficient Information Access	Institutional-domain knowledge	GPT-3.5-turbo
[47]	A Client-Server Based Educational Chatbot for Academic Institutions	Any	Unspecified
[48]	Comparing the Performance of LLMs in RAG-Based Question-Answering: A Case Study in Computer Science Literature	Computer science	Mistral-7b-instruct, LLaMa2-7b-chat, Falcon-7b-instruct and Orca-mini-v3-7b, GPT-3.5
[49]	Developing a Retrieval-Augmented Generation (RAG) Chatbot App Using Adaptive Large Language Models (LLM) and LangChain Framework	Institutional-domain knowledge	Gemma (local version)
[50]	Development of a liver disease–specific large language model chat interface using retrieval-augmented generation	Health sciences	GPT-3.5-turbo
[51]	Enhancing textual textbook question answering with large language models and retrieval augmented generation	Various	LLaMA-2
[52]	Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications	Health sciences	GPT-4
[53]	Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice's Cooperative Principles and Trust	Any	GPT-3.5, GPT-4
[54]	Let us not squander the affordances of LLMS for the sake of expedience: using retrieval-augmented generative AI chatbots to support and evaluate student reasoning	Any	Unspecified
[55]	Moroccan legal assistant Enhanced by Retrieval-Augmented Generation Technology	Law	LLama 70-B
[56]	PaperQA: Retrieval-Augmented Generative Agent for Scientific Research	Biomedical	Claude-2, GPT-3.5, GPT-4
[57]	Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation	Health sciences	Unspecified
[58]	Retrieval-Augmented Generation Based Large Language Model Chatbot for Improving Diagnosis for Physical and Mental Health	Health sciences	Unspecified
[59]	TaxTajweez: A Large Language Model-based Chatbot for Income Tax Information In Pakistan Using Retrieval Augmented Generation (RAG)	Finance	GPT-3.5-turbo

3.2. Target of Support

LLM-based chatbots have found numerous points of application in the educational domain [53]. Here, for the sake of analysis, we grouped them into the following types:

1. Learning— chatbots whose sole purpose is to support the learning process, combining the access to knowledge on the studied subject with various learner support tech-

niques, such as tailoring the responses to individual learning styles and pacing or arousing the learner's engagement;
2. Access to source knowledge—chatbots which provide convenient access to some body of source knowledge (the purpose of which is not necessarily learning);
3. Access to information on assessment—this category has been distinguished as, in contrast to access to source knowledge, it is not the learning content that is accessed here but the information on how to assess the learned content;
4. Organizational matters—chatbots which help students to learn the rules and customs at their alma mater;
5. Local culture—applicable when the chatbot helps (international) students to become acquainted with the rules and customs of the specific nation or community rather than the educational institution;
6. Course selection—chatbots helping candidates and/or students in choosing their educational path;
7. Question generation—chatbots whose sole purpose is to support the teachers in developing questions about which they will ask the students;
8. Learning analytics support—chatbots helping the teachers in analyzing the recorded effort of their students.

There was no work in the analyzed set that did not belong to any of the categories listed above. Few works addressed more than one purpose.

In Figure 2, we indicate the number of works assigned to respective application types, grouping them depending on the target user (student or teacher). There is one category (access to source knowledge) that is equally useful for both students (learning from the provided sources) and teachers (using these sources to prepare educational materials for their students). Note that, even though teachers can use chatbots assigned to the learning category to augment their own knowledge on the specific topic, we consider their role then as students not teachers; therefore, the learning category is attached to students only.

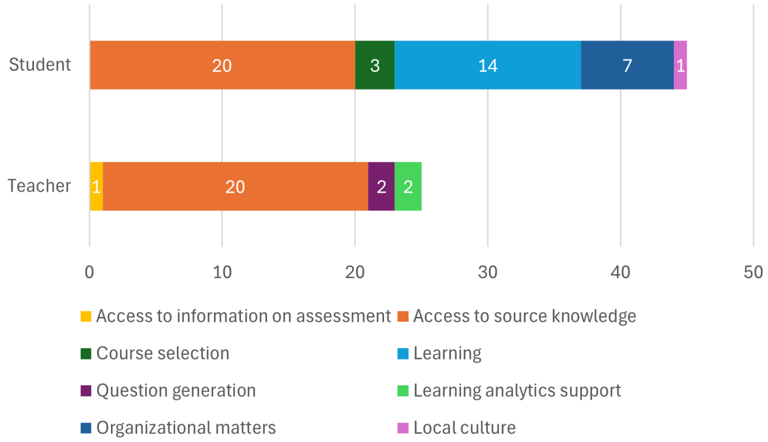

Figure 2. Target of support.

Looking at Figure 2, it is clear that the access to source knowledge category is the most numerous one, with 20 relevant publications ([40–59]). This may stem from the fact that these chatbots are the simplest to develop, if the collection of source data is available, as they need no additional features (e.g., tailoring the responses to meet individual students' learning styles and pacing, as is the case with the learning category chatbots). Moreover, some of these chatbots have been developed to also serve other purposes than education and other target groups than students and/or teachers (e.g., supporting the practice of legal professionals [55], medical personnel [58], or tax consultants [59]).

Closely following the first category in the number of publications (14) is learning [13–24,30,31]. We distinguish these chatbots as providing some value added to the mere access to the source knowledge on a specific topic. This may consist of integration with Learning Management Systems (e.g., ref. [21]), fetching or generating relevant questions or exercises and evaluating their solutions to let students self-check their learning progress (e.g., ref. [21]), providing personalized and/or adaptive learning support (e.g., refs. [13,18,23]), introducing AI-based agents to ask further questions to let the students learn about things about which they were not knowledgeable enough to ask (e.g., ref. [20]), generating animated three-dimensional avatars or short movies that show speaking humans to provide students with human-like conversation experiences (e.g., refs. [16,19]), generating mind maps (e.g., ref. [30]), generating a comprehensive report from the discussion at its end (e.g., ref. [20]) or ensuring response traceability (e.g., ref. [15]). We have also included in this category papers that do not clearly indicate the specific value added, yet compare various solutions (chatbots or LLMs that they are based on) in their suitability for the specific purpose of students' learning (e.g., refs. [14,17]).

The third category with respect to the number of publications (7) comprises chatbots that support students in organizational matters, explaining various aspects of operation (including the rules and customs) of their educational institution [25–29,31,35].

Only a few papers were identified that belong to the other categories (and, for some of them, they are not even the only assigned categories). These include chatbots generating questions [33] and also other testing materials (such as quizzes or Anki flashcards [30]) to be used by teachers (note, we do not include in this category chatbots generating questions to be directly answered by students as a part of the dialogue [21]); chatbots supporting students in their selection of courses to take and plan their educational path [36–38]; a chatbot helping international students to adapt to the local culture [35]; and chatbots helping the teachers in performing learning analytics on their students' data [34,39].

3.3. Thematic Scope

As RAG chatbots are used in various fields of education, for the purpose of this study, we have arranged them into the following four types:

1. Institutional-domain knowledge—chatbots focused on the knowledge about a given institution, including its policies and procedures, regardless of its field(s) of education.
2. Area-specific—chatbots designed to support education in a particular area of education, including the following:
    (a) Biomedical;
    (b) Computer science;
    (c) Culture;
    (d) Education;
    (e) Finance;
    (f) Health sciences;
    (g) Language;
    (h) Law;
    (i) Math.
3. Various—chatbots designed to support education in more than one field.
4. Any—chatbots that may be used in any field of education.

Figure 3 shows the number of chatbots reported in the analyzed papers that belong to specific types and areas according to the classification presented above.

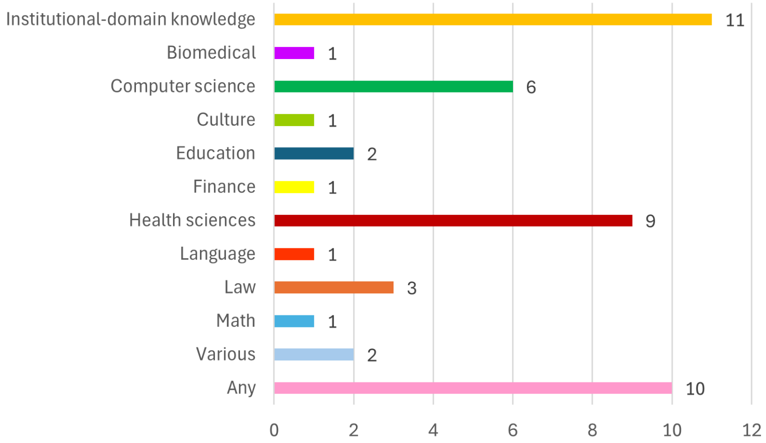

Figure 3. Thematic scope.

As shown in Figure 3, the most commonly reported were chatbots providing access to the institutional-domain knowledge, with 11 publications [25–29,35–38,46,49]. The majority of them concern the organization and policies of universities [25,27–29,46,49].

Among the chatbots providing area-specific knowledge, the most popular domain of knowledge is health sciences, with nine papers [14,15,24,42,45,50,52,57,58].

Closely following is computer science, with six publications. In this group, four chatbots support the learning of specific subjects: databases [21], data science and artificial intelligence [18], and programming [17,31], whereas the remaining two [44,48] provide access to area-specific knowledge sourced from the literature.

Next were law, with three publications [41,43,55]; education, with two publications [32,39]; and the special case of chatbots providing knowledge from various areas of science [40,51].

The remaining areas were addressed by only one paper: biomedical [56], finance [59], language [13], culture [35], and math [22].

Lastly, there are 10 publications [16,19,20,23,30,33,34,47,53,54] that were assigned to the "Any" type. Some of them, however, mention targeting a specific level of education, most often higher education [23,47,54], with one paper that focused particularly on secondary-school education [33].

3.4. Used Large Language Models

The seamless integration of the retrieved information with the LLM is essential for the effective operation of a RAG-based chatbot. Not every LLM is known to be equally fit for this role [60]. It is therefore interesting in this context to see which models were chosen by the authors of the chatbots in the analyzed papers. This information can be found in Figure 4, where the respective models are shown as columns, whereas their different versions are indicated by different shades of the column color. Note that the presented version identifiers are exactly those reported by the respective authors; therefore, they are given at different levels of detail. Note also that publications that compare the use of different models or different versions of the same model are counted multiple times.

Looking at Figure 4, it is evident that the vast majority of described chatbots are based on various versions of OpenAI's GPT (36 instances [13–15,18,20–24,28–33,35,36,38–40,45,46,48,50,52,53,56,59]). There is a long gap to the competition: the second in popularity, LLaMa, was reported in 15 instances [16,19,25,26,31,33,38,42,48,51,55]. Only eight chatbots were powered by other LLMs [17,18,41,43,48,49,56], though, in seven papers, we failed to identify the used LLM [27,37,44,47,54,57,58].

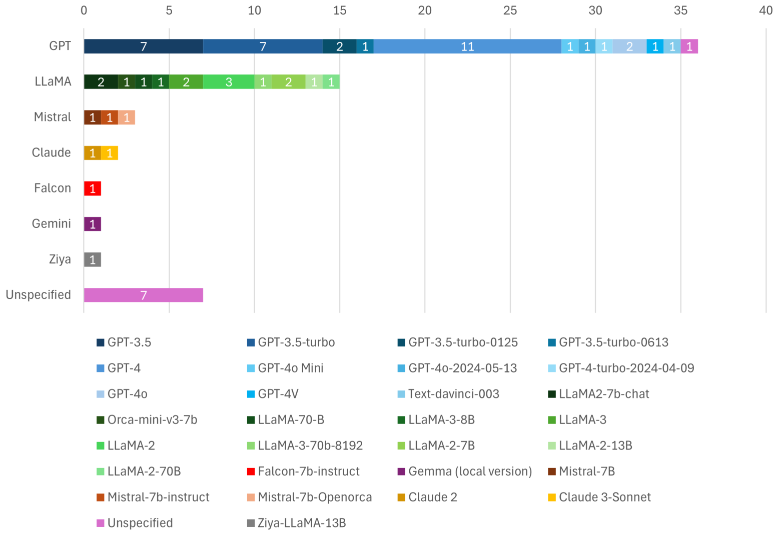

Figure 4. Used large language models.

3.5. Performed Evaluation

The fact that the performance of RAG-based chatbots depends on multiple factors, including the choice of retrieval model, the large language model, the considered corpus, the prompt formulation, and the tuning of all these components [61], their evaluation is of primary importance, as even high-quality components do not guarantee satisfactory results.

The problem is that there is no widely accepted standard for the evaluation of RAG-based chatbots. As they have several distinct quality facets, the concepts for their evaluation are adopted from different areas of research:

- Because of the retrieval component, we have examples of evaluation based on classic information retrieval measures [62], such as Precision, Recall, F1 Score, Accuracy, Sensitivity, and Specificity; these are instrumental in comparing the performance of different approaches that can be used to generate vector representations of the queries and responses—for instance, F1 Score reported in [25] (p. 5697) ranges from 0.35 for GloVe [63] to 0.67 for TF-IDF [64];
- Because of the LLM component, we have examples of evaluation based on LLM benchmarks, such as AGIEval [65], measuring the ability to pass human-centric standardized exams; this is instrumental in comparing the performance of different LLMs that can be used as the base of the RAG chatbots—for instance, AGIEval measurements reported in [55] range from 21.8 for Falcon-7B [66] to 54.2 for LLaMa2-70B [67];
- Because chatbots are made for conversation, we have examples of evaluation based on conversational quality measures, in particular based on Grice's Maxims of Conversation [68], such as Correctness (assessing whether the chatbot's answer is factually correct), Completeness (assessing whether the answer includes all relevant facts and information), and Communication (assessing whether the answer is clear and concise) [32], or (in Wölfel et al.'s interpretation) Information Quantity (assessing whether the provided amount of information is neither too much nor too little), Relevance (assessing whether the answer is pertinent to the ongoing topic or the context of the conversation), Clarity (assessing whether the answer is clear, concise, and in order), and Credibility (assessing whether the recipient trusts the chatbot) [53]; these could be used for, e.g., comparing the users' trust in the answers generated with the help of different LLMs that can power the RAG chatbots (e.g., Wölfel et al. compared the trust level between two versions of GPT: 3.5 and 4.0 [69], and reported 5.50 for answers generated by GPT 3.5 and of 5.59 for answers generated by GPT 4.0 [53]), but also the

change in the users' trust due to incorporating RAG—interestingly and unexpectedly, the same source reports a decrease in the average trust level for answers generated by both GPT 3.5 (from 5.98 to 5.50) and GPT 4.0 (from 6.34 to 5.59) after combining them with the content extracted from lecture slides and their transcripts [53];

- Because chatbots are instances of generative Natural Language Processing, we have examples of computer-generated text quality measures such as the similarity to human-written text (measured with metrics such as K-F1++ [70], BERTscore [71], BLEURT [72], or ROUGE [73]) [28,46], text readability (measured with, e.g., Coleman–Liau [74] and Flesch–Kincaid [75] metrics) [44], and human-assessed text quality aspects such as Fluency, Coherence, Relevance, Context Understanding, and Overall Quality [17]; these, again, can be instrumental in comparing the performance of different LLMs that can be used as the base of the RAG chatbots—for instance, Šarčević et al. compared five models (LLaMa-2-7b [67], Mistral-7b-instruct [76], Mistral-7b-openorca [77], Zephyr-7b-alpha [78], and Zephyr-7b-beta [79]) in terms of human-assessed text quality to detect sometimes notable differences: the measured Fluency ranged from 4.81 for Zephyr-7b-alpha to 4.92 for Mistral-7b-openorca; Coherence from 4.58 for Mistral-7b-instruct to 4.88 for Mistral-7b-openorca; Relevance from 4.15 for Llama-2-7b to 4.43 for Zephyr-7b-alpha; Context Understanding from 3.73 for Mistral-7b-instruct to 4.44 for Mistral-7b-openorca; and Overall Quality from 3.67 for Zephyr-7b-alpha to 4.48 for Mistral-7b-openorca [17];
- Because chatbots are a kind of computer software, we have examples of evaluation based on software usability, including technical measures of Operational Efficiency (such as the average handling time [25] or response time [48]) and survey-based measures of general User Acceptance (the extent to which the user accepts to work with the software, assessed using, e.g., the Technology-Acceptance model constructs [80]) [21] and User Satisfaction (the extent to which the user is happy with a system, assessed using, e.g., System Usability Scale [81]) [29], as well as specific usability aspects, in particular, Safety—ensuring that the answer is appropriate and compliant with protocols, which is especially relevant for domains where wrong guidance could cause a serious material loss or increase the risk of health loss, e.g., in medical education [15,50]; the reported usability measurements can be used for comparisons not only with other chatbots but also with any kind of software evaluated using the same measurement tool, for instance, the average SUS score of 67.75 reported by Neupane et al. [29] for their chatbot providing campus-related information, such as academic departments, programs, campus facilities, and student resources, is little above the average score of 66.25 measured for 77 Internet-based educational platforms (LMS, MOOC, wiki, etc.) by Vlachogianni and Tselios [82] but is far below the score of 79.1 reported for the FGPE interactive online learning platform [83]; and
- Because chatbots' development and maintenance incur costs, we have examples of economic evaluation based on measures such as cost per query or cost per student [21]; for instance, Neumann et al. compared the average cost per student of their RAG chatbot (USD 1.65 [21]) to that reported by Liu et al., who developed a set of AI-based educational tools featuring a GPT4-based chatbot without the RAG component (USD 1.90) [84].

Moreover, there are schemes designed specifically for the evaluation of RAG-based chatbots. The relatively most successful of them (according to the number of implementing research papers, though this amounts to an unimpressive 4 out of 47 analyzed papers [29,33,44,59]) is RAGAS, which considers three measures: Faithfulness, which assesses whether the generated answer is grounded in the given context and free of hallucinations; Answer Relevance, which measures the extent to which the answer addresses the actual question that was provided; and Context Relevance, which measures the extent to which the answer contains as little irrelevant information as possible [61]. Another one [85], still waiting to gain adoption, measures Correctness, Completeness, and Honesty, which are then aggregated into a single evaluation score.

There is also one scheme proposed for the evaluation of educational chatbots [86], which considers seven evaluation objectives (Acceptance & Adoption, Usability, Learning Success, Increased Motivation, Psychological Factors, Technical Correctness, and Further Beneficial Effects), and suggests four procedures to assess these factors (Wizard-of-Oz Experiment, Laboratory Study, Field Study, and Technical Validation) and five measurement instruments (Quantitative Survey, Quantitative Survey, Qualitative Interview, Discourse Analysis, and Analysis of Technical Log Files). None of the papers in the analyzed set, however, used it.

While an obvious purpose of such dedicated schemes is the multi-aspectual comparison of performance between various chatbots, with so-far few researchers having adopted them, they can still be used to compare the performance of different models that can power a given chatbot (e.g., Oldensand [33] compares GPT 3.5, GPT 4, and Llama 3 in terms of the respective RAGAS components, though the measured difference does not exceed the negligible level of 0.01 in any case) or the performance of a given chatbot in different usage scenarios (or types of questions)— for instance, Neupane et al. [29] compared the harmonic mean of RAGAS component measures across four use cases: Engineering Programs, General Inquiry, Research Opportunities, and University Resources (here, again, the measured difference is no more than 0.01 in all cases).

In their strive for a wide, multi-faceted chatbot evaluation, the analyzed studies very often combine measures of different character. Sometimes, they propose their own specific evaluation dimensions, e.g., Novelty, indicating that the conversation is providing fresh insights or perspectives that the user might not have considered; Engaging, measuring how interesting and captivating the conversation is for the user; Consistency, assessing whether the conversation turn contradicts previous statements or established facts; Repetition, indicating the degree to which the conversation turn repeats information that has already been provided) [20]; or Formulation, assessing whether the generated output fits the style expected for it [33].

In some papers, the Qualitative evaluation has been performed involving users (e.g., ref. [36]) or in the form of a case study by the authors themselves (e.g., ref. [52]). There were also papers containing no evaluation results whatsoever (e.g., ref. [40]) or barely mentioning them without providing concrete numbers (e.g., “promising results” in [19]).

In order to provide a comprehensible overview of the types of evaluation reported in the analyzed set of papers, we have aggregated the numbers of publications featuring similar measures, combining not only those having consistent definitions (e.g., Guidance and Suitability; Completeness and Comprehensiveness; or Faithfulness, Groundedness, and Truthfulness), but also those having similar interpretation, although addressing different concepts, e.g., User Acceptance and User Satisfaction or Readability and Clarity. The numbers are aggregated under the label of the most frequently found measure in Figure 5. Various measures used to assess the similarity of the generated text and natural text were aggregated into Quality of Generated Text. Answer Relevance was combined with Context Relevance into Relevance, as studies usually either report both measures or just one— General Relevance. Note that the numbers do not sum up to 47, as most papers used more than one measure.

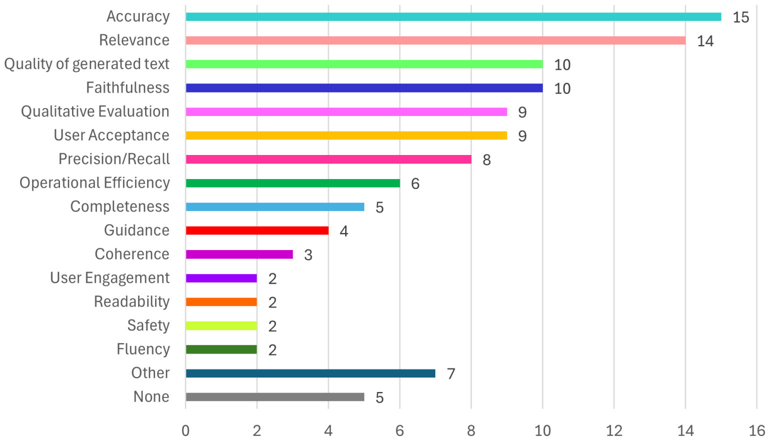

Figure 5. Evaluation scope.

As can be observed in Figure 5, there is a notable variety among the most frequently encountered evaluation criteria, which include measures typical to information-retrieval studies (Accuracy), chatbot studies (Relevance and Faithfulness), generative NLP (Quality of Generated Text), software usability (User Acceptance), and Qualitative Evaluation. The Other category includes measures mentioned in only one of the analyzed papers, including Consistency, Credibility, Formulation, Increased Motivation, Novelty, Repetition, and Formulation.

The evaluation may be performed automatically (by calculating metrics or running dedicated software tools) or with the involvement of humans (users, subject matter experts, or the authors themselves), with many studies combining these two approaches. In Figure 6, we can observe the share of papers in the analyzed set that reported the respective kind of evaluation.

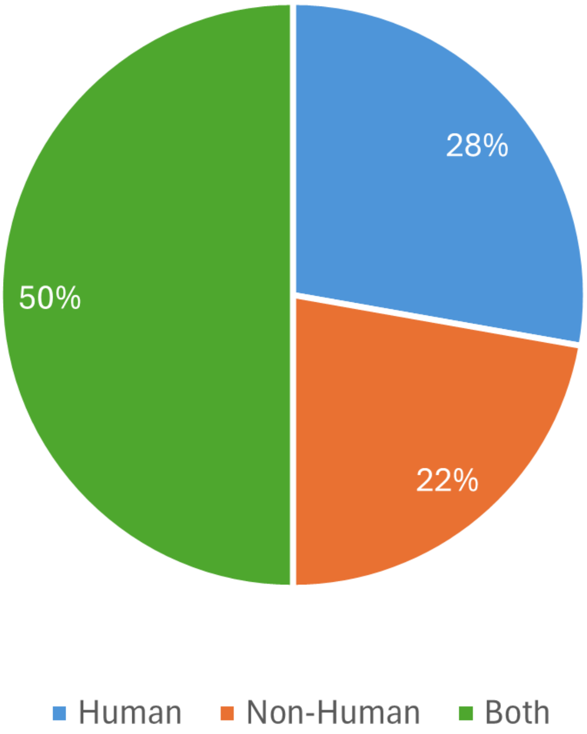

Figure 6. Human / non-human evaluation.

4. Discussion and Recommendations

The practical advantages of RAG make it a capable technological solution for the development of chatbots for education [53]. Although the number of their reported applications keeps growing, so far there has been no survey exploring what has been achieved so far in applying RAG chatbots in the educational domain, like there is for, e.g., medical sciences [9]. The presented work addresses this gap.

By a careful analysis of the identified body of the literature, we were able to answer the five key questions, helping us to picture the current state of the field. Regarding RQ1, the number of 47 identified relevant publications indicates that the application of RAG chatbots to the educational domain remains a fresh topic. Interestingly, this number is close to the number of relevant papers on the use of RAG in healthcare established by Amugongo et al. [87]—37, based on two bibliographic sources (PubMed and Google Scholar), showing a similar speed of RAG adoption in both fields. For a comparison, the survey of educational chatbots in general that was performed by Wollny et al. using five bibliographic sources already in 2021 reported 80 relevant papers [11]. The novelty of the area is confirmed by the character of the identified works, with only 13 refereed journal papers but 23 conference proceeding papers (having a much shorter submission-to-publication time span) and 9 preprints (having no publication delay at all).

With regard to RQ2, the majority of described chatbots (25) aim to support the students, with only 5 clearly aimed to support the teachers, and the remainder (20) are equally useful for both target user groups (these are the chatbots giving access to some subject-specific databases). The purpose of more than a half of the chatbots developed for students is supporting their learning process, whereas the remainder supports other aspects of their education, in most cases, their acclimatization in the educational institution (7) or their choice of courses (3). To put these numbers in context, roughly similar ratios were reported by Wollny et al., where, among the 80 analyzed papers, 36 were dedicated to supporting learning and 15 were providing assistance to students [11].

Regarding RQ3, the largest share of chatbots (11) deals with the theme of the institutional-domain knowledge. Focusing at the areas of education, the one most often supported with RAG chatbots is health sciences (nine instances), with computer science coming second (six instances). The interest in the use of chatbots in these two areas is well-established in the literature (see, e.g., refs. [9,88]).

In response to RQ4, we can clearly point to OpenAI's GPT as the range of models by far most frequently used in constructing RAG chatbots in the educational domain. This dominance is consistent with the findings of other authors, e.g., Raihan et al., who investigated various uses of LLMs in computer science education, reported 82 papers directly or indirectly using GPT among the 125 papers that they analyzed [88].

Regarding RQ5, we have identified a plethora of methods and criteria used in the evaluation of the educational RAG chatbots. Although several solutions specifically aimed at this particular area have been proposed, many of them highly relevant, they seem to lack either comprehensiveness in their consideration of the analyzed dimensions [61,85] or a detailed specification of evaluation measures and guidelines [86]; therefore, they cannot replace the other evaluation methods in use in their entirety. We also cannot indicate some methods as objectively better or worse than others, as in reality they all have some merits, covering different aspects of evaluation and having different purposes. Nonetheless, in an attempt to synthesize our findings on the evaluation methods in use and provide some recommendations for future evaluators of RAG chatbots in educational contexts, we have constructed Table 5, matching the aspect or purpose of evaluation with most adequate methods and measures.

Table 5. Recommended application area for evaluation methods in use.

Approaches (Measures)\Purposes	Response Content	Response Form	User Guidance	Value Added	User Experience	Specific Issues
Information retrieval (e.g., Precision, Recall, F1 Score)	R	-	-	-	-	-
LLM benchmarks (e.g. AGIEval)	r	r	R	-	-	-
Conversational quality (e.g. Correctness, Completeness, Communication)	r	R	-	-	-	-
Generated text quality (e.g., K-F1++, BLEURT, ROUGE)	r	R	-	-	-	-
Human-assessed text quality (e.g., Fluency, Coherence, Relevance)	r	R	-	-	r	-
Technical performance (e.g., operational efficiency, response time)	-	-	-	R	-	-
Economic performance (e.g., cost per query, cost per user)	-	-	-	R	-	-
User acceptance and satisfaction (e.g., SUS, Perceived Usability)	r	r	r	-	R	-
Overall RAG chatbot quality (e.g., RAGAS)	r	r	r	-	r	r
Qualitative evaluation (e.g., user interviews)	-	-	-	-	-	R

R—recommended; r—recommended with limitations of scope or depth.

Our most profound finding in this aspect is, however, the lack of even a single study in the analyzed set that would report the actual effects regarding the very purpose for which the given educational RAG chatbot has been developed. We consider this an appalling weakness in the evaluation of RAG chatbots in the educational domain calling for urgent attention. This allows us to propose some recommendations regarding the character of such goal-specific evaluation. In our opinion, its key properties should be the relevance to the target of support provided by the evaluated chatbot:

1. For chatbots supporting learning, we recommend comparing learning outcomes attained by students using a chatbot and the control group. Competency tests specific to the learned subject should be performed before the chatbot is presented to the students and after a reasonable time of learning with its support. The higher test results in the experimental group indicate the positive effect of the chatbot;
2. For chatbots providing access to source knowledge, we recommend comparing the project or exam grades received by students using a chatbot and the control group. The higher grades in the experimental group indicate the positive effect of the chatbot;
3. For chatbots providing access to information on assessment, we recommend comparing the number of questions regarding assessment asked by students using a chatbot and the control group. The lower number of questions asked by members of the experimental group indicate the positive effect of the chatbot;
4. For chatbots supporting students with organizational matters, we recommend comparing the number of help requests from students using a chatbot and the control group. The lower number of help requests by members of the experimental group indicate the positive effect of the chatbot;
5. For chatbots educating students on local culture, we recommend comparing the number of questions related to local customs asked by international students using a chatbot and the control group. The lower number of questions asked by members of the experimental group indicate the positive effect of the chatbot;
6. For chatbots supporting students in their course selection, we recommend comparing the students' drop-out ratio from courses selected after chatbot recommendation and the control group. The lower drop-out ratio in the experimental group indicates the positive effect of the chatbot;
7. For chatbots supporting teachers in question generation, we recommend comparing both the time of question development and the quality of the questions generated by a chatbot versus manually developed questions. This should consist in selecting a number of learning topics, then asking a small group of teachers to develop a partial set of questions relevant to these topics, one half without (and the other half with) the use of the chatbot, measuring their development time. Next, this should expose students to tests consisting of all questions, eventually either asking the

involved students to directly assess the individual test questions (in terms of question comprehensibility, relevance, difficulty, and fairness) or, provided that a sufficiently large group of respondents is available, performing a statistical analysis of test results, e.g., using one of the models based on Item Response Theory [89] (pp. 3–4). The shorter development time and/or the higher quality of questions developed with the help of the chatbot indicates its positive effect;

8. For chatbots supporting teachers in learning analytics, we recommend comparing the teacher-assessed ease and speed of obtaining specific insights regarding the effort of their students with the help of a chatbot and with the baseline learning analytics tool. This could be carried out by preparing task scenarios, then letting a small group of teachers perform them, measuring their time and success ratio. Task scenarios should have some context provided to make them engaging for the users, realistic, actionable, and free of clues or suggested solution steps [90]. The shorter execution time and/or the higher success ratio of tasks performed with the help of the chatbot indicate its positive effect.

In all the cases involving students listed above (1–6), it should be ensured that only students who actually used (not only knew of or logged in to) the chatbot are included in the experimental group results and that only students who did not use the chatbot are included in the control group results.

To finalize the discussion on the educational applications of RAG chatbots, the following general statements can be made:

1. The three key elements that affect the output of RAG chatbots are the underlying LLM, the source of content to augment the responses, and the approach chosen to generate vector representations of the queries and responses. A number of papers provide insights regarding the choice of the first [17,28,32,33,43,55], but only a few help with choosing the second [53] and the third [25,39,41];
2. As RAG chatbots are based on LLMs, they share their vulnerabilities with regard to data privacy and potential plagiarism [33]. The design and development of RAG chatbots should always strive to address any possible ethical and legal concerns;
3. While the availability of ready-to-use components makes the development of new RAG chatbots relatively easy, it is still a project burdened with many risks. A number of papers provide more-or-less detailed step-by-step descriptions of the development process, which can be followed by future implementers [25,27,29,40,51,54,56];
4. There are many aspects in which RAG chatbots can be evaluated, but the most important one is the extent to which they achieve their goal. Despite being unpopular in existing publications, this can be measured with some effort, as described in the recommendations listed above. Nonetheless, real-world RAG chatbot implementations are also constrained by technical and financial limitations, so pursuing the best available solution is not always a viable option.

5. Conclusions

In the presented work, we have investigated the quickly-developing area of educational applications of Retrieval-Augmented Generation chatbots. As these tools are believed to deal well with hallucinations, addressing the main barrier for the adoption of LLM-based chatbots in education, and are relatively easy to implement to address the various needs of students and teachers, we expect that the current scientific interest in the topic, evidenced by the 47 identified research papers, is just the beginning of much larger wave of publications to come. This indicates the primary contribution of our work, which is mapping the research veins on different kinds of the educational uses of RAG chatbots, which will help future authors to better position their work. We have also reviewed the evaluation criteria applied to educational RAG chatbots and indicated the lack of studies verifying the achievement of the specific aims of the respective types of chatbots in education. We hope that our recommendations in this regard could help future authors to prove the usefulness of their chatbots in addressing various educational purposes.

There are some limitations that may have affected this study. The first limitation stems from the use of just three bibliographic data sources. Although comprehensive and, as established by comparing their results, complementary, they may have possibly failed to identify all relevant papers. The second stems from failing to capture works that did not mention RAG chatbots in their titles, keywords, or abstracts in Scopus—though we have striven to mitigate this by both using Google Scholar employing a different search method and applying the snowballing procedure on the initially found set of publications. A third limitation is caused by the publication bias: while we addressed this by including preprints identified by Google Scholar in the performed analysis, the authors might still have been reluctant to publish papers about unsuccessful applications of RAG chatbots in education, even as preprints. Another limitation stems from considering only publications written in English. Papers written in low-resource languages may be more probable to describe RAG chatbots using such languages, for which LLMs may perform significantly worse, negatively impacting the chatbot output quality and limiting the range of its possible applications.

The last limitation indicates an interesting future research direction, i.e., comparing educational-domain RAG chatbots performance with regard to the used language of communication. A much wider future research direction would be performing a meta-analysis of RAG chatbots serving various purposes that would measure and compare their effectiveness in attaining the intended outcomes (e.g., improving learning effects or making students adapt faster to new educational environments). This, however, has to wait for the field to mature and a substantial number of relevant papers reporting the outcomes of using the educational RAG chatbots (following our recommendations in this regard or not) to become available for the analysis.

Author Contributions: Conceptualization, J.S.; methodology, J.S.; validation, J.S. and M.G.; investigation, M.G. and J.S.; resources, J.S. and M.G.; data curation, M.G. and J.S.; writing—original draft preparation, J.S.; writing—review and editing, J.S. and M.G.; visualization, M.G.; supervision, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Uyen Tran, L. Dialogue in education. *J. Sci. Commun.* **2008**, *7*, C04. [CrossRef]
2. Roumeliotis, K.I.; Tselikas, N.D. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet* **2023**, *15*, 192. [CrossRef]
3. Han, J.; Yoo, H.; Myung, J.; Kim, M.; Lee, T.Y.; Ahn, S.Y.; Oh, A.; Answer, A.N. Exploring Student-ChatGPT Dialogue in EFL Writing Education. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems. Neural Information Processing Systems Foundation, New Orleans, LA, USA, 10–16 December 2023.
4. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
5. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]

6. Yu, H.; Gan, A.; Zhang, K.; Tong, S.; Liu, Q.; Liu, Z. Evaluation of Retrieval-Augmented Generation: A Survey. In *Proceedings of the Big Data*; Zhu, W., Xiong, H., Cheng, X., Cui, L., Dou, Z., Dong, J., Pang, S., Wang, L., Kong, L., Chen, Z., Eds.; Springer: Singapore, 2025; pp. 102–120.
7. Gupta, S.; Ranjan, R.; Singh, S.N. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. *arXiv* **2024**, arXiv:2410.12837.
8. Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; Cui, B. Retrieval-Augmented Generation for AI-Generated Content: A Survey. *arXiv* **2024**, arXiv:2402.19473.
9. Bora, A.; Cuayáhuitl, H. Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2355–2374. [CrossRef]
10. Pérez, J.Q.; Daradoumis, T.; Puig, J.M.M. Rediscovering the use of chatbots in education: A systematic literature review. *Comput. Appl. Eng. Educ.* **2020**, *28*, 1549–1565. [CrossRef]
11. Wollny, S.; Schneider, J.; Di Mitri, D.; Weidlich, J.; Rittberger, M.; Drachsler, H. Are We There Yet?—A Systematic Literature Review on Chatbots in Education. *Front. Artif. Intell.* **2021**, *4*, 654924. [CrossRef]
12. Okonkwo, C.W.; Ade-Ibijola, A. Chatbots applications in education: A systematic review. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100033. [CrossRef]
13. Nouzri, S.; EL Fatimi, M.; Guerin, T.; Othmane, M.; Najjar, A. Beyond Chatbots: Enhancing Luxembourgish Language Learning Through Multi-Agent Systems and Large Language Model. In Proceedings of the PRIMA 2024: Principles and Practice of Multi-Agent Systems 25th International Conference, Kyoto, Japan, 18–24 November 2024; Arisaka R., Ito T., Sanchez-Anguix V., Stein S., Aydoğan R., van der Torre L., Eds.; Lecture Notes in Artificial Intelligence (LNAI); Springer: Cham, Switzerland, 2025; Volume 15395 , pp. 385–401. [CrossRef]
14. Arun, G.; Perumal, V.; Urias, F.; Ler, Y.; Tan, B.; Vallabhajosyula, R.; Tan, E.; Ng, O.; Ng, K.; Mogali, S. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: A comparative pilot study. *Anat. Sci. Educ.* **2024**, *17*, 1396–1405. [CrossRef] [PubMed]
15. Kelly, A.; Noctor, E.; Van De Ven, P. Design, architecture and safety evaluation of an AI chatbot for an educational approach to health promotion in chronic medical conditions. In Proceedings of the 12th Scientific Meeting on International Society for Research on Internet Interventions, ISRII-12 2024, Limerick, Ireland, 9–14 October 2023; Anderson P.P., Teles A., Esfandiari N., Badawy S.M., Eds.; Procedia Computer Science; Elsevier B.V.: Amsterdam, The Netherlands, 2024, Volume 248, pp. 52–59. [CrossRef]
16. Zhao, Z.; Yin, Z.; Sun, J.; Hui, P. Embodied AI-Guided Interactive Digital Teachers for Education. In Proceedings of the SIGGRAPH Asia 2024 Educator's Forum, SA 2024, Tokyo, Japan, 3–6 December 2024; Spencer, S.N., Ed.; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
17. Šarčević, A.; Tomičić, I.; Merlin, A.; Horvat, M. Enhancing Programming Education with Open-Source Generative AI Chatbots. In Proceedings of the 47th ICT and Electronics Convention, MIPRO 2024, Opatija, Croatia, 20–24 May 2024; Babic S., Car Z., Cicin-Sain M., Cisic D., Ergovic P., Grbac T.G., Gradisnik V., Gros S., Jokic A., Jovic A., et al., Eds.; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024; pp. 2051–2056. [CrossRef]
18. Thway, M.; Recatala-Gomez, J.; Lim, F.S.; Hippalgaonkar, K.; Ng, L.W. Harnessing GenAI for Higher Education: A Study of a Retrieval Augmented Generation Chatbot's Impact on Human Learning. *arXiv* **2024**, arXiv:2406.07796.
19. Abraham, S.; Ewards, V.; Terence, S. Interactive Video Virtual Assistant Framework with Retrieval Augmented Generation for E-Learning. In Proceedings of the 3rd International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2024, Salem, India, 5–7 June 2024; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024, pp. 1192–1199. [CrossRef]
20. Jiang, Y.; Shao, Y.; Ma, D.; Semnani, S.J.; Lam, M.S. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. *arXiv* **2024**, arXiv:2408.15232.
21. Neumann, A.; Yin, Y.; Sowe, S.; Decker, S.; Jarke, M. An LLM-Driven Chatbot in Higher Education for Databases and Information Systems. *IEEE Trans. Educ.* **2025**, *68*, 103–116. [CrossRef]
22. Levonian, Z.; Li, C.; Zhu, W.; Gade, A.; Henkel, O.; Postle, M.E.; Xing, W. Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference. In Proceedings of the NeurIPS'23 Workshop: Generative AI for Education (GAIED), New Orleans, LA, USA, 15 December 2023.
23. Soliman, H.; Kravcik, M.; Neumann, A.; Yin, Y.; Pengel, N.; Haag, M. Scalable Mentoring Support with a Large Language Model Chatbot. In Proceedings of the 19th European Conference on Technology Enhanced Learning, EC-TEL 2024, Krems, Austria, 16–20 September 2024; Ferreira Mello, R., Rummel, N., Jivet, I., Pishtari, G., Ruipérez Valiente, J.A., Eds.; Lecture Notes in Computer Science (LNCS); Springer Science and Business Media: Berlin/Heidelberg, Germany, 2024; Volume 15160, pp. 260–266. [CrossRef]

24. Al Ghadban, Y.; Lu, H.Y.; Adavi, U.; Sharma, A.; Gara, S.; Das, N.; Kumar, B.; John, R.; Devarsetty, P.; Hirst, J.E. Transforming Healthcare Education: Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation. In Proceedings of the NeurIPS'23 Workshop: Generative AI for Education (GAIED), New Orleans, LA, USA, 15 December 2023.
25. Islam, M.; Warsito, B.; Nurhayati, O. AI-driven chatbot implementation for enhancing customer service in higher education: A case study from Universitas Negeri Semarang. *J. Theor. Appl. Inf. Technol.* **2024**, *102*, 5690–5701.
26. Saad, M.; Qawaqneh, Z. Closed Domain Question-Answering Techniques in an Institutional Chatbot. In Proceedings of the International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2024, Male, Maldives, 4–6 November 2024; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024. [CrossRef]
27. Olawore, K.; McTear, M.; Bi, Y. Development and Evaluation of a University Chatbot Using Deep Learning: A RAG-Based Approach. In Proceedings of the CONVERSATIONS 2024—The 8th International Workshop on Chatbots and Human-Centred AI, 2025, Thessaloniki, Greece, 4–5 December 2024.
28. Chen, Z.; Zou, D.; Xie, H.; Lou, H.; Pang, Z. Facilitating university admission using a chatbot based on large language models with retrieval-augmented generation. *Educ. Technol. Soc.* **2024**, *27*, 454–470. [CrossRef]
29. Neupane, S.; Hossain, E.; Keith, J.; Tripathi, H.; Ghiasi, F.; Golilarz, N.A.; Amirlatifi, A.; Mittal, S.; Rahimi, S. From Questions to Insightful Answers: Building an Informed Chatbot for University Resources. *arXiv* **2024**, arXiv:2405.08120.
30. Bimberg, R.; Quibeldey-Cirkel, K. AI assistants in teaching and learning: Compliant with data protection regulations and free from hallucinations! In Proceedings of the EDULEARN24, IATED, 2024, Palma, Spain, 1–3 July 2024.
31. Hicke, Y.; Agarwal, A.; Ma, Q.; Denny, P. AI-TA: Towards an Intelligent Question-Answer Teaching Assistant Using Open-Source LLMs. *arXiv* **2023**, arXiv:2311.02775.
32. Zhang, T.; Patterson, L.; Beiting-Parrish, M.; Webb, B.; Abeysinghe, B.; Bailey, P.; Sikali, E. Ask NAEP: A Generative AI Assistant for Querying Assessment Information. *J. Meas. Eval. Educ. Psychol.* **2024**, *15*, 378–394. [CrossRef]
33. Oldensand, V. Developing a RAG System for Automatic Question Generation: A Case Study in the Tanzanian Education Sector. Master's Thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2024.
34. Klymkowsky, M.; Cooper, M.M. The end of multiple choice tests: Using AI to enhance assessment. *arXiv* **2024**, arXiv:2406.07481.
35. Saha, B.; Saha, U. Enhancing International Graduate Student Experience through AI-Driven Support Systems: A LLM and RAG-Based Approach. In Proceedings of the 2024 International Conference on Data Science and Its Applications, ICoDSA 2024, Kuta, Bali, Indonesia, 10–11 July 2024; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024; pp. 300–304. [CrossRef]
36. Wong, L. Gaita: A RAG System for Personalized Computer Science Education. Master's Thesis, Johns Hopkins University, Baltimore, MD, USA, 2024.
37. Amarnath, N.; Nagarajan, R. An Intelligent Retrieval Augmented Generation Chatbot for Contextually-Aware Conversations to Guide High School Students. In Proceedings of the 4th International Conference on Sustainable Expert Systems, ICSES 2024, Lekhnath, Nepal, 15–17 October 2024; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024; pp. 1393–1398. [CrossRef]
38. Rao, J.; Lin, J. RAMO: Retrieval-Augmented Generation for Enhancing MOOCs Recommendations. In Proceedings of the Educational Datamining '24 Human-Centric eXplainable AI in Education and Leveraging Large Language Models for Next-Generation Educational Technologies Workshop Joint Proceedings, Atlanta, GA, USA, 13 July 2024; CEUR-WS; Volume 3840.
39. Yan, L.; Zhao, L.; Echeverria, V.; Jin, Y.; Alfredo, R.; Li, X.; Gašević , D.; Martinez-Maldonado, R. VizChat: Enhancing Learning Analytics Dashboards with Contextualised Explanations Using Multimodal Generative AI Chatbots. In Proceedings of the 25th International Conference on Artificial Intelligence in Education, AIED 2024, Recife, Brazil, 8–12 July 2024; Olney A.M., Chounta I.-A., Liu Z., Santos O.C., Bittencourt I.I., Eds.; Lecture Notes in Artificial Intelligence (LNAI); Springer Science and Business Media: Berlin/Heidelberg, Germany, 2024; Volume 14830 , pp. 180–193. [CrossRef]
40. Forootani, A.; Aliabadi, D.; Thraen, D. Bio-Eng-LMM AI Assist chatbot: A Comprehensive Tool for Research and Education. *arXiv* **2024**, arXiv:2409.07110.
41. Wiratunga, N.; Abeyratne, R.; Jayawardena, L.; Martin, K.; Massie, S.; Nkisi-Orji, I.; Weerasinghe, R.; Liret, A.; Fleisch, B. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. In Proceedings of the Case-Based Reasoning Research and Development, Merida, Mexico, 1–4 July 2024; Recio-Garcia, J.A., Orozco-del Castillo, M.G., Bridge, D., Eds.; Springer: Cham, Switzerland, 2024; pp. 445–460.
42. Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; Zhang, Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus* **2023**, *15*, e40895. [CrossRef] [PubMed]
43. Cui, J.; Ning, M.; Li, Z.; Chen, B.; Yan, Y.; Li, H.; Ling, B.; Tian, Y.; Yuan, L. Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model. *arXiv* **2024**, arXiv:2306.16092.

44. Dean, M.; Bond, R.; McTear, M.; Mulvenna, M. ChatPapers: An AI Chatbot for Interacting with Academic Research. In Proceedings of the 2023 31st Irish Conference on Artificial Intelligence and Cognitive Science, AICS 2023, Letterkenny, Ireland, 7–8 December 2023; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2023. [CrossRef]
45. Bhavani Peri, S.; Santhanalakshmi, S.; Radha, R. Chatbot to chat with medical books using Retrieval-Augmented Generation Model. In Proceedings of the NKCon 2024—3rd Edition of IEEE NKSS's Flagship International Conference: Digital Transformation: Unleashing the Power of Information, Bagalkot, India, 21–22 September 2024; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024. [CrossRef]
46. Maryamah, M.; Irfani, M.; Tri Raharjo, E.; Rahmi, N.; Ghani, M.; Raharjana, I. Chatbots in Academia: A Retrieval-Augmented Generation Approach for Improved Efficient Information Access. In Proceedings of the KST 2024—16th International Conference on Knowledge and Smart Technology, Krabi, Thailand, 28 February–2 March 2024; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024; pp. 259–264. [CrossRef]
47. Richard, R.; Veemaraj, E.; Thomas, J.; Mathew, J.; Stephen, C.; Koshy, R. A Client-Server Based Educational Chatbot for Academic Institutions. In Proceedings of the 2024 4th International Conference on Intelligent Technologies, CONIT 2024, Bangalore, India, 21–23 June 2024; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024. [CrossRef]
48. Dayarathne, R.; Ranaweera, U.; Ganegoda, U. Comparing the Performance of LLMs in RAG-Based Question-Answering: A Case Study in Computer Science Literature. In Proceedings of the Artificial Intelligence in Education Technologies: New Development and Innovative Practices, Barcelona, Spain, 29–31 July 2024; Schlippe, T., Cheng, E.C.K., Wang, T., Eds.; Springer: Singapore, 2025; pp. 387–403.
49. Burgan, C.; Kowalski, J.; Liao, W. Developing a Retrieval Augmented Generation (RAG) Chatbot App Using Adaptive Large Language Models (LLM) and LangChain Framework. *Proc. West Va. Acad. Sci.* **2024**, *96*. [CrossRef]
50. Ge, J.; Sun, S.; Owens, J.; Galvez, V.; Gologorskaya, O.; Lai, J.C.; Pletcher, M.J.; Lai, K. Development of a liver disease–specific large language model chat interface using retrieval-augmented generation. *Hepatology* **2024**, *80*, 1158–1168. [CrossRef]
51. Alawwad, H.A.; Alhothali, A.; Naseem, U.; Alkhathlan, A.; Jamal, A. Enhancing textual textbook question answering with large language models and retrieval augmented generation. *Pattern Recognit.* **2025**, *162*, 111332. [CrossRef]
52. Miao, J.; Thongprayoon, C.; Suppadungsuk, S.; Garcia Valencia, O.A.; Cheungpasitporn, W. Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications. *Medicina* **2024**, *60*, 445. [CrossRef]
53. Wölfel, M.; Shirzad, M.; Reich, A.; Anderer, K. Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice's Cooperative Principles and Trust. *Big Data Cogn. Comput.* **2024**, *8*, 2. [CrossRef]
54. Cooper, M.; Klymkowsky, M. Let us not squander the affordances of LLMS for the sake of expedience: Using retrieval augmented generative AI chatbots to support and evaluate student reasoning. *J. Chem. Educ.* **2024**, *101*, 4847–4856. [CrossRef]
55. Amri, S.; Bani, S.; Bani, R. Moroccan legal assistant Enhanced by Retrieval-Augmented Generation Technology. In Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security, NISS 2024, Meknes, Morocco, 18–19 April 2024; ACM International Conference Proceeding Series; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
56. Lála, J.; O'Donoghue, O.; Shtedritski, A.; Cox, S.; Rodriques, S.G.; White, A.D. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research. *arXiv* **2023**, arXiv:2312.07559.
57. Wang, C.; Ong, J.; Wang, C.; Ong, H.; Cheng, R.; Ong, D. Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation. *Ann. Biomed. Eng.* **2024**, *52*, 1115–1118. [CrossRef] [PubMed]
58. Sree, Y.; Sathvik, A.; Hema Akshit, D.; Kumar, O.; Pranav Rao, B. Retrieval-Augmented Generation Based Large Language Model Chatbot for Improving Diagnosis for Physical and Mental Health. In Proceedings of the 6th International Conference on Electrical, Control and Instrumentation Engineering, ICECIE 2024, Pattaya, Thailand, 23 November 2024; Institute of Electrical and Electronics Engineers: New York, NY, USA, 2024. [CrossRef]
59. Habib, M.; Amin, S.; Oqba, M.; Jaipal, S.; Khan, M.; Samad, A. TaxTajweez: A Large Language Model-based Chatbot for Income Tax Information In Pakistan Using Retrieval Augmented Generation (RAG). In Proceedings of the 37th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2024, Miramar Beach, FL, USA, 19–21 May 2024; Volume 37.
60. Chen, J.; Lin, H.; Han, X.; Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 17754–17762. [CrossRef]
61. ES, S.; James, J.; Anke, L.E.; Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024—System Demonstrations, St. Julians, Malta, 17–22 March 2024; Aletras, N., Clercq, O.D., Eds.; Association for Computational Linguistics: St. Stroudsburg, PA, USA, 2024; pp. 150–158.
62. Giner, F. An Intrinsic Framework of Information Retrieval Evaluation Measures. *In Proceedings of the Intelligent Systems and Applications*; Arai, K., Ed.; Springer: Cham, Switzerland, 2024; pp. 692–713.
63. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [CrossRef]

64. Lan, F. Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method. *Adv. Multimed.* **2022**, *2022*, 1–11. [CrossRef]
65. Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; Duan, N. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, 16–21 June 2024; pp. 2299–2314. [CrossRef]
66. Aridoss, M.; Bisht, K.S.; Natarajan, A.K. Comprehensive Analysis of Falcon 7B: A State-of-the-Art Generative Large Language Model. In *Generative AI: Current Trends and Applications*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 147–164.
67. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
68. Grice, H.P. *Studies in the Way of Words*; Harvard University Press: Cambridge, MA, USA, 1991.
69. Koubaa, A. GPT-4 vs. GPT-3.5: A Concise Showdown. *Preprints* **2023**. [CrossRef]
70. Chiesurin, S.; Dimakopoulos, D.; Sobrevilla Cabezudo, M.A.; Eshghi, A.; Papaioannou, I.; Rieser, V.; Konstas, I. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. *arXiv* **2023**, arXiv:2305.16519. [CrossRef]
71. Hanna, M.; Bojar, O. A Fine-Grained Analysis of BERTScore. In Proceedings of the Sixth Conference on Machine Translation, Online, 10–11 November 2021; Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M.R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., et al., Eds.; pp. 507–517.
72. Sellam, T.; Das, D.; Parikh, A. BLEURT: Learning Robust Metrics for Text Generation. *arXiv* **2020**, arXiv:2004.04696. [CrossRef]
73. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
74. Liau, T.L.; Bassin, C.B.; Martin, C.J.; Coleman, E.B. Modification of the Coleman readability formulas. *J. Read. Behav.* **1976**, *8*, 381–386. [CrossRef]
75. Solnyshkina, M.; Zamaletdinov, R.; Gorodetskaya, L.; Gabitov, A. Evaluating text complexity and Flesch-Kincaid grade level. *J. Soc. Stud. Educ. Res.* **2017**, *8*, 238–248.
76. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
77. Lian, W.; Goodson, B.; Wang, G.; Pentland, E.; Cook, A.; Vong, C. "Teknium". MistralOrca: Mistral-7B Model Instruct-Tuned on Filtered OpenOrcaV1 GPT-4 Dataset. 2023. Available online: https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca (accessed on 4 April 2025).
78. Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; von Werra, L.; Fourrier, C.; Habib, N.; et al. Zephyr: Direct Distillation of LM Alignment. *arXiv* **2023**, arXiv:2310.16944.
79. Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; von Werra, L.; Fourrier, C.; Habib, N.; et al. Zephyr 7B Beta. 2024. Available online: https://huggingface.co/HuggingFaceH4/zephyr-7b-beta (accessed on 4 April 2025).
80. Lin, C.C. Exploring the relationship between technology acceptance model and usability test. *Inf. Technol. Manag.* **2013**, *14*, 243–255. [CrossRef]
81. Lewis, J.R. The system usability scale: Past, present, and future. *Int. J. Human Comput. Interact.* **2018**, *34*, 577–590. [CrossRef]
82. Vlachogianni, P.; Tselios, N. Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review. *J. Res. Technol. Educ.* **2022**, *54*, 392–409. [CrossRef]
83. Montella, R.; De Vita, C.G.; Mellone, G.; Di Luccio, D.; Maskeliūnas, R.; Damaševičius, R.; Blažauskas, T.; Queirós, R.; Kosta, S.; Swacha, J. Re-Assessing the Usability of FGPE Programming Learning Environment with SUS and UMUX. In Proceedings of the Information Systems Education Conference, Virtual, 19 October 2024; Foundation for IT Education: Chicago, IL, USA, 2024; pp. 128–135.
84. Liu, R.; Zenke, C.; Liu, C.; Holmes, A.; Thornton, P.; Malan, D.J. Teaching CS50 with AI: Leveraging Generative Artificial Intelligence in Computer Science Education. In Proceedings of the 55th ACM Technical Symposium on Computer Science Education, Portland, OR, USA, 20–23 March 2024; Volume 1, pp. 750–756. [CrossRef]
85. Wang, Y.; Hernandez, A.G.; Kyslyi, R.; Kersting, N. Evaluating Quality of Answers for Retrieval-Augmented Generation: A Strong LLM Is All You Need. *arXiv* **2024**, arXiv:2406.18064.
86. Hobert, S. How Are You, Chatbot? Evaluating Chatbots in Educational Settings—Results of a Literature Review. In *DELFI 2019*; Gesellschaft für Informatik e.V.: Bonn, Germany, 2019; pp. 259–270. [CrossRef]
87. Amugongo, L.M.; Mascheroni, P.; Brooks, S.G.; Doering, S.; Seidel, J. Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review. 2024. Available online: https://www.preprints.org/manuscript/202407.0876/v1 (accessed on 4 April 2025).
88. Raihan, N.; Siddiq, M.L.; Santos, J.C.; Zampieri, M. Large Language Models in Computer Science Education: A Systematic Literature Review. In Proceedings of the 56th ACM Technical Symposium on Computer Science Education, Pittsburgh, PA, USA, 26 February–1 March 2025; Volume 1, pp. 938–944. [CrossRef]

89. Brown, G.T.L.; Abdulnabi, H.H.A. Evaluating the Quality of Higher Education Instructor-Constructed Multiple-Choice Tests: Impact on Student Grades. *Front. Educ.* **2017**, *2*, 24. [CrossRef]

90. McCloskey, M. Turn User Goals into Task Scenarios for Usability Testing. 2014. Available online: https://www.nngroup.com/articles/task-scenarios-usability-testing/ (accessed on 4 April 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.