mathematics

MDPI

Article

An Integrated AHP–Fuzzy AHP Evaluation Framework for Large Language Models in Software Engineering Education

Jovana Lj. Jović [1], Dragan S. Domazet [1], Nenad O. Vesić [2], Branislav M. Ranđelović [3,4,*] and Dušan J. Simjanović [1]

[1] Faculty of Information Technology, Belgrade Metropolitan University, Tadeuša Košćuška 63, 11000 Belgrade, Serbia; jovana.jovic@metropolitan.ac.rs (J.L.J.); dragan.domazet@metropolitan.ac.rs (D.S.D.); dusan.simjanovic@metropolitan.ac.rs (D.J.S.)
[2] Mathematical Institute of Serbian Academy of Sciences and Arts, University of Belgrade, 11000 Belgrade, Serbia; n.o.vesic@turing.mi.sanu.ac.rs
[3] Faculty of Electronic Engineering, University of Nis, 18000 Nis, Serbia
[4] Faculty of Teachers Education, University of Kosovska Mitrovica, 38218 Leposavic, Serbia
* Correspondence: bane@elfak.ni.ac.rs

Abstract

The use of large language models (LLMs) in higher education has increased significantly, and their potential for supporting teaching and learning is considerable. However, their reliability and suitability for generating educational content remain open questions, particularly in technically demanding fields such as software engineering. This paper proposes a multi-criteria framework for assessing the quality of educational content generated by LLMs. The framework is based on existing open educational resource (OER) evaluation rubrics, which were adapted for the assessment of LLM-generated content and further refined based on expert evaluation and consultation. The evaluation was conducted by a panel of eight experts from software engineering, artificial intelligence, education, and related fields, using predefined criteria and pairwise comparisons. The framework was applied to five contemporary LLMs across three selected topics in software engineering. The relative importance of the criteria was determined using the Analytic Hierarchy Process (AHP) and its fuzzy extension (FAHP). The results show that accuracy and professional correctness represent the most important criterion, while visual presentation and language style have the least influence. The findings also indicate differences across models and a high level of agreement between AHP and FAHP rankings.

Keywords: large language models (LLMs); software engineering education; multi-criteria decision making; AHP; fuzzy AHP; educational content evaluation; AI in education

MSC: 68T50; 03E72; 68T05

Academic Editor: Ignacio Javier Perez Galvez

Received: 6 April 2026
Revised: 29 April 2026
Accepted: 9 May 2026
Published: 12 May 2026

1. Introduction

The use of large language models (LLMs) in education has changed the way students learn and the way instructors prepare teaching materials and organize courses [1,2]. LLMs can help students gain a better understanding of the learning content, obtain additional explanations, and access additional learning resources. However, their impact on learning is not straightforward. Empirical evidence shows that increased reliance on LLMs, particularly for tasks requiring critical thinking such as code generation and debugging, may negatively affect student performance, while their use for explanations appears more beneficial, indicating a complex and non-trivial role in educational contexts [3].

Therefore, it is necessary to examine whether the generated content is accurate, pedagogically useful, and aligned with course learning objectives [4]. The evaluation of LLM-generated educational content represents a multi-dimensional problem that involves multiple criteria such as accuracy, depth of explanation, pedagogical clarity, structure, and practical applicability. These aspects cannot be adequately evaluated using single-metric approaches. Therefore, the evaluation of LLM-generated educational content can be formulated as a multi-criteria decision-making problem, where multiple criteria must be considered simultaneously [5–7]. This is particularly important in disciplines that require a high level of conceptual and practical precision, such as software engineering, which combines theoretical foundations with practical application and requires students not only to understand fundamental concepts but also to develop analytical thinking, design skills, and professional judgment [8,9].

In the domain of multi-criteria decision-making (MCDM), various methods such as TOPSIS, VIKOR, and ELECTRE have been widely applied. However, the selection of an appropriate method depends on the nature of the decision problem and the characteristics of the available data. In this study, the evaluation of large language models in software engineering education is inherently expert-driven, qualitative, and hierarchically structured, involving criteria whose relative importance is not known a priori and cannot be directly quantified. Under such conditions, the Analytic Hierarchy Process (AHP) is particularly suitable, as it enables the decomposition of complex decision problems into hierarchical levels and supports intuitive pairwise comparisons for eliciting expert judgments. In addition, AHP provides a consistency verification mechanism (the consistency ratio), which is essential when working with subjective inputs. While alternative methods such as TOPSIS and VIKOR are effective in scenarios with well-defined quantitative performance indicators, they typically rely on predefined weights or distance-based ranking approaches. In contrast, the present study requires deriving criteria weights directly from expert knowledge, making AHP a more appropriate choice. To further enhance the robustness of the evaluation, the AHP framework is extended using fuzzy logic (FAHP), which allows the modeling of uncertainty and vagueness inherent in human judgment. This extension contributes to more balanced weight distributions and improves the interpretability and stability of the results without altering the overall ranking structure. Therefore, the selection of AHP/FAHP in this study is based on its alignment with the problem structure and data characteristics, ensuring a reliable and transparent evaluation framework. This "fit-for-purpose" selection ensures that the chosen method is consistent with both the decision context and the type of available information.

While research on AI in education has expanded in recent years, fewer studies have focused on systematic evaluations of LLM performance in actual educational contexts [10,11]. The current literature often deals with broader questions, such as: the general role of LLMs in teaching and learning [12], issues of academic integrity related to AI-generated writing [13], and technical challenges such as hallucinations and factual reliability in generated responses [14]. Only a small number of studies have applied structured, multi-criteria methodological frameworks for evaluation of the pedagogical suitability of LLM responses within specific educational domains [15]. While the Delphi–AHP approach has been used to evaluate AI systems in education [15], the problem of evaluating the quality of LLM-generated educational content using fuzzy multi-criteria decision-making methods remains insufficiently explored. Most existing research focuses on the general capabilities, opportunities, and risks of using LLMs in education rather than systematically evaluating the quality of the learning content they generate [4,10,16]. In software engineering education, the quality of educational content depends on several interrelated factors, such as content accuracy, depth of explanation, clarity of presentation, content organization, and practical

applicability [9,17]. Therefore, evaluating the quality of such content requires an approach that can determine the relative importance of evaluation criteria and support decision-making under uncertainty. Although some studies have proposed structured frameworks for evaluating educational AI systems and LLMs, these approaches are typically based on general performance metrics, rubric-based assessment, or expert judgment without a formal multi-criteria mathematical model. In addition, uncertainty and subjectivity in expert evaluations are often not explicitly modeled. Therefore, there is a need for a hierarchical multi-criteria decision-making model that incorporates fuzzy logic to handle uncertainty in the evaluation of LLM-generated educational content, particularly in technical and engineering education [5,18].

Based on the identified research gap, this paper proposes a hierarchical multi-criteria evaluation framework for assessing and ranking the quality of educational content generated by large language models. The proposed framework represents an integration and application of established MCDM methods rather than the development of a fundamentally new methodology. The proposed framework is based on the Analytic Hierarchy Process (AHP) and its fuzzy extension (Fuzzy AHP), which enable the determination of criteria weights and ranking of alternatives under uncertainty and subjective expert judgment. Existing OER rubrics provide a solid basis for evaluating educational content, but they do not fully capture the specific characteristics and challenges of content generated by LLMs, especially in software engineering education. In this paper, existing OER rubrics were used as a starting point, but their criteria were systematically adapted and extended to reflect key aspects of LLM-generated content, including technical and professional correctness, control of hallucinations, variability in the quality of explanations, and alignment with curricula, academic and instructional requirements. Particular emphasis was also placed on the integration of theory and practice, which is a key requirement in software engineering education.

Accordingly, this study addresses the following research questions:

- RQ1: Can the proposed framework support the evaluation of LLM-generated content in software engineering education?
- RQ2: Does a highly influencing sub-criteria exist?
- RQ3: Are there significant changes in the sub-factor rankings when the five values of an optimism index in the FAHP method are used?
- RQ4: Can the combination of AHP and Fuzzy AHP provide a robust framework for evaluating and ranking the quality of educational content generated by large language models?

The evaluation framework developed in this study is based on existing open educational resource (OER) evaluation rubrics, primarily the Open Textbook rubric [19] and the Achieve OER rubric [20,21], which define the quality of educational content through dimensions such as content accuracy, coverage and depth, clarity of explanation, organization, and structure, as well as linguistic and technical correctness. These dimensions are adapted to the context of software engineering and the evaluation of content generated by LLMs, resulting in a multi-criteria evaluation framework consisting of six groups of criteria and twenty-five sub-criteria. The proposed framework is applied to the evaluation of educational content generated by five contemporary large language models: GPT-4o, Claude 3.5 Sonnet, Gemini 2.5 Flash, DeepSeek-R1 and Llama 3.3 70B. The models are used to generate educational content for three selected topics in software engineering.

The increasing use of LLMs in education, together with concerns regarding the quality and reliability of generated content, highlights the need for systematic evaluation approaches. To address the identified research gap, this study develops a structured and integrated framework for evaluating the quality of educational content generated by LLMs,

particularly in the context of software engineering education. The main contribution lies in the application of multi-criteria decision-making (MCDM) to educational content evaluation, enabling a systematic, transparent, and reproducible assessment process. The proposed approach is based on the AHP and Fuzzy AHP methods, which support the determination of criteria importance while accounting for uncertainty in expert judgments. The main contributions of this paper are as follows:

- A multi-criteria evaluation framework for assessing the quality of educational content generated by LLMs is developed.
- Evaluation criteria are defined and extended with new criteria specific to LLM-generated educational content.
- The AHP and Fuzzy AHP methods are integrated to model uncertainty in expert-based evaluation.
- The stability of the ranking results is analyzed for different values of the optimism index in the FAHP method.
- An empirical comparison of five LLMs is conducted based on the defined evaluation criteria.

This paper is organized as follows: Section 2 presents the literature review. Section 3 describes the proposed multi-criteria evaluation framework. Section 4 presents the results of empirical evaluation, followed by a discussion. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Literature Review

LLMs are used in higher education as tools for learning support, explanation generation, writing assistance, programming help, and preparation of teaching materials [22,23]. The complexity of higher education environments has been highlighted in previous studies, where multiple interacting factors influence educational processes and outcomes, indicating the need for comprehensive and structured evaluation approaches [24].

Systematic literature reviews show that these systems are used for tutoring, content generation, automated feedback, and research support, but that there are also concerns related to the accuracy of information, academic integrity, and student over-reliance on AI systems [22,25]. In addition, recent research shows that artificial intelligence can be used to track student progress and generate personalized learning recommendations using clustering and explainable artificial intelligence methods, providing transparent and understandable feedback to students [26]. For this reason, the literature emphasizes that LLMs should be used as supplementary learning tools rather than as a replacement for teachers, with a focus on responsible and transparent use of artificial intelligence in education [27].

Several studies have examined the use of LLMs as tutoring systems in software engineering and programming education. For example, in [28], LLMs are used as tutors in software engineering education where the results showed that the models can provide fast support, explanations, and help during problem solving, especially when instructor support is limited. Another study developed a ChatGPT-based intelligent tutor for Python, and an eleven-week experiment with an experimental and a control group showed improved student engagement, higher task completion rates, and better results for students with weaker prior knowledge [29]. In another study, an AI tutor for programming was developed and evaluated in a university course, and the results showed lower cognitive load and improved learning outcomes, including better knowledge transfer [30]. However, these studies also show that the usefulness of LLMs does not automatically mean better understanding of the learning material. An experimental study with students in an introductory programming course showed that students using ChatGPT solved tasks faster, but this did

not necessarily lead to better understanding of programming concepts [31]. Another study evaluated worked examples generated by LLMs using expert evaluation criteria and found that students considered these examples useful for revision and for overcoming difficulties during problem solving, but that the explanations were often not detailed enough for beginners [32]. These results are important because they show that speed, fluency, and user satisfaction cannot be the only criteria for evaluating LLM-generated educational content. On the other hand, the literature clearly identifies several risks. One of the most important problems is hallucination in natural language generation, which refers to the generation of incorrect information that may appear convincing [14]. In technical fields such as software engineering, hallucinations and incorrect model outputs may lead to misunderstanding of concepts and adoption of incorrect solutions, which has been reported in several studies [14,28,31]. Another important problem is student over-reliance on AI systems, which can reduce student engagement in problem solving and negatively affect the development of analytical thinking skills [31,33]. A third issue is academic integrity, since LLMs can generate text and code that appear to be original student work, which makes it more difficult to assess students' actual knowledge [13]. For this reason, the literature emphasizes the need to develop methods for evaluating the quality of educational content generated by LLM systems [25].

To evaluate the educational content generated by large language models (LLMs), it is first necessary to define what high-quality educational content is. Research in the field of open educational resources (OERs) provides useful quality rubrics for instructional materials. These rubrics include criteria such as accuracy, clarity, organization, quality of explanations, practice exercises, and opportunities for deeper learning [20,21]. These criteria are important for evaluating LLM-generated content because they allow evaluation not only of factual correctness but also of pedagogical value. In software engineering education, these criteria also include technical accuracy, quality of examples, and the correct connection between theory and practice, which is defined in the SWEBOK framework [9]. However, studies show that existing rubrics differ from each other and that many of them are not empirically validated, which may affect the reliability of evaluation results [21]. Therefore, defining the evaluation criteria is not sufficient on its own, and it is also necessary to define how these criteria will be evaluated. The evaluation of AI-generated educational content is a complex task because quality does not depend only on accuracy, but also on clarity of explanation, structure, and pedagogical value. For this reason, recent research has explored the use of large language models as evaluators (LLM-as-a-judge). Studies have shown that LLM-based evaluation can achieve a high correlation with human judgments and can be used to evaluate explanation quality, coherence, and overall usefulness of generated content [34]. However, other studies show that LLM evaluators may be biased toward longer and more fluent answers, even when those answers are not necessarily more accurate or pedagogically better [35]. In addition to these biases, the reliability of LLM-based evaluation can vary depending on the evaluation criteria, prompt design, and model configuration. Lower agreement among human evaluators is often associated with reduced consistency in LLM assessments, indicating sensitivity to ambiguous evaluation tasks. Therefore, in educational contexts, a structured evaluation approach is necessary to ensure more stable and reliable results [36].

This shows that LLM-based evaluation is useful, but it still requires clearly defined evaluation criteria in educational contexts. For this reason, recent studies propose multi-dimensional evaluation of LLM outputs, where multiple quality dimensions such as accuracy, robustness, bias, efficiency, and usefulness are evaluated [5]. In educational contexts, evaluation is often based on expert judgment or rubric-based assessment, where criteria such as clarity of explanation, structure, and pedagogical value are considered [32]. In [32],

worked examples generated by LLMs were evaluated by experts using criteria such as clarity, structure, and pedagogical usefulness. Other studies show that LLMs can also be used for automatic grading of programming assignments, but the reliability of grading depends on clearly defined evaluation criteria [15]. At the same time, rubric-guided approaches have been shown to improve consistency and transparency in LLM-based evaluation by structuring the assessment process through explicitly defined criteria. However, differences in scoring and limited agreement between evaluators still remain, indicating that additional methodological support is needed for reliable evaluation [37].

Since the quality of educational content depends on multiple criteria that may have different levels of importance, this problem can be formulated as a multi-criteria decision-making problem. Traditional evaluation approaches are usually based on single metrics or single evaluation methods, which are not sufficient for evaluating educational content because educational quality depends on multiple criteria. Therefore, the evaluation of LLM-generated educational content should be treated as a multi-criteria decision-making problem rather than a single-metric evaluation problem. Multi-criteria decision-making methods allow the evaluation of alternatives based on multiple criteria and the determination of the relative importance of each criterion. Recent studies have demonstrated the effectiveness of MCDM approaches in evaluating LLMs by enabling structured, multi-criteria assessment and incorporating expert judgment. In particular, the use of hesitant Fuzzy AHP allows for handling uncertainty in expert evaluations and for determining the relative importance of evaluation criteria. Such approaches have shown that trustworthiness assessment of LLMs requires the integration of multiple dimensions, including fairness, robustness, integrity, explainability, and safety, while also providing interpretable rankings of alternative models [38]. Similarly, Fuzzy AHP-based frameworks have been developed to define and rank evaluation criteria in domain-specific contexts such as healthcare, highlighting the importance of systematic criteria structuring [39]. In addition, neutrosophic MCDM approaches, namely ARAS, have been applied to decision-making problems under uncertainty, showing strong capability in dealing with incomplete and imprecise information [40]. These findings confirm the suitability of MCDM methods for complex evaluation tasks such as assessing LLM-generated content.

One of the most commonly used multi-criteria decision-making methods is the Analytic Hierarchy Process (AHP), which is used to determine criteria weights and rank alternatives based on expert judgment [41]. The AHP method has been widely used in education and educational system evaluation, teaching quality assessment, and the selection of educational technologies, where multiple criteria such as content quality, usability, interaction, and system reliability must be considered [7,42,43]. In the context of higher education, recent studies have applied hybrid MCDM models to evaluate learning management systems, combining structured criteria weighting and value-based aggregation to support transparent and systematic decision-making [43]. However, classical AHP requires precise numerical judgments, which can be difficult when evaluation is based on subjective expert opinions. For this reason, Fuzzy AHP was developed, which uses fuzzy numbers to model uncertainty and imprecision in expert judgments [44]. Recent studies show that AHP can be used to evaluate AI-based educational systems, where the goal is to evaluate the system as an educational tool using multiple criteria such as usability, reliability, interaction, and pedagogical value [15]. In these studies, the focus is on system performance and functionality. However, in the case of large language models, an additional problem appears, which is the evaluation of the quality of the educational content generated by the system, rather than the evaluation of the system itself. The quality of generated educational content depends on criteria such as accuracy, clarity of explanation, structure, pedagogical value, and quality of examples, which makes this a multi-criteria evaluation problem.

Existing studies typically use traditional evaluation methods, single metrics, expert judgment, or rubric-based approaches, but these approaches usually evaluate only one aspect of quality and do not provide a comprehensive evaluation framework. In particular, there is a lack of multi-criteria evaluation models for assessing the quality of LLM-generated educational content in software engineering education. Therefore, there is a need for a structured multi-criteria evaluation model that can consider multiple quality criteria and determine their relative importance when evaluating LLM-generated educational content. To address this problem, this study proposes a multi-criteria evaluation model based on the Analytic Hierarchy Process (AHP) and Fuzzy AHP methods for evaluating the quality of LLM-generated educational content in software engineering education. In particular, there is a lack of studies that combine hierarchical multi-criteria decision-making models and fuzzy logic for evaluating the quality of LLM-generated educational content in technical education.

3. Multi-Criteria Evaluation Framework

In this paper, we propose a structured multi-criteria evaluation framework for assessing the quality of educational content generated by LLMs in software engineering. The proposed framework consists of several phases, as shown in Figure 1. The process begins with the definition of the course context and the selection of evaluation topics, ensuring alignment with learning outcomes and different cognitive levels. Standardized system and user prompts are then designed so that all models generate content in a comparable format. The selected large language models generate educational content for the predefined topics, which is then evaluated by experts using a predefined set of criteria and sub-criteria. In the final phase, multi-criteria decision-making (MCDM) methods, namely the Analytic Hierarchy Process (AHP) and Fuzzy AHP (FAHP), are applied to determine the criteria weights and to rank the evaluated models. The AHP method was selected because it enables the hierarchical structuring of evaluation criteria and determination of criteria weights based on expert judgments, while Fuzzy AHP was used to model uncertainty and imprecision in expert evaluations.

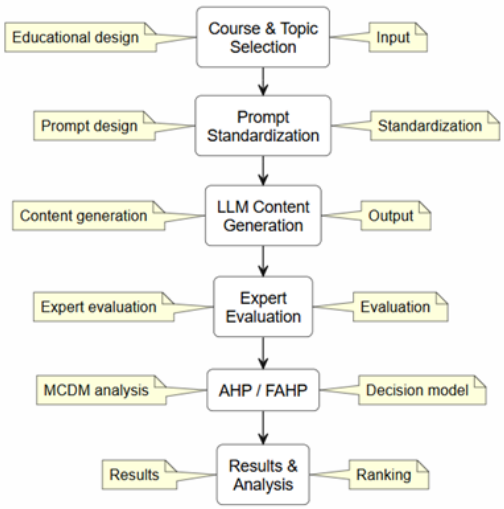

Figure 1. Research framework for the multi-criteria evaluation of LLM-generated educational content in software engineering education.

The framework is standardized so that it can be applied not only to software engineering courses, but also to other educational domains where LLMs are used for generating educational content. The following sections describe each phase of the framework in detail.

3.1. Course Context and Topic Selection

As the evaluation context, the undergraduate course SE201 Introduction to Software Engineering from the Faculty of Information Technology, Belgrade Metropolitan University was used. The course was designed following a competence-based and hybrid personalized learning approach supported by generative AI [45], which integrates competence development, personalized learning, and the use of LLMs as learning support tools. This provided a realistic educational context for evaluating LLM-generated instructional content. The course covers key areas of software engineering, including requirements analysis, software architecture, testing, and quality assurance, which is consistent with the standard software engineering literature and curricula [8,9].

The selection of topics was conducted based on three defined criteria: (1) alignment with the learning outcomes and competencies of the course, (2) sufficient technical complexity to allow differences in the quality of model responses to be observed, and (3) coverage of different cognitive levels according to the revised Bloom's taxonomy. The goal of topic selection was to include different areas of software engineering and different levels of complexity in order to enable a realistic evaluation of the models' ability to generate educational content. Three topics were selected: microservices architecture, requirements analysis and specification, and software testing with a test-driven development approach in a DevOps environment.

The cognitive levels of the selected topics were determined according to the revised Bloom's taxonomy of educational objectives [46], which includes the levels of understanding, application, analysis, evaluation, and synthesis. This ensured that the evaluation covered different levels of cognitive complexity and different types of tasks. An overview of the selected topics, their connection to course competencies, their cognitive levels, and their role in the evaluation process is presented in Table 1.

Table 1. Summary of evaluation topics, competency alignment and selection rationale.

Topic	Title	Competency Domain	Cognitive Level (Bloom's Taxonomy)	Differentiating Value	Role in Evaluation
T1	Microservices Architecture	SWF-DES-03 Architectural Design	Analysis & Evaluation	Trade-off reasoning; hallucination risk	Evaluation of ability to explain architectural concepts, compare architectural styles and discuss trade-offs
T2	Requirements Analysis & Specification	SWF-REQ Requirements Analysis	Comprehension & Application	Theory and practice bridge; terminology precision	Evaluation of understanding of key concepts, terminology and ability to connect theory and practice
T3	Software Testing & TDD in DevOps	SWF-VAV-03 Testing	Synthesis (Integration)	Multi-concept integration; practical depth	Evaluation of ability to integrate multiple concepts such as testing levels, TDD and DevOps practices

Table 1 presents the selected topics, the software engineering areas they cover, their cognitive levels according to the revised Bloom's taxonomy, and their role in the evaluation process. Each topic corresponds to a different area of software engineering and a different level of cognitive complexity. The microservices architecture topic focuses on analysis

and evaluation of architectural solutions and comparison of different approaches. The requirements analysis and specification topic focuses on understanding and applying key concepts and terminology. The software testing and TDD topic requires the synthesis and integration of knowledge from multiple software engineering areas and represents the highest cognitive level among the selected topics. This structure ensures that the evaluation covers different cognitive levels, which increases the relevance and reliability of the research results.

3.2. Prompt Design

For the purposes of this study, a predefined system prompt was used to assign all models the role of a university lecturer in the field of software engineering. The system prompt defined the structure of the response and the way the content should be presented, so that all models would generate content in the same format, which enabled comparability of the results. The system prompt required that the responses be organized as a structured lecture and include the following sections: introduction, explanation of key concepts, visualization of relevant parts of the content, overview of advantages and disadvantages, a practical example, a code example where appropriate, and a short summary. In this way, all models generated content with a similar structure, which enabled a more reliable evaluation of the quality of the generated educational material.

In addition to the system prompt, three user prompts were used, corresponding to the selected topics in the field of software engineering. The following prompts were used:

- Prompt 1—Microservices Architecture: "Create a structured lecture on the microservices architectural style: its key characteristics, advantages, and disadvantages compared to monolithic architecture, with emphasis on scalability and maintainability. Include architecture diagrams and practical examples".
- Prompt 2—Requirements Analysis and Specification: "Create a structured lecture on requirements analysis and specification in software engineering: distinguish between functional and non-functional requirements, describe elicitation techniques, and explain how user stories are used in agile development. Include process diagrams and examples".
- Prompt 3—Software Testing Techniques and TDD in DevOps: "Create a structured lecture on software testing techniques: unit, integration, system, and acceptance testing. Explain TDD and how automated testing supports DevOps and CI/CD pipelines. Include flow diagrams and code examples".

The same system prompt and the same user prompts were used for all evaluated LLM models. No additional instructions or additional context were provided. In this way, a realistic scenario was simulated in which a student uses a LLM as a general learning tool without additional model configuration.

3.3. Selection of Large Language Models for Evaluation

The selection of LLMs included in the evaluation was an important methodological step in order to ensure comparability of the results and the possibility of generalizing the research findings. The models were selected based on their presence in relevant research and their use in education, the representativeness of different types of models in the current LLM ecosystem, and a high level of performance that allows for meaningful multi-criteria evaluation.

Five models representing different approaches to the development and use of LLMs were selected in this study: GPT-4o (M1), Claude 3.5 Sonnet (M2), Gemini 2.5 Flash (M3), DeepSeek-R1 (M4) and Llama 3.3 70B (M5). The selected models include proprietary models available via API (M1, M2, and M3), an open-source model (M4), and an open-

weight model that can be run locally (M5). This selection allows the comparison of models that differ in terms of access, availability, local deployment options, and data protection considerations in educational settings.

Some models that appear in recent LLM research were not included in this evaluation. Older-generation models and lower-performance models were excluded because the performance differences would be too large and would reduce the relevance of the multi-criteria evaluation. In addition, model variants optimized for lower cost or lower resource usage were not included, since the goal of this study was not to compare models from different performance classes, but to compare modern high-performance models in the context of generating educational content.

3.4. Definition of Evaluation Criteria and Expert Panel

The evaluation framework used in this study is based on established open educational resource (OER) evaluation rubrics, primarily the Open Textbook rubric and the Achieve OER rubric [20,21]. These rubrics define the quality of educational materials through dimensions such as content accuracy, comprehensiveness and depth of explanation, clarity of presentation, organization and structure, quality of presentation, and language correctness, as well as the presence of examples and elements that support the learning process. For this reason, they provide an appropriate foundation for evaluating educational content generated by large language models. The Open Textbook rubric was used to define criteria related to content quality, such as accuracy, depth and coverage, clarity of text, organization and structure of content, visual and structural presentation, and language quality. The Achieve OER rubric was used to define criteria related to the pedagogical value of the material, particularly in the areas of examples, practical application, and the connection between theory and practice. The criteria derived from the OER rubrics (Open Textbook and Achieve OER) were adapted and extended, and further operationalized through a set of sub-criteria, to capture the specific characteristics of LLM-generated content. The accuracy dimension was refined to explicitly account for technical and professional correctness, as well as the presence of hallucinations. Criteria related to the depth, clarity, and structure of explanations were further developed to reflect variability in model responses. In addition, criteria addressing examples, practical applicability, and the integration of theory and practice were expanded in line with the requirements of software engineering as an applied discipline.

Since the focus of this study is the evaluation of content generated by large language models in the field of software engineering, the existing OER rubrics were adapted to reflect the specific characteristics of this domain and the nature of AI-generated content. In particular, the applied nature of software engineering was taken into account, where the quality of instructional material depends not only on the accuracy and clarity of explanations, but also on the presence of practical examples, real-world scenarios, diagrams, and code samples. Based on the OER rubrics and the specific requirements of software engineering education, six main groups of evaluation criteria were defined: accuracy and professional correctness (A), depth and coverage of content (D), examples and practical application (E), pedagogical clarity and didactic quality (P), visual and structural presentation (V), and language style and academic tone (L). The criteria of accuracy and professional correctness, depth and coverage, pedagogical clarity, visual and structural presentation, and language style are based on the Open Textbook rubric, while the criterion of examples and practical application is primarily based on the Achieve OER rubric and was further developed due to the applied nature of software engineering. The criterion of accuracy and professional correctness was further extended to address issues specific to large language models, particularly the problem of hallucinations, i.e., the generation of plausible but

incorrect information. Therefore, in this study, accuracy is considered not only as factual correctness, but also as the technical and professional correctness of the content.

The definition of sub-criteria was based on the literature review presented in Section 2 and further refined through expert consultation. The debate and consultation among eight experts from the areas of education, artificial intelligence, software engineering, pedagogy, and mathematics led to the selection of 25 sub-criteria, classified into the six defined criteria groups. The experts were selected to include different professional backgrounds and perspectives from academia, industry, and the public sector. All participants declared any potential conflicts of interest before taking part in the study. Basic information about the experts involved in the evaluation process is presented in Table 2.

Table 2. The basic information on the experts.

ID	Area of Expertise	Years of Experience	Institution/Organization
Expert 1	Software engineering	9	Faculty of Electronic Engineering
Expert 2	Software engineering and education	5	Faculty of Information Technology
Expert 3	Software engineering	8	Faculty of Information Technology
Expert 4	Artificial intelligence	13	Faculty of Electronic Engineering and private company
Expert 5	Software engineering and artificial intelligence	18	IT sector
Expert 6	Informatics and pedagogy	6	Faculty of Education
Expert 7	Software engineering and education	15	Faculty of Information Technology
Expert 8	Applied mathematics	16	Faculty of Science and Mathematics

Experts were selected based on their expertise and professional or academic education and experience in relevant fields. A minimum of five years of experience was considered as a baseline criteria. The panel was composed to ensure a balance between theory, pedagogical and practical perspectives.

The selection of sub-criteria was conducted through several rounds of expert consultation. During this process, experts evaluated the importance of each sub-criterion using predefined factors such as its relevance to the educational process, measurability, and suitability for the educational context. Based on expert agreement and analysis, the most relevant sub-criteria were selected and organized into six criteria groups to form a structured evaluation framework.

Each sub-criterion is defined through qualitative evaluation levels, which represent an evaluation scale used by experts during the assessment process. The defined criteria groups and corresponding sub-criteria used in the evaluation process are presented in Figure 2.

Each sub-criterion was assigned to a corresponding evaluation dimension and evaluated according to its contribution to instructional quality. All sub-criteria in this study are defined as benefit-type criteria, meaning that higher values correspond to a higher quality of LLM-generated educational content. This assumption is consistent with the AHP and FAHP frameworks, in which criteria are evaluated according to their relative contribution to the overall evaluation goal.

In the evaluation phase, experts assessed each generated response by selecting the single sub-criterion level that best described the quality of the content within each main criterion group. For each topic and each model, all experts evaluated the same generated output, which ensured the consistency and comparability of judgments across alternatives. The resulting expert selections were then used in combination with the AHP and FAHP-derived weights to calculate the final ranking of models using the formula:

$\frac{1}{k}\sum_{i=1}^{k} m_i$. Each LLM generated one response for each of the three selected topics using the same predefined system and user prompts, resulting in a total of 15 generated responses (3 topics $\times$ 5 models). In the criteria weighting phase, individual expert pairwise comparison matrices were aggregated into a group decision matrix using the arithmetic mean. The aggregated matrices were then used to compute criteria weights in both the AHP and FAHP methods.

Figure 2. The adopted criteria and sub-criteria for assessing the quality of LLM-generated educational content.

After deriving the criterion weights using the AHP and Fuzzy AHP (FAHP) procedures, the final scoring and ranking of alternatives (large language models) are performed. The evaluation framework consists of six main groups of criteria with each group including several sub-criteria, for which normalized weights w_i^* are obtained using the AHP/FAHP methodology. In the evaluation phase, experts assess the performance of each alternative with respect to each sub-criterion by assigning qualitative levels (e.g., A1–A5, D1–D4, E1–E5, P1–P5, V1–V3, and L1–L3). These levels indicate the degree to which a given large language model satisfies the corresponding evaluation requirement. To enable quantitative

analysis, the qualitative levels are mapped onto a numerical scale. In this study, a linear transformation is applied, where the lowest level corresponds to the minimum value and the highest level corresponds to the maximum value. This mapping ensures consistency across all criterion groups and allows integration with the derived weights. When multiple experts are involved, their evaluations are aggregated using the arithmetic mean, resulting in a representative performance value for each alternative under each sub-criterion. Letting w_i^* denote the normalized weight of sub-criterion i, and letting v_{ki} denote the aggregated numerical performance value of alternative k with respect to sub-criterion i, the overall score of each alternative is computed using the following weighted aggregation model $S_k = \sum_{i=1}^{n} w_i^* \cdot v_{ki}$, where S_k represents the final score of alternative k, and n is the total number of sub-criteria across all groups (A, D, E, P, V, and L). This formulation corresponds to the weighted sum model, which is widely used in multi-criteria decision-making due to its interpretability and ability to integrate multiple evaluation dimensions into a single performance measure.

Finally, the alternatives are ranked in ascending order according to their computed scores S_k, where higher values indicate better overall performance across all evaluation criteria.

This approach ensures that all criterion groups (A–L) contribute proportionally to the final evaluation, in accordance with their relative importance derived from the AHP/FAHP analysis.

3.5. AHP/FAHP Methodology

For more than half a century, fuzzy set theory has been recognized as an effective framework for addressing the uncertainty and imprecision inherent in linguistic assessments, thereby providing substantial support for decision-making processes [47]. As a generalization of classical (crisp) set theory, the primary objective of fuzzy sets is to enable a mathematical representation of linguistic variables, allowing decision-makers to model systems characterized by partially known or incomplete information [48]. In classical set theory, an element either belongs to a set or does not, implying a binary membership relationship. In contrast, fuzzy set theory introduces a membership function (MF), usually denoted by μ, which assigns to each element of the universal set a value within the interval $[0, 1]$. This value represents the degree to which the element belongs to a given fuzzy set, thus capturing gradual transitions between full membership and non-membership. Letting all fuzzy sets defined on the set of real numbers $\mathbb{R}$ be denoted as FS ($\mathbb{R}$), the number $G \in FS(\mathbb{R})$ is a fuzzy number if there exists $x_0 \in \mathbb{R}$ so it holds $\mu_G(x_0) = 1$, and for every $\lambda \in [0, 1]$, $G_\lambda = \left[x, \mu_{G_\lambda}(x) \geq \lambda\right]$ is a closed interval [49].

3.5.1. Triangular Fuzzy Sets: Preliminaries

The fundamental component of a triangular fuzzy number (TFN) is its membership function, which describes how each real value is associated with a degree of membership. This function is defined as follows:

$$\mu_{TFN}(x) = \begin{cases} \frac{x-l}{m-l}, & l \leq x \leq m \\ \frac{u-x}{u-m}, & m \leq x \leq u \\ 0, & otherwise, \end{cases} \quad (1)$$

where the inequality $l \leq m \leq u$ holds. Numbers l, m and u serve as the lower, middle, and upper value of G, respectively, while for $l = m = u$, TFN becomes a crisp number. The usual notation of the triangular fuzzy number can be expressed as $\widetilde{G} = (l, m, u)$.

The left and right sides of the membership function $\mu_{TFN}(x)$ of TFN $\widetilde{G} = (l, m, u)$, $\mu^{l}_{\widetilde{G}}$ and $\mu^{r}_{\widetilde{G}}$, and their corresponding inverse functions $(\mu^{l}_{\widetilde{G}})^{-1}$ and $(\mu^{l}_{\widetilde{G}})^{-1}$, are defined as $\mu^{l}_{\widetilde{G}}=\frac{x-l}{m-l}$, $\mu^{r}_{\widetilde{G}}=\frac{u-x}{u-m}$, $(\mu^{l}_{\widetilde{G}})^{-1} = l + (m-l)y$, $(\mu^{r}_{\widetilde{G}})^{-1} = u + (m-u)y$, $y \in [0,1]$, yielding the formula for the total integral value [50]:

$$\begin{aligned} I^{\lambda}_{T}(\widetilde{G}) &= \lambda I_R(\widetilde{G}) + (1-\lambda) I_L(\widetilde{G}) = \lambda \textstyle\int_0^1 (\mu^{r}_{\widetilde{G}})^{-1} dy + (1-\lambda) \int_0^1 (\mu^{l}_{\widetilde{G}})^{-1} dy = \\ &= \tfrac{1}{2}\lambda(m+u) + \tfrac{1}{2}(1-\lambda)(m+l) = \tfrac{1}{2}(\lambda u + m + (1-\lambda) l), \end{aligned} \tag{2}$$

where λ represents an optimism index, i.e., the attitude of an expert during the decision-making process. The pessimistic point of view is presented taking the value $\lambda = 0$, from where it is obtained that $I^{0}_{T}(\widetilde{G}) = I_L(\widetilde{G})$; for the value $\lambda = 1$, the optimistic point of view is given, and $I^{1}_{T}(\widetilde{G}) = I_R(\widetilde{G})$. For $\lambda = 0.5$, the balanced (moderate) attitude of the decision-maker is granted, and $I^{0.5}_{T}(\widetilde{G}) = \frac{1}{2}(I_L(\widetilde{G}) + I_G(\widetilde{G}))$. In addition to the pessimistic and optimistic perspectives, recently introduced semi-pessimistic and semi-optimistic points of view have been proposed, corresponding to $\lambda = 0.25$ and $\lambda = 0.75$, respectively, providing a more nuanced interpretation of uncertainty [51].

The main unary (scalar multiplication and inverse) and binary (addition, subtraction and multiplication) operations for TFNs $G_1 = (l_1, m_1, u_1)$ and $G_2 = (l_2, m_2, u_2)$ and scalar $k > 0$, $k \in \mathbb{R}$ are shown below:

$$\widetilde{G}_1 \oplus \widetilde{G}_2 = (l_1, m_1, u_1) \oplus (l_2, m_2, u_2) = (l_1 + l_2, m_1 + m_2, u_1 + u_2), \tag{3}$$

$$\widetilde{G}_1 \ominus \widetilde{G}_2 = (l_1, m_1, u_1) \ominus (l_2, m_2, u_2) = (l_1 - u_2, m_1 - m_2, u_1 - l_2), \tag{4}$$

$$\widetilde{G}_1 \otimes \widetilde{G}_2 = (l_1, m_1, u_1) \otimes (l_2, m_2, u_2) = (l_1 \cdot l_2, m_1 \cdot m_2, u_1 \cdot u_2), \tag{5}$$

$$k \cdot \widetilde{G}_1 = k \cdot (l_1, m_1, u_1) = (k \cdot l_1, k \cdot m_1, k \cdot u_1), \tag{6}$$

$$\widetilde{G}_1^{-1} = (l_1, m_1, u_1)^{-1} = \left(\frac{1}{u_1}, \frac{1}{m_1}, \frac{1}{l_1} \right). \tag{7}$$

Triangular fuzzy numbers are employed to represent expert judgments due to their simplicity and effectiveness in handling uncertainty in decision-making problems. TFNs are particularly suitable when evaluations are expressed in linguistic terms, as they allow each assessment to be modeled as a triplet (l, m, u), representing the lower, most probable, and upper values. The use of TFNs enables the modeling of different sources of uncertainty inherent in expert evaluations, including imprecision, vagueness, and subjectivity. In contrast to single-point (crisp) values, this representation captures the range of possible interpretations of a given assessment, thereby providing a more flexible and realistic description of expert preferences.

3.5.2. Algorithm

In the sequel, the steps of the Fuzzy Analytic Hierarchy Process are summarized [49]:

Step 1: Establishing the main goal and hierarchical appearance of criteria.

The hierarchical structure is organized in a vertical manner, where the main goal represents the highest level of importance. The criteria and sub-criteria that influence the achievement of the goal occupy the intermediate levels, whereas the set of alternatives is positioned at the lowest level of the hierarchy.

Step 2: Setting the matrix $\widetilde{H}$ in terms of triangular fuzzy numbers.

The criteria and sub-criteria are used during the pairwise comparisons, enabling the creation of matrix $\widetilde{H} = (\widetilde{h}_{ij})_{n \times n}$. A total of $n(n-1)/2$ comparisons of elements from a

higher level with elements from a lower level are made. Using triangular fuzzy numbers (TFNs), the hierarchy and comparison are given, where $\widetilde{h}_{ij}$ is a fuzzy value representing the relative importance of one criterion to another. It holds that $\widetilde{h}_{ii} = (1,1,1)$, when comparing criteria to itself, and $\widetilde{h}_{ij} = 1/\widetilde{h}_{ji}$ for $i \neq j$.

The fuzzy scale, TFNs, and their explanations used to enable pairwise comparisons are given:

TFN $\widetilde{1} = (1,1,3)$: "Two criteria are equally important."

TFN $\widetilde{3} = (1,3,5)$: "One criteria is slightly more important than another."

TFN $\widetilde{5} = (3,5,7)$: "One criteria is strongly more important than another."

TFN $\widetilde{7} = (5,7,9)$: "One criteria is very strongly more important than another."

TFN $\widetilde{9} = (7,9,9)$: "One criteria is absolutely strongly more important than another."

$\widetilde{2} = (1,2,3)$, $\widetilde{4} = (3,4,5)$, $\widetilde{6} = (5,6,7)$, and $\widetilde{8} = (7,8,9)$ are intermediate values used when compromise is needed. The graphic representation of the used FAHP scale with lower, median, and upper values is presented in Figure 3.

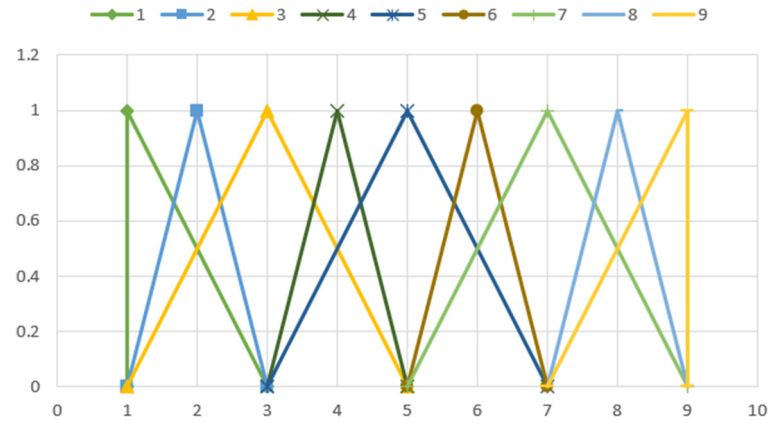

Figure 3. Graphic representation of triangular fuzzy numbers.

Step 3: Matrix consistency calculation.

For matrix $H = (h_{ij})_{n\times n}$, we calculate the consistency index CI and consistency ratio CR using formulas

$$CI = \frac{\lambda_{max} - n}{n-1}, CR = \frac{CI}{RI},$$

where λ_{max} represents the maximal eigenvalue of matrices H. The random index RI determined by the matrix size and corresponding value is shown as

$$RI = \{(3, 0.58), (4, 0.9), (5, 1.12), (6, 1.24)\}.$$

The value $CR < 0.1$ verifies the matrix H's consistency; otherwise, the reason for inconsistency should be determined, and all calculations redone.

Step 4: The fuzzification process.

Applying formulas

$$D = \sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{h}_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n}(l_{ij}, m_{ij}, u_{ij}) \tag{8}$$

and

$$D^{-1} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{h}_{ij}\right)^{-1} = \left(\frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{n}u_{ij}}, \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{n}m_{ij}}, \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{n}l_{ij}}\right) \tag{9}$$

on triangular fuzzy numbers from the matrix $H = \left(h_{ij}\right)_{n\times n}$, the Chang synthetic fuzzy number $\widetilde{S}_i = (l_i, m_i, u_i) = \sum_{j=1}^{n} \widetilde{h}_{ij} \otimes D^{-1}$, $i = \overline{1,n}$ is obtained [44].

Step 5: The defuzzification process.

Applying the formula

$$w_i = I_T^{\lambda}(\widetilde{S}_i) = 0.5(\lambda u_i + m_i + (1-\lambda)l_i), i = \overline{1,n}, \lambda \in [0,1], \tag{10}$$

on obtained TFNs $\widetilde{S}_i$, $\overline{1,n}$, the total integral value is calculated.

Step 6: Vector normalization and criteria weight calculation.

The weight vector $w = (w_1, w_2, \ldots, w_n)^T$ is normalized using formula

$$w_i^* = w_i \left(\sum_{i=1}^{n} w_i \right)^{-1} \tag{11}$$

After this, criteria ranking is performed.

To enable quantitative analysis, qualitative assessments provided by experts were transformed into numerical values using a predefined linguistic scale. Each qualitative level (very low, low, medium, high, and very high) was assigned a corresponding numerical score (1–9). In cases involving multiple experts, individual evaluations were aggregated using the arithmetic mean to obtain a representative score for each criterion. These numerical values were then used as inputs for the subsequent weighting procedure.

4. Results and Discussion

In this section, an algorithm presented in Section 3.5 has been applied. Crisp AHP and five degrees of optimism in the FAHP cases (pessimistic, λ = 0; semi-pessimistic, λ = 0.25; balanced, λ = 0.5; semi-optimistic, λ = 0.75; and optimistic, λ = 1) are used to compare the obtained results and sub-criteria ranking [47]. A group of experts from the areas of education, artificial intelligence, software engineering, informatics, pedagogy, and mathematics selected groups of criteria and sub-criteria, and expressed their opinions based on the meaning of TFNs presented in Figure 3. The assessments the experts gave were aggregated based on the first step of the presented algorithm.

Firstly, the main criteria group rankings were determined. According to the aggregated experts' opinions, the most important group is accuracy and professional correctness, denoted by A, followed by the equally ranked depth and coverage of the content (D), examples and practical application (E), and pedagogical clarity and didactic quality (P). The criteria visual and structural presentation (V) and language style and academic tone (L) are placed at the bottom.

The values CI = 0.037 and CR = 0.03, enabling the consistency of the main criteria comparison matrix, are calculated as follows:

$$\lambda_{max} = \frac{6.289 + 6.245 + 6.245 + 6.245 + 6.027 + 6.059}{6} = 6.185 \tag{12}$$

$$CI = \frac{6.185 - 6}{6 - 1} = 0.037, CR = \frac{0.037}{1.24} = 0.03. \tag{13}$$

In the AHP case, the leading criteria has weight $w(A)$ = 0.405, while $w(D)$ = $w(E)$ = $w(P)$ = 0.1735. In the balanced FAHP case, the most significant criteria have weight $w(A)$ = 0.319, having 1.53 times higher weight than criteria D, and 1.6 and 1.69 times more importance than criteria E and P. In the pessimistic FAHP case, the previous quotients are equal to 1.6, 1.64, and 1,67, while for $\lambda = 1$ they respectively are 1.5, 1.59, and 1.7. Considering the weights of lastly ranked criteria, in all five cases of the FAHP they

are equal: $w(V) = 0.057$, $w(V) = 0.06$, $w(V) = 0.062$, $w(V) = 0.063$, and $w(V) = 0.064$; $w(L) = 0.026$, $w(L) = 0.026$, $w(L) = 0.025$, $w(L) = 0.025$, and $w(L) = 0.025$.

The ranking of sub-criteria is performed in the same manner as the ranking of the main criteria, and the fuzzy comparison matrices and corresponding weights are given in Tables 3–7 and Figures 2–4 below. According to the experts' opinions, the fuzzy matrices for the sub-criteria of groups A, E, and P are the same, as well as the matrices for the sub-criteria of groups V and L. All comparison matrices are consistent.

Table 3. Fuzzy comparison matrix and weights for the sub-criteria from group A.

A	A1	A2	A3	A4	A5
A1	$\tilde{1}$	$\tilde{3}$	$\tilde{5}$	$\tilde{7}$	$\tilde{9}$
A2	$\tilde{3}^{-1}$	$\tilde{1}$	$\tilde{3}$	$\tilde{5}$	$\tilde{7}$
A3	$\tilde{5}^{-1}$	$\tilde{3}^{-1}$	$\tilde{1}$	$\tilde{3}$	$\tilde{5}$
A4	$\tilde{7}^{-1}$	$\tilde{5}^{-1}$	$\tilde{3}^{-1}$	$\tilde{1}$	$\tilde{3}$
A5	$\tilde{9}^{-1}$	$\tilde{7}^{-1}$	$\tilde{5}^{-1}$	$\tilde{3}^{-1}$	$\tilde{1}$

Table 4. The weights for the sub-criteria from group A in the AHP and FAHP cases.

A	AHP	FAHP				
		λ = 0	λ = 0.25	λ = 0.5	λ = 0.75	λ = 1
A1	0.502819	0.445249	0.429636	0.419718	0.412859	0.407833
A2	0.260232	0.283019	0.286063	0.287996	0.289334	0.290314
A3	0.13435	0.159886	0.167171	0.171799	0.174999	0.177344
A4	0.067778	0.076977	0.082977	0.086789	0.089425	0.091356
A5	0.034821	0.03487	0.034153	0.033698	0.033383	0.033153

Table 5. Fuzzy comparison matrix and weights for the sub-criteria from group D.

D	D1	D2	D3	D4
D1	$\tilde{1}$	$\tilde{3}$	$\tilde{7}$	$\tilde{9}$
D2	$\tilde{3}^{-1}$	$\tilde{1}$	$\tilde{5}$	$\tilde{7}$
D3	$\tilde{7}^{-1}$	$\tilde{5}^{-1}$	$\tilde{1}$	$\tilde{3}$
D4	$\tilde{9}^{-1}$	$\tilde{7}^{-1}$	$\tilde{3}^{-1}$	$\tilde{1}$

Table 6. The weights for the sub-criteria from group D in the AHP and FAHP cases.

D	AHP	FAHP				
		λ = 0	λ = 0.25	λ = 0.5	λ = 0.75	λ = 1
D1	0.573947	0.513364	0.502292	0.494638	0.489031	0.484746
D2	0.291316	0.340528	0.343762	0.345997	0.347635	0.348887
D3	0.090263	0.101364	0.109448	0.115037	0.119132	0.122261
D4	0.044474	0.044744	0.044498	0.044327	0.044202	0.044106

Table 7. The weights for the sub-criteria from group V in the AHP and FAHP cases.

V	V1	V2	V3
V1	$\tilde{1}$	$\tilde{3}$	$\tilde{5}$
V2	$\tilde{3}^{-1}$	$\tilde{1}$	$\tilde{3}$
V3	$\tilde{5}^{-1}$	$\tilde{3}^{-1}$	$\tilde{1}$

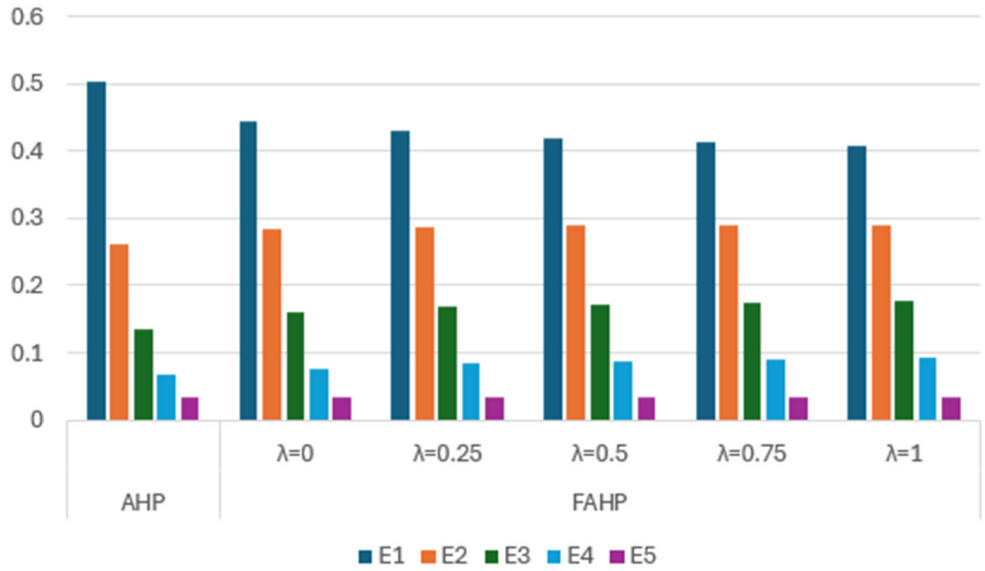

Figure 4. Weights for the sub-criteria belonging to group E in the AHP and FAHP cases.

As can be seen in Table 3, the leading sub-criteria from group A deals with a completely accurate concept and terminology, with no conceptual or technical errors (A1) with a corresponding weight equal to $w(A1) = 0.503$ in the AHP case, and an averaged weight in all FAHP cases of $w(A1) = 0.423$. The highest weights of the sub-criteria regarding to minor inaccuracies is obtained in the optimistic FAHP case, $w(A2) = 0.29$, while the lowest value is reached in the AHP case, $w(A2) = 0.26$. Predominantly inaccurate or incorrect content is the least important sub-criteria from this group, being 12.456, 12.6, and 12.367 times less important than the leading one in the balanced, semi-pessimistic, and semi-optimistic FAHP cases.

In the same way, the sub-criteria from the groups examples and practical application (E) and pedagogical clarity and didactic quality (P) were ranked. The weights for sub-criteria from group E and TFNs for sub-criteria from group P are presented in Figures 4 and 5, respectively.

The sub-criteria belonging to the group depth and coverage of the content are ranked in the following way. Comprehensive and detailed coverage of the topic with an academic level of presentation (D1) is at the top, followed by good coverage of most key aspects without deeper elaboration of advanced components (D2), superficial treatment with limited depth (D3), and minimal or trivial content (D4), as can be seen in Tables 4 and 5. The highest value of the sub-criteria D1 is obtained in the AHP case, $w(D1) = 0.574$, while the averaged weight in the FAHP cases is equal to $w(D1) = 0.497$. The second sub-criteria in the rank has weights $w(D2) = 0.341$, $w(D2) = 0.346$, and $w(D2) = 0.349$ in the pessimistic, balanced, and optimistic FAHP cases. Comparing the results in the semi-pessimistic and semi-optimistic points of view (FAHP), $w(D3) = 0.109$, and $w(D3) = 0.119$, while trivial content has weights $w(D4) = 0.044 = w(D4)$, being 11.288 and 11.064 times less important than D1.

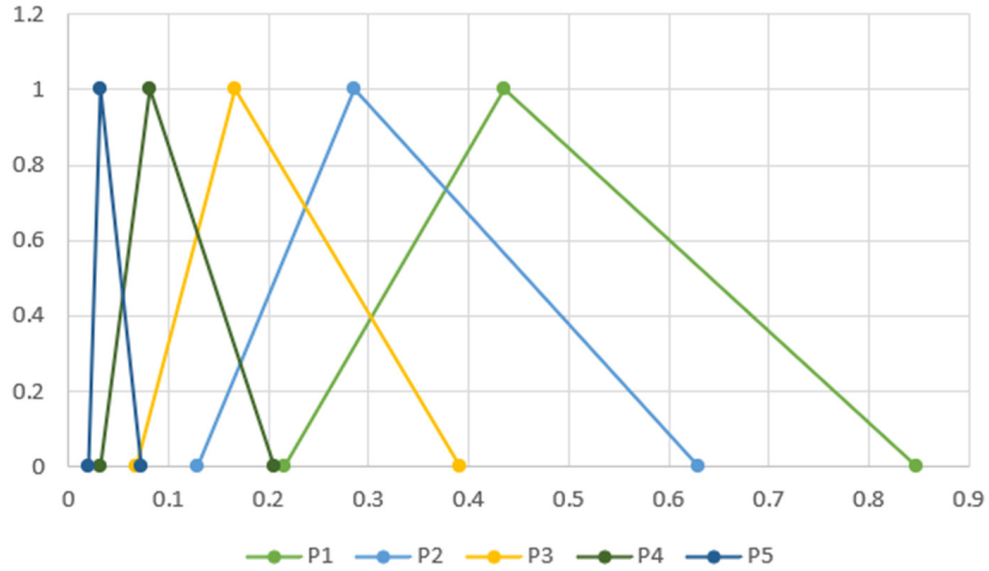

Figure 5. Triangular fuzzy numbers for the sub-criteria related to group P.

Speaking of the visual and structural presentation and language style and academic tone groups, clear and relevant visual elements and consistent academic style with no linguistical and grammatical flaws rank first, with $w(V1) = w(L1) = 0.633$, followed by basic but acceptable structure (V2) and a mixture of academic and informal style (L2), $w(V2) = w(L2) = 0.26$, being 5.966, 5.098, 5.479, and 5.653 times less important than L1 (V1) in the AHP, pessimistic, balanced, and optimistic FAHP cases, respectively. The comparison matrix (Table 7) and corresponding TFNs (Figure 6) for these two groups are given below.

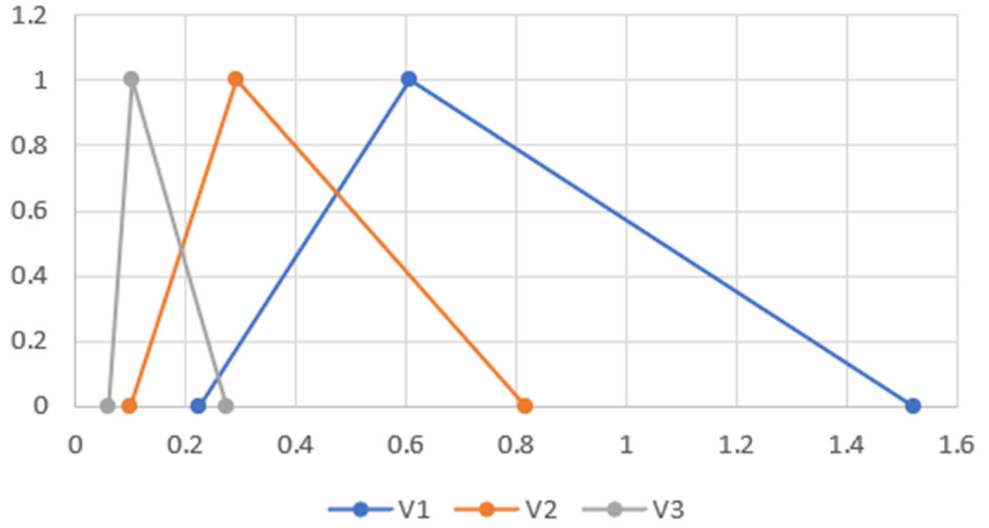

Figure 6. Triangular fuzzy numbers for the sub-criteria related to groups V and L.

Now, using the six different ways, we performed the sub-criteria final ranking. It can be observed (Figure 7) that, while the sub-criterions named A1 and A2 are still ranked highest, all the remaining sub-criteria from A (accuracy and professional correctness), A3, A4, and A5, are ranked sixth, eleventh, and sixteenth, respectively, in the AHP, with some differences in the FAHP cases, placing A2 third and the rest of the sub-criteria from A three places lower than in the crisp case. The sub-criteria from group D, D1–D4, are ranked second, sixth, fourteenth, and nineteenth in all five cases of the FAHP. A similar situation is within the group E sub-criteria, ranking E1, E2, and E3 fourth, seventh, and eleventh in the FAHP, while E4 ranks seventeenth in the pessimistic case and sixteenth in other cases.

E5 is ranked twenty-second for $\lambda = 0$, $\lambda = 0.25$, and $\lambda = 0.5$, while in the semi-optimistic and optimistic case is ranked one place lower. Exceptionally clear and logically structured presentation, the leading sub-criteria from the group P, is placed fifth in all six cases, while P2 is ranked ninth in crisp and three FAHP cases ($\lambda = 1$, $\lambda = 0.75$, and $\lambda = 0.5$) and eighth in the rest of the possibilities. Understandable with increased effort from the reader (P3) is ranked thirteenth in the AHP, and one place higher in five FAHP cases. Furthermore, the leading sub-criteria from groups V and L, V1 and L1, are ranked tenth (all six cases) and fourteenth (AHP) and four places below (FAHP). The sub-criteria placed at the bottom of the ladder are related to the non-existence of useful examples, low pedagogical quality, disorganized content and inappropriate style for academic use, named E5, P5, V3, and L3.

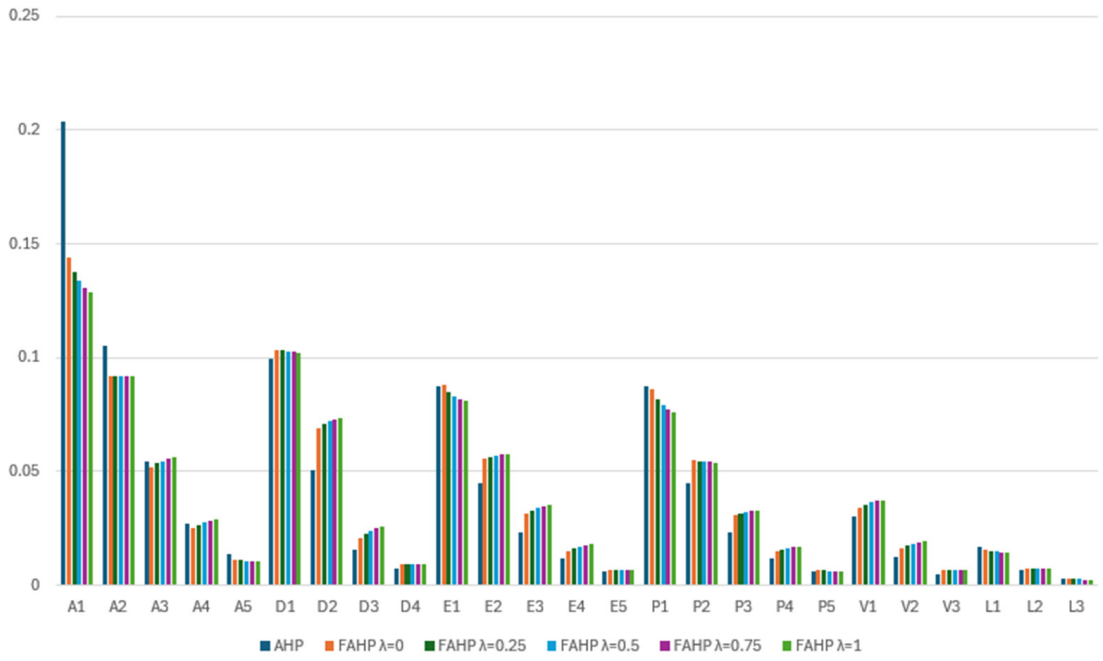

Figure 7. Final ranking of all sub-criteria.

The findings verify that there are some influential sub-criteria (A1 and D1) identified in the evaluation framework. These sub-criteria have consistently had larger weights than other sub-criteria across several experts and methods (AHP and FAHP), indicating they contribute more toward the overall decision-making process. Additionally, the sensitivity of the overall rankings to these sub-criteria suggests that these are highly important factors in generating the evaluation. Their prominent status indicates how important those sub-criteria are, as well as suggesting that experts have reached convergence regarding the important qualities associated with these sub-criteria. Accordingly, defining and validating the sub-criteria in the evaluation of LLM-generated content in educational software engineering will be paramount, confirming the research question RQ2. Therefore, the results confirm RQ2, showing that highly influential sub-criteria do exist, with accuracy and professional correctness (A1) and depth and coverage of content (D1) identified as the most influential factors in the evaluation process.

In this paper, several different rankings were obtained. In the process of analyzing ranking similarities in the crisp AHP and five different parts of FAHP algorithms, the authors most often use Spearman's rank correlation coefficient (n stands for the number of elements in ranking, and R_{x_i} and R_{y_i} mean the i-th element in the rankings used for comparison) [52]:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad d_i = R_{x_i} - R_{y_i} \tag{14}$$

As presented in Figure 8, the highest similarity (100%) in ranking sub-criteria is obtained when comparing the semi-optimistic and optimistic FAHP, while the lowest value of the index r_s = 0.977 occurs when the semi-pessimistic FAHP is compared with the crisp AHP and pessimistic FAHP. Important to mention is that r_s = 0.999 when comparing the balanced FAHP with the semi-pessimistic, semi-optimistic and optimistic FAHP. According to the obtained results, one can conclude that all rankings have high similarity.

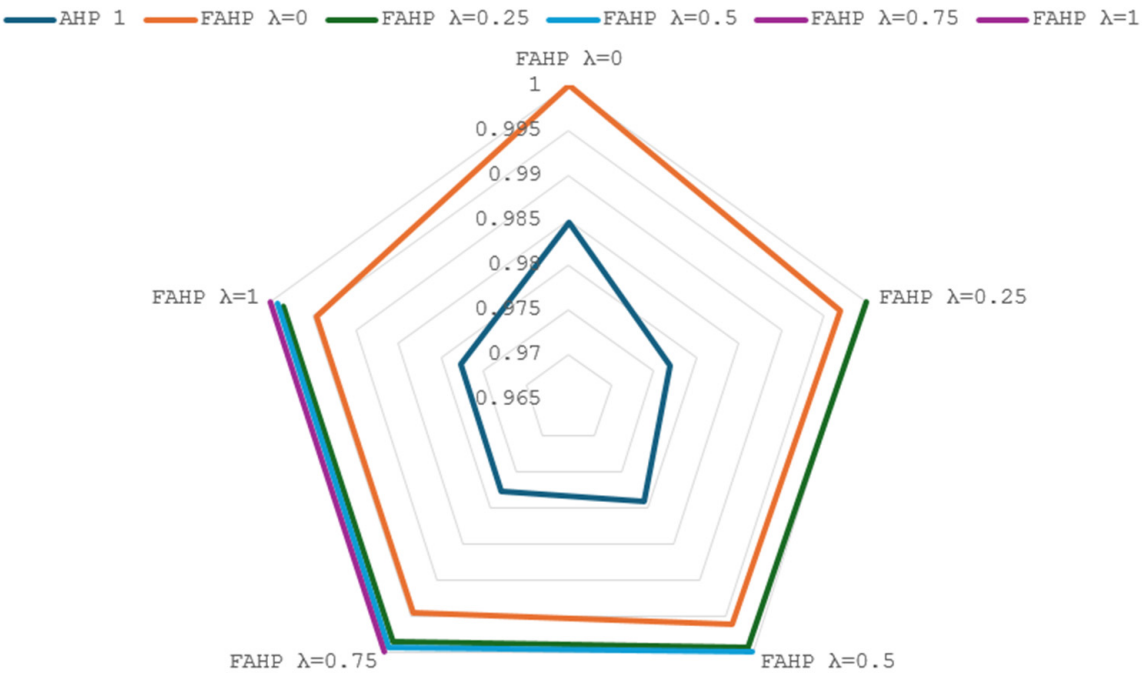

Figure 8. The sub-criteria ranking similarity (crisp AHP and five FAHP cases).

The high similarity of rankings obtained for different values of the optimism index indicates that changes in the optimism index do not significantly affect the ranking order of sub-criteria. This suggests that the proposed FAHP-based evaluation model is stable with respect to variations in decision-maker attitude and fuzzy parameters. From a methodological perspective, this result demonstrates the potential robustness of the proposed multi-criteria decision-making model. Therefore, these findings provide a clear answer to RQ3, confirming that there are no significant changes in sub-criteria ranking when different optimism index values are used.

A comparative analysis of the rankings obtained using both AHP and FAHP approaches indicates a high level of agreement between the two methods, particularly with respect to the top-ranked alternatives, which remain unchanged across both models. This consistency suggests a strong structural stability of the proposed evaluation framework. At the same time, the FAHP-based results exhibit a more balanced distribution of criteria weights, reflecting the incorporation of uncertainty and imprecision in expert judgments. The aggregation of inputs from eight experts, combined with acceptable consistency levels, further supports the robustness of the findings. Overall, these results demonstrate that the proposed framework provides reliable and interpretable rankings while maintaining resilience to minor variations in expert input.

A sensitivity analysis was conducted to assess the robustness of the obtained results. Specifically, the normalized weights of the leading criteria from all groups were varied within a predefined range (±10–20%), in all six cases. The resulting changes in the final ranking of alternatives were then analyzed.

The results indicate that the overall ranking remains stable under moderate variations of the leading criteria, as presented in Figure 9 below, with only minor changes observed among closely ranked alternatives. This suggests that the proposed model demonstrates a satisfactory level of robustness despite the simplified treatment of uncertainty.

Sensitivity Analysis of Criteria Under Weight Variations - AHP

Group	Criterion	Weight variation (%)								
		-20%	-15%	-10%	-5%	0% (Base case)	5%	10%	15%	20%
Accuracy and professional correctness (A)	A1	+0	+0	+0	+0	+0	+0	+0	+0	+0
	A2	+0	+0	+0	+0	+0	+0	−1	−1	−1
	A3	+0	+0	+0	+0	+0	+0	+0	+0	+0
	A4	+1	+1	+0	+0	+0	+0	+0	+0	+0
	A5	+1	+0	+0	+0		+0	+0	+0	+0
Depth and coverage of the content (D)	D1	+0	+0	+0	+0	+0	+0	+1	+1	+1
	D2	+0	+0	+0	+0	+0	+0	+0	+0	+0
	D3	+1	+1	+1	+0	+0	+0	+0	+0	+0
	D4	+0	+0	+0	+0	+0	+0	+0	+0	+0
Examples and practical application (E)	E1	+0	+0	+0	+0	+0	+0	+0	+0	+0
	E2	+0	+0	+0	+0	+0	+0	+0	+0	+0
	E3	+0	+0	+0	+0	+0	+0	+0	+0	+0
	E4	+0	+0	+0	+0	+0	+0	+0	+0	+0
	E5	+0	+0	+0	+0	+0	+0	+0	+0	+0
Pedagogical clarity and didactic quality (P)	P1	+0	+0	+0	+0	+0	+0	+0	+0	+0
	P2	+0	+0	+0	+0	+0	+0	+0	+0	+0
	P3	+0	+0	+0	+0	+0	+0	+0	+0	+0
	P4	+0	+0	+0	+0	+0	+0	+0	+0	+0
	P5	+0	+0	+0	+0	+0	+0	+0	+0	+0
Visual and structural presentation (V)	V1	−1	−1	+0	+0	+0	+0	+0	+0	+0
	V2	+0	+0	+0	+0	+0	+0	+0	+0	+0
	V3	+0	+0	+0	+0	+0	+0	+0	+0	+0
Language style and academic tone (L)	L1	−2	−1	−1	+0	+0	+0	+0	+0	+0
	L2	+0	+0	+0	+0	+0	+0	+0	+0	+0
	L3	+0	+0	+0	+0	+0	+0	+0	+0	+0

−2 Strong negative | −1 Negative | +0 No change | +1 Positive | +2 Strong positive

- Positive values (green): criterion's increase in weight improves the alternative's relative performance.
- Negative values (red): criterion's increase in weight worsens the alternative's relative performance.
- Zero (white): no change in the alternative's relative position.

Figure 9. The sensitivity analysis of the criteria under weight variations, AHP case.

Figure 9 presents the sensitivity analysis results under variations of criterion weights ranging from −20% to +20% in the AHP case. The majority of criteria exhibit stable behavior, with no significant impact on the final ranking, indicating the robustness of the model. However, certain criteria (A2 and L1) show noticeable sensitivity, suggesting that they play a more influential role in the evaluation process. Overall, the limited variation in most criteria confirms the stability and reliability of the proposed decision-making framework.

A similar sensitivity analysis was conducted for the FAHP method considering different values of the optimism index (λ = 0, 0.25, 0.5, 0.75, and 1). For brevity, the results corresponding to the balanced case (λ = 0.5) are presented in Figure 10. The obtained results show a similar overall pattern as in the AHP case, with most criteria exhibiting stable behavior under weight variations. However, slightly more pronounced changes can be observed for certain criteria (A2, D2, V1, and L1), reflecting the influence of uncertainty incorporated through fuzzy modeling. Despite these local variations, the overall ranking remains largely unchanged, indicating the potential robustness and consistency of the proposed FAHP-based evaluation framework.

After determining the relative weights of the criteria using the AHP and FAHP methods, the evaluation and ranking of the considered alternatives is approached, representing the central element of the proposed framework.

According to Expert 1's evaluation, for T1 and M1, the corresponding sub-criteria are A1, D2, E2, P1, V1, and L1, with the resulting weights $w_{T1}(M1) = 0.433814$ in the

AHP case, and $w_{T1}(M1) = 0.404877$, $w_{T1}(M1) = 0.393184$, and $w_{T1}(M1) = 0.388057$ in the pessimistic, balanced and optimistic FAHP cases. The lowest value for the model M1, 0.240739, is obtained in the AHP case selected by Expert 2. The average values of all experts for all six cases of algorithms, as can be seen in Table 8, yield interesting results, presenting the same weight for M1 in the results of Expert 1, Expert 3, and Expert 5. Average results in the semi-pessimistic, semi-optimistic, and optimistic FAHP cases are equal to 0.335212, 0.3902, and 0.333007, making a decreasing sequence of values as the optimism index rises.

Sensitivity Analysis of Criteria Under Weight Variations - FAHP λ = 0.5

Group	Criterion	Weight variation (%)								
		-20%	-15%	-10%	-5%	0% (Base case)	5%	10%	15%	20%
Accuracy and professional correctness (A)	A1	+0	+0	+0	+0	+0	+0	+0	+0	+0
	A2	+1	+1	+0	+0	+0	+0	+0	−1	−2
	A3	+0	+0	+0	+0	+0	+0	+0	+0	+0
	A4	+0	+0	+0	+0	+0	+0	+0	+0	+0
	A5	+0	+0	+0	+0	+0	+0	+0	+0	+0
Depth and coverage of the content (D)	D1	−1	−1	+0	+0	+0	+0	+0	+0	+0
	D2	+2	+2	+1	+0	+0	+0	+0	+0	+0
	D3	+0	+0	+0	+0	+0	+0	+0	+0	+0
	D4	+0	+0	+0	+0	+0	+0	+0	+0	+0
Examples and practical application (E)	E1	−1	−1	+0	+0	+0	+0	+0	+1	+1
	E2	+0	+0	+0	+0	+0	+0	+0	+0	+0
	E3	+1	+1	+1	+0	+0	+0	+0	+0	+0
	E4	+0	+0	+0	+0	+0	+0	+0	+0	−1
	E5	+0	+0	+0	+0	+0	+0	+0	+0	+0
Pedagogical clarity and didactic quality (P)	P1	−1	−1	−1	+0	+0	+0	+0	+0	+1
	P2	+0	+0	+0	+0	+0	+0	+0	+0	+0
	P3	+1	+1	+0	+0	+0	+0	+0	+0	+0
	P4	+0	+0	+0	+0	+0	+0	−1	−1	−1
	P5	+0	+0	+0	+0	+0	+0	+0	+0	+0
Visual and structural presentation (V)	V1	−2	−2	−1	+0	+0	+0	+0	+0	+0
	V2	+0	+0	+0	+0	+0	+0	+0	+0	+0
	V3	+0	+0	+0	+0	+0	+0	+0	+0	+0
Language style and academic tone (L)	L1	+0	+0	+0	+0	+0	+0	+1	+1	+2
	L2	+0	+0	+0	+0	+0	+0	+0	+0	+0
	L3	+0	+0	+0	+0	+0	+0	+0	+0	+0

−2 Strong negative | −1 Negative | +0 No change | +1 Positive | +2 Strong positive

- Positive values (green): criterion's increase in weight improves the alternative's relative performance.
- Negative values (red): criterion's increase in weight worsens the alternative's relative performance.
- Zero (white): no change in the alternative's relative position.

Figure 10. The sensitivity analysis of the criteria under weight variations, FAHP balanced case λ = 0.5.

Table 8. The evaluation grades for the model M1 in the T1 case (AHP and five FAHP cases).

Ex1	Ex2	Ex3	Ex4	Ex5	Ex6	Ex7	Ex8
A1	A2	A1	A2	A1	A2	A1	A2
D2	D3	D2	D2	D2	D2	D3	D1
E2	E2	E2	E3	E2	E2	E3	E2
P1	P2	P1	P2	P1	P2	P2	P1
V1	V2	V1	V2	V1	V2	V1	V2
L1	L1	L1	L2	L1	L2	L1	L2

Comparing the following results, it can be observed that Model M2 has identical weights assigned by Experts 1 to 6, amounting to $w_{T1}(M2) = 0.524942$ in the AHP case, and 0.471839, 0.450208, and 0.440029 in the FAHP ($\lambda = 0$, $\lambda = 0.5$, and $\lambda = 1$) cases, indicating that it is the most important alternative within this group of evaluations. In contrast, the weights assigned by Expert 7 are lower, being 1.146, 1.126, 1.122, and 1.118 times smaller than the corresponding values reported above. A similar trend is observed for Expert 8, with the respective ratios equal to 1.23, 1.125, 1.103, and 1.092. This deviation suggests that

Experts 7 and 8 adopt a more conservative assessment of Model M2 compared to the rest of the expert group, although the overall ranking of the model remains unchanged. Such consistency in ranking, despite differences in absolute values, indicates the robustness of the evaluation results.

Similar results can also be seen in the answers of Expert 1, Expert 3, and Expert 5, regarding the model M3, with weights $w_{T1}(M3) = 0.20892$ in the AHP, and 0.226875 and 0.232303 in the FAHP ($\lambda = 0.25$ and $\lambda = 0.75$) cases, as can be seen in Figure 11. The highest value for M3 is obtained from the Expert 8 opinion and is equal to 0.412821 (AHP), and 0.381959 and 0.377208 (FAHP: $\lambda = 0$; $\lambda = 0.25$). The second highest weight is obtained from Expert 6, with an averaged value of all six cases equal to 0.346992, while the lowest value in general is equal to 0.124604 and is obtained from Expert 4 (AHP).

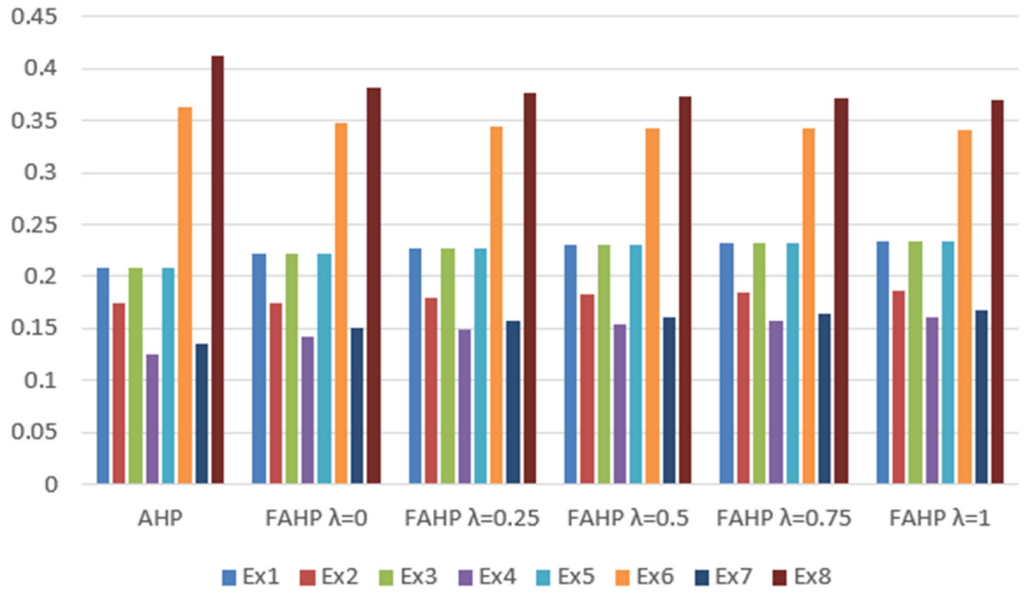

Figure 11. The weights of the model M3 (T1 case).

The highest value for Model 4 is obtained from Expert 2's opinion with the corresponding weights 0.524942, 0.450208, and 0.440029 (AHP, FAHP: $\lambda = 0.5$; $\lambda = 1$), while the lowest value is given by Experts 1, 3, and 5, with weight 0. 131063 (AHP). The average weights of Model 4 are equal to 0.261286, 0.268921, 0.270258, and 0.270961 (AHP, FAHP: $\lambda = 0$; $\lambda = 0.5$; $\lambda = 1$), which demonstrates a relatively stable evaluation across different methodological settings. This stability indicates that the application of fuzzy parameters (λ) does not significantly affect the overall ranking of this model. Furthermore, Experts 6 and 7 provided evaluations whose corresponding weights are the closest to the average values (Ex6: 0.265645, 0.294545, 0.301130, and 0.304171; Ex7: 0.258209, 0.285248, 0.289288, and 0.291196; AHP, FAHP: $\lambda = 0$; $\lambda = 0.5$; $\lambda = 1$), suggesting that their judgments can be considered the most representative or consensus-based within the expert group. Overall, the observed dispersion of weights implies a moderate level of variability in expert opinions, while the consistency of average values across the AHP and FAHP approaches shows promising results for the proposed framework robustness, as presented in Table 9.

Comparing the points of view for Model 5, it can be observed that the pessimistic perspective of Expert 2 yields the highest score (0.440432), followed by the semi-pessimistic and balanced cases. In contrast, the lowest weight is assigned by Expert 6 within the AHP framework (0.214527), indicating a more conservative evaluation of this model. Furthermore, a notable consistency in the evaluations can be observed among Experts 1, 3, and 5, whose assessments result in identical weights. This alignment suggests a shared perception

of Model 5 among these experts, contributing to the overall stability of the evaluation. Overall, the distribution of weights indicates a certain degree of variability across expert opinions, with Expert 2 demonstrating a more favorable stance, while Expert 6 reflects a more critical viewpoint. Such differences highlight the importance of incorporating multiple perspectives when assessing LLM-based models.

Table 9. The sub-criteria and weights corresponding to Model 4 and Model 5 (T1).

								AHP	**FAHP**				
									λ = 0	**λ = 0**	**λ = 0**	**λ = 0**	**λ = 0**
M4	Ex1	A4	D3	E3	P2	V2	L2	0.131	0.155	0.162	0.166	0.169	0.171
	Ex2	A1	D1	E1	P1	V1	L1	0.525	0.472	0.459	0.45	0.444	0.44
	Ex3	A4	D3	E3	P2	V2	L2	0.131	0.155	0.162	0.166	0.169	0.171
	Ex4	A1	D2	E1	P1	V1	L1	0.476	0.437	0.426	0.419	0.415	0.411
	Ex5	A4	D3	E3	P2	V2	L2	0.131	0.155	0.162	0.166	0.169	0.171
	Ex6	A2	D2	E2	P2	V2	L2	0.266	0.295	0.299	0.301	0.303	0.304
	Ex7	A2	D2	E2	P2	V3	L2	0.258	0.285	0.288	0.289	0.29	0.291
	Ex8	A3	D3	E2	P2	V3	L2	0.172	0.197	0.201	0.204	0.206	0.208
M5	Ex1	A2	D2	E3	P2	V1	L2	0.262	0.288	0.293	0.296	0.298	0.3
	Ex2	A1	D1	E1	P2	V1	L1	0.483	0.44	0.431	0.425	0.421	0.418
	Ex3	A2	D2	E3	P2	V1	L2	0.262	0.288	0.293	0.296	0.298	0.3
	Ex4	A2	D2	E3	P3	V2	L1	0.232	0.255	0.26	0.264	0.266	0.268
	Ex5	A2	D2	E3	P2	V1	L2	0.262	0.288	0.293	0.296	0.298	0.3
	Ex6	A2	D2	E3	P3	V3	L2	0.215	0.237	0.242	0.244	0.246	0.248
	Ex7	A2	D2	E3	P2	V2	L1	0.254	0.279	0.283	0.285	0.287	0.289
	Ex8	A2	D2	E2	P3	V2	L2	0.244	0.271	0.276	0.279	0.281	0.283

For the topic T2—Requirements Analysis and Specification, Model M2 is again identified as the highest-ranked alternative (consistent with the results for T1), with an average weight of 0.465038 in the AHP case, while in the balanced FAHP scenario the weight amounts to 0.413013. Six experts assigned the highest evaluations to M2, yielding weights of 0.524942 (AHP), and 0.471839, 0.450208, and 0.440029 in the FAHP cases ($\lambda = 0$; $\lambda = 0.5$; $\lambda = 1$), confirming a strong consensus regarding the importance of this model. In contrast, Experts 4 and 5 provided less high evaluations, particularly within the first, second, and fourth groups of criteria. Their assessments resulted in weights of 0.384713 and 0.185936 in the AHP framework, and 0.385261 and 0.215304, as well as 0.382052 and 0.222348, in the semi-pessimistic and semi-optimistic FAHP cases, respectively. Overall, these results indicate a divergence in expert opinions. However, the dominance of Model M2 remains consistent across all considered scenarios, supporting the robustness of the ranking, despite variability in individual judgments and criteria group sensitivities.

The notable change in the ranking process is how Model M3 made a significant jump from fifth (in T1) to second place overall. This change is primarily attributable to strong evaluation from an average of all three sets of evaluation criteria as well as by a number of experts who provided consistent evaluations over time. The average weights from the AHP and three FAHP cases, $\lambda = 0$, $\lambda = 0.5$ and $\lambda = 1$, from the group of experts were 0.369175, 0.359609, 0.354688 and 0.352513, respectively. The highest individual weight (0.524942) is assigned to Model M2 by Expert 2, while the lowest weight is observed for Experts 6 and 8 (AHP: 0.243084), indicating a considerable spread in expert judgments. In addition, Experts 1 and 3 evaluated Model M2 the same across all of the FAHP cases, as well as Experts 6 and 8. The weights assigned to those experts in the balanced FAHP case were 0.393184 and 0.278096, respectively, whereas their corresponding weights in the $\lambda = 0.25$ case were 0.397617 and 0.275044, and were 0.3902 and 0.280204 in the $\lambda = 0.75$ case. The results highlight that the improved ranking position of Model M3 is driven not only by

stronger performance across key criteria groups, but also by the reduced variability in expert assessments, which contributes to a more stable and reliable evaluation outcome.

In the middle of the ranking lies Model M1, with identical evaluations provided by Experts 1, 3, and 5. The corresponding weights are 0.436953, 0.426050, 0.419296, 0.414705, and 0.411383 across all FAHP cases, indicating a high level of agreement among these experts. Expert 2 assigns the highest scores to Model M2 (AHP: 0.524942; FAHP, λ = 0.5: 0.450208), whereas Experts 6 and 8 provide the lowest evaluations among all experts (AHP: 0.156542; FAHP, λ = 0.5: 0.163544), reflecting a more critical perspective. Furthermore, the results obtained from Expert 7 are higher than those of Expert 4, as can be seen in Figure 12, by factors of 2.082, 1.940, 1.806, and 1.745 (AHP, FAHP: λ = 0; λ = 0.5; λ = 1), respectively. This notable discrepancy highlights the variability in expert judgments, although it does not significantly affect the relative positioning of Model M1 within the overall ranking.

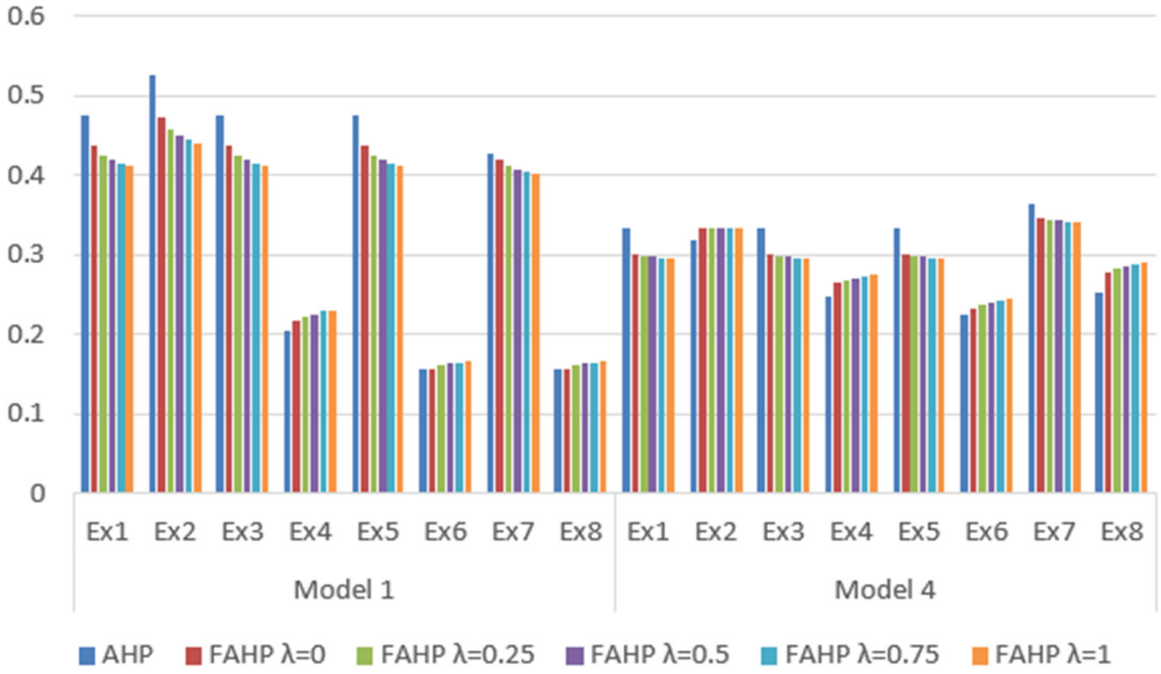

Figure 12. The weights of the models M1 and M4 (T2 topic) for the AHP and five FAHP cases.

Model M4 (see Figure 12) has AHP weights of 0.334999, 0.317712, 0.334999, 0.248742, 0.334999, 0.225310, 0.363783, and 0.253781, and all experts gave it similar scores. This distribution shows that experts mostly agree, even though there are some differences in how they rate things. The highest scores are assigned by Expert 7, amounting to 0.344618 and 0.342063 in the semi-pessimistic and semi-optimistic FAHP cases, respectively. In contrast, Expert 6 provides the lowest evaluation (0.225310), which is lower by a factor of 1.615 compared to the highest value, reflecting a more critical viewpoint. The average weights for the different FAHP scenarios range from 0.295099 to 0.296535, which shows that there is minimal variation. These averages are lower than the highest assigned weight by factors of 1.15 to 1.76, presenting a stable central tendency despite small differences in expert opinions. The fairly small range of average values shows that Model M4 stays in the same place in the ranking, supporting the idea that the evaluation results are strong even when FAHP parameters change.

At the bottom of the ladder, two places lower in rank than in T1, lies model M5, with averaged weights in all six cases of algorithms of 0.201537, 0.208423, 0. 213105, 0.216114, 0.218217, and 0.219774, respectively, showing the high degree of experts' opinion similarity. The results obtained from Expert 5 (FAHP: λ = 0; λ = 0.5; and λ = 1: 0.282495, 0.281302, and

0.280807) are higher than those from other experts, especially Expert 7, whose lowest grades in the same cases correspond to the weights 0.148925, 0.1593, and 0.164251, as can be seen in Figure 13. Despite the observed differences between individual expert assessments, the relatively narrow range of average weights confirms the stability of Model M5's position at the lower end of the ranking. This consistency further supports the robustness of the evaluation results across different methodological settings.

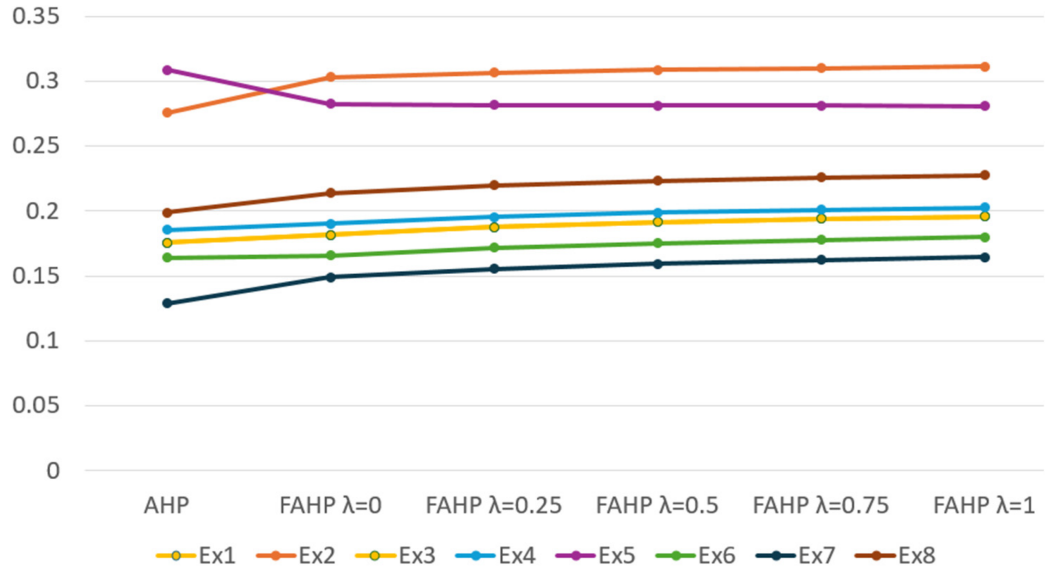

Figure 13. The weights of the model M5 (T2 topic) for the AHP and five FAHP cases.

With the same ranking order, as it was presented in T2, model M2 remains highly graded, the leading one in topic T3, followed by M3, M1, M4, and at the bottom, M5, supporting the ranking robustness and proposed framework justification (see Figure 14).

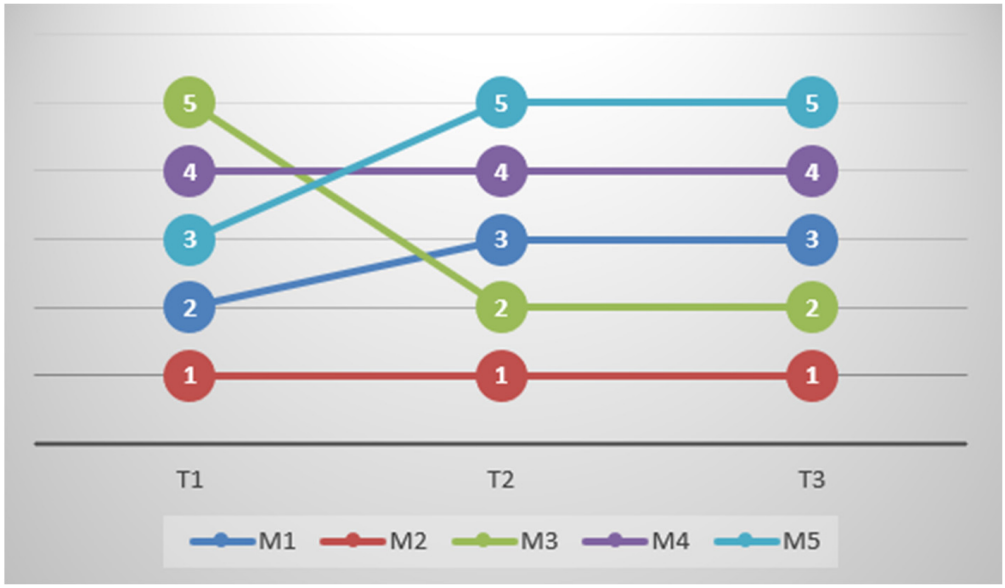

Figure 14. The ranking order of models M1–M5 in topics T1–T3.

According to the points of view of Experts 1 and 3, their same weights to the model M1 of 0.388057, 0.393184, and 0.404877 (FAHP optimistic, balanced, and pessimistic points of view) represent the highest grades, while Expert 6, with some differences in judgements, yields smaller weights: 0.166418, 0.16476, and 0.160969. The tendency of similar experts'

opinions can be seen in the results of Experts 2 and 7, whose determined weights are only 1.016 and 1.043 times less important than the leading ones in the AHP case. Those quotients in the FAHP cases (λ = 0.25, λ = 0.75) are equal to 0.964 and 1.047; and 0.963 and 1.048, respectively. The average weights for M1, $w_{T3}(M1)$, are equal to 0.34001 (AHP), 0.325443, 0.3204, and 0.318141 (FAHP: λ = 0, λ = 0.5, and λ = 1).

The majority of experts (five out of eight) assigned the highest grades for model M2 with the corresponding weights 0.524942, 0.471839, 0.45864, 0.450208, 0.444345, and 0.440029 in all six cases (AHP and FAHP), respectively. Expert 6, with the lowest, but relatively uniform values, yields weights in the range 0.336 to 0.351, remaining consistent across different methodological settings, further contributing to the overall stability of the decision-making process.

Expert 2 favors model M3, assigning it near-maximum evaluations with weight 0.424096 in the balanced FAHP case, and 0.430207 and 0.419841 when λ = 0.25 and λ = 0.75. A similar evaluation pattern is observed for Experts 1 and 3, who provide identical assessments (AHP: 0.475904; FAHP: λ = 0.25 and λ = 0.75: 0.426050 and 0.414705). Likewise, Experts 6 and 8 also demonstrate consistent judgments, with weights of 0.231940 (AHP) and 0.260102 and 0.265892 (FAHP: λ = 0.25 and λ = 0.75). One can say that consistency within expert pairs suggests the presence of coherent evaluation patterns, making the reliability of the obtained results stronger despite variations in individual scoring levels.

The lowest ranked in T3, as in the previous topic, are models M4 and M5. The consensus of the first three experts, as can be seen in Table 10, yields the following weights for M4, 0.433814 (AHP) and 0.393814 for balanced FAHP, higher than the corresponding average values of 0.12 and 0.083, respectively, indicating a more favorable assessment within this subgroup of experts. The weights obtained from Expert 4, the lowest of all, determined by some differences in judgements, are 0.207347 (smallest weight, AHP), and 0.213914, 0.22054, and 0.223811 (FAHP: λ = 0., λ = 0.5, and λ = 1).

At the bottom of the ladder is placed, according to the expert's assessments, model M5, with average weights of 0.306293, 0.301373, 0.301069, 0.301001, 0.301015, and 0.301059 in all six cases of algorithms, respectively. The highest grades were obtained from Experts 1 and 2 with 1.416, 1.343, 1.306, and 1.289 times higher weights than the averaged ones (AHP, FAHP: λ = 0., λ = 0.5, and λ = 1), while decision variations make Expert 4's grades the lowest ones: 0.163978, 0.17514, and 0.177918 (AHP, FAHP: λ = 0.25 and λ = 75).

The consistency of the ranking results across multiple topics and evaluation scenarios demonstrates that the proposed framework can be reliably applied to the evaluation of LLM-generated educational content in software engineering education. The framework enables systematic comparison of different models across different topics, criteria, and cognitive levels, confirming its practical applicability in real educational evaluation contexts. Furthermore, the high level of agreement between the results obtained using the crisp AHP method and multiple FAHP scenarios indicates that the integration of AHP and Fuzzy AHP suggests the effectiveness of the multi-criteria decision-making framework under uncertainty. The stability of rankings and the small variations in calculated weights across different methodological settings demonstrate the reliability of the proposed approach. These findings provide answers to RQ1 and RQ4, confirming that the proposed framework can be successfully applied to the evaluation of LLM-generated educational content and that the combined AHP–FAHP approach provides a robust evaluation and ranking framework.

Table 10. The sub-criteria and weights corresponding to the models M1–M5 (T3) and all six cases of algorithms.

								AHP	FAHP				
									λ = 0	λ = 0	λ = 0	λ = 0	λ = 0
M1	Ex1	A1	D2	E2	P1	V1	L1	0.434	0.405	0.398	0.393	0.39	0.388
	Ex2	A2	D1	E1	P1	V1	L1	0.427	0.419	0.413	0.408	0.405	0.403
	Ex3	A1	D2	E2	P1	V1	L1	0.434	0.405	0.398	0.393	0.39	0.388
	Ex4	A1	D3	E3	P3	V2	L2	0.285	0.251	0.25	0.25	0.25	0.25
	Ex5	A2	D1	E2	P1	V1	L1	0.385	0.387	0.384	0.382	0.381	0.38
	Ex6	A2	D4	E4	P3	V3	L2	0.16	0.161	0.163	0.165	0.166	0.166
	Ex7	A1	D2	E2	P1	V2	L1	0.416	0.387	0.38	0.375	0.372	0.37
	Ex8	A2	D3	E3	P3	V3	L2	0.18	0.189	0.193	0.196	0.198	0.2
M2	Ex1	A1	D1	E1	P1	V1	L1	0.525	0.472	0.459	0.45	0.444	0.44
	Ex2	A1	D1	E1	P1	V1	L1	0.525	0.472	0.459	0.45	0.444	0.44
	Ex3	A1	D1	E1	P1	V1	L1	0.525	0.472	0.459	0.45	0.444	0.44
	Ex4	A1	D1	E1	P1	V1	L1	0.525	0.472	0.459	0.45	0.444	0.44
	Ex5	A2	D2	E1	P1	V1	L1	0.378	0.384	0.38	0.377	0.376	0.374
	Ex6	A2	D2	E2	P1	V1	L1	0.336	0.352	0.352	0.351	0.351	0.351
	Ex7	A1	D1	E1	P1	V1	L1	0.525	0.472	0.459	0.45	0.444	0.44
	Ex8	A1	D2	E2	P1	V2	L1	0.416	0.387	0.38	0.375	0.372	0.37
M3	Ex1	A1	D2	E1	P1	V1	L1	0.476	0.437	0.426	0.419	0.415	0.411
	Ex2	A1	D1	E2	P1	V1	L1	0.483	0.44	0.43	0.424	0.42	0.417
	Ex3	A1	D2	E1	P1	V1	L1	0.476	0.437	0.426	0.419	0.415	0.411
	Ex4	A2	D2	E2	P1	V1	L1	0.336	0.352	0.352	0.351	0.351	0.351
	Ex5	A1	D2	E2	P1	V1	L1	0.434	0.405	0.398	0.393	0.39	0.388
	Ex6	A2	D2	E3	P3	V2	L1	0.232	0.255	0.26	0.264	0.266	0.268
	Ex7	A2	D2	E2	P2	V2	L1	0.276	0.303	0.306	0.308	0.31	0.311
	Ex8	A2	D2	E3	P3	V2	L1	0.232	0.255	0.26	0.264	0.266	0.268
M4	Ex1	A1	D2	E2	P1	V1	L1	0.434	0.405	0.398	0.393	0.39	0.388
	Ex2	A1	D2	E2	P1	V1	L1	0.434	0.405	0.398	0.393	0.39	0.388
	Ex3	A1	D2	E2	P1	V1	L1	0.434	0.405	0.398	0.393	0.39	0.388
	Ex4	A2	D3	E4	P2	V2	L1	0.207	0.214	0.218	0.221	0.222	0.224
	Ex5	A2	D2	E2	P1	V1	L1	0.336	0.352	0.352	0.351	0.351	0.351
	Ex6	A2	D3	E3	P2	V2	L1	0.219	0.23	0.235	0.237	0.239	0.241
	Ex7	A2	D3	E3	P2	V2	L2	0.209	0.222	0.227	0.23	0.232	0.234
	Ex8	A2	D2	E3	P2	V3	L2	0.236	0.261	0.264	0.266	0.268	0.269
M5	Ex1	A1	D2	E2	P1	V1	L1	0.434	0.405	0.398	0.393	0.39	0.388
	Ex2	A2	D2	E2	P2	V1	L1	0.294	0.321	0.324	0.326	0.328	0.329
	Ex3	A1	D2	E2	P1	V1	L1	0.434	0.405	0.398	0.393	0.39	0.388
	Ex4	A2	D3	E4	P4	V2	L2	0.164	0.166	0.172	0.175	0.178	0.18
	Ex5	A1	D3	E2	P1	V1	L1	0.399	0.357	0.349	0.345	0.342	0.34
	Ex6	A2	D3	E3	P2	V1	L1	0.237	0.248	0.253	0.255	0.257	0.259
	Ex7	A2	D3	E4	P3	V2	L2	0.176	0.182	0.188	0.191	0.194	0.196
	Ex8	A2	D2	E3	P1	V1	L1	0.314	0.328	0.328	0.328	0.328	0.328

5. Conclusions

This study has demonstrated the feasibility of using a combined AHP and Fuzzy AHP approach for evaluating the quality of educational content generated by LLMs in software engineering education. The proposed framework integrates structured multi-criteria methods with fuzzy logic in order to model both the complexity of evaluation criteria and the uncertainty inherent in expert judgments. In this way, the evaluation process becomes more systematic, transparent, and suitable for complex educational evaluation scenarios involving subjective assessments. It is important to emphasize that this study does not evaluate LLMs as software systems, but rather evaluates the quality of the educational content generated by these models in a specific educational context. Therefore, the focus of

the proposed framework is not on model performance in general, but on the pedagogical value, accuracy, structure, and practical applicability of the generated instructional content in software engineering education.

The results of the study indicate that the proposed framework can successfully support the evaluation and ranking of LLM-generated educational content across different topics and cognitive levels. The findings show that accuracy and professional correctness, as well as depth and coverage of content, represent the most influential criteria in the evaluation process. In addition, the results demonstrate a high level of consistency in rankings across different FAHP optimism index values, indicating that the proposed model is stable and not sensitive to variations in decision-maker attitudes or fuzzy parameters. The high similarity of ranking results obtained using the crisp AHP method and multiple FAHP scenarios confirms the robustness and reliability of the proposed multi-criteria evaluation framework. The integration of classical AHP and Fuzzy AHP enables both clear prioritization of the evaluation criteria and the incorporation of uncertainty, vagueness, and subjectivity present in expert-based evaluations. Furthermore, the framework allows the identification of variability among expert judgments while maintaining stable overall ranking results, which confirms its suitability for complex evaluation tasks such as the assessment of LLM-generated educational content.

The results of this study provide answers to the research questions. The proposed framework can support the evaluation of LLM-generated educational content in software engineering education (RQ1), highly influential sub-criteria do exist (RQ2), the ranking of sub-criteria does not significantly change for different optimism index values (RQ3), and the combination of AHP and Fuzzy AHP provides a robust and reliable framework for evaluating and ranking the quality of educational content generated by large language models (RQ4).

The main contribution of this research is a structured multi-criteria decision-making framework under uncertainty for evaluating the quality of LLM-generated educational content. The study contributes both to the field of software engineering education and to the field of multi-criteria decision-making by demonstrating that the integration of classical and fuzzy approaches can provide stable and reliable results in complex evaluation problems involving subjective expert judgments. However, despite the promising results, this study has several limitations. The evaluation was conducted on a limited dataset, which may not fully capture the variability of real-world conditions. Additionally, no external validation on independent datasets was performed. Therefore, the generalizability of the proposed framework should be interpreted with caution and remains an open direction for future research. Future work will focus on evaluating the proposed framework on larger and more diverse datasets, as well as conducting cross-domain validation to better assess its generalization capabilities. Additionally, the proposed framework relies on expert-based evaluations, which are inherently subjective and may introduce bias depending on the selection and experience of the experts involved. Second, although a fuzzy representation using triangular fuzzy numbers (TFNs) is employed, the initial judgments are collected in a crisp form, which limits the extent to which uncertainty is genuinely captured. In this sense, the adopted fuzzy approach should be interpreted as a simplified representation of vagueness rather than a comprehensive uncertainty modeling framework. Furthermore, the parameter λ is applied only during the defuzzification stage, which restricts its role to influencing the final numerical scores rather than enabling full uncertainty propagation throughout the decision-making process. As a result, the model may not fully reflect the variability and ambiguity present in real-world decision environments. In addition, the structure of the criteria and sub-criteria, although carefully designed, may still involve a degree of overlap or dependency, which is not explicitly modeled within the current

AHP-based framework. Finally, the results are context-dependent and may vary with different expert groups, evaluation settings, or alternative methodological choices. Future research could address these limitations by incorporating more advanced uncertainty modeling approaches, such as interval-valued judgments, type-2 fuzzy sets, or probabilistic frameworks, as well as by applying robustness.

Future research should include a larger number of experts, a broader range of educational topics, and additional LLMs in order to further validate and generalize the proposed framework. Future studies may also explore the integration of automated evaluation metrics and LLM-as-a-judge approaches, as well as the application of other multi-criteria decision-making methods such as TOPSIS, ANP, or VIKOR. Additionally, future work may consider the use of other types of fuzzy numbers, such as trapezoidal fuzzy numbers, spherical fuzzy sets, or Z-numbers, to further improve the modeling of uncertainty in expert evaluations.

Author Contributions: Conceptualization, D.J.S. and J.L.J.; methodology, J.L.J. and D.J.S.; software, J.L.J. and D.J.S.; validation, J.L.J., D.J.S. and N.O.V.; formal analysis, D.J.S., N.O.V. and B.M.R.; investigation, J.L.J. and D.J.S.; resources, D.S.D. and B.M.R.; data curation, J.L.J.; writing—original draft preparation, D.J.S. and J.L.J.; writing—review and editing, N.O.V., B.M.R. and D.S.D.; visualization, D.J.S. and J.L.J.; supervision, D.S.D.; project administration, D.S.D.; funding acquisition, N.O.V., B.M.R., D.J.S. and D.S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science, Innovations and Technological Development of Serbia, through the grants 451-03-34/2026-03/200251 and 451-03-34/2026-03/200102 and funded by the Faculty of Teacher Education, Leposavic, through the grant IMP-003. Nenad Vesić was supported by project O-40-26 of the Serbian Academy of Sciences and Arts, Branch in Niš.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: This manuscript was prepared by the authors. Generative artificial intelligence (GPT-5) was used to assist with language editing, stylistic refinement, translation, and improvement of clarity during the preparation of this manuscript. The tool was accessed between February and April 2026. The selection of studies, screening, data extraction, coding, analysis, and interpretation of the results were conducted exclusively by the authors. AI tools did not contribute to data generation, analytical decisions, or substantive interpretation of the results. The authors assume full responsibility for the content of this manuscript. All authors have read and agreed to the published version of the manuscript.

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

References

1. Xu, H.; Gan, W.; Qi, Z.; Wu, J.; Yu, P.S. Large Language Models for Education: A Survey. *arXiv* **2024**, arXiv:2405.13001. [CrossRef]
2. Peláez-Sánchez, I.C.; Velarde-Camaqui, D.; Glasserman-Morales, L.D. The Impact of Large Language Models on Higher Education: Exploring the Connection between AI and Education 4.0. *Front. Educ.* **2024**, *9*, 1392091. [CrossRef]
3. Jošt, G.; Taneski, V.; Karakatič, S. The Impact of Large Language Models on Programming Education and Student Learning Outcomes. *Appl. Sci.* **2024**, *14*, 4115. [CrossRef]
4. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
5. Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. Holistic Evaluation of Language Models. *arXiv* **2022**, arXiv:2211.09110. [CrossRef]
6. Saaty, T.L. How to Make a Decision: The Analytic Hierarchy Process. *Eur. J. Oper. Res.* **1990**, *48*, 9–26. [CrossRef]
7. Wang, N.; Ren, Z.; Zhang, Z.; Fu, J. Evaluation and Prediction of Higher Education System Based on AHP-TOPSIS and LSTM Neural Network. *Appl. Sci.* **2022**, *12*, 4987. [CrossRef]

8. Sommerville, I. *Engineering Software Products*; Pearson: Upper Saddle River, NJ, USA, 2019; ISBN 9780135210642.
9. Washizaki, H. *Guide to the Software Engineering Body of Knowledge*; IEEE Computer Society: Washington, DC, USA, 2024.
10. Shi, Y.; Yu, K.; Dong, Y.; Chen, F. Large Language Models in Education: A Systematic Review of Empirical Applications, Benefits, and Challenges. *Comput. Educ. Artif. Intell.* **2026**, *10*, 100529. [CrossRef]
11. Zhang, X.; Zhang, P.; Shen, Y.; Liu, M.; Wang, Q.; Gašević, D.; Fan, Y. A Systematic Literature Review of Empirical Research on Applying Generative Artificial Intelligence in Education. *Front. Digit. Educ.* **2024**, *1*, 223–245. [CrossRef]
12. Baidoo-Anu, D.; Owusu Ansah, L. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *J. AI* **2023**, *7*, 52–62. [CrossRef]
13. Perkins, M. Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond. *J. Univ. Teach. Learn. Pract.* **2023**, *20*, 7. [CrossRef]
14. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2022**, *55*, 1–38. [CrossRef]
15. Lin, P.; Deng, Q.; Zhou, Y. Towards Responsible AI in Education: A Delphi-AHP-Based Framework for Evaluating Educational Large Language Models. *Comput. Educ. Artif. Intell.* **2026**, *10*, 100534. [CrossRef]
16. Zhai, X. ChatGPT for next Generation Science Learning. *SSRN Electron. J.* **2023**, *29*, 42–46. [CrossRef]
17. Garousi, V.; Giray, G.; Tuzun, E.; Catal, C.; Felderer, M. Closing the Gap between Software Engineering Education and Industrial Needs. *IEEE Softw.* **2020**, *37*, 68–77. [CrossRef]
18. Lu, H.-A.; Sung, C.-Y. A Delphi-AHP Model of Consultants Evaluation for Airline Operation Systems. *Int. J. Appl. Phys. Sci.* **2017**, *3*, 5–12. [CrossRef]
19. BCcampus. Open Textbook Review Criteria. 2016. Available online: https://open.umn.edu/opentextbooks/reviews/rubric (accessed on 8 May 2026).
20. Achieve, I. *Rubrics for Evaluating Open Educational Resources (OER) Objects*; Achieve, Inc.: Washington, DC, USA, 2011.
21. Yuan, M.; Recker, M. Not All Rubrics Are Equal: A Review of Rubrics for Evaluating the Quality of Open Educational Resources. *Int. Rev. Res. Open Distrib. Learn.* **2015**, *16*, 16–38. [CrossRef]
22. Baig, M.I.; Yadegaridehkordi, E. ChatGPT in the Higher Education: A Systematic Literature Review and Research Challenges. *Int. J. Educ. Res.* **2024**, *127*, 102411. [CrossRef]
23. Abdallah, N.; Katmah, R.; Khalaf, K.; Jelinek, H.F. Systematic Review of ChatGPT in Higher Education: Navigating Impact on Learning, Wellbeing, and Collaboration. *Soc. Sci. Humanit. Open* **2025**, *12*, 101866. [CrossRef]
24. Brzaković, A.; Brzaković, T.; Karabašević, D.; Popović, G.; Činčikaitė, R. The Interface between the Brand of Higher Education and the Influencing Factors. *Sustainability* **2022**, *14*, 6151. [CrossRef]
25. Yan, L.; Sha, L.; Zhao, L.; Li, Y.; Martinez-Maldonado, R.; Chen, G.; Li, X.; Jin, Y.; Gašević, D. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. *Br. J. Educ. Technol.* **2024**, *55*, 90–112. [CrossRef]
26. Kisić, E.; Raspopović Milić, M.; Jović, J.; Zdravković, N. Tracking Student Progress and Generating Personalized Recommendations Using Clustering and Explainable Artificial Intelligence. *Univers. Access Inf. Soc.* **2026**, *25*, 26. [CrossRef]
27. Holmes, W.; Miao, F. *Guidance for Generative AI in Education and Research*; Unesco Publishing: Paris, France, 2023.
28. Frankford, E.; Sauerwein, C.; Bassner, P.; Krusche, S.; Breu, R. AI-Tutoring in Software Engineering Education. *arXiv* **2024**, arXiv:2404.02548. [CrossRef]
29. Yang, A.C.M.; Lin, J.-Y.; Lin, C.-Y.; Ogata, H. Enhancing Python Learning with PyTutor: Efficacy of a ChatGPT-Based Intelligent Tutoring System in Programming Education. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100309. [CrossRef]
30. Zhao, H.; Wu, Y.; Lu, Z.; Yu, X.; Miao, W.; Chen, L. SageJavon: A Scalable AI Tutor for Personalized Programming Learning. *Inf. Process. Manag.* **2026**, *63*, 104605. [CrossRef]
31. Xue, Y.; Chen, H.; Bai, G.R.; Tairas, R.; Huang, Y. Does ChatGPT Help with Introductory Programming? An Experiment of Students Using ChatGPT in CS1. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training*; ACM: New York, NY, USA, 2024.
32. Jury, B.; Lorusso, A.; Leinonen, J.; Denny, P.; Luxton-Reilly, A. Evaluating LLM-Generated Worked Examples in an Introductory Programming Course. In *Proceedings of the 26th Australasian Computing Education Conference*; ACM: New York, NY, USA, 2024; pp. 77–86.
33. Jafari, F.; Keykha, A. Identifying the Opportunities and Challenges of Artificial Intelligence in Higher Education: A Qualitative Study. *J. Appl. Res. High. Educ.* **2024**, *16*, 1228–1245. [CrossRef]
34. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment. *arXiv* **2023**, arXiv:2303.16634. [CrossRef]
35. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv* **2023**, arXiv:2306.05685.
36. Seo, H.; Hwang, T.; Jung, J.; Kang, H.; Namgoong, H.; Lee, Y.; Jung, S. Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy. *Appl. Sci.* **2025**, *15*, 671. [CrossRef]

37. Anghel, C.; Craciun, M.V.; Pecheanu, E.; Cocu, A.; Anghel, A.A.; Iacobescu, P.; Maier, C.; Andrei, C.A.; Scheau, C.; Dragosloveanu, S. CourseEvalAI: Rubric-Guided Framework for Transparent and Consistent Evaluation of Large Language Models. *Computers* **2025**, *14*, 431. [CrossRef]
38. Aksoy, M.; Adem, A.; Dagdeviren, M. Trustworthiness Evaluation of Large Language Models Using Multi-Criteria Decision Making. *IEEE Access* **2025**, *13*, 168183–168201. [CrossRef]
39. Alabool, H.M. Large Language Model Evaluation Criteria Framework in Healthcare: Fuzzy MCDM Approach. *SN Comput. Sci.* **2025**, *6*, 57. [CrossRef]
40. Kalaivani, K.; Kaliyaperumal, P.; Cebi, S. Selection of Third Party Logistics (3PL) to Transport Cold Manufactured Items Using MCDM under Neutrosophic Approach. In *Proceedings of the 2023 Innovations in Power and Advanced Computing Technologies (i-PACT)*; IEEE: Piscataway, NJ, USA, 2023.
41. Saaty, R.W. The Analytic Hierarchy Process—What It Is and How It Is Used. *Math. Model.* **1987**, *9*, 161–176. [CrossRef]
42. Sančanin, B.; Penjišević, A.; Simjanović, D.J.; Ranđelović, B.M.; Vesić, N.O.; Mladenović, M. A Fuzzy AHP and PCA Approach to the Role of Media in Improving Education and the Labor Market in the 21st Century. *Mathematics* **2024**, *12*, 3616. [CrossRef]
43. Veljić, A.; Viduka, D.; Ilić, L.; Karabasevic, D.; Šijan, A.; Papić, M. Sustainable Decision-Making in Higher Education: An AHP-NWA Framework for Evaluating Learning Management Systems. *Sustainability* **2025**, *17*, 10130. [CrossRef]
44. Chang, D.-Y. Applications of the Extent Analysis Method on Fuzzy AHP. *Eur. J. Oper. Res.* **1996**, *95*, 649–655. [CrossRef]
45. Domazet, D. Competences-Driven and GenAI-Supported Hybrid Personalized Learning. In *Proceedings of the 16th International Conference on eLearning (ELEARNING2025)*; Belgrade Metropolitan Univesity: Belgrade, Serbia, 2025.
46. Anderson, L.W.; Krathwohl, D.R. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition*; Addison Wesley Longman, Inc.: Boston, MA, USA, 2001.
47. Zadeh, L.A. The Concept of a Linguistic Variable and Its Application to Approximate Reasoning—II. *Inf. Sci.* **1975**, *8*, 301–357. [CrossRef]
48. Zadeh, L.A. The Concept of a Linguistic Variable and Its Application to Approximate Reasoning—I. *Inf. Sci.* **1975**, *8*, 199–249. [CrossRef]
49. Milošević, D.M.; Milošević, M.R.; Simjanović, D.J. Implementation of Adjusted Fuzzy AHP Method in the Assessment for Reuse of Industrial Buildings. *Mathematics* **2020**, *8*, 1697. [CrossRef]
50. Kulak, O.; Durmuşoğlu, M.B.; Kahraman, C. Fuzzy Multi-Attribute Equipment Selection Based on Information Axiom. *J. Mater. Process. Technol.* **2005**, *169*, 337–345. [CrossRef]
51. Simjanović, D.J.; Zdravković, N.; Vesić, N.O. On the Factors of Successful E-Commerce Platform Design during and after COVID-19 Pandemic Using Extended Fuzzy AHP Method. *Axioms* **2022**, *11*, 105. [CrossRef]
52. Ceballos, B.; Lamata, M.T.; Pelta, D.A. A Comparative Analysis of Multi-Criteria Decision-Making Methods. *Prog. Artif. Intell.* **2016**, *5*, 315–322. [CrossRef]