International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Evaluating the Accuracy and Usability of an Academic Digital Information Assistant Chatbot Based on Retrieval-Augmented Generation for Academic Learning Grant Services

Umi Widyastuti[1], Suryadi Budi Utomo[2], Retno Wulan Damayanti[3]*, Hasna Wahidaturrasyidah Anshori[4], Sigit Prastowo[5], Agus Dwi Priyanto[6], Haryono Setiadi[7], Dwi Wahyudi[8], Adip Safiudin[9], Galih Wahyu Pambudi[10], Nunung Sulistyawati[11]

[1,8,11] Directorate of Academic Affairs, Universitas Sebelas Maret, Surakarta, Indonesia
[2] Chemistry Education Study Program, Faculty of Teacher Training and Education, Universitas Sebelas Maret, Surakarta, Indonesia
[3,4] Industrial Engineering Study Program, Faculty of Engineering, Universitas Sebelas Maret, Surakarta, Indonesia
[5] Animal Science Study Program, Universitas Sebelas Maret, Surakarta, Indonesia
[6] English Study Program, Vocational School, Universitas Sebelas Maret, Surakarta, Indonesia
[7] Informatics Study Program, Universitas Sebelas Maret, Surakarta, Indonesia
[9,10] Directorate of Information and Communication Technology, Universitas Sebelas Maret, Surakarta, Indonesia

ABSTRACT: Information services for student grant programmes require answers that are fast, consistent, and grounded in official documents, while the rules themselves change with every programme cycle. This study evaluates AIDA (Academic Digital Information Assistant), a university academic service chatbot, within the domain of an academic learning grant programme. Following a Design Science Research framework, the evaluation proceeded in two cycles. The first measured answer accuracy through the assessment of three programme administrators across 32 test questions on the dimensions of relevance, correctness, and completeness. The second measured usability through the Chatbot Usability Questionnaire administered to 76 student users. The first cycle yielded 90.6% relevance, 81.3% correctness, 90.6% completeness, and an overall pass rate of 81.3%, with inter-rater agreement between 82% and 85%. All six failing items shared a single root cause, namely timeline information that had not been updated to the current programme period; three of them were nonetheless judged relevant and complete, rendering the error difficult for users to detect. The second cycle produced a usability score of 67.72 (SD 11.93) with a reliability of 0.865, highest on ease of use and lowest on conversational naturalness and error handling. The study concludes that the principal constraint on administrative service chatbots lies not in language understanding but in the governance of content freshness, and proposes a knowledge base update mechanism tied to the programme calendar.

KEYWORDS: academic chatbot, BM25, chatbot usability questionnaire, knowledge base governance, retrieval-augmented generation

INTRODUCTION

Policies on learning outside the study programme allow students to complete part of their study load beyond the classroom, with an equivalent recognition of semester credit units [1], [2]. The university that forms the setting of this study implements the policy through an academic learning grant. The programme provides competitive incentives for student groups to design their own learning activities, ranging from internships, teaching assistantships, and research to village development projects, entrepreneurship, and independent study.

The characteristics of the grant make its information service demands intense. Each proposal is prepared by a group of five to ten students, so a single administrative question is often raised repeatedly by several members of the same team. Proposed activities must be recognised as 10 to 20 semester credit units in accordance with the provisions on activity-hour recognition [3], and their financial accountability follows detailed rules on expenditure categories, tax deductions, and the format of supporting evidence. Beyond this, the programme runs in cycles: the provisions applying in one period do not automatically apply in the next. As a result,

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

the service unit faces a high volume of questions during peak periods, together with the risk of providing answers that are no longer valid.

Artificial intelligence chatbots have been widely adopted by higher education institutions to handle such workloads. Systematic reviews show that chatbots are most frequently used for student service support and for handling frequently asked questions, with benefits including round-the-clock availability, answer consistency, and a reduced staff workload [4], [5]. Other reviews note that these benefits depend heavily on the design, the quality of the knowledge base, and the evaluation mechanism [6], [7], [8]. The approach now commonly used to keep answers anchored to official documents is retrieval-augmented generation (RAG), which combines document retrieval with a language model that only summarises the retrieved documents [9], [10]. This approach reduces, although it does not eliminate, the tendency of language models to produce unsupported information [11].

Nevertheless, three gaps remain in the existing research. First, evaluations of campus chatbots are generally conducted on general academic service domains such as registration or payment, whose provisions are relatively stable, so the distinctive character of grant domains with rules that change each period has received little attention. Second, most evaluations rely on user perceptions, while the correctness of answer content is rarely assessed by those who master the underlying policy. Third, implementation reports in Indonesia tend to stop at measuring technical performance without tracing the error patterns that emerge and their implications for service governance [12].

This study addresses those gaps by evaluating AIDA, the university's academic service chatbot, specifically within the academic learning grant domain. Three research questions are posed. First, what is the design and specification of the system actually running in that service? Second, how accurate are the answers it produces according to the assessment of programme administrators? Third, how usable is it according to students as end users?

This study offers three contributions. Methodologically, it presents a dual evaluation combining expert assessment of content correctness with user-based usability measurement, two elements rarely reported in a single study. In terms of findings, it identifies a temporal failure pattern that has received little discussion in the educational chatbot literature. Practically, it proposes a content governance mechanism that academic service units can apply directly.

LITERATURE REVIEW

A. The Use of Chatbots in Higher Education Services

Research on chatbots in higher education has grown rapidly since 2020. Systematic reviews of dozens of studies find that the most common use is not in instruction but in student service support, particularly in handling repetitive administrative questions [13], [14]. Consistently reported benefits include response speed, availability outside working hours, and information consistency [4], [15].

The same reviews, however, also note recurring limitations, namely knowledge base maintenance, answer accuracy, and shallow evaluation [5], [6]. Many studies report success on the basis of user satisfaction without examining whether the answers given are correct according to the policy documents in force. Peyton et al. [7] specifically highlight that university chatbots are often assessed at the level of mere availability rather than on the quality of service resolution. A similar pattern appears in customer service outside education, where chatbots are considered effective for routine questions but remain weak in cases requiring human judgement [16].

B. Retrieval-Augmented Generation and Document Retrieval Methods

RAG was introduced as a way of combining language models with a retrieval mechanism over external sources, so that answers can be linked to reference documents and updated without retraining the model [9]. Surveys of RAG development divide its components into three parts, namely document retrieval, context augmentation, and answer generation [10].

Two approaches are recognised at the retrieval stage. Lexical retrieval assesses relevance on the basis of term occurrence, with BM25 as the most established formulation: document chunks containing rare query terms receive high scores, with adjustments so that long chunks do not automatically prevail [17]. Dense retrieval converts text into numerical representations so that semantic similarity can be computed even when different words are used, as developed in dense passage retrieval models for open-domain question answering [18].

A common assumption places dense retrieval as the more advanced approach. Empirical evidence presents a more balanced picture. Cross-domain testing on the BEIR benchmark found that BM25 remains competitive and frequently outperforms dense retrieval models on domains for which they have not been trained [19]. A similar finding has been reported in the Indonesian context:

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Wulandari and Yamasari [12] tested classical BM25-based RAG in an academic chatbot and concluded that the approach is fast and adequate for procedural information, although it is weak when faced with paraphrased questions.

C. Measuring Chatbot Performance and Usability

The evaluation of service chatbots encompasses three measures that need to be distinguished. The response rate indicates how often the system produces a response, without assessing its content. Accuracy indicates how correct the content of the response is against the reference sources. Usability indicates how users perceive the interaction experience.

These three measures need to be distinguished so that system performance is not overstated. A language model can produce answers that are fluent, relevant, and convincing yet factually wrong. This phenomenon is known as hallucination, namely model output that appears plausible but is not supported by valid sources, and it has been extensively studied in recent years [20], [21]. Testing of several language models in RAG scenarios also shows that the ability to reject incorrect information and to handle mutually contradictory documents remains a weak point [22]. In the context of administrative services, errors of this kind have direct consequences because users make decisions on the basis of those answers.

For the usability dimension, Holmes et al. [23] developed the Chatbot Usability Questionnaire (CUQ) because general usability instruments were judged to capture conversational aspects inadequately. The CUQ consists of 16 statements arranged alternately as positive and negative items, covering chatbot personality, answer quality, ease of navigation, and error handling. The use of such a specific instrument is consistent with the finding that chatbot user experience is shaped not only by answer correctness but also by the tone of communication and by recovery when misunderstandings occur [24], [25].

MATERIALS AND METHODS

A. Research Design

The study adopts the Design Science Research (DSR) framework, which treats the artefact as the object that is both designed and evaluated, with evaluation serving as feedback for artefact improvement [26]. This framework was chosen because the study does not test hypotheses about a population but rather assesses the performance of a system operating in a real service setting.

The evaluation was carried out in two sequential cycles. The first cycle consisted of answer validation by programme administrators on 11 August 2026. The findings of the first cycle became the basis for improving the knowledge base. The second cycle consisted of a user trial with students from 1 to 13 October 2026.

B. Research Object and Scope

The object of the study is AIDA (Academic Digital Information Assistant), the chatbot of the university's Directorate of Academic Affairs, accessible through a web interface and a WhatsApp channel. AIDA serves several academic domains, including study plan cards, tuition fees, academic leave, student identity cards, and registration. This study limits the evaluation to a single domain, namely JARPAK the academic learning grant. The limitation is deliberate so that answer correctness can be traced against one clearly bounded set of policy documents. Consequently, the results are not generalised to AIDA's other service domains.

C. Knowledge Base

The knowledge base for this domain draws on the official grant guidelines, the programme's financial administration manual, and a compilation of questions actually raised by students with the service unit. When the first cycle was conducted, the internal question-and-answer compilation contained 25 entries. After improvement, the compilation grew to 101 entries, with additions concentrated on the provisions of the current period, proposal structure, assessment rubrics, budget details, and financial accountability. As of 13 October 2026, the knowledge base for this domain in the system was recorded as 10 topic blocks containing 136 question-and-answer pairs.

D. Instrument for Answer Validation

The first instrument consists of a set of 32 test questions constructed to follow the structure of the guideline document and covering seven categories, as presented in Table 1.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Table 1. Composition of test items by category

Category	Number of items
Basic concepts and policy	4
Eligibility and proposal submission	6
Credit recognition including community service	6
Schedule, reporting, and assessment	5
Financial accountability	7
Budget provisions and restrictions	2
Official contacts and references	2
Total	**32**

Each item was submitted to AIDA through the web interface, and the resulting answer was recorded without editing. Three raters then assessed each answer independently. All three are administrative staff of the Directorate of Academic Affairs who manage the grant programme on a daily basis, so they both master the programme provisions and are accustomed to answering student questions directly.

The assessment used three dimensions with binary yes or no responses. Relevance means that the answer addresses the question asked. Correctness means that the content of the answer is accurate according to the official documents. Completeness means that the answer fully satisfies the enquirer's information need. Raters also wrote brief notes for every item judged not to meet a dimension.

E. Instrument for Usability Measurement

The second instrument is the Indonesian-adapted version of the CUQ (Chatbot Usability Questionnaire), consisting of 16 alternating positive and negative statements on a five-point scale [23]. Respondents were students who tried AIDA in the academic learning grant domain through a trial link. The questionnaire also collected respondent profiles, the device used, frequency of use, and one open question on suggestions for improvement.

F. Data Analysis

Validation results were computed using a majority rule: an item is declared to meet a dimension if at least two of the three raters answered yes. The percentage per dimension was computed as the number of items meeting that dimension divided by the total number of items. The overall pass rate was computed as the number of items meeting all three dimensions simultaneously. Inter-rater agreement was computed as the percentage of matching assessments across all rater pairs over 96 assessment cells, that is, 32 items multiplied by three dimensions.

CUQ scores were computed following the instrument's standard procedure. Positive statements are scored as the response minus one, negative statements as five minus the response; all values are then summed and multiplied by 100 divided by 64, yielding a range from 0 to 100. Internal reliability was tested using Cronbach's alpha coefficient. Open responses were analysed through inductive thematic grouping. System operational data were obtained from AIDA's production database through read-only access as of 13 October 2026, in the form of aggregate counts without reading the content of user messages.

G. Research Ethics and Data Protection

Participation in the trial was voluntary, and respondents were informed that their responses would be used for system evaluation. Analysis was conducted at the aggregate level, and respondent identities are not presented in the reporting. The service design also applies the principle of human oversight, whereby questions the system cannot answer are referred to staff, in line with guidance on the use of generative artificial intelligence in education [27].

RESULTS

A. System Design and Specification

AIDA answers questions through five layers attempted in sequence, as illustrated in Figure 1. The first layer handles greetings with templates. The second layer matches the question against question-and-answer pairs in the knowledge base on the basis of

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

keywords. The third layer reuses temporarily stored answers for similar questions. The fourth layer retrieves document chunks and then asks the language model to compose an answer solely from those chunks. The fifth layer calls the language model with the entire knowledge block as its material. Earlier and cheaper layers are always attempted first, so that paid layers are invoked only when required.

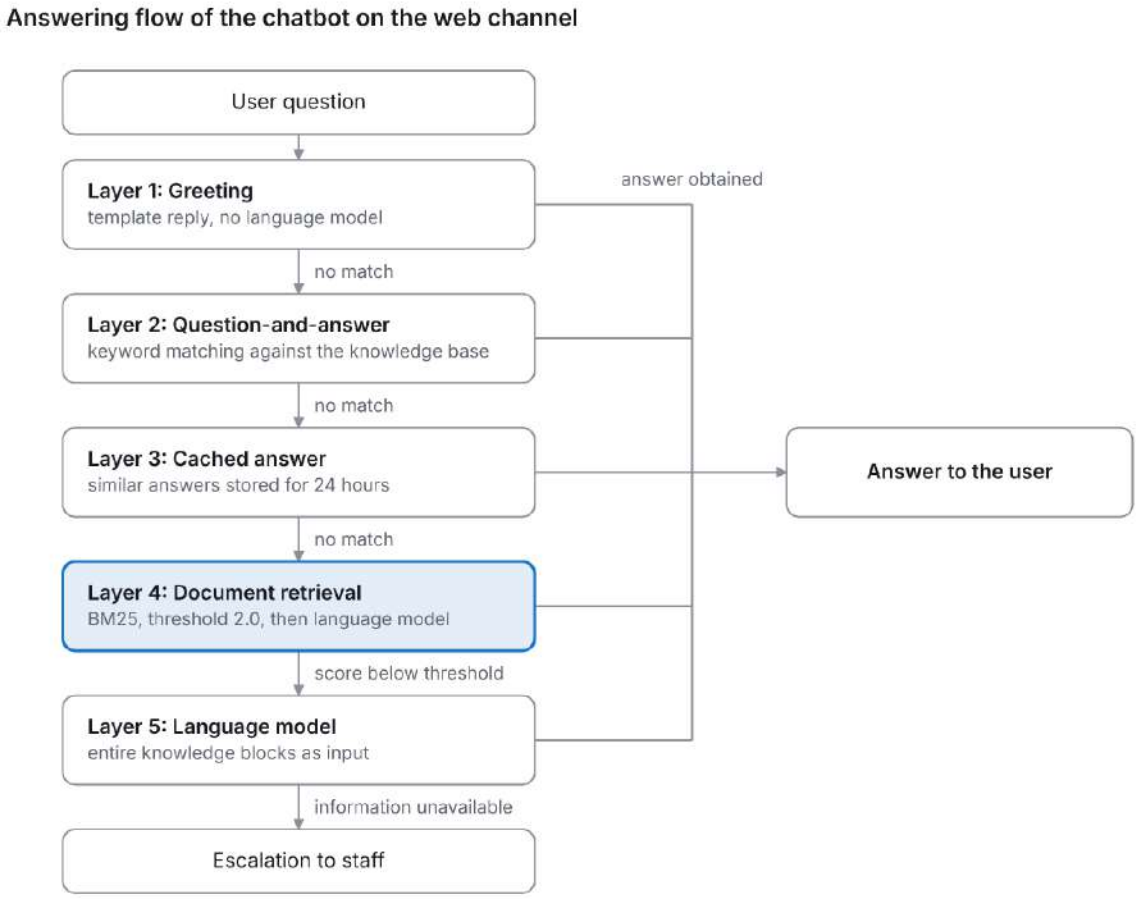


Figure 1. Five-layer answering architecture of AIDA on the web channel

On the web channel, follow-up questions are first rewritten into self-contained sentences by a local model so that pronouns do not obscure retrieval. The system then scores the document chunks, checks the eligibility threshold, and composes an answer. If the information is unavailable, the model inserts a special marker that makes the interface display a link for contacting the Directorate of Academic Affairs. The interface also provides a link that displays the list of documents used as references for the answer, allowing users to trace the sources of the information. The full technical specification is presented in Table 2.

Table 2. Technical specification of AIDA as of 13 October 2026

Component	Value
Primary language model	DeepSeek (deepseek-chat), accessed October 2026
Supporting language model	gemma4:12b via Ollama on an institutional server
Embedding model	Not used
Vector database	Not used
Document retrieval method	Okapi BM25 (k1 = 1.5; b = 0.75) over a MySQL 8 inverted index
Query pre-processing	Stopword removal and light stemming

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Component	Value
Indexed chunks	5,243 chunks from 275 documents
Chunk length	237 words on average (about 1,927 characters), 300-word target, 40-word overlap
Results retrieved	Top 10 chunks, maximum 2 chunks per document
Eligibility threshold	Minimum BM25 score of 2.0 and at least 2 unique keywords
Maximum context to the language model	12,000 characters
Intent classification	Not used in the answering pipeline
Framework	Laravel 13 (PHP 8.4), MySQL 8; WhatsApp gateway on Node.js 22
Grant domain knowledge base	10 topic blocks, 136 question-and-answer pairs

It should be emphasised that although the initial development proposal described a design based on semantic similarity retrieval and intent classification, the system actually in operation uses neither. Document retrieval is entirely lexical. This discrepancy is reported as it stands because it forms the basis for interpreting the evaluation results in the following sections.

Two interaction examples are presented to illustrate the two output paths of the system. Figure 2 shows an answer composed from the knowledge base, including an explicit statement that the expansion of the programme acronym is not recorded in the source documents and therefore requires confirmation from the service unit. Figure 3 shows the escalation mechanism, whereby the system directs the user to the WhatsApp channel of the Directorate of Academic Affairs when the requested information is unavailable.

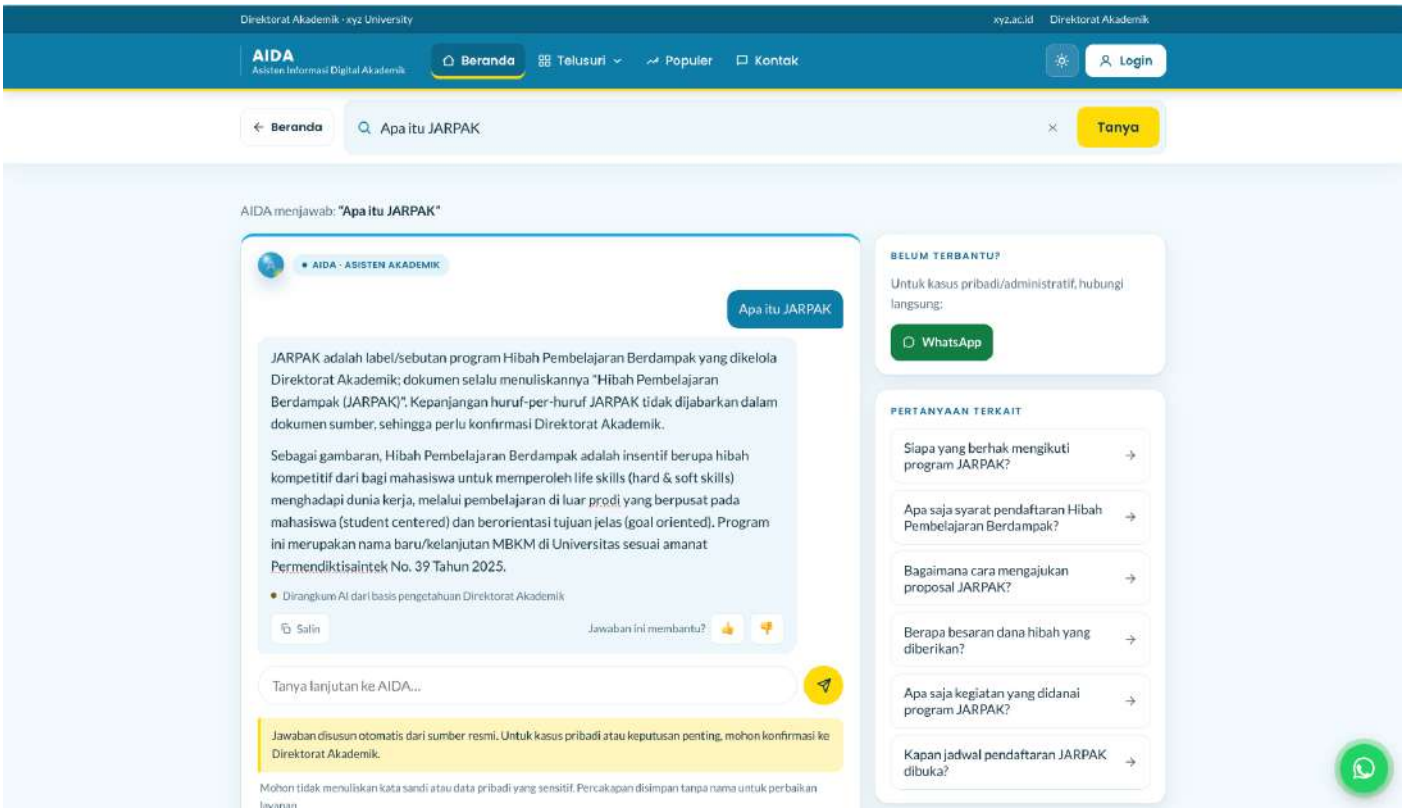


Figure 2. Example of an answer meeting all three assessment dimensions

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

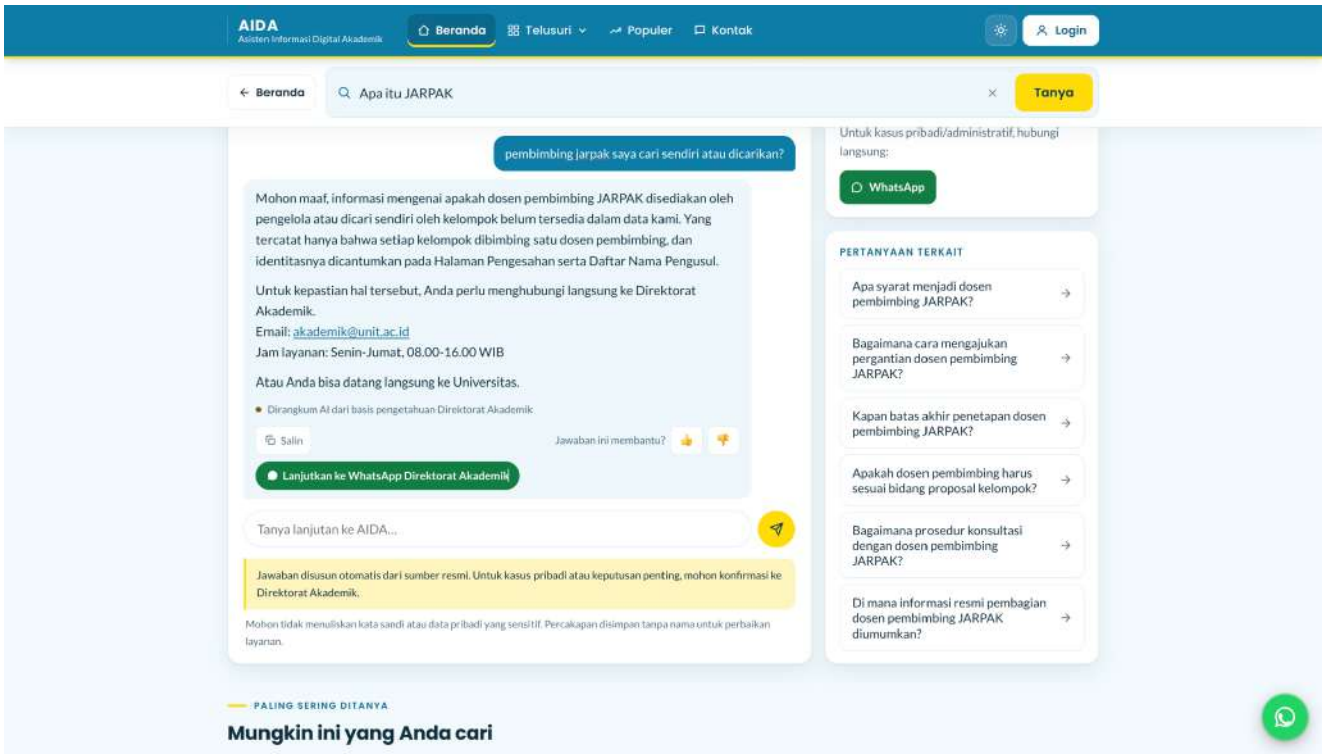


Figure 3. Example of the escalation mechanism when information is unavailable in the knowledge base

B. System Operational Data

System logs for the period from 4 July to 13 October 2026 recorded 1,268 questions across 507 conversations. The distribution of answering layers is presented in Table 3.

Table 3. Distribution of answering layers, 4 July to 13 October 2026

Answering layer	Count	Mean time
Document retrieval and answer generation	735	3.95 seconds
Language model with knowledge blocks	373	4.18 seconds
Instant question-and-answer search	71	—
Single question outside a conversation	41	9.89 seconds
Fallback question-and-answer without the language model	21	0.32 seconds
Failed to answer	15	35.1 seconds
Not found	11	—
Greeting	1	—
Total	**1,268**	

The system produced a response for 1,242 of 1,268 questions, or 98%. This figure indicates the response rate, not accuracy, and therefore cannot be read as a measure of answer correctness. Examination of the database revealed a problem not visible from the user side. A total of 272 web pages retrieved automatically on 10 July 2026 were assigned a 90-day validity period and all expired on 8 October 2026. From that date, of the 5,243 indexed chunks only 47 chunks belonging to three PDF files remained active in retrieval. The impact is visible in the system logs: the proportion of answers sourced from document retrieval fell from about 67% before that date to about 38% afterwards, although the number of observations afterwards remains small at 24 questions. The system did not stop answering because questions were diverted to the fallback layer, which still draws on the knowledge blocks, but the content of the web pages no longer served as a reference.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

C. Answer Accuracy

The assessment results for the 32 test items are presented in Table 4. Of the 32 items, 26 met all three dimensions simultaneously.

Table 4. Validation results by assessment dimension, majority rule across three raters

Dimension	Items meeting	Percentage
Relevance	29 of 32	90.6%
Correctness	26 of 32	81.3%
Completeness	29 of 32	90.6%
Meeting all three dimensions	26 of 32	81.3%

Inter-rater agreement lies within an adequate range. Across 96 assessment cells, agreement between the first and second raters reached 84.4%, between the first and third raters 85.4%, and between the second and third raters 82.3%. This range indicates that the three raters applied the criteria relatively consistently, so the aggregate results are usable. The distribution of passing items by category is presented in Table 5.

Table 5. Items meeting all three dimensions by category

Category	Items passing	Percentage
Basic concepts and policy	4 of 4	100%
Credit recognition including community service	6 of 6	100%
Financial accountability	7 of 7	100%
Budget provisions and restrictions	2 of 2	100%
Eligibility and proposal submission	4 of 6	66.7%
Official contacts and references	1 of 2	50%
Schedule, reporting, and assessment	2 of 5	40%

Four categories passed all of their items, including financial accountability, which contains seven items on expenditure categories, tax deductions, and the format of supporting evidence. The schedule and reporting category passed only two of five items. Details of the six items that did not pass are presented in Table 6.

Table 6. Items not meeting the assessment, with rater notes

Item	Category	Dimensions not met	Core rater note
7	Eligibility and proposal submission	Relevance, correctness, completeness	The question asked about submission steps; the answer presented the timeline of the previous period
8	Eligibility and proposal submission	Correctness	The period stated had not been adjusted to the current period
17	Schedule, reporting, and assessment	Relevance, correctness, completeness	The timeline given is not the timeline of the current period
18	Schedule, reporting, and assessment	Relevance, correctness, completeness	The duration in months is correct, but the period stated has not been updated
19	Schedule, reporting, and assessment	Correctness	The timeline given is not the timeline of the current period
31	Official contacts and references	Correctness	The period referred to is not yet appropriate

The six failures share a single root cause. All of them concern time-bound information, namely timelines and references to the programme period. At the time of validation, the knowledge base still contained the provisions of the February to July 2026 period, whereas the period in force was August 2026 to January 2027. None of the failures resulted from misunderstanding the intent of the question, from reasoning errors, or from information invented by the model.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Figure 4 presents an example of this error. The answer displayed states an implementation period of 23 February to 3 July 2026, whereas the guideline serving as the reference covers the period from August 2026 to January 2027. The screenshot was taken on 14 October 2026, indicating that this timeline information had not been updated at the time of writing.

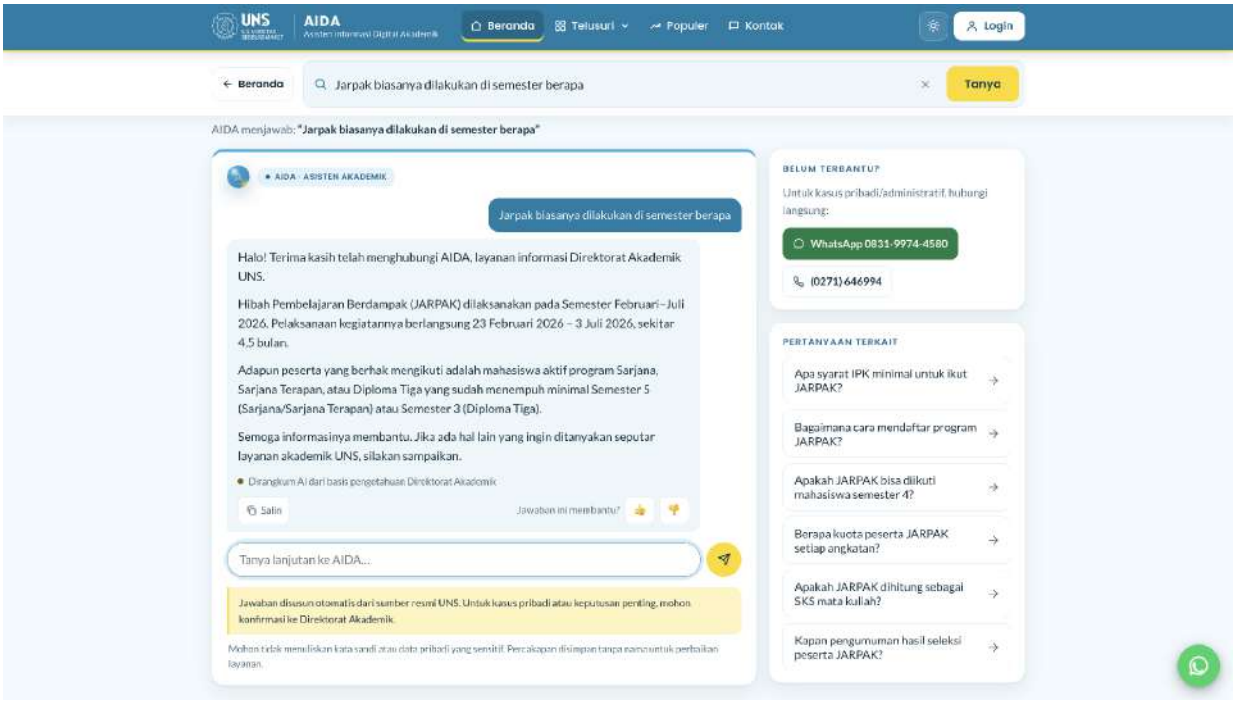


Figure 4. Example of an answer that is relevant and complete yet states a programme period no longer in force

Three of the six items still met the relevance and completeness dimensions and failed only on correctness. The answers to those items matched the questions and contained complete information, even though their content was no longer valid.

D. System Usability

The trial involved 76 respondents, all students, with the profile presented in Table 7. There were no missing data across the sixteen instrument items. Six of the 76 respondents completed the questionnaire on or after 8 October 2026, that is, after some source documents had passed their validity period as described in the system operational data above.

Table 7. Respondent profile

Characteristic	Category	Count
Study programme	Agribusiness	46
	Indonesian Literature	18
	Industrial Engineering	7
	English Language Education	2
	Others (Civic Education, Educational Technology, Agricultural Extension and Communication)	3
Device	Mobile phone	69
	Laptop or desktop	6
	Tablet	1
Frequency of use	First time	63
	Two to three times	13

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

The mean CUQ score was 67.72 with a standard deviation of 11.93, a median of 66.4, and a range from 46.9 to 96.9. A Cronbach's alpha coefficient of 0.865 indicates good internal consistency, so the aggregate score can be relied upon.

Values across items show considerable variation. Table 8 presents the mean adjusted score for each item on a scale from 0 to 4, ordered from the lowest.

Table 8. Mean adjusted score per CUQ item, range 0 to 4

Statement	Polarity	Mean
AIDA feels too rigid, like a robot	Negative	1.74
AIDA handles errors and mistakes well	Positive	2.39
AIDA has an engaging communication style that feels like interacting with a human	Positive	2.45
AIDA struggles to handle errors and mistakes	Negative	2.53
AIDA does not explain its purpose and usefulness	Negative	2.54
AIDA often fails to understand the questions or input I provide	Negative	2.63
AIDA's answers are often not aligned with the questions I ask	Negative	2.75
AIDA provided a pleasant first-time experience	Positive	2.79
AIDA understands the questions or input I provide well	Positive	2.87
AIDA gives unfriendly responses	Negative	2.88
AIDA's answers are useful, relevant, and informative	Positive	2.89
AIDA is easy to navigate	Positive	2.92
I easily became confused when using AIDA	Negative	2.92
AIDA explains its purpose and usefulness well	Positive	2.93
AIDA feels very complex	Negative	2.99
AIDA is very easy to use	Positive	3.12

Items relating to ease of use, clarity of purpose, and ease of navigation rank highest. Conversely, items relating to the naturalness of the communication style and the handling of misunderstandings rank lowest, with the item on rigidity of expression standing well apart from the rest. Respondents judged the system easy to use but considered its manner of expression rigid and its handling inadequate when misunderstandings occurred.

All 76 respondents answered the open question on suggestions for improvement. Inductive thematic grouping produced four clusters. The largest concern concerns the interface and display, which respondents proposed be made more attractive and more legible on mobile screens. The second concerns response speed. The third concerns information coverage, namely requests that the knowledge base be expanded because some questions remain unanswered. The fourth concerns the style in which answers are presented, with requests that answers be more concise, more direct, and accompanied by examples.

DISCUSSION

A. Characteristics of Answer Errors

The validation results show that answer errors are concentrated in a single type of information. The six items that did not meet the assessment concern timelines and references to the programme period, whereas categories whose content does not change between periods achieved full results. The financial accountability category, which contains the most detailed rules, passed all seven of its items, while the schedule and reporting category passed only two of five.

This pattern differs from what is commonly reported. Several reviews identify limitations in natural language understanding as the principal cause of failure in educational chatbots [5], [14]. In this study, all questions were understood correctly and the documents retrieved were appropriate, but the content of those documents still carried the provisions of the previous period.

The system logs provide supporting evidence. A total of 272 web pages serving as answer sources passed their validity period on 8 October 2026, so that the document chunks actively retrieved fell from 5,243 to 47, and the proportion of answers sourced from document retrieval fell from about 67% to 38%. Two different data sources, namely human assessment and system logs, point to the same problem, namely unscheduled content updating.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

B. Errors in the Correctness Dimension

Three of the six failing items still met the relevance and completeness dimensions. The answers to those items matched the questions and contained complete information, but the timelines stated were no longer valid.

This condition is consistent with studies of language model hallucination, which state that fluency and completeness of an answer do not guarantee its correctness [20], [21]. The difference lies in the source of the error. The model did not produce information beyond the documents but conveyed the content of documents that had expired, so the error arose even though the system operated as designed. This document-grounded behaviour is also visible in Figure 2, where the system states that the expansion of the programme acronym is not recorded in the source documents instead of composing an answer of its own.

The consequences warrant attention in administrative services. Users receive no indication that the information they obtain is no longer valid, so the error may only become apparent after a deadline has passed. Chatbot design in policy domains therefore needs to display the period of validity and the date on which the information was last updated.

C. Performance of Keyword-Based Document Retrieval

The system evaluated uses neither an embedding model nor a vector database, yet achieved a relevance score of 90.6%, and no failure was caused by a mismatch between the question and the documents retrieved.

This result is consistent with cross-domain testing showing that BM25 remains competitive with dense retrieval models on domains for which they have not been trained [19]. The fit can be explained by the characteristics of the domain: policy documents use standardised terminology that enquirers also use, such as recognition, outputs, and accountability.

The result must be read alongside its limitations. Wulandari and Yamasari [12] report that a similar approach declines in performance on paraphrased questions, whereas the test items in this study were constructed with terminology close to that of the source documents. Even so, the result indicates that information services based on policy documents can be built without vector retrieval infrastructure.

D. Layered Architecture and Operational Efficiency

The five-layer design places cost-free paths before paths that invoke the language model. According to the system logs, 93 of 1,268 questions, or about 7%, were resolved at the greeting and question-and-answer layers with a response time of 0.32 seconds, compared with about four seconds on paths invoking the language model.

That proportion is not yet large, but the design offers two benefits, namely savings on model invocation costs and faster answers to the most frequently asked questions. Cost efficiency is rarely reported in studies of chatbot implementation in higher education, even though service sustainability depends on it.

E. The Relationship between Accuracy and Usability

An accuracy value of 81.3% is accompanied by a usability score of 67.72, which falls in the moderate range. The two figures measure different aspects, and the difference can be explained through the ranking across items.

Items relating to ease of use, clarity of purpose, and ease of navigation obtained the highest values, whereas items on the naturalness of expression and the handling of misunderstandings obtained the lowest, with the item on rigidity of expression scoring 1.74 out of 4. The aspect respondents judged lacking is therefore interaction quality rather than answer correctness.

This result supports the view that chatbot evaluation is inadequate when it relies on a single measure [7], [24]. Interpretation of the usability score must also take into account that 63 of the 76 respondents were using the system for the first time.

F. Implications for Service Governance

Because the source of the error lies in content updating, the improvement required is a matter of governance. Four steps are proposed on the basis of the findings.

First, knowledge base updating should be scheduled according to the programme calendar, so that every change of period triggers a review of all entries containing dates and deadlines. Second, every entry should be marked with its period of validity and the date it was last updated. Third, the validity period of automatically retrieved documents should be monitored regularly to prevent an unnoticed reduction in answer sources. Fourth, responsibility for content should be formally assigned within the service unit and separated from responsibility for the technical system.

These four steps indicate that managing a service chatbot does not end at the implementation stage but requires regular content maintenance.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

G. Limitations and Directions for Further Research

This study has three main limitations.

First, the usability respondents were all students with an uneven distribution across study programmes, with 46 of 76 respondents coming from a single programme and 63 using the system for the first time. The usability score therefore reflects the first impressions of student users and does not yet encompass the assessments of supervising lecturers or administrative staff who also use the service.

Second, the test items were constructed to follow the structure and terminology of the guideline document, so the system's ability to handle questions expressed in students' everyday language, including paraphrases and non-standard abbreviations, has not been measured.

Third, the study included neither a comparison group nor before-and-after measurement of implementation, so the impact of the chatbot on the workload of the service unit cannot yet be concluded.

On the basis of these three limitations, further research is advised to involve lecturers and administrative staff as respondents, to construct test items from questions actually asked by students, and to measure question volume and service waiting time before and after system implementation.

CONCLUSION

This study evaluated the AIDA chatbot in the academic learning grant domain through two cycles, namely programme administrators' assessment of answer accuracy and usability measurement by student users.

In response to the first research question, the system in operation uses keyword-based document retrieval with BM25 combined with answer generation by a language model, organised into five answering layers that prioritise the least costly path. The system uses neither an embedding model, nor a vector database, nor intent classification.

In response to the second research question, the 32 test items yielded 90.6% relevance, 81.3% correctness, 90.6% completeness, and 81.3% of items meeting all three dimensions simultaneously, with inter-rater agreement between 82% and 85%. All failures stem from timeline information not adjusted to the current period, and none originated from misunderstanding the questions.

In response to the third research question, the usability score reached 67.72 with a reliability of 0.865 across 76 respondents. Ease of use was rated highest, while naturalness of expression and handling of misunderstandings were rated lowest.

The principal conclusion of this study is that the greatest constraint on administrative service chatbots lies in the governance of content freshness rather than in the capacity of the technology to understand language. Two independent sources of evidence, human assessment and system logs, point to the same root cause. Keyword-based retrieval proved adequate for a closed policy domain with standardised terminology, so the resources of the service unit are more usefully directed towards content curation and updating than towards upgrading retrieval technology.

In practical terms, the study recommends tying knowledge base updates to the programme calendar, marking each entry with its period of validity, monitoring document validity periods regularly, and assigning responsibility for content separately from responsibility for technical matters.

ACKNOWLEDGEMENTS

This research was funded through the Policy Governance Research scheme (PTK-UNS) of Universitas Sebelas Maret for the 2026 budget year, registration number 460/UN27.22/PT.01.03/2026. The authors thank the Directorate of Academic Affairs for access to service data, the AIDA development team for preparing the technical specification report, and the raters and students who participated in the trial.

REFERENCES

1. Republic of Indonesia, Regulation of the Minister of Higher Education, Science, and Technology Number 39 of 2025 on Higher Education Quality Assurance. Jakarta, 2025.
2. Republic of Indonesia, Regulation of the Minister of Education, Culture, Research, and Technology Number 53 of 2023 on Higher Education Quality Assurance. Jakarta, 2023.
3. Republic of Indonesia, Decree of the Minister of Education and Culture Number 74/P/2021 on the Recognition of Semester Credit Units for Kampus Merdeka Programme Learning. Jakarta, 2021.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

4. M. A. Kuhail, N. Alturki, S. Alramlawi, and K. Alhejori, "Interacting with educational chatbots: A systematic review," Educ. Inf. Technol. (Dordr)., vol. 28, no. 1, pp. 973–1018, Jan. 2023, doi: 10.1007/s10639-022-11177-3.
5. L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," International Journal of Educational Technology in Higher Education, vol. 20, no. 1, p. 56, Oct. 2023, doi: 10.1186/s41239-023-00426-1.
6. D. Ramandanis and S. Xinogalos, "Investigating the Support Provided by Chatbots to Educational Institutions and Their Students: A Systematic Literature Review," Multimodal Technologies and Interaction, vol. 7, no. 11, p. 103, Nov. 2023, doi: 10.3390/mti7110103.
7. K. Peyton, S. Unnikrishnan, and B. Mulligan, "A review of university chatbots for student support: FAQs and beyond," Discover Education, vol. 4, no. 1, p. 21, Jan. 2025, doi: 10.1007/s44217-025-00397-7.
8. G.-J. Hwang and C.-Y. Chang, "A review of opportunities and challenges of chatbots in education," Interactive Learning Environments, vol. 31, no. 7, pp. 4099–4112, Oct. 2023, doi: 10.1080/10494820.2021.1952615.
9. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2005.11401
10. Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," Mar. 2024.
11. L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Trans. Inf. Syst., vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.
12. K. Wulandari, "Analisis kinerja Retrieval Augmented Generation (RAG) klasik dalam chatbot akademik berbasis multimodal [Performance analysis of classical Retrieval Augmented Generation (RAG) in multimodal-based academic chatbots]," Journal of Informatics and Computer Science (JINACS), vol. 7, no. 3, pp. 628–632, 2026.
13. S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler, "Are We There Yet? - A Systematic Literature Review on Chatbots in Education," Front. Artif. Intell., vol. 4, Jul. 2021, doi: 10.3389/frai.2021.654924.
14. C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," Computers and Education: Artificial Intelligence, vol. 2, p. 100033, 2021, doi: 10.1016/j.caeai.2021.100033.
15. D. S. M. Pereira, F. Falcão, L. Costa, B. S. Lunn, J. M. Pêgo, and P. Costa, "Here's to the future: Conversational agents in higher education- a scoping review," Int. J. Educ. Res., vol. 122, p. 102233, 2023, doi: 10.1016/j.ijer.2023.102233.
16. L. Nicolescu and M. T. Tudorache, "Human-Computer Interaction in Customer Service: The Experience with AI Chatbots—A Systematic Literature Review," Electronics (Basel)., vol. 11, no. 10, p. 1579, May 2022, doi: 10.3390/electronics11101579.
17. S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends® in Information Retrieval, vol. 4, no. 1–2, pp. 1–174, Sep. 2009, doi: 10.1561/1500000019.
18. V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.
19. N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models." [Online]. Available: http://www.ukp.tu-darmstadt.de
20. L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Trans. Inf. Syst., vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.
21. Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Comput. Surv., vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.
22. J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 16, pp. 17754–17762, Mar. 2024, doi: 10.1609/aaai.v38i16.29728.
23. S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, and M. Mctear, "Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?," in Proceedings of the 31st European Conference on Cognitive Ergonomics, New York, NY, USA: ACM, Sep. 2019, pp. 207–214. doi: 10.1145/3335082.3335094.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 10 October 2026
DOI: 10.47191/ijcsrr/V9-i10-26, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

24. A. Rapp, L. Curti, and A. Boldi, "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots," Int. J. Hum. Comput. Stud., vol. 151, p. 102630, Jul. 2021, doi: 10.1016/j.ijhcs.2021.102630.
25. A. Følstad and P. B. Brandtzaeg, "Users' experiences with chatbots: findings from a questionnaire study," Qual. User Exp., vol. 5, no. 1, p. 3, Dec. 2020, doi: 10.1007/s41233-020-00033-2.
26. K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," Journal of Management Information Systems, vol. 24, no. 3, pp. 45–77, Dec. 2007, doi: 10.2753/MIS0742-1222240302.
27. UNESCO, "Guidance for generative AI in education and research," Paris, 2023.

Cite this Article: Widyastuti, U., Utomo, S.B., Damayanti, R.W., Anshori, H.W., Prastowo, S., Priyanto, A.D., Setiadi, H., Wahyudi, D., Safiudin, A., Pambudi, G.W., Sulistyawati, N. (2026). Evaluating the Accuracy and Usability of an Academic Digital Information Assistant Chatbot Based on Retrieval-Augmented Generation for Academic Learning Grant Services. International Journal of Current Science Research and Review, 9(10), pp. 5471-5484. DOI: https://doi.org/10.47191/ijcsrr/V9-i10-26