Preprints.org

Article

Extension of Interval-Valued Hesitant Fermatean Fuzzy TOPSIS for Evaluating and Benchmarking of Generative AI Chatbots

Galina Ilieva [*]

Article

Extension of Interval-Valued Hesitant Fermatean Fuzzy TOPSIS for Evaluating and Benchmarking of Generative AI Chatbots

Galina Ilieva

Department of Management and Quantitative Methods in Economics, University of Plovdiv Paisii Hilendarski, 4000 Plovdiv, Bulgaria; galili@uni-plovdiv.bg

Abstract: To aid in the selection of generative Artificial Intelligence (GAI) chatbots, this paper introduces a fuzzy multi-attribute decision-making framework based on their key features and performance. The proposed framework includes a new modification of the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), adapted for an interval-valued hesitant Fermatean fuzzy (IVHFF) environment. This TOPSIS extension addresses the limitations of classical TOPSIS in handling complex and uncertain data capturing detailed membership degrees and representing hesitation more precisely. The framework is applicable for both static and dynamic evaluations of GAI chatbots in crisp or fuzzy assessments. Results from a practical example demonstrate the effectiveness of the proposed approach for comparing and ranking GAI chatbots. Finally, recommendations are provided for selecting and implementing these conversational agents in various applications.

Keywords: MCDM; interval-valued hesitant Fermatean fuzzy (IVHFF) sets; distance measure; Technique for Order Preference by Similarity to Ideal Solution (TOPSIS); IVHFFs TOPSIS; generative AI chatbots; evaluation framework; benchmarking criteria

MSC: 68T07; 03E72; 90B50

1. Introduction

Generative Artificial Intelligence (GAI) chatbots, also known as conversational agents, are becoming an increasingly prevalent type of chatbot worldwide for several reasons. Advancements in AI and natural language processing have significantly enhanced their capabilities [1]. These intelligent chatbots leverage Large Language Models (LLMs) to generate human-like responses in natural language, enabling more dynamic and contextually appropriate interactions compared to their traditional rule-based predecessors [2,3].

Additionally, the rise of digital communication platforms and the growing demand for instant customer service have driven businesses to adopt GAI chatbots. These systems provide a cost-effective way to deliver customer support, answer queries and even facilitate transactions.

The COVID-19 pandemic further accelerated the adoption of remote work and virtual communication, increasing the demand for AI chatbots [4]. They have proven instrumental in managing the surge of online interactions, such as handling customer inquiries, scheduling appointments and disseminating information.

The flexibility and scalability of GAI chatbots make them suitable for diverse industries, including construction [5], healthcare [6], finance [7], and e-commerce [8]. These chatbots can be tailored to specific use cases and integrated into existing systems, enhancing productivity and user experiences [9].

Market research predictions confirm a growing adoption of AI chatbots in the coming years. According to a Statista forecast [10], the global chatbot market is projected to reach approximately $1.25 billion by 2025—an increase of nearly fivefold from $190.8 million in 2016. Meanwhile, Gartner

estimates [11] that over 80% of enterprises will leverage GAI APIs or applications by 2026, highlighting the rapid and widespread adoption of advanced AI technologies for enhancing business efficiency, innovation and customer experiences. However, this growth also presents challenges, particularly in areas such as data privacy, ethics and addressing skill gaps.

Despite their growing popularity, GAI chatbots face several challenges to widespread adoption. Key obstacles include:

- Lack of trust – Users may hesitate to fully trust AI-powered chatbots, particularly when dealing with sensitive information or complex interactions. Building trust in the accuracy, security and reliability of these systems is critical for their broader acceptance.

- Limited understanding and awareness – Many users are unfamiliar with the capabilities and benefits of GAI chatbots. This lack of knowledge or understanding about how they function and what they offer may hinder adoption.

- User experience and satisfaction – Poorly designed chatbots can lead to unsatisfactory user experiences. Frustrating interactions or failure to resolve queries effectively may discourage continued use.

- Cost and ROI – Developing and maintaining GAI chatbots can be expensive, particularly for small and medium-sized enterprises. Organizations must carefully assess the return on investment (ROI) and weigh costs against potential benefits.

- Ethical and bias concerns – GAI chatbots are only as reliable and fair as the data they are trained on, which can sometimes perpetuate biases or unfair practices. Ensuring chatbots are ethical, unbiased and inclusive is important for their acceptance and broader implementation.

Overcoming these barriers will require advancements in technology, increased transparency, education and a focus on user-centric design. To address the first three challenges, multi-criteria decision-making (MCDM) methods can be employed. These techniques enable organizations to compare a finite set of decision alternatives across various criteria, helping them select the most feasible option. MCDM methods have been successfully applied in several GAI-related fields, such as technology selection [12] and cloud system prioritization [13].

While conventional MCDM methods are reliable, they often struggle to address the complexities associated with imprecise and ambiguous evaluations. In contrast, fuzzy-based methods are specifically designed to manage such uncertainties, making them more effective in identifying the most suitable alternatives.

Various MCDM techniques have been enhanced through the integration of fuzzy sets and their advanced extensions [14]. By incorporating fuzzy assessments, these methods provide a more accurate representation of real-world conditions, thereby improving the reliability of rankings in scenarios characterized by subjectivity and evaluation uncertainties.

The key advantage of fuzzy multi-criteria algorithms lies in their ability to produce more realistic and dependable rankings, enhancing the overall decision-making process.

Key contributions of this paper include:

1. Analysis and categorization of existing multi-criteria approaches for AI chatbot selection, classified by the techniques used and the types of estimates employed (numeric, interval, linguistic values, as well as crisp and fuzzy numbers). These approaches are then grouped into three main categories based on complexity (number of multi-criteria techniques), flexibility (type of fuzziness) and iterativeness (single or repeated data processing).

2. Development of a theoretical framework for ranking GAI chatbots using both single and hybrid methods with crisp and fuzzy estimates. Single methods rely on one weight determination or ranking approach, while hybrid methods combine multiple approaches. The framework also incorporates complementary techniques such as fuzzy interval arithmetic operations, robustness analysis and sensitivity analysis to enhance decision-making and benchmark rankings. Additionally, it proposes a newly developed 3D distance metric to improve the efficiency of the hesitant Fermatean uzzy group TOPSIS method, enabling more effective multi-criteria comparisons of chatbot features.

3. Creation of static and dynamic rankings of an AI chatbot dataset via single or repeated multi-criteria decision analysis. In static rankings, experts' opinions serve as inputs for the decision matrices, whereas dynamic rankings measure user attitudes—potentially informed by behavior

or survey data. Comparative analyses with other multi-criteria baselines underscore both the effectiveness and reliability of the proposed methods.

The paper begins with a literature review in Section 2, discussing the motivation behind exploring fuzzy ranking for GAI chatbots. Next, Section 3 details the proposed theoretical decision-making framework for GAI chatbot selection, emphasizing the role of Interval-Valued Hesitant Fermatean Fuzzy Numbers (IVHFFNs) and a modified TOPSIS method tailored for the IVHFF environment. Practical example and result analysis are provided in Section 4, showcasing the application of the framework. The final section concludes the research by summarizing the key findings, offering insights and proposing directions for future studies.

2. Related Work

2.1. Literature Review on MCDM Methods for GAI Chatbot Evaluation

GAI chatbots, despite being a relatively recent development, have garnered significant attention in both academic research and practical applications. Approaches to their study vary widely: some researchers focus on technical aspects, offering descriptive or general analyses that often emphasize feature comparisons while omitting advanced computational methods. Conversely, other studies adopt modern model-driven techniques, such as machine learning, optimization, and MCDM methods.

MCDM methods present distinct advantages in the evaluation and selection of AI chatbots. One key benefit is that they do not rely on extensive datasets or computationally intensive procedures, making them accessible and efficient. These methods simplify the decision-making process by facilitating comprehensive evaluations across multiple criteria, ensuring objectivity through a systematic analysis of both the criteria and stakeholder preferences.

Additionally, MCDM approaches are well-suited for diverse decision-making scenarios and can manage the complexities inherent in chatbot evaluations. By incorporating stakeholder preferences, these methods enable informed decision-making and improve the likelihood of selecting the most appropriate chatbot for a given context.

Drawing on data from previous studies, interviews, questionnaires, and surveys, Chakrabortty et al. [15] constructed a comparison matrix with eight alternatives and nine criteria: Empathy, Engagement, Tangibility, Assurance, Reliability, Satisfaction, Responsiveness, Speed, and Security. These criteria were derived from established service quality models alongside AI- and chatbot-specific considerations. A survey was conducted to gather expert opinions and the Single-Valued Neutrosophic (SVN) Analytic Hierarchy Process (AHP) was employed to determine their relative weights. The Combined Compromise Solution (CoCoSo) method was then used within the SVN environment to rank the options, ultimately identifying the optimal chatbot.

Santa Barletta et al. [16] proposed a novel clinical chatbot selection model using the AHP technique, assessing chatbot assistants based on the "Quality in Use" concept from the ISO/IEC 25010 standard. Two healthcare-oriented chatbots were evaluated against five criteria groups: Effectiveness, Efficacy, Satisfaction, Freedom from risk, and Context coverage across three dimensions—providing information, prescriptions and process management.

Singh et al. [17] identified twelve acceptance factors for Conversational Digital Assistants (CDAs) through a literature review and expert input. These factors were analyzed for their cause-and-effect relationships using the grey-Decision-Making Trial and Evaluation Laboratory (DEMATEL) method. The study highlighted key cause factors, including Humanness, Social influence, Social presence, Social capability and Ease of use, which significantly impact CDA adoption and provide insights for managerial and policy decisions in online shopping contexts.

Pandey et al. [18] addressed concerns about the impact of GAI tools, particularly ChatGPT, by examining twelve challenges related to its adoption. These challenges were analyzed using the intuitionistic fuzzy DEMATEL approach, which proved more effective than classical and fuzzy DEMATEL methods in terms of Mean Absolute Error (MAE). By categorizing challenges into cause-and-effect relationships, the study provides valuable guidance for experts and project managers in identifying areas for improvement.

Pathak and Bansal [19] mapped twenty factors to the Technology-Organization-Environment-Individual (T-O-E-I) framework, derived from the Technology-Organization-Environment (T-O-E) and Human-Organization-Technology fit (H-O-T fit) frameworks. After ranking these factors, the

global ranking was computed using the Rough Stepwise Weight Assessment Ratio Analysis Method (R-SWARA). The top seven factors included Perceived benefits of AI, AI system capabilities, Organizational data ecosystem, Perceived compatibility of AI systems, Ease of use, IT infrastructure and Top management support. Sensitivity analysis confirmed the robustness of these rankings.

Wiangkham and Vongvit [20] applied both MCDM and artificial neural network (ANN) methods to prioritize factors influencing ChatGPT adoption in higher education. Fourteen criteria were grouped into Usage, Agent, Technical and Trust-related categories. Using a Likert-scale questionnaire, criteria importance was assessed, and Weighted Sum Model (WSM) and ANN methods were applied, alongside SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). The study systematically prioritized factors affecting ChatGPT adoption.

Ojo et al. [21] employed a fuzzy TOPSIS-based method to evaluate six AI alternatives for mental health treatment planning: rule-based systems, logistic regression, neural networks, evolutionary algorithms, hybrid models and benchmark algorithms. The evaluation considered criteria such as Privacy protection, Treatment effectiveness, Explainability, Healthcare costs, Regulatory compliance, and Ethical implications. Rule-based systems and benchmark algorithms emerged as the preferred approaches.

The key characteristics of the methodologies used to investigate factors influencing the selection and ranking of conversational digital assistants are summarized in Table 1.

After analysis of previous studies on GAI chatbot selection, we categorize them based on their distinctive features. According to the specificity of input data, the utilized models can be divided into two groups: crisp and fuzzy estimates. The crisp group, exemplified by Santa Barleta et al. [16] and Wiangkham and Vongvit [20], is designed for arithmetic calculations and distance metrics using precise input values. In contrast, the fuzzy group includes multi-criteria methods operating in various fuzzy environments, as demonstrated by Chakrabortty et al. [15], Singh et al. [17], Pandey et al. [18], Pathak and Bansal [19] and Ojo et al. [21].

Table 1. Comparison of existing studies on chatbots evaluation and ranking.

Reference	Methodology	Application area	Alternatives	Criteria (number)	Ranking validation
Chakrabortty et al. 2023 [15]	SVN AHP-CoCoCo	Telecommunication business	Eight chatbots for customer service	Security, Speed, Responsiveness, Satisfaction, Reliability, Assurance, Tangibility, Engagement, Emphaty (9)	SVN MABAC, Pythagorean fuzzy CoCoCo, Interval valued Neutrosophic TOPSIS
Santa Barleta et al. 2023 [16]	AHP	Medical care	Two clinical chatbots	Effectiveness, Efficacy, Satisfaction, Freedom from risk, Context coverage in three fuctional dimensions (5 criteria groups)	Superdecision software
Singh et al. 2023 [17]	Grey-DEMATEL	Online retail	Only criteria weights	Social influence, Enjoyment, Performance, Ease of use, Usefulness, Social presence, Anxiety, Trust, Rapport, Privacy risk, Social isolation, Sense of control, Compatibility (12)	Sensitivity analysis in three scenarious

Pandey at al. 2024 [18]	Intuitionistic fuzzy DEMATEL	GAI chatbots challanges	Only criteria weights	Hallucination*, Bias, Language learning, Real-world harm*, Proprietary LLMs*, AI problems*, Disruption, Jobs at risk, Educational system problems, Training data amount*, Unknown threats, Ethical and legal implications (12)	MAE of issues using Classical DEMATEL and fuzzy DEMATEL
Pathak and Bansal 2024 [19]	Rough SWARA	Insurance	Only criteria weights	Technology (7), Organization (6), Environment (3), Individual (4) criteria groups	Sensitivity analysis
Wiangkha m and Vongvit 2024 [20]	WSM, ANN with SHAP and LIME	Higher education	Only criteria weights	Usage (4), Agent (3), Technical (4), Trust (3) related criteria groups	WMAPE for ANN models
Ojo et al. 2024 [21]	Fuzzy triangular TOPSIS	Medical care	Six AI alternatives	Privacy protection, Treatment effectiveness, Explainability, Costs, Regulatory compliance, Ethical implications (6)	Comparative analysis

Remark: The symbol '*' denotes the most impacful factors for ChatGPT adoption.

In terms of complexity, existing models can be classified into single and hybrid multi-criteria techniques. Single methods, used by Singh et al. [17], Pandey et al. [18], Pathak and Bansal [19], Wiangkham and Vongvit [20] and Ojo et al. [21], apply only one MCDM method. Hybrid approaches, employed by Chakrabortty et al. [15] and Santa Barleta et al. [16], combine two methods: one for determining relative criteria weights and another for ranking alternatives.

The literature review reveals the absence of a universal approach for addressing the GAI chatbot selection problem. While previous studies offer valuable insights into comparing conversational chatbots, they exhibit several shortcomings:

- Lack of holistic multi-criteria solutions – Many proposed solutions focus on specific aspects, such as determining the relative importance of features within the criteria system [17–20] or generating chatbot rankings using a single multi-criteria method [15,21].

- Limited handling of inaccurate attribute estimates – Few studies, such as those by Chakrabortty et al. [15], Pandey et al. [18] and Ojo et al. [21], effectively address imprecise attribute estimates. Since AI chatbot evaluations often depend on subjective factors, assessments should involve expert groups utilizing classic fuzzy numbers or their advanced variants.

- Non-iterative fuzzy solutions – Existing fuzzy methodologies typically implement only one or two MCDM methods in a single, non-iterative procedure.

Evaluation should adopt a holistic process that considers various factors, including technological, economic and organizational parameters, which are often expressed through imprecise, unclear and uncertain estimates. To address these drawbacks, we propose a new fuzzy methodology for GAI chatbot selection.

Selecting a specific chatbot assistant aligned with organizational strategies or individual preferences is a complex process influenced by numerous factors. At the organizational level, the preferred intelligent chatbot depends on considerations such as data security requirements, regulatory compliance, subscription costs and seamless integration with existing systems. At the individual level, preferences may be shaped by use cases, ease of use, domain-specific capabilities or community recommendations and reviews. The optimal solution is the GAI chatbot that best meets the requirements of the organization or the preferences of the individual user.

2.2. Chatbot Evaluation Criteria

Despite the availability of practical tools and platforms for chatbot benchmarking and user testing—such as those offered by Hugging Face, Chatbot Arena (formerly LMSYS) [22] and Artificial Analysis [23]—these solutions often lack the flexibility needed to accommodate specific study goals

and use cases. Evaluating GAI chatbots requires a systematic approach that integrates diverse attributes to address their multifaceted roles and applications.

In this subsection, we review the criteria proposed in prior studies to identify relevant attributes for developing a multi-attribute evaluation system specifically tailored to GAI chatbots.

The literature review reveals that previous studies on developing multi-criteria systems for evaluating GAI chatbots have primarily adopted a combined approach, integrating multiple criteria, indices and metrics derived from various theoretical models and software quality standards.

For example, Chakrabortty et al. proposed a system based on nine criteria: Security, Speed, Responsiveness, Satisfaction, Reliability, Assurance, Tangibility, Engagement and Empathy. These criteria were drawn from SERVQUAL [24] (Responsiveness, Reliability, Assurance, Tangibility, Empathy), ISO/IEC 25010 [25] (Security, Speed), the Technology Acceptance Model (TAM) [26] (Engagement), and Customer Experience theory [27] (Satisfaction) [15].

Santa Barleta et al. focused on five criteria groups: Effectiveness, Efficacy, Satisfaction, Freedom from Risk, and Context Coverage. These were derived from ISO/IEC 25010 [25] and applied across three functional dimensions [16].

Singh et al. [17] developed a system incorporating 12 criteria, including Social Influence, Enjoyment, Performance, Ease of Use, Usefulness, Trust, and Privacy Risk, based on TAM [26] and UTAUT [28].

Pandey et al. [18] introduced 12 evaluation criteria emphasizing ChatGPT-related issues such as Hallucination, Bias, Proprietary LLMs, Ethical Implications, and broader AI-related problems.

Pathak and Bansal [19] utilized the T-O-E-I framework [29,30], organizing criteria into four groups: Technology (7 criteria), Organization (6), Environment (3), and Individual (4).

Wiangkham and Vongvit [20] adopted a system comprising Usage (4 criteria), Agent (3), Technical (4), and Trust-Related (3) categories, primarily based on TAM and UTAUT.

Ojo et al. [21] focused on six criteria: Privacy Protection, Treatment Effectiveness, Explainability, Costs, Regulatory Compliance, and Ethical Implications. These criteria, designed for evaluating medical chatbots, stem from healthcare technology frameworks, ISO standards, AI ethics and health economics models. Their system ensures that medical chatbots are safe, effective, transparent, and legally compliant while addressing critical aspects such as patient data security, cost-effectiveness, and ethical concerns.

The compared evaluation systems for GAI chatbot selection emphasize a multi-criteria approach, integrating elements from SERVQUAL framework, ISO standards, TAM, UTAUT and AI ethics models. These assessment indices are designed to address specific contexts, including functionality, user experience and ethical considerations, enabling effective comparisons by evaluating both technical capabilities and societal impacts.

However, existing evaluation systems have limitations, including their domain-specific focus, insufficient attention to rapidly evolving GAI challenges and reliance on subjective criteria weighting. To address these gaps, we have developed a GAI chatbot evaluation system, ensuring a comprehensive and holistic evaluation approach.

The proposed system includes four key criteria – Conversational ability, User experience, Integration capability and Price:

- Conversational ability evaluates the chatbot's capacity to understand and generate natural language responses, ensuring context-aware, coherent, and human-like interactions.

- User experience measures ease of use, intuitiveness, and satisfaction, focusing on design, accessibility, and the chatbot's ability to meet user needs effectively.

- Integration capability assesses how seamlessly the chatbot integrates with existing tools, platforms, or workflows, enhancing usability and productivity.

- Price considers the affordability of the chatbot, evaluating its cost relative to its features, functionality and overall value.

Our evaluation system aligns with the TAM [26] and UTAUT [28] models. Conversational ability corresponds to Perceived ease of use in TAM and Performance expectancy in UTAUT, reflecting user expectations for accurate, natural communication. User experience relates to Perceived usefulness and Effort expectancy, where intuitive and enjoyable interactions drive adoption. Integration capability aligns with Facilitating conditions in UTAUT and external variables in TAM, as compatibility with existing systems enhances utility. The Price criterion captures the cost-value

relationship, where users weigh the chatbot's cost against its utility and benefits. Framing these criteria through TAM and UTAUT provides organizations with a comprehensive means of evaluating chatbots from both technical and user-acceptance perspectives.

New chatbot evaluation system employs a combining approach that integrates metrics, indices and factors from diverse theoretical models. This multidimensional system provides a compex assessment that addresses functional, experiential, technical and economic dimensions. By tailoring it to the specific requirements of corporate and individual users, our approach ensures an effective evaluation of chatbots across varied use cases and priorities.

2.3. State-of-the-Art of the Most Widely Used GAI Chatbots

In this subsection, we present a comparative overview of the most popular GAI chatbots recognized by the global AI community for their transformative role in enhancing human-machine interaction: ChatGPT, Copilot, Gemini, Claude, and Perplexity AI.

OpenAI ChatGPT (https://chatgpt.com) is a state-of-the-art GAI chatbot renowned for its advanced conversational capabilities. Powered by Generative Pre-trained Transformer (GPT) models, it excels in understanding and generating human-like responses, making it suitable for a wide range of applications, from casual conversations to professional tasks. Its intuitive interface and versatility have made it widely adopted, offering features like text summarization, content generation, and creative writing. Available in free and premium versions, ChatGPT is accessible to individuals, educators, and businesses alike [31]. Recent enhancements include the launch of ChatGPT Pro, a $200/month subscription that provides unlimited access to advanced models such as GPT-o1 and GPT-4o, along with features like Advanced Voice Mode. OpenAI has also expanded ChatGPT's functionality to include web-based search capabilities for up-to-date information. Additional updates include the introduction of the Projects tool, which simplifies managing multiple chats and group files, and Canvas, an interface for collaborative writing and coding.

Microsoft Copilot is a GAI-powered assistant integrated into the Microsoft 365 ecosystem, designed to enhance productivity across office tools. Built on OpenAI's LLMs models, it provides contextual suggestions, automates repetitive tasks, and supports content generation tailored to user needs [32]. Recent updates include general availability for Microsoft 365 Copilot (with new pricing), the introduction of Windows Copilot with deep OS integration, enhancements to GitHub Copilot (Copilot Chat), and ongoing improvements in Microsoft products like Dynamics 365 and the Power Platform. These updates offer intuitive assistance with writing code, analyzing data, generating content, and automating routine tasks.

Google Gemini (https://gemini.google.com), formerly Bard, is a GAI chatbot that combines conversational AI with the knowledge base of Google's search engine. It delivers accurate, contextually relevant answers and supports tasks such as brainstorming, drafting, and question answering. Integrated into Google's ecosystem, Gemini works with tools like Google Workspace, making it a reliable assistant for personal and professional use [33]. Recent advancements include access to experimental models like Gemini Exp-1206, designed for complex tasks such as coding, mathematics, reasoning, and instruction following. The Gemini 2.0 Flash model improves academic benchmarks and speed, while Gemini Deep Research offers a personal research assistant capable of generating comprehensive reports. Additionally, new Gems for Google Workspace enhance workflow efficiency, and the Gemini app now provides enterprise-grade data protection for business and education customers.

Anthropic Claude (https://claude.ai) is an AI chatbot designed to deliver safe, ethical, and contextually aware conversations. Claude handles complex queries and supports tasks like content creation and data analysis for personal and professional use. Its user-friendliness and accessibility have made it popular, especially in educational and research settings [34]. In 2024, Anthropic introduced the upgraded Claude 3.5 Sonnet model, enhancing capabilities in coding, reasoning, and instruction following. The Claude 3.5 Haiku model offers state-of-the-art performance with improved speed and affordability. Additionally, a new "computer use" feature enables Claude to interact with computer interfaces, automating tasks by simulating human actions like moving a cursor and typing text.

Perplexity AI (https://www.perplexity.ai/) is a search-driven chatbot that combines GAI with real-time information retrieval to generate concise and accurate answers. Known for its minimalistic interface and focus on transparency, Perplexity is relatively inexpensive, making it appealing to

individuals and small organizations. Although it lacks deep integration capabilities, it emphasizes precision and real-time information [35]. Recent updates include Internal Knowledge Search, allowing Pro and Enterprise Pro users to search public web content and internal knowledge bases simultaneously, and Spaces, an AI-powered collaboration hub for organizing research, connecting internal files, and customizing AI assistants for specific tasks, enhancing teamwork and productivity.

These five chatbots demonstrate the diverse capabilities of GAI technology, excelling in areas such as professional productivity, ethical AI, real-time information, and integration. Table 2 provides a detailed comparison of the utilized LLMs, functionality, applicability, integration capability, real-time access and pricing for these leading GAI chatbots.

Table 2. Comparison of the most widely used GAI chatbots.

Feature	ChatGPT	Copilot	Gemini	Claude	Perplexity
Foundation LLM(s)	GPT-o1, GPT-4o	GPT-4o	Gemini 2.0 Flash, Gemini 1.5 Pro	Claude 3.5 Sonnet, Claude 3.5 Haiku	Sonar Small, Sonar Large
Features	Web browsing, code execution, image generation, custom GPTs for tailored interactions	Coding assistance, task automation, integration with MS product	Multimodal data processing, integration with Google services, advanced reasoning capabilities	Safety, ethical considerations, handling extensive context for in-depth analyses	Information retrieval, real-time web search capabilities, user-friendly interfaces
Advantages	Versatile tasks, including content creation, coding assistance, and data analysis.	Deep integration with MS's ecosystem, excelling in coding support and task automation within MS applications	Handling large context reasoning, multimodal data processing, and integration with Google services	Managing extensive context windows, suitable for processing large documents and complex conversations	Quick information retrieval and concise answers, functioning as an AI-powered search assistant
Context length	128K	128K	1M, 2M	200K	131K
Integration	Available as an API, browser and mobile app	Integrated into MS produts (web, Windows, mobile) and code editors (Visual Studio, GitHub)	Integrated into Google Workspace and other Google services	Available via API and standalone applications	Accessible through web interface and browser extensions
Price	Free tier available; Plus and Pro subscriptions at $20/month and $200/month for priority access and additional features	Integrated into MS's ecosystem; pricing varies based on specific application and subscription model, Pro at $20/month	Offers free and premium versions, advanced plan priced at $19/month	Free tier with limited daily messages; Pro plan at $20/month offering enhanced capabilities	Free access with basic functionalities; Pro version at $20/month for advanced features

Real-time access	Yes, can browse the internet to provide current information	Yes, accesses real-time data from the web	Yes, designed for real-time interactions and data retrieval	Limited, primarily relies on training data with some real-time capabilities in advanced versions	Yes, provides up-to-date information from the web
Price	Free tier available; Plus and Pro subscriptions at $20/month and $200/month for priority access and additional features	Integrated into MS's ecosystem; pricing varies based on specific application and subscription model, Pro at $20/month	Offers free and premium versions, advanced plan priced at $19/month	Free tier with limited daily messages; Pro plan at $20/month offering enhanced capabilities	Free access with basic functionalities; Pro version at $20/month for advanced features

According to the collected data (Table 2), the five GAI chatbots demonstrate unique strengths and capabilities tailored to diverse user needs.

ChatGPT offers a context window of up to 128K tokens, making it suitable for tasks requiring extended interactions. Its features include web browsing, code execution, image generation and the ability to create custom GPTs for tailored applications. This makes ChatGPT particularly suited for content creation, coding assistance, and data analysis within flexible and interactive use cases.

Copilot, integrated within Microsoft's ecosystem, shares similar context window capabilities with ChatGPT due to its foundation on the same model. Its strengths lie in coding assistance, task automation and seamless integration with Microsoft Office applications. This tight integration makes it a powerful productivity tool for users working within Microsoft's suite of tools, offering efficiency for enterprise and professional workflows.

Gemini excels in processing multimodal data and handling extensive context, with the ability to manage up to 1 million tokens. This capability positions it as a leader for tasks involving large datasets, advanced reasoning, and integration with Google services. Its rapid processing and support for multimodal data, combined with Google Workspace integration, make it particularly strong in professional and research-oriented environments.

Claude, with a context window of approximately 200K tokens, is ideal for tasks requiring extensive document processing and in-depth analyses. Its emphasis on safety, ethical considerations, and privacy measures positions it as a preferred choice for applications where ethical AI use and robust security are critical, particularly in education, research and data-sensitive industries.

Perplexity, with a context window of around 131K tokens, is designed for quick information retrieval and concise answers. Its focus on real-time web search capabilities and user-friendly interfaces makes it highly effective as an AI-powered search assistant, catering to users who prioritize fast, precise and up-to-date information.

In summary, while all five chatbots are effective tools for various AI-driven tasks, their characteristics vary depending on the specific application considered. ChatGPT and Copilot are well-suited for tasks requiring extensive context and integration within specific ecosystems. Gemini excels in handling large datasets and multimodal processing. Claude is ideal for managing extensive context windows with a focus on safety and ethics. Perplexity is best for quick information retrieval and comprehensive responses. ChatGPT and Copilot are more versatile, with features like image generation and internet access, while Gemini, Claude and Perplexity offer cheaper API access and a larger context windows.

In summary, while all five chatbots are effective tools for various AI-driven tasks, their features and strengths vary depending on the specific application. ChatGPT and Copilot are well-suited for tasks that require extensive context handling and integration within specific ecosystems, such as Microsoft Office. Gemini stands out in managing large datasets and multimodal processing, making it ideal for advanced reasoning and complex data tasks. Claude is best suited for applications requiring extensive context windows, with a strong emphasis on safety, ethics and privacy. Perplexity excels in quick information retrieval and concise, accurate responses. ChatGPT and

Copilot offer greater versatility with features like image generation and internet access, making them valuable for diverse use cases. On the other hand, Gemini, Claude and Perplexity provide larger context windows and more affordable API access, catering to users with specific technical or budgetary requirements.

Given the rapidly evolving nature of the GAI chatbot field and the continuous emergence of new players, our review represents only a snapshot of the current landscape.

3. Methodological Framework for GAI Chatbot Selection

This section outlines the theoretical foundations of Interval-Valued Hesitant Fermatean Fuzzy Numbers (IVHFFNs), introduces a modified TOPSIS approach utilizing IVHFFNs, and proposes a conceptual framework for decision analysis of GAI chatbot data.

To address the challenge of GAI chatbot selection, we employ the classic Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [36], complemented by recently developed fuzzy sets modification. As a distance-based multi-criteria decision-making method, TOPSIS determines the relative closeness of each alternative to the ideal solution (best outcome) and the anti-ideal solution (worst outcome) for each criterion. The alternative with the highest coefficient of relative proximity to the ideal solution is selected as the most suitable.

3.1. Interval-Valued Hesitant Fermatean Fuzzy Numbers–Some Basic Definitions and Operations

To enhance the TOPSIS methodology, we integrate Interval-Valued Hesitant Fermatean Fuzzy Sets (IVHFFSs) [14]. This subsection provides an overview of the key concepts and arithmetic operations associated with IVHFFNs, which are essential for implementing this modification.

IVHFFSs extend earlier models such as Interval-Valued Fuzzy Sets (IVFSs) (1975) [37], Hesitant Fuzzy Sets (HFSs) (2010) [38] and Fermatean Fuzzy Sets (FFSs) (2020) [39]. Represented in a three-dimensional space, IVHFFSs use interval values within the range [0, 1] to describe Belongingness Degree (BD), Non-Belongingness Degree (NBD) and Indeterminacy Degree. A notable feature of IVHFFSs is the use of interval values for BD and NBD, with the constraint that the cube of the upper bounds for these intervals must not exceed 1. Compared to FFSs, IVHFFSs provide a more complex representation of uncertainty.

When crisp BD and NBD values are challenging to obtain—due to imprecise or uncertain data—IVHFFNs, with their interval-valued flexibility and the ability to accommodate multiple intervals, offer a practical solution for decision-makers and researchers. This flexibility ensures more accurate assessments of alternatives in situations where precise evaluations are unattainable.

In this section, some basic concepts of IVHFFSs are described.

Definition 1. [14]

The IVHFFS T in a universe U is defined by

$$T = \{u_i, \langle (\alpha_T(u_i), \beta_T(u_i)) \rangle \mid u_i \in U\}$$

where

$$\alpha_T(u_i) = \cup_{[\mu_T^l(u_i), \mu_T^u(u_i)] \in \alpha_T(u_i)} \{[\mu_T^l(u_i), \mu_T^u(u_i)]\} \text{ and}$$

$$\beta_T(u_i) = \cup_{[v_T^l(u_i), v_T^u(u_i)] \in \beta_T(u_i)} \{[v_T^l(u_i), v_T^u(u_i)]\}$$

represent two sets of interval values in [0, 1] signifying the possible BD and NBD of an object $u_i \in U$ to T, with the constraints:

$$0 \le \mu_T^l(u_i) \le \mu_T^u(u_i) \le 1, \ 0 \le v_T^l(u_i) \le v_T^u(u_i)1 \text{ and}$$

$$0 \le \left((\mu_T^u(u_i))^+\right)^3 + \left((v_T^u(u_i))^+\right)^3 \le 1,$$

such that

$$[\mu_T^l(u_i), \mu_T^u(u_i)] \in \alpha_T(u_i), \ [v_T^l(u_i), v_T^u(u_i)] \in \beta_T(u_i),$$

$$(\mu_T^u(u_i))^+ \in \alpha_T^+(u_i) = \cup_{[\mu_T^l(u_i), \mu_T^u(u_i)] \in \alpha_T(u_i)} \max\{\mu_T^u(u_i)\},$$

$$\left(v_T^u(u_i)\right)^+ \in \beta_T^+(u_i) = \bigcup_{\left[v_T^l(u_i),v_T^u(u_i)\right]\in\beta_T(u_i)} \max\{v_T^u(u_i)\} \quad \text{for all } u_i \in U.$$

The pair $((\alpha_T(u_i), \beta_T(u_i))$ is called an Interval-Valued Hesitant Fermatean Fuzzy Number (IVHFFN), denoted by $\xi = (\alpha, \beta)$.

Definition 2. [14] Suppose that $\xi = (\alpha, \beta)$ is an IVHFFN. Then the score function s for ξ can be defined as:

$$s(\xi) = \frac{1}{2}\left(\frac{1}{\#a}\sum_{[\mu_T^l(u_i),\mu_T^u(u_i)]\in\alpha_T(u_i)} \left(\mu_T^l(u_i)\right)^3 - \left(\frac{1}{\#b}\sum_{[v_T^l(u_i),v_T^u(u_i)]\in\beta_T(u_i)} \left(v_T^l(u_i)\right)^3\right) + \right.$$
$$+$$
$$\left. \left(\frac{1}{\#a}\sum_{[\mu_T^l(u_i),\mu_T^u(u_i)]\in\alpha_T(u_i)} \left(\mu_T^u(u_i)\right)^3\right) - \left(\frac{1}{\#b}\sum_{[v_T^l(u_i),v_T^u(u_i)]\in\beta_T(u_i)} \left(v_T^u(u_i)\right)^3\right) \right),$$

(2)

where $\#a$ and $\#b$ represent the number of interval values in α and β, respectively.

The larger the score value $s(\xi)$, the greater the IVHFFN ξ.

Since $s(\xi) \in [-1, 1]$, an improved score function for an IVHFFN ξ in described in the following definition:

Definition 3. [Mishra et al. 2022] Assume $\xi = (\alpha, \beta)$ be an IVHFFN. Then an improved score function is defined by

$$s^*(\xi) = \frac{1}{2}(s(\xi) + 1),$$

(3)

such that $s^*(\xi) \in [0, 1]$.

In case of different numbers of intervals in BD and NBD of an IVHFFN, a preprocessing step should be added. We assume to add the mean value of BD or the NBD for given object.

The arithmetic operations on IVHFFNs are given by the next definition.

Definition 4. [Mishra et al. 2022] Let $\xi_1 = (\alpha_1, \beta_1)$ and $\xi_2 = (\alpha_2, \beta_2)$ be two IVHFFNs. Then we have:

$$\xi_1 \oplus \xi_2 =$$
$$\bigcup_{\substack{[\mu_{\xi_1}^l,\mu_{\xi_1}^u]\in\alpha_1,[v_{\xi_1}^l,v_{\xi_1}^u]\in\beta_1 \\ [\mu_{\xi_2}^l,\mu_{\xi_2}^u]\in\alpha_2,[v_{\xi_2}^l,v_{\xi_2}^u]\in\beta_2}} \left\{ \left\{ \left[\sqrt[3]{\left(\mu_{\xi_1}^l\right)^3 + \left(\mu_{\xi_2}^l\right)^3 - \left(\mu_{\xi_1}^l\right)^3\left(\mu_{\xi_2}^l\right)^3}, \sqrt[3]{\left(\mu_{\xi_1}^u\right)^3 + \left(\mu_{\xi_2}^u\right)^3 - \left(\mu_{\xi_1}^u\right)^3\left(\mu_{\xi_2}^u\right)^3} \right] \right\}, \left\{ \left[v_{\xi_1}^l v_{\xi_2}^l, v_{\xi_1}^u v_{\xi_2}^u \right] \right\} \right\}$$

(4.1)

$$\xi_1 \otimes \xi_2 =$$
$$\bigcup_{\substack{[\mu_{\xi_1}^l,\mu_{\xi_1}^u]\in\alpha_1,[v_{\xi_1}^l,v_{\xi_1}^u]\in\beta_1 \\ [\mu_{\xi_2}^l,\mu_{\xi_2}^u]\in\alpha_2,[v_{\xi_2}^l,v_{\xi_2}^u]\in\beta_2}} \left\{ \left\{ \left[\mu_{\xi_1}^l \mu_{\xi_2}^l, \mu_{\xi_1}^u \mu_{\xi_2}^u \right] \right\}, \left\{ \left[\sqrt[3]{\left(v_{\xi_1}^l\right)^3 + \left(v_{\xi_2}^l\right)^3 - \left(v_{\xi_1}^l\right)^3\left(v_{\xi_2}^l\right)^3}, \sqrt[3]{\left(v_{\xi_1}^u\right)^3 + \left(v_{\xi_2}^u\right)^3 - \left(v_{\xi_1}^u\right)^3\left(v_{\xi_2}^u\right)^3} \right] \right\} \right\}$$

(4.2)

$$\lambda\xi = \bigcup_{[\mu_\xi^l,\mu_\xi^u]\in\alpha,[v_\xi^l,v_\xi^u]\in\beta} \left\{ \left\{ \left[\sqrt[3]{1 - \left(1 - \left(\mu_\xi^l\right)^3\right)^\lambda}, \sqrt[3]{1 - \left(1 - \left(\mu_\xi^u\right)^3\right)^\lambda} \right] \right\}, \left\{ \left[\left(v_\xi^l\right)^\lambda, \left(v_\xi^u\right)^\lambda \right] \right\} \right\},$$

(4.3)

where $\lambda\,(\geq 0) \in \mathbb{R}$.

$$\xi^\lambda = \bigcup_{[\mu_\xi^l,\mu_\xi^u]\in\alpha,[v_\xi^l,v_\xi^u]\in\beta} \left\{ \left\{ \left[\left(\mu_\xi^l\right)^\lambda, \left(\mu_\xi^u\right)^\lambda \right] \right\}, \left\{ \left[\sqrt[3]{1 - \left(1 - \left(v_\xi^l\right)^3\right)^\lambda}, \sqrt[3]{1 - \left(1 - \left(v_\xi^u\right)^3\right)^\lambda} \right] \right\} \right\},$$

(4.4)

where $\lambda\,(\geq 0) \in \mathbb{R}$.

Definition 5. (based on [14]) Let $\xi_1 = (\alpha_1, \beta_1)$ and $\xi_2 = (\alpha_2, \beta_2)$ be two IVHFFNs. Then the distance between ξ_1 and ξ_2 is defined as follows:

$$d(\xi_1, \xi_2) = \left(\frac{1}{4} \left(\begin{array}{c} \left| (\varphi_{1\mu}^l)^3 - (\varphi_{2\mu}^l)^3 \right|^\lambda + \left| (\varphi_{1\mu}^u)^3 - (\varphi_{2\mu}^u)^3 \right|^\lambda + \\ \left| (\varphi_{1\nu}^l)^3 - (\varphi_{2\nu}^l)^3 \right|^\lambda + \left| (\varphi_{1\nu}^u)^3 - (\varphi_{2\nu}^u)^3 \right|^\lambda + \\ \left| (\pi_1^l)^3 - (\pi_2^l)^3 \right|^\lambda + \left| (\pi_1^u)^3 - (\pi_2^u)^3 \right|^\lambda \end{array} \right) \right)^{1/\lambda}$$

where $\varphi_{s\mu}^l = \frac{1}{\#a_s}\sum_{i=1}^{\#a_s}(\mu_i^l)^3$, $\varphi_{s\mu}^u = \frac{1}{\#a_s}\sum_{i=1}^{\#a_s}(\mu_i^u)^3$, $\varphi_{s\nu}^l = \frac{1}{\#b_s}\sum_{i=1}^{\#b_s}(\nu_i^l)^3$, $\varphi_{s\nu}^u = \frac{1}{\#b_s}\sum_{i=1}^{\#b_s}(\nu_i^u)^3$, $\#a_s$ and $\#b_s$ denote the number of BD and NBD intervals in ξ_1 and ξ_2 respectively, s = 1,2, $\lambda > 0$ and

$$\pi_1^l = \sqrt[3]{1 - \left(\frac{1}{\#a_1}\sum_{[\mu_1^l,\mu_1^u]\in\alpha_1}(\mu_1^u)^3 + \frac{1}{\#b_1}\sum_{[\nu_1^l,\nu_1^u]\in\beta_1}(\nu_1^u)^3 \right)},$$

$$\pi_1^u = \sqrt[3]{1 - \left(\frac{1}{\#a_1}\sum_{[\mu_1^l,\mu_1^u]\in\alpha_1}(\mu_1^l)^3 + \frac{1}{\#b_1}\sum_{[\nu_1^l,\nu_1^u]\in\beta_1}(\nu_1^l)^3 \right)},$$

$$\pi_2^l = \sqrt[3]{1 - \left(\frac{1}{\#a_2}\sum_{[\mu_2^l,\mu_2^u]\in\alpha_2}(\mu_2^u)^3 + \frac{1}{\#b_2}\sum_{[\nu_2^l,\nu_2^u]\in\beta_2}(\nu_2^u)^3 \right)},$$

$$\pi_2^u = \sqrt[3]{1 - \left(\frac{1}{\#a_2}\sum_{[\mu_2^l,\mu_2^u]\in\alpha_2}(\mu_2^l)^3 + \frac{1}{\#b_2}\sum_{[\nu_2^l,\nu_2^u]\in\beta_2}(\nu_2^l)^3 \right)}.$$

(5)

Definition 6. [14] Let $\xi_i = \left\{ \{(\mu_i^l, \mu_i^u)\}, \{(\nu_i^l, \nu_i^u)\} \right\}$ (i = 1, 2, …, m) be a collection of IVHFFNs and $w = (w_1, w_2, \dots, w_m)^T$ such that $w_i \geq 0, \sum_{i=1}^m w_i = 1$, then an Interval-Valued Hesitant Fermatean Fuzzy Weighted Average (IVHFFWA) operator is mapping $IVHFFWA: T^n \rightarrow T$, where

$$IVHFFWA(\xi_1, \xi_2, \dots, \quad \xi_m) = \oplus_{i=1}^m w_i\xi_i =$$

$$\bigcup_{[\mu_i^l,\mu_i^u]\in\alpha_i,[\nu_i^l,\nu_i^u]\in\beta_i} \left\{ \left\{ \left[\sqrt[3]{1 - \prod_{i=1}^m(1-(\mu_i^l)^3)^{w_i}}, \sqrt[3]{1 - \prod_{i=1}^m(1-(\mu_i^u)^3)^{w_i}} \right] \right\}, \left\{ \left[\prod_{i=1}^m((\nu_i^l)^3)^{w_i} \right], \prod_{i=1}^m((\nu_i^u)^3)^{w_i} \right\} \right\}. \quad (6.1)$$

Specifically, if $w = (1/m, 1/m, \dots, 1/m)^T$, then IVHFFWA operator is converted into the following formula:

$$IVHFFWA(\xi_1, \xi_2, \dots, \quad \xi_m) = \frac{1}{m}\oplus_{i=1}^m \xi_i =$$

$$\bigcup_{[\mu_i^l,\mu_i^u]\in\alpha_i,[\nu_i^l,\nu_i^u]\in\beta_i} \left\{ \left\{ \left[\sqrt[3]{1 - \prod_{i=1}^m(1-(\mu_i^l)^3)^{1/m}}, \sqrt[3]{1 - \prod_{i=1}^m(1-(\mu_i^u)^3)^{1/m}} \right] \right\}, \left\{ \left[\prod_{i=1}^m((\nu_i^l)^3)^{1/m} \right], \prod_{i=1}^m((\nu_i^u)^3)^{1/m} \right\} \right\}. \quad (6.2)$$

In summary, the space of Interval-Valued Hesitant Fermatean Fuzzy Numbers (IVHFFNs) is broader than that of Interval-Valued Fermatean Fuzzy Numbers (IVFFNs). With a less restrictive constraint, IVHFFSs provide greater precision in addressing complex and uncertain MCDM problems compared to IVFFSs.

3.2. TOPSIS in IVHFFNs Environment

TOPSIS evaluates alternatives by measuring their closeness to an ideal solution and their distance from a negative-ideal solution. To adapt this method for IVHFFNs, we propose calculating the distances between alternatives using Eq. (5). The pseudocode for the modified TOPSIS approach within the IVHFFN framework is presented in Algorithm 1.

Let $A[i], i = 1, 2, \dots, N$ represent the given set of alternatives, $C[j], j = 1, 2, \dots, M$ denote the set of identified criteria for A evaluation and $\omega[j]$ be the set of relative weights of criteria C.

Algorithm 1. IVHFFNs TOPSIS.
Step 1. Gather the linguistic evaluations provided by expert k in the decision matrix
$X^k[i,j] \leftarrow A[i], C[j], k = 1, 2, \dots K,$
where K is the number of experts. Convert the X matrices into values represented by IVHFFNs values.
Step 2. Compute the aggregated matrix $\tilde{X}$ for all experts according to Eq. (6.1). Assume equal weighting for all experts ($1/K$) and apply the averaging formula provided:
$\tilde{X}[i,j] \leftarrow IVHFFWA\left(\tilde{X}^1[i,j], \tilde{X}^2[i,j], \dots, \tilde{X}^k[i,j] \right).$
Step 3. Identify the minimizing criteria, referred to as the cost criteria and denoted by $\mathbb{C}$, while the remaining criteria are categorized as benefit criteria and denoted by $\mathbb{B}$.

Step 4. Determine the normalized values of the decision matrix $\tilde{X}$ using its score function as described in Eq. (3):

$$\tilde{r}[i,j] \leftarrow \frac{\tilde{x}[i,j]}{\sqrt{\tilde{x}[i,j]^2}}$$

Step 5. Derive the weighted values of assessments for each criterion:

$$\tilde{a}[i,j] \leftarrow w_j \tilde{r}[i,j]$$

according to Eq. (4.3).

Step 6. Establish the ideal $\tilde{A}^*$ and negative ideal $\tilde{A}^-$ solutions for each criterion:

$$\tilde{A}^* = \{\tilde{a}_1^*, \tilde{a}_2^*, \ldots, \tilde{a}_M^*\} = \begin{cases} \max_j \tilde{a}[i,j] | j \in \mathbb{B} \\ \min_j \tilde{a}[i,j] | j \in \mathbb{C} \end{cases}$$

$$\tilde{A}^- = \{\tilde{a}_1^-, \tilde{a}_2^-, \ldots, \tilde{a}_M^-\} = \begin{cases} \min_j \tilde{a}[i,j] | j \in \mathbb{B} \\ \max_j \tilde{a}[i,j] | j \in \mathbb{C} \end{cases}$$

for beneficial $(\mathbb{B})$ and cost criteria $(\mathbb{C})$.

Step 7. Measure the distances from each alternative to the ideal and negative ideal solutions using Eq. (5):

$$D^*[i] = \sum_{j=1}^{M} D_G(\tilde{a}[i,j], \tilde{a}^*[j])$$
$$D^-[i] = \sum_{j=1}^{M} D_G(\tilde{a}[i,j], \tilde{a}^-[j])$$

Step 8. Calculate the coefficients of relative closeness of each alternative to the ideal solution:

$$RC_i = \frac{D^-}{D^- + D^+}.$$

Order the alternatives in descending order based on their coefficients of relative closeness to the ideal solution RC_i and select the alternative with the highest coefficient as the optimal choice.

The proposed modification of TOPSIS, which integrates a new flexible IVHFFNs distance metric from Eq. (5), involves a greater computational effort compared to the traditional fuzzy TOPSIS approach. Nonetheless, this enhancement enables a more objective and accurate assessment of alternatives, leading to a more thorough comparison and improved ranking results.

3.3. Theoretical Framework for GAI Chatbot Selection

Selecting an appropriate Generative AI (GAI) chatbot involves a structured, multi-stage decision-making process to ensure alignment with organizational needs and user expectations. The new framework for unified decision analysis of GAI chatbot data consists of eight stages (Figure 1).

Stage 1: Needs Assessment

The decision-making process begins with clearly identifying the specific requirements and expectations for a GAI chatbot. This involves collecting data on available chatbots and understanding the current state of chatbot technology. Relevant information can be gathered from industry reports, user reviews and technical specifications. The goal is to determine which chatbots are available, their capabilities and how well they align with the organization's needs. If the assessment confirms a need for a GAI chatbot, the process advances to the next stage.

Stage 2: User Requirements Specification

In this stage, surveys or interviews are conducted to collect feedback from potential users about their expectations and preferences. This input helps define the desired features and functionalities of the chatbot, such as natural language understanding, integration capabilities and user interface design.
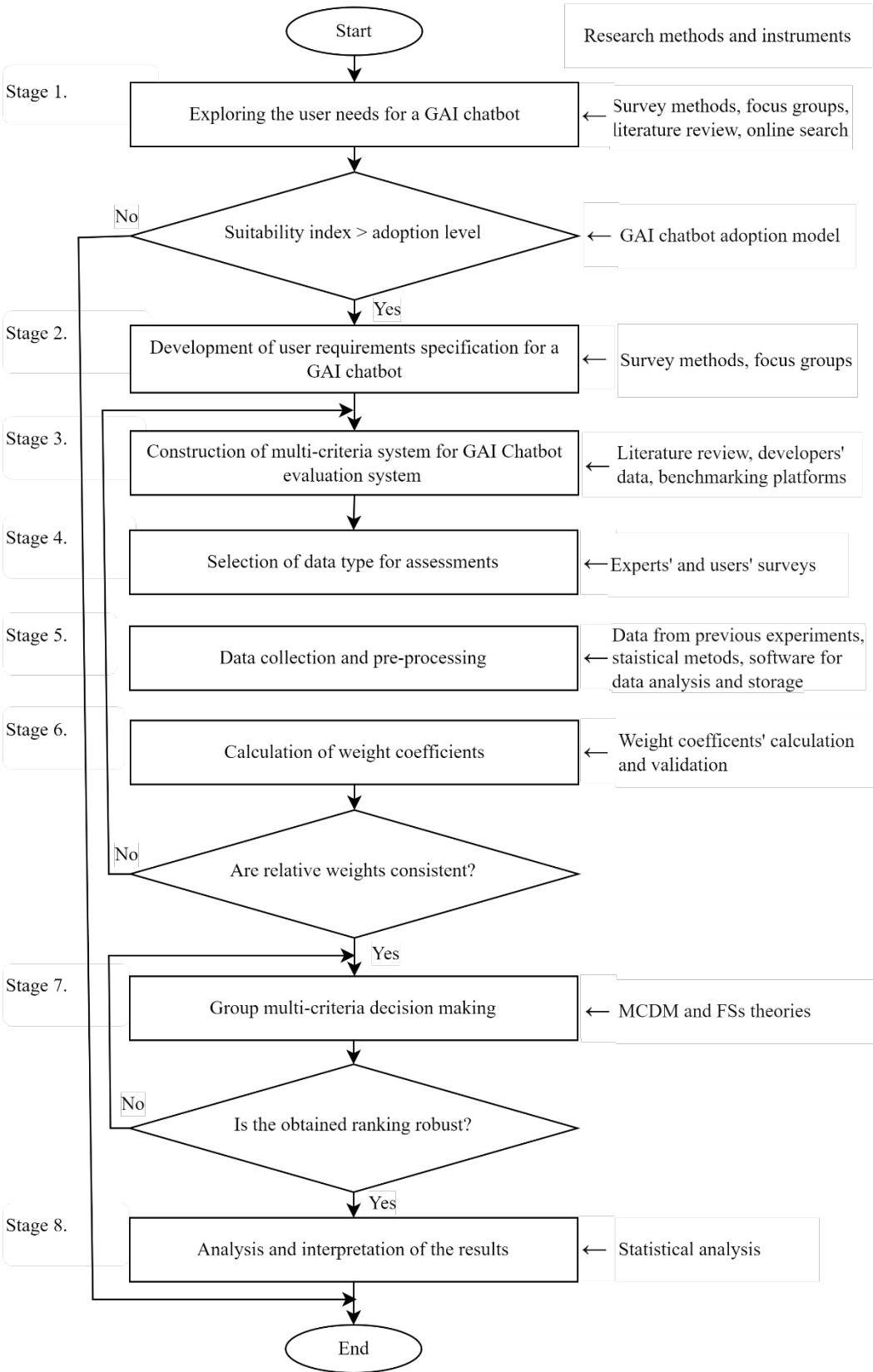
Figure 1. The flowchart of proposed framework for decision analysis of GAI chatbots.

Stage 3: Development of Evaluation Criteria

A multi-criteria evaluation system is created to facilitate a systematic comparison of chatbots. This system is based on user requirements and the organizational importance of specific chatbot features. Key criteria may include technological specifications, ease of integration, user-friendliness, scalability and cost.

Stage 4. Selection of data types

The choice of data types and decision-making methods depends on the resources available and the data collected in Stage 3. If resources are limited, decision-makers may select traditional data types and algorithms with lower computational complexity respectively. For more precise results, advanced data types and MCDM methods can be employed, though they may require greater resources. Data collection methods may include expert evaluations, user testing, and market analysis.

Stage 5. Data reprocessing and storage

Collected data is processed and stored appropriately for further analysis. This step includes coding qualitative assessments into numerical forms, identifying and resolving duplicates or errors, addressing missing values and ensuring overall data integrity. Once processed, the data is stored in a database or dataset for subsequent stages.

Stage 6. Determination of criteria weights

Based on the evaluation criteria and collected data, weight coefficients are assigned to each criterion to reflect their relative importance. These weights can either be predetermined or calculated using methods such as AHP or other weighting techniques.

Stage 7. Multi-criteria analysis

In this stage, MCDM algorithm is applied to rank chatbot alternatives according to the weighted criteria. Using multiple MCDM methods or hybrid combinations can yield a more robust and comprehensive analysis.

Stage 8. Results analysis and interpretation

Decision-makers analyze the rankings to identify the top chatbot alternatives. If the highest-ranked option satisfies organizational requirements, it is selected. If not, additional data may be collected and the process iterated from Stage 4. The final selection should align with long-term organizational goals and user expectations.

This structured approach ensures a comprehensive and objective selection process for GAI chatbots, customized to meet specific organizational needs.

4. A Case Study of Quality-Based Evaluation of GAI Chatbots

Let S be an organization faced with a GAI chatbot selection problem. The benefits of implementation of GAI chatbot in Organization S workflow are numerous. The problem is how to find the best GAI chatbot for the organizational specifics.

The execution of Stage 1 of the proposed framework shows that there are several available GAI chatbots and the process of chatbot selection can start. In this illustrative example, we utilize our own chatbot dataset, collected from benchmarking websites such as [23]. The dataset consists of four assessment criteria $C_1, C_2, ..., C_4$ (Subsection 2.2) and five GAI chatbots $A_1, A_2, ..., A_5$ (Subsection 2.3). The criteria are related to the following aspects of GAI chatbot features: C_1 – Conversational ability, C_2 – User experience, C_3 – Integration capability and C_4 – Price. The GAI chatbots are as follows: A_1 – ChatGPT, A_2 – Copilot, A_3 – Gemini, A_4 – Claude and A_5 – Perplexity.

In Stage 2, experts from Organization S fill in the questionnaire about their GAI chatbot requirements. Respondents evaluate the chabot features via a five-point Likert scale ranging from "Extremely important" (corresponding to 5) to "Unimportant" (corresponding to 1).

In the next stage, experts from Organization S complete a questionnaire outlining their requirements for Generative AI (GAI) chatbots. Participants assess the chatbot features using a five-point Likert scale, ranging from "Unimportant" (1) to "Extremely Important" (5).

In Stage 3, a multi-attribute criteria index is developed, consisting of variables
$C_i, i = \overline{1,4}$.

In the next stage, decision makers decide that the data type is IVHFFNs and employ the proposed new IVHFFNs TOPSIS modification. The values of decision matrix are converted into five-point Likert scale (Table 3). For transforming every linguistic variable into its corresponding IVHFFNs, the conversion table (Table 4) is applied.

In Stage 5, we decide that the data type is IVHFFNs and implement the proposed IVHFFNs TOPSIS modification. The decision matrix values are converted into linguistic variables as shown in Table 3. Each linguistic variable is then transformed into its corresponding IVHFFN using the conversion rules provided in Table 4.

Table 3. Input decision matrix for GAI chatbots selection.

Criteria Alternative	C_1	C_2	C_3	C_4
A_1	VH	H	VH	H
A_2	H	H	VH	H
A_3	H	H	M	L
A_4	M	M	M	L
A_5	M	M	L	H
Criterion type	𝔹	𝔹	𝔹	ℂ

Table 4. Linguistic variables and their corresponding IVHFFNs.

Linguistic term	IVHFFN
Very Low (VL)	{(0.1, 0.2) (0.3, 0.4)}, {(0.7, 0.8) (0.75, 0.85)}
Low (L)	{(0.3, 0.4) (0.5, 0.6)}, {(0.5, 0.6) (0.55, 0.65)}
Medium (M)	{(0.5, 0.6) (0.7, 0.8) (0.75, 0.9)}, {(0.3, 0.4) (0.35, 0.45)}
High (H)	{(0.7, 0.8) (0.8, 0.9)}, {(0.1, 0.2)}
Very High (VH)	{(0.9, 0.95) (0.9, 0.99)}, {(0.01, 0.1) (0.06, 0.15)}

The weight coefficients for the criteria are equal, such that $w_1 = w_2 = w_3 = w_4 = 0.25$.

The obtained overall scores and rankings of given GAI chatbots by using IVHFFNs and crisp TOPSIS method are displayed in Table 5.

Table 5. Scores and their corresponding rankings – TOPSIS method in IVHFFNs.

		A_1	A_2	A_3	A_4	A_5
IVHFFNSs	Score	0.45	0.39	0.34	0.30	0.30
TOPSIS	Rank	1	2	3	4	4

The problem has also been solved using several other MCDM methods (Table 6) – Weighted Sum Method (WSM), Triangular Fuzzy Numbers' (TFNs) WSM, Evaluation Based on Distance from Average Solution (EDAS) and TOPSIS. In order to show that IVHFF TOPSIS solution is feasible, we compare the obtained ranking with those obtained with crisp and triangular fuzzy estimates.

Table 6. Overall scores and their corresponding ranking.

	WSM		TFNs WSM		EDAS		TOPSIS	
Alternative	Score	Rank	Score	Rank	Score	Rank	Score	Rank
A_1	0.40	1	0.19	1	0.67	1	0.65	1
A_2	0.36	2	0.17	3	0.58	3	0.60	3
A_3	0.36	2	0.18	2	0.64	2	0.54	2
A_4	0.20	4	0.10	4	0.42	5	0.22	5
A_5	0.16	5	0.08	5	0.50	4	0.0	4
Spearman's ρ	Benchmark		0.95		0.85		0.95	

The final rankings are as follows:

WSM (Benchmarking method): $A_1 \succ A_2 \approx A_3 \succ A_4 \succ A_5$,

TFNs WSM: $A_1 \succ A_3 \succ A_2 \succ A_4 \succ A_5$, $\rho = 95\%$,

EDAS: $A_1 \succ A_3 \succ A_2 \succ A_5 \succ A_4$, $\rho = 85\%$,

TOPSIS: $A_1 \succ A_2 \succ A_3 \succ A_4 \succ A_5$, $\rho = 95\%$,

IVFFNs TOPSIS: $A_1 \succ A_2 \succ A_3 \succ A_4 \approx A_4$, $\rho = 90\%$.

Spearman's rank correlation coefficient was utilized to assess the agreement between the benchmark ranking (WSM) and the rankings produced by other four MCDM methods. The analysis demonstrated high reliability of the alternative methods, with TFNs WSM and TOPSIS both achieving a Spearman's ρ of 95% and EDAS reaching a ρ of 85%. These substantial correlation coefficient of proposed IVHFFNs TOPSIS (ρ = 90%) confirms that the proposed method aligns closely with the benchmark and alternative methods, ensuring dependable and consistent ranking outcome.

Analysis of the obtained rankings categorizes the GAI chatbots into two primary groups:

Group 1 (Leading GAI chatbots) includes the leading GAI chtabots – ChatCPT (A_1), Copilot (A_2) and Gemini (A_3). ChatGPT consistently secures the top position across all methods, highlighting its superior Conversational ability (C_1) and robust User experience (C_2). Copilot and Gemini follow closely, demonstrating strong performance in Integration capability (C_3) and competitive Price (C_4). While Gemini maintains a comparable standing in most methods, Copilot showcases enhanced strengths in specific criteria, particularly in Integration capability.

Group 2 (Lower-ranked GAI Chatbots) with Claude (A_4) and Perplexity (A_5) consistently occupy the lower ranks across all methods. Claude exhibits moderate performance but lags in Conversational ability (C_1), User experience (C_2) and Integration capability (C_3), whereas Perplexity AI falls behind primarily due to its less competitive Integration capability (C_3).

The ranking analysis across multiple MCDM methods consistently identifies ChatGPT as the leading AI chatbot, followed by Copilot and Gemini. Claude and Perplexity are positioned in the lower tier, highlighting the need for further enhancements to improve their performance in areas such as Conversational ability, User experience and Integration capability. The high correlation coefficient shows the robustness of the proposed TOPSIS modifikation, ensuring that the ranking reflects the underlying performance metrics.

In can be concluded, that the proposed framework is reliable and properly reflects the requirements of organization S.

Selecting the appropriate chatbot is crucial for enhancing user engagement and operational efficiency. To streamline this selection process, a comprehensive approach is essential. The proposed methodology enables experts to evaluate various technological, integration, and performance characteristics, establish specific requirements, utilize fuzzy assessments, and objectively identify the most suitable chatbot for a particular organization. Decision-makers can further refine the evaluation system by incorporating factors such as anticipated interaction volumes, scalability, maintenance and support, error handling and revovery, and customization capabilities.

The proposed methodology offers benefits to both end-users and organizational decision-makers. For end-users, aligning chatbot functionalities with user preferences and requirements enhances satisfaction and engagement. A chatbot selected through this process delivers precise and efficient assistance, thereby elevating the overall user experience. For organizational decision-makers, the new MCDM approach provides a clear and unbiased framework for evaluating chatbots against the organization's strategic goals and operational needs. This leads to informed investment choices and the smooth integration of AI technologies into business processes.

5. Conclusions

The rapid advancement of LLMs has significantly increased the prominence of GAI chatbots in various sectors. Many organizations are integrating these conversational assistants into their workflows to enhance workflow efficiency and user engagement. However, there is currently no unified algorithmic approach for selecting suitable intelligent assistants.

In response to this challenge, we have developed an integrated framework for GAI chatbot selection. This framework introduces an extension of TOPSIS within an IVHFFNs environment, enabling objective evaluation of generative chatbots. The fuzzy nature of this method effectively addresses uncertainty and vagueness in expert assessments. Moreover, the framework is versatile, accommodating both single and repeated data processing for chatbot selection.

The key advantages of the IVHFF TOPSIS include:

- Incorporation of interval-valued membership and non-membership grades, along with interval-valued hesitancy degrees in the evaluation process.
- Integration of Minkowski distance-based family of metrics, enabling flexible and accurate distance calculations tailored to various data types.

- Consideration of the lengths of belongingness, non-belongingness, and hesitancy intervals in distance calculations, ensuring a comprehensive assessment of each criterion's impact.

To demonstrate the effectiveness of this new framework, we applied it to a practical scenario involving the selection of five GAI chatbots: ChatGPT, Copilot, Gemini (formerly Bard), Claude and Perplexity. To capture the performance of the chatbots, we selected four critical criteria that align with user needs and technological capabilities. The analysis of the results indicates that new methodology reliably reflects the features of the chatbots in the final rankings.

This evaluation process can be conducted periodically to account for the rapid advancements in GAI technologies and the evolving needs of organizations. Implementing an iterative procedure allows for continuous refinement of the selection criteria and adaptation to new developments, ensuring that the chosen chatbot solutions remain optimal over time.

In future work, we aim to enhance this conceptual framework by integrating recently developed multi-criteria decision-making methods. Additionally, we intend to develop a new hybrid method for chatbot evaluation that combines innovative weight determination algorithms with advanced multi-criteria decision-making techniques. We also plan to expand the ranking mechanism to address uncertainties using various classic and interval fuzzy sets, including interval type-3 and T-spherical fuzzy numbers.

References

1. Bulchand-Gidumal, J. Impact of artificial intelligence in travel, tourism, and hospitality. In *Handbook of e-Tourism*, pp. 1943–1962. Cham: Springer International Publishing, 2022.

2. Obaid, A.J.; Bhushan, B.; Rajest, S.S., Eds. *Advanced Applications of Generative AI and Natural Language Processing Models*. IGI Global, 2023.

3. Al-Amin, M.; Ali, M.S.; Salam, A.; Khan, A.; Ali, A.; Ullah, A.; ...; Chowdhury, S.K. History of generative Artificial Intelligence (AI) chatbots: Past, present, and future development. *arXiv preprint* arXiv:2402.05122, 2024. Available online: https://arxiv.org/abs/2402.05122 (accessed on 1 January 2025).

4. Yenduri, G.; Srivastava, G.; Maddikunta, P.K.R.; Jhaveri, R.H.; Wang, W.; Vasilakos, A.V.; Gadekallu, T.R. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv preprint* arXiv:2305.10435, 2023.

5. Saka, A.; Taiwo, R.; Saka, N.; Salami, B.A.; Ajayi, S.; Akande, K.; Kazemi, H. GPT models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment* **2023**, 100300.

6. Dwivedi, Y.K.; Pandey, N.; Currie, W.; Micu, A. Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: Practices, challenges and research agenda. *International Journal of Contemporary Hospitality Management* **2024**, 36(1), 1–12.

7. Chen, B.; Wu, Z.; Zhao, R. From fiction to fact: The growing role of generative AI in business and finance. *Journal of Chinese Economic and Business Studies* **2023**, 21(4), 471–496.

8. Ghaffari, S.; Yousefimehr, B.; Ghatee, M. Generative-AI in E-Commerce: Use-Cases and Implementations. In *Proceedings of the 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*, February 2024, pp. 1–5. IEEE.

9.   Al Naqbi, H.; Bahroun, Z.; Ahmed, V. Enhancing Work Productivity through Generative Artificial Intelligence: A Comprehensive Literature Review. *Sustainability* **2024**, 16(3), 1166.

10.   Statista. Chatbot market worldwide 2016–2025. Available online: https://www.statista.com/statistics/656596/worldwide-chatbot-market/ (accessed on 30 June 2024).

11.   Gartner. Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026. Available online: https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026 (accessed on 30 June 2024).

12.   Wang, K.; Ying, Z.; Goswami, S.S.; Yin, Y.; Zhao, Y. Investigating the role of artificial intelligence technologies in the construction industry using a Delphi-ANP-TOPSIS hybrid MCDM concept under a fuzzy environment. *Sustainability* **2023**, 15(15), 11848.

13.   Alshahrani, R.; Yenugula, M.; Algethami, H.; Alharbi, F.; Goswami, S.S.; Naveed, Q.N.; Zahmatkesh, S. Establishing the fuzzy integrated hybrid MCDM framework to identify the key barriers to implementing artificial intelligence-enabled sustainable cloud system in an IT industry. *Expert Systems with Applications* **2024**, 238, 121732.

14.   Mishra, A.R.; Liu, P.; Rani, P. COPRAS method based on interval-valued hesitant Fermatean fuzzy sets and its application in selecting desalination technology. *Applied Soft Computing* **2022**, 119, 108570.

15.   Chakrabortty, R.K.; Abdel-Basset, M.; Ali, A.M. A multi-criteria decision analysis model for selecting an optimum customer service chatbot under uncertainty. *Decision Analytics Journal* **2023**, 6, 100168.

16.   Santa Barletta, V.; Caivano, D.; Colizzi, L.; Dimauro, G.; Piattini, M. Clinical-chatbot AHP evaluation based on "quality in use" of ISO/IEC 25010. *International Journal of Medical Informatics* **2023**, 170, 104951.

17.   Singh, C.; Dash, M.K.; Sahu, R.; Singh, G. Evaluating Critical Success Factors for Acceptance of Digital Assistants for Online Shopping Using Grey–DEMATEL. *International Journal of Human–Computer Interaction* **2023**, 1–15.

18.   Pandey, M.; Litoriya, R.; Pandey, P. Indicators of AI in Automation: An Evaluation Using Intuitionistic Fuzzy DEMATEL Method with Special Reference to Chat GPT. *Wireless Personal Communications* **2024**, 134, 445–465. Available online: https://link.springer.com/article/10.1007/s11277-024-10917-7 (accessed on 30 June 2024).

19.   Pathak, A.; Bansal, V. Factors Influencing the Readiness for Artificial Intelligence Adoption in Indian Insurance Organizations. In *Transfer, Diffusion and Adoption of Next-Generation Digital Technologies*, S.K. Sharma; Y.K. Dwivedi; B. Metri; B. Lal; A. Elbanna, Eds.; IFIP Advances in Information and Communication Technology, vol 698, Springer, Cham, 2024, pp. 384–397. Available online: https://link.springer.com/chapter/10.1007/978-3-031-50192-0_5 (accessed on Day Month Year).

20.   Wiangkham, A.; Vongvit, R. Comparative Analysis of MCDM Methods for Prioritizing Influential Factors of Chatgpt Adoption in Higher Education. Available online: https://ssrn.com/abstract=5040810 (2024).

21.   Ojo, Y.; Davids, V.; Oni, O.; Odoemene, M.; Idowu-Collin, P.; Eyeregba, U. A multi-criteria approach for evaluating the use of AI for matching patients to optimal mental health treatment plans. *Reading Time* **2024**, 05–09. Available online: https://worldscientificnews.com/wp-content/uploads/2024/04/WSN-1932-2024-201-222.pdf (accessed on Day Month Year).

22.   Chatbot Arena. Available online: https://lmarena.ai (accessed on 01 January 2025).

23.   Artificial Analysis. Available online: https://artificialanalysis.ai/ (accessed on 01 January 2025).

24.   Parasuraman, A.; Zeithaml, V.A.; Berry, L.L. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing* **1988**, 64(1), 12–40.

25.   ISO/IEC. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model. International Organization for Standardization (ISO), Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/78176.html (accessed on 01 January 2025).

26.   Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* **1989**, 319–340.

27. Verhoef, P.C.; Lemon, K.N.; Parasuraman, A.; Roggeveen, A.; Tsiros, M.; Schlesinger, L.A. Customer experience creation: Determinants, dynamics and management strategies. *Journal of Retailing* **2009**, 85(1), 31–41.

28. Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* **2003**, 27(3), 425–478. DOI:10.2307/30036540.

29. Tornatzky, L.G.; Fleischer, M. *The Processes of Technological Innovation*. Lexington Books, 1990.

30. Yusof, M.M.; Kuljis, J.; Papazafeiropoulou, A.; Stergioulas, L.K. An Evaluation Framework for Health Information Systems: Human, Organization and Technology-Fit Factors (HOT-Fit). *International Journal of Medical Informatics* **2008**, 77(6), 386–398.

31. Pan, C.; Banerjee, J.S.; De, D.; Sarigiannidis, P.; Chakraborty, A.; Bhattacharyya, S. ChatGPT: A OpenAI platform for society 5.0. In *Proceedings of the Doctoral Symposium on Human Centered Computing*, Singapore, February 2023, pp. 384–397. Springer Nature Singapore.

32. Stratton, J. An Introduction to Microsoft Copilot. In *Copilot for Microsoft 365: Harness the Power of Generative AI in the Microsoft Apps You Use Every Day*, pp. 19–35. Berkeley, CA: Apress, 2024.

33. Saeidnia, H.R. Welcome to the Gemini era: Google DeepMind and the information industry. *Library Hi Tech News* 2023, ahead-of-print.

34. Priyanshu, A.; Maurya, Y.; Hong, Z. AI Governance and Accountability: An Analysis of Anthropic's Claude. *arXiv preprint* arXiv:2407.01557, 2024.

35. Deike, M. Evaluating the performance of ChatGPT and Perplexity AI in Business Reference. *Journal of Business & Finance Librarianship* **2024**, 29(2), 125–154.

36. Hwang, C.L.; Yoon, K. *Multiple Attribute Decision Making: Methods and Applications A State-of-the-Art Survey*. Springer, 1981, vol. 186.

37. Zadeh, L.A. The concept of a linguistic variable and its application to approximate reasoning—I. *Information Sciences* **1975**, 8(3), 199–249.

38. Torra, V. Hesitant fuzzy sets. *International Journal of Intelligent Systems* **2010**, 25(6), 529–539.

39. Senapati, T.; Yager, R.R. Fermatean fuzzy sets. *Journal of Ambient Intelligence and Humanized Computing* **2020**, 11, 663–674.