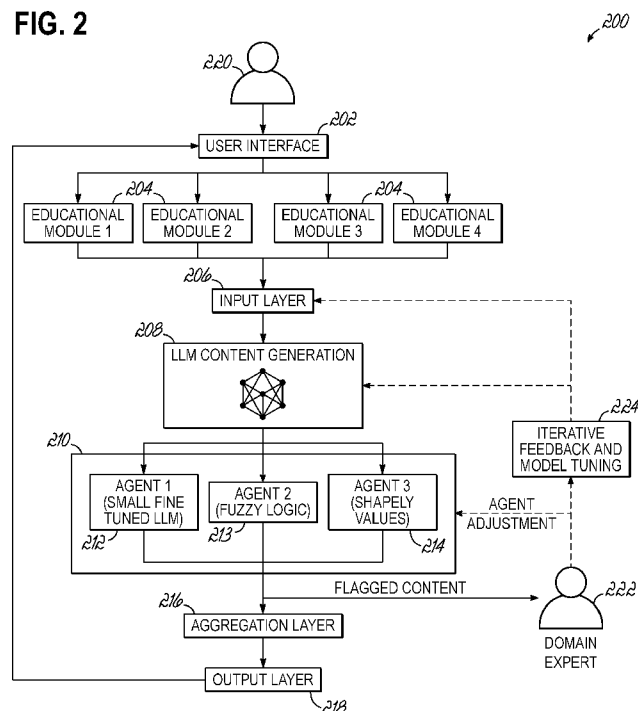




- (51) International Patent Classification:
G09B 7/00 (2006.01) *G06N 20/00* (2019.01)
- (21) International Application Number:
PCT/US2024/036486
- (22) International Filing Date:
02 July 2024 (02.07.2024)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
63/524,759 03 July 2023 (03.07.2023) US
63/620,383 12 January 2024 (12.01.2024) US
- (71) Applicant: **UNIVERSITY OF CINCINNATI** [US/US];
2900 Reading Road, Suite 460N, Cincinnati, Ohio 45206 (US).
- (72) Inventors: **TURNER, Laurah**; 3003 Fairfield Avenue, Cincinnati, Ohio 45206 (US). **OVERLA, Seth**; 1504 Red-sunset Drive, Dayton, Ohio 45458 (US).
- (74) Agent: **JEFFERIES, David E.** et al.; Wood Herron & Evans LLP, 600 Vine Street, Suite 2800, Cincinnati, Ohio 45202 (US).
- (81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,

(54) Title: AGENTIC ARTIFICIAL INTELLIGENCE SYSTEM FOR EDUCATIONAL ENVIRONMENTS

FIG. 2



(57) Abstract: Methods, systems, and computer program products for providing an educational environment using agentic artificial intelligence. A prompt is received from a system user (220), and a response to the prompt is generated using a large language model (208). The response includes raw content defining a clinical scenario consistent with the prompt. The raw content is interpreted using fuzzy logic (213) to generate filtered content based on the first raw content, and the filtered content is displayed to the user (220). In response to the user (220) inputting additional prompts responsive to the filtered content, the large language model (208) generates another response including raw content that defines an updated clinical scenario consistent with both the initial clinical scenario and the additional prompts received from the user. The fuzzy logic (213) generates new filtered content based on the new raw content, and this filtered content is displayed to the system user (220).

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS,
ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

AGENTIC ARTIFICIAL INTELLIGENCE SYSTEM FOR EDUCATIONAL ENVIRONMENTS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims the filing benefit of co-pending U.S. Provisional Application Serial No. 63/524,759 entitled "Training Method Using Artificial Intelligence" filed July 3, 2023, and Provisional Application Serial No. 63/620,383 filed on January 12, 2024 and entitled "Fuzzy Logic Integrated with Ontogenic Neural Networks", the disclosures of which are incorporated by reference herein in their entireties.

FIELD OF THE INVENTION

[0002] This invention generally relates to artificial intelligence and, in particular, to methods, systems, and computer program products that leverage agentic artificial intelligence to provide one-on-one instruction to students.

BACKGROUND

[0003] Conventional "one-size-fits-all" methodologies in traditional classroom education often overlook the unique learning needs, pace, and style of each individual student. This oversight can lead to significant educational gaps by failing to fully cater to individual student needs at different stages of the learning process. This issue may be further accentuated by disparities in student access to resources, and in particular, access to personalized tutoring. Traditional education paradigms also typically enforce a rigid curricular timeline, which can lead to unnecessary pressure, stress, and suboptimal learning experiences. These characteristics collectively contribute to significant educational gaps within conventional educational environments.

[0004] Bloom's 2-Sigma problem refers to an educational phenomenon originally observed by Benjamin Bloom. Bloom's study of university students found that when average students receive one-on-one tutoring, they performed two standard deviations better than similar students educated in a classroom environment. This suggests that one-on-one tutoring could eliminate the aforementioned education gaps for students who are not well served by conventional instructional methods. However, due to resource constraints, and in particular, the lack of qualified instructors, it would be virtually impossible to hire the necessary academic coaches or faculty staff to provide every student with a personal tutor.

[0005] Artificial Intelligence (AI) has witnessed significant advancements in recent years, with technologies such as large language models, fuzzy logic inference systems, ontologies, and knowledge graphs emerging as pivotal tools in the quest for more intelligent and responsive AI systems. These developments have not only revolutionized how machines interpret and process information, but have also opened new avenues for innovation and application across diverse sectors.

[0006] Broad general knowledge and exceptional processing capabilities enable large language models to generate coherent and contextually relevant text responses. However, large language models also have certain limitations. One notable problem is their tendency to struggle with domain-specific or rapidly evolving information. Additionally, large language models may exhibit training data artifacts (e.g., biases) that can impact the neutrality and accuracy of their output data. Another significant limitation is their inherent lack of explainability. The complex “black-box” nature of large language models often makes it difficult to understand or trace the reasoning behind their generated outputs. This lack of transparency poses challenges in contexts where interpretability of results is crucial, such as in educational settings.

[0007] Thus, there is a need for improved systems, methods, and computer program products that enable large language models to provide one-on-one instruction to students by addressing the shortcomings of existing large language model technology.

SUMMARY

[0008] In an aspect of the invention, a method for providing an educational environment using an agentic artificial intelligence system is provided. The method includes receiving a first prompt and generating a first response based on the first prompt using a first large language model. The first response includes first raw content which defines a first clinical scenario consistent with the first prompt. The method interprets the first raw content by using fuzzy logic to generate at least a first portion of a first filtered content based on the first raw content and displays the first filtered content to a system user.

[0009] In an embodiment of the method, the method may further include receiving an input from the system user responsive to the first filtered content, generating a second prompt based on the input from the system user, providing the second prompt to the first large language model, generating a second response to the

second prompt using the first large language model, the second response including second raw content that defines a second clinical scenario consistent with both the first clinical scenario and the second prompt, interpreting the second raw content by using the fuzzy logic to generate a second filtered content based on the second raw content, and displaying the second filtered content to the system user.

[0010] In another embodiment of the method, generating the at least the first portion of the first filtered content may include tokenizing the first raw content to generate a plurality of raw content tokens, and applying a set of fuzzy logic rules to the raw content tokens to generate the at least the first portion of the first filtered content.

[0011] In another embodiment of the method, generating the first response may include tokenizing the prompt to generate a plurality of prompt tokens, and the method may further include determining a Shapley value for each prompt token that quantifies a contribution of the prompt token to the raw content.

[0012] In another embodiment of the method, the method may further include forwarding the first raw content to a domain agent for human review in response to the fuzzy logic interpreting the first raw content as including at least some uncertain content or some irrelevant content.

[0013] In another embodiment of the method, generating the first filtered content may include using a second large language model to interpret the first raw content and generate at least a second portion of the first filtered content.

[0014] In another embodiment of the method, the second large language model may interpret the first raw content by tokenizing the first raw content into a plurality of raw content tokens, and generating the at least the second portion of the first filtered content by statistically analyzing the raw content tokens.

[0015] In another embodiment of the method, each of the first large language model and the second large language model may be configured to analyze strings of tokens, and a maximum length of the strings of tokens analyzed by the first large language model may be larger than the maximum length of the strings of tokens analyzed by the second large language model.

[0016] In another embodiment of the method, the first large language model may be a strong large language model, and the second large language model may be a weak large language model.

[0017] In another embodiment of the method, the method may further include aggregating the first and second portions of the first filtered data.

[0018] In another aspect of the invention, the agentic artificial intelligence system for providing the educational environment is provided. The system includes one or more processors and a memory coupled to the one or more processors that contains program code. When executed by at least one of the one or more processors, the program code causes the system to receive the first prompt and generate the first response based on the first prompt using the first large language model. The first response includes first raw content that defines the first clinical scenario consistent with the first prompt. The program code further causes the system to use an interpretability agent including the fuzzy logic that interprets the first raw content to generate at least the first portion of the first filtered content based on the interpretation of the first raw content by the fuzzy logic, and display the first filtered content to the system user.

[0019] In an embodiment of the system, the program code may further cause the system to receive the input from the system user responsive to the first filtered content, generate the second prompt based on the input from the system user, provide the second prompt to the first large language model, generate the second response to the second prompt using the first large language model, the second response including second raw content that defines a second clinical scenario consistent with both the first clinical scenario and the second prompt, use the interpretability agent to generate the second filtered content based on the interpretation of the second raw content by the fuzzy logic, and display the second filtered content to the system user.

[0020] In another embodiment of the system, the interpretability agent may interpret the first raw content and generate the at least the first portion of the first filtered content by tokenizing the first raw content to generate the plurality of raw content tokens, and applying the set of fuzzy logic rules to the raw content tokens to generate the at least the first portion of the first filtered content.

[0021] In another embodiment of the system, generating the first response may include tokenizing the prompt to generate the plurality of prompt tokens, and the program code may further cause the system to use an oversight agent to determine the Shapley value for each prompt token, the Shapley value quantifying the contribution of the prompt token to the raw content.

[0022] In another embodiment of the system, the program code may further cause the system to forward the first raw content to the domain agent for human review in response to the interpretability agent interpreting the first raw content as including at least some uncertain content or some irrelevant content.

[0023] In another embodiment of the system, generating the first filtered content may further include using a supervision agent to interpret the first raw content and generate at least the second portion of the first filtered content.

[0024] In another embodiment of the system, the supervision agent may interpret the first raw content by tokenizing the first raw content into a plurality of raw content tokens, and providing the raw content tokens to the second large language model that generates the at least the second portion of the first filtered content.

[0025] In another embodiment of the system, each of the first large language model and the second large language model may be configured to analyze strings of tokens, and the maximum length of the strings of tokens analyzed by the first large language model may be larger than the maximum length of the strings of tokens analyzed by the second large language model.

[0026] In another embodiment of the system, the first large language model may be a strong large language model, and the second large language model may be a weak large language model.

[0027] In another aspect of the invention, a computer program product for providing the educational environment using the agentic artificial intelligence system is provided. The computer program product includes a non-transitory computer-readable storage medium, and program code stored on the non-transitory computer-readable storage medium. The program code is configured so that, when executed by one or more processors, the program code causes the one or more processors to receive the first prompt and generate the first response based on the first prompt using the first large language model. The first response includes first raw content that defines the first clinical scenario consistent with the first prompt. The program code further causes the system to use the interpretability agent including the fuzzy logic that interprets the first raw content to generate the at least the first portion of the first filtered content based on the interpretation of the first raw content by the fuzzy logic, and display the first filtered content to the system user.

[0028] The above summary presents a simplified overview of some embodiments of the invention to provide a basic understanding of certain aspects of the invention

discussed herein. The summary is not intended to provide an extensive overview of the invention, nor is it intended to identify any key or critical elements, or delineate the scope of the invention. The sole purpose of the summary is merely to present some concepts in a simplified form as an introduction to the detailed description presented below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0029] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate various embodiments of the invention and, together with the general description of the invention given above, and the detailed description of the embodiments given below, serve to explain the embodiments of the invention.

[0030] FIG. 1 is a flowchart illustrating a process for interacting with an agentic AI system.

[0031] FIG. 2 is a schematic view of the agentic AI system.

[0032] FIG. 3 is a flowchart illustrating a process for providing interactive training using the agentic AI system of FIG. 2.

[0033] FIG. 4 is a schematic view of a process for evaluating the effectiveness of the agentic AI system of FIG. 4.

[0034] FIG. 5 is a diagrammatic view of a computer that may be used to implement one or more of the components or processes shown in FIGS 1-4.

[0035] It should be understood that the appended drawings are not necessarily to scale, and may present a somewhat simplified representation of various features illustrative of the basic principles of the invention. The specific design features of the sequence of operations disclosed herein, including, for example, specific dimensions, orientations, locations, and shapes of various illustrated components, may be determined in part by the particular intended application and use environment.

DETAILED DESCRIPTION

[0036] The below specification discloses an agentic AI system that provides precision education, with a specific focus on medical education. In particular, the disclosed agentic AI system addresses the inherent challenges of standardization and personalization in physician training. Embodiments of agentic AI system use a large language model to generate content, and one or more of fuzzy logic to interpret the generated content, and Shapley values to quantify input contributions. This

novel approach provides a dynamic, responsive educational environment that mitigates the limitations of current AI models (e.g., bias, transparency, and the propensity for errors) by enhancing interpretability and ensuring content reliability. The agentic AI system integrates its various components in a way that significantly improves the alignment, oversight, and interpretability of educational content as compared to known AI systems. Embodiments of the agentic AI system may be used to provide education in a way that is personalized, adaptable, and efficient, thereby fostering a deeper, more intuitive learning experience for students, such as physician trainees. Although embodiments of the agentic AI system are generally described herein in the context of educating medical students, embodiments of the invention are not limited to use in any particular field of study.

Educational Environment Using Artificial Intelligence

[0037] Ontologies and knowledge graphs offer a structured, explainable approach to knowledge representation that may be used to provide decision-making and interpretability within AI systems. Knowledge graphs are well adapted for organizing and linking information. However, knowledge graphs may present challenges in terms of completeness and staying current with the latest information. Knowledge graphs are also generally not suitable for processing natural language and unstructured text.

[0038] Fuzzy logic systems also offer a mechanism for enhancing explainability in AI decision-making. Where large language models often operate as “black boxes” with outputs that are difficult to interpret, fuzzy logic systems provide a transparent approach to reasoning under uncertainty. Fuzzy logic systems may be used to process and interpret complex, ambiguous, or subjective information that defies binary categorization to handle data and concepts that span across spectrums. This capacity to make sense of nuanced and overlapping categories makes fuzzy logic suitable for critical tasks requiring precision and reliability. Fuzzy logic systems can thus bring clarity to AI decision-making processes. However, fuzzy logic systems may also require well-defined rules and may face challenges in integration with other AI technologies, especially those handling large-scale structured data.

[0039] Embodiments of the present invention include a large language model trained to generate educational module content, evaluate student responses to this content, as well as provide feedback on and assessments of these responses to both the student and human instructors. The large language model is integrated with

one or more agents each including one or more of a small fine-tuned large language model, a fuzzy logic system, Shapley values, ontologies, or knowledge graphs. The resulting agentic AI system combines the generative and interpretive power of the large language model with the structured, explicit knowledge representation provided by the one or more agents. This synergy provides an agentic AI system that is both more accurate and context-aware, but also inherently more explainable and transparent than known large language model-based educational systems. The resulting AI solutions can be finely tuned to specific tasks to enhance reliability and improve safety, particularly in domains that require critical decision-making.

[0040] Embodiments of the disclosed agentic AI system harness the advanced language processing of large language models in conjunction with one or more of the precise knowledge organization of ontologies and knowledge graphs, the nuanced decision-making of fuzzy logic and Shapley values, and the supervision of fine-tuned weak large language models. The resulting integrated system enhances AI performance in complex tasks like entity recognition, relation extraction, and knowledge engineering. This synergy both facilitates the interpretation of varied and intricate data and optimizes the efficiency of knowledge consolidation. The resulting agentic AI system can be uniquely tailored for specific tasks and has unparalleled reliability and adaptability. Embodiments of the agentic AI system may evolve and respond effectively in dynamic information environments, ensuring both enhanced accuracy and deeper contextual awareness than known AI systems. The integration of these diverse AI technologies also provides an agentic AI system that is not only more capable but also inherently safer and more aligned with the demanding requirements of critical decision-making domains than known systems.

[0041] Embodiments of the agentic AI system may be used to enhance traditional educational paradigms and bridge training gaps to allow students to progress at a pace that suits them, providing personalized, interactive, and adaptive learning experiences to every student, regardless of their background or resources. Although specific examples of the agentic AI system are described below in the context of educating medical students, embodiments of the invention are not limited to just this application. Accordingly, embodiments of the agentic AI system may be used to educate students in other fields of endeavor by selecting suitable training datasets.

[0042] Embodiments of the agentic AI system may increase access to premium medical education by significantly reducing educational disparities and minimizing costs associated with traditional medical education. The disclosed technology may reduce the need for multiple study resources or a large numbers of preceptors and tutors, thereby providing a cost-effective and efficient solution. The disclosed agentic AI system may thereby enable institutions to provide personalized medical education for all students irrespective of their socio-economic status or geographic location.

[0043] Potential users of the agentic AI system include medical students, residents, fellows, and attending physicians striving to refine their clinical skills and expand their medical knowledge. The agentic AI system may also be used to support medical institutions committed to eradicating educational disparities and enhancing the quality of their training programs. By offering a personalized, AI-powered, self-paced learning experience, embodiments of the agentic AI system can assist these institutions in achieving their educational goals and cultivating highly skilled medical professionals at a fraction of the traditional cost.

[0044] The approach to medical training provided by the agentic AI system essentially creates an advanced “AI Coach” for medical students, residents, fellows, and attending physicians. This AI Coach enables users to develop, refine, and reinforce clinical skills and medical knowledge in a safe, low-pressure, and highly engaging environment. The agentic AI system can provide personalized, one-on-one academic coaching tailored to a student’s unique pace and style, thereby bridging the education gap in medical training.

[0045] Features of the agentic AI system may include, but are not limited to, AI coaching sessions and natural language interaction, chat interfaces, performance monitoring and feedback, and adaptive learning.

[0046] With regard to AI coaching sessions and natural language interaction, the system may enable students to engage in scenario-based clinical encounters, one-on-one tutoring, question banks, or flashcard reviews, all tailored to their needs and preferences through interactive, adaptable sessions. Embodiments of the agentic AI system may employ advanced natural language interaction, facilitating seamless communication and enhancing learning experience. In some embodiments, generative AI may be used, enabling unlimited versions of the encounters so that students do not repeat sessions. As used herein, generative AI refers to AI that is

configured to generate content (e.g., text, images, audio, videos, etc.) in response to input from a user. The generative AI may be configured to identify patterns and structures in training data, and then generate new content that has characteristics similar to the training data based on these patterns and structures.

[0047] Embodiments of the agentic AI system may also include advanced chat interfaces that deliver personalized content based on the dialogue between the system and the student. This may not only enhance one-to-one tutoring experiences, but also provides instant, accurate, and verified responses to the student's queries.

[0048] After each coaching session, students may receive detailed, actionable performance monitoring feedback, supplemented with relevant academic resources to encourage self-directed learning. Performance analytics and session summaries may be generated and tracked in the student's personal dashboard.

[0049] Embodiments of the agentic AI system may also utilize the performance analytics and reinforcement machine learning (adaptive learning) from the personal dashboard to personalize the learning experience by continuously adapting coaching sessions to the student's strengths, areas needing improvement, and skill level.

[0050] Features provided to users by the agentic AI system may include, but are not limited to, goal setting, resource libraries, community forums, and profile customization. With regard to goal setting, students may set and manage their learning goals directly from a home screen, fostering a guided, goal-oriented learning experience. A comprehensive collection of additional educational materials may be available as a resource library to supplement the student's studies and deepen their understanding. Embodiments of the system may also facilitate peer-to-peer learning through an interactive community forum, encouraging discussion, collaboration, and knowledge sharing. Students may also be allowed to personalize their profile settings to align the agentic AI system with their unique learning needs and preferences or to align the agentic AI system with their experience and level of training (e.g., a third-year medical student, a first-year pediatrics resident, a neurology fellow, etc.).

[0051] In some embodiments, the agentic AI system may use a large language model trained for a specific area of study (e.g., medical) and reinforcement learning to provide a more efficient and individualized learning environment. Unlike conventional educational approaches, the agentic AI system may be configured to

provide immediate, precise, and customized feedback to a diverse range of queries, significantly enhancing educational quality and effectiveness. By improving both the interface and the learning experience, the agentic AI system may transcend the limitations of existing technologies, making learning more intuitive, engaging, and responsive.

[0052] The agentic AI system may include a robust performance monitoring and feedback mechanism that includes reinforcement learning refined using individual student data. Following each coaching session, students may receive detailed, actionable feedback along with relevant academic resources to encourage self-directed learning. Performance analytics and session summaries may be generated and tracked within the student's personal dashboard. Leveraging these performance analytics, the agentic AI system may continuously adapt the coaching sessions to align with the student's strengths, areas for improvement, and progression level, further personalizing the learning experience.

[0053] The agentic AI system may be used, for example, to transform the way medical professionals learn, evolve, and excel in their careers. Although embodiments of the agentic AI system may be used to provide educational environments for many fields of study, there may be a particular need for this paradigm shift in the medical field for several reasons. First, an efficient, personalized, and readily accessible approach to medical education has become increasingly necessary to address the global surge in demand for healthcare professionals. The agentic AI system can serve a wide array of stakeholders, including medical students and medical institutions, thereby leading to improved patient care and outcomes. The agentic AI system also aligns seamlessly with the broader goal of lowering barriers to receiving a medical education by providing broad access to top-tier, personalized coaching. Embodiments of the agentic AI system are also well adapted for nurturing a culture of lifelong learning and continuous improvement among medical professionals, ensuring they stay current with the latest medical advancements and best practices. The one-on-one instructional approach enabled by the agentic AI system has the potential to reshape the future of medical education, benefiting not only the medical community but also patients and the healthcare system at large.

[0054] FIG. 1 depicts a flowchart of an exemplary process 100 that may be implemented by an embodiment of the invention. In block 102, the user (e.g., a

student) logs into the system. In response to successfully authenticating the user in block 104, the process 100 may proceed to block 106 and display a home screen to the user. The home screen may allow the user to set learning goals and provide one or more options to the user. For example, the user may select “start AI coaching session” in block 108 (thereby causing the process 100 to proceed to block 110), search for additional learning materials by selecting “access resource library” in block 112 (thereby causing the process 100 to proceed to block 114), select “interact with community forum” in block 116 to participate in discussions and peer learning (thereby causing the process 100 to proceed to block 118), or select “update profile settings” in block 120 (thereby causing the process 100 to proceed to block 122). Regardless of which option the user selects, they may be eventually led back to the home screen, thereby allowing for a continuous learning session.

[0055] In block 110, the process 100 may present one or more (e.g., four) AI coaching session modules 124-127 for selection by the user. These modules 124-127 may include, for example, a scenario-based clinical encounter, one-on-one tutoring, question banks, flash cards, or any other type of AI coaching session. Although the process 100 is depicted as having four selectable AI coaching session modules 124-127, embodiments of the invention are not so limited. Accordingly, it should be understood that other embodiments of the process 100 may present more or fewer AI coaching session modules 124-127. In response to the user selecting a module 124-127, the process 100 may proceed to a respective block 128-131 and begin the selected AI coaching session.

[0056] At the end of each type of coaching session (blocks 140-143 of process 100), the process 100 may proceed to a respective block 150-153 and provide one or more of feedback on the user’s performance and further study resources. The feedback may be comprehensive, specific, and actionable. The resources may include, for example, academic papers to encourage and support the user’s self-directed learning. The process 100 may then proceed to a respective block 160-163 and display a summary to the user before taking the user to a dashboard in block 170. The process 100 may use the session summary to populate the dashboard. The process 100 may then interact with the user to set learning goals in block 172 before returning to block 106 and displaying the home screen.

[0057] The dashboard may be directly linked to the home screen, and may be configured to track analytics on user performance over time. The user performance

analytics may be fed back into the AI model used by the system to generate the coaching sessions through a reinforcement learning process. The reinforcement learning process may cause the AI model to adapt to the user's strengths, areas of improvement, and skill level as they change over time. Advantageously, this feature may enable the system to continuously adapt to individual users based on their data.

[0058] Embodiments of the invention may harness the power of advanced AI technologies to facilitate personalized, precision education, and may be configured to adapt to the individual strengths and improvement areas of each user, e.g., medical students, residents, and attending physicians. The disclosed system may be a continuous educational companion, providing a secure, low-stress environment that allows users to enhance their clinical skills, stimulate in-depth medical discussions, and fortify their knowledge retention. Each interaction within the system, whether an AI-generated clinical encounter, an interactive quiz, or a detailed conversation via the intelligent chat interface, may be designed to enrich the user's understanding of complex medical concepts and foster an effective and engaging learning environment.

[0059] The agentic AI system may be configured to align with the Association of American Medical Colleges (AAMC) Core Entrustable Professional Activities (EPAs) for entering residency. The agentic AI system may also prepare students for the realities and challenges of medical practice by accommodating their unique learning pace and style, thereby equipping them with the skills and confidence to excel in their medical careers.

Fuzzy LIONN system

[0060] Certain embodiments of the agentic AI system may be based on a framework that integrates fuzzy logic with ontological neural networks. These systems may be referred to herein as Fuzzy LIONN systems. Fuzzy LIONN systems combine the broad, intuitive processing capabilities of large language models with the structured, explicit data representation of ontologies and knowledge graphs, and the nuanced decision-making ability of fuzzy logic systems. This synthesis may address gaps in known data processing and decision-support systems, providing a solution that addresses the complexities and ambiguities inherent in real-world data. As used herein, the term "ontology" refers to a structured representation of domain specific knowledge. Fuzzy LIONN systems may use ontologies to complement large language models and knowledge graphs. This integration may provide the Fuzzy

LIONN system with access to a wide array of interconnected information while adhering to well-defined domain-specific concepts and relationships. Ontologies may serve as a blueprint, guiding the structures of knowledge graphs and enhancing the system's ability to precisely interpret complex medical education data.

[0061] One advantage of Fuzzy LIONN systems is their ability to handle ambiguity. Fuzzy LIONN systems may be specifically designed to address scenarios where data is incomplete, unclear, or inherently ambiguous. Such situations are commonplace in real-world datasets, where neat and precise information is more the exception than the rule. Known AI models often falter in these scenarios, as they predominantly rely on precise, well-defined data inputs. Fuzzy LIONN systems integrate fuzzy logic in such a way that they are adept at interpreting imprecise and uncertain information and making decisions based thereon. Fuzzy LIONN systems thereby offer a significant advantage over conventional AI models.

[0062] Fuzzy LIONN systems are an amalgamation of the predictive language capabilities of large language models, the precision and clarity of fuzzy logic, and the structured knowledge representation of knowledge graphs. This integration may enable a Fuzzy LIONN system to process and analyze a diverse range of data types, from structured datasets to unstructured textual information. This approach may enable agentic AI systems in a manner not previously unattainable with traditional AI models, which lack the versatility to bridge different data formats and interpretative requirements.

[0063] Another advantage of fuzzy LIONN systems is their dynamic nature and adaptive learning features, which enable continual evolution and adaptation to new information and changing conditions. This adaptive learning ability may be particularly advantageous in rapidly evolving domains where static AI models quickly become outdated. The adaptability of fuzzy LIONN systems may improve their ability to provide relevant and accurate analysis and recommendations, regardless of shifts in the underlying data landscape.

[0064] Fuzzy LIONN systems are inherently versatile and flexible by design, and thus lend themselves to customization across a broad spectrum of industries and applications. This adaptability allows fuzzy LIONN systems to cater to the specific needs and nuances of different domains. This feature may be particularly valuable in specialized fields like healthcare, where the data and decision-making processes may be markedly distinct from those in other industries.

[0065] The integration of ontologies into fuzzy LIONN systems may elevate their decision-support capabilities above those provided by mere data analysis, particularly in complex medical training scenarios. By adhering to established medical education ontologies, fuzzy LIONN systems may ensure that their analyses and recommendations are grounded in comprehensive and standardized medical knowledge bases and professional standards. This integration may be advantageous for applications where precise language and adherence to medical standards are required. This capability may enhance informed and nuanced decision-making, especially in complex and high stakes scenarios.

[0066] Fuzzy LIONN systems integrate large language models, knowledge graphing, ontologies, and fuzzy logic, with each playing a role in the system's overall functionality. Together, they create a sophisticated AI system capable of handling complex, ambiguous, and unstructured data with higher efficiency and accuracy than known AI systems.

[0067] The large language model components may provide a broad and in-depth interpretation and generation of language (e.g., text), and may be trained on extensive datasets that encompass a wide array of topics and contexts to develop a comprehensive linguistic and contextual language processing ability. The main role of the large language models of the fuzzy LIONN system may be to extract and interpret unstructured data. This functionality may be achieved using advanced neural network architectures that are designed for both accuracy and efficiency in natural language processing. The large language models may be used for deciphering the nuances of human language, enabling the fuzzy LIONN system of which they are a part to process and analyze textual data with a high degree of sophistication.

[0068] A dynamic knowledge graph may be used to provide a structured representation of interconnected data points in the fuzzy LIONN system. The dynamic knowledge graph is not a static knowledge graph, and thus may be periodically or continuously updated by one or more algorithms. Ontologies may be used to complement the large language models and knowledge graphs so that the fuzzy LIONN system not only has access to a wide array of interconnected information, but also adheres to well-defined domain-specific concepts and relationships. Ontologies may provide a blueprint, guiding the knowledge graph's structure and enhancing the fuzzy LIONN system's ability to interpret complex

medical data with greater precision. The knowledge graph may be defined by categorizing and linking data entities, thereby forming a network that represents relationships and associations among different concepts. This structured approach may enable fuzzy LIONN systems to maintain a comprehensive and evolving ability to process domain-specific knowledge, thereby enabling accurate data interpretation and decision-making.

[0069] The fuzzy logic component of fuzzy LIONN systems may be configured to handle ambiguity and uncertainty in data, which is a common challenge in real-world scenarios. The fuzzy logic may operate using a set of rules that effectively translates quantitative data into qualitative insights. This translation may improve the system's ability to correctly interpret data that is not black and white, enabling the system to make probabilistic evaluations where known AI systems would provide an oversimplified analysis or otherwise fail. This feature may be particularly advantageous in scenarios where data is incomplete or inherently vague by enabling nuanced and context-aware interpretations of data.

[0070] Integration of the linguistic processing of large language models, the structured knowledge base of knowledge graphs, and the interpretative ability of fuzzy logic makes Fuzzy LIONN systems uniquely capable of handling the complexities of modern data environments. This synergy may enable fuzzy LIONN systems to generate insights and decisions that are both data-driven and contextually nuanced, addressing the needs of diverse applications, such as medical education, healthcare, and business intelligence, to name a few. The integration process may involve a seamless data flow among the large language models, knowledge graphs, and fuzzy logic. The large language models may preprocess and categorize input data. The preprocessed and categorized data may then be provided to the knowledge graph for structural organization. This organized data may then be analyzed by fuzzy logic to generate output data. A Fuzzy LIONN system may be configured, for example, to provide personalized learning, predictive analytics, or adaptive content delivery. With regard to medical education, configuring a fuzzy LIONN system could involve personalizing learning paths based on individual performance and preferences.

[0071] One step in the configuration process of a fuzzy LIONN system may be to determine the objectives for the system, and the scope of the knowledge areas to be covered by the system, which may be tailored to specific fields as needed. The

objectives and scope may in turn drive functional requirements (like processing speed and accuracy) and non-functional requirements (such as scalability and data security).

[0072] Another step in the fuzzy LIONN system configuration process may be to develop the knowledge graph. This step may include data collection, e.g., gathering domain-specific data, such as academic content, research papers, or expert opinions, to define the underlying knowledge base for the knowledge graph. An ontology for the knowledge graph may then be defined based on the knowledge base. Large language models may be used to identify an ontology consistent with the knowledge base, and to enrich the knowledge graph by processing and categorizing content according to the ontology by identifying key concepts, relationships, and hierarchies. The knowledge graph may then be defined to represent and link different concepts and provide a structured understanding of the domain.

[0073] Fuzzy logic may be integrated into a fuzzy LIONN system by using a knowledge graph to establish categories and subcategories for a fuzzy inference system. Fuzzy logic rules may then be developed to handle ambiguity and imprecise information, e.g., based on human expertise and feedback. The fuzzy logic inference system may be integrated with the knowledge graph by ensuring the fuzzy logic system's interpretations and decisions are consistent with the ontological representations of the knowledge graph, thereby enhancing the accuracy and relevancy of the knowledge graph's output data.

[0074] Integrating a large language model into a fuzzy LIONN system may involve collecting and categorizing data, such as user interactions or performance metrics, and using the large language model to map the collected data onto a knowledge graph. The large language model may be used to interpret outputs from the fuzzy inference system to ensure that these interpretations are in line with the ontological structure of the knowledge graph. This approach may ensure that feedback and recommendations are not only personalized but also medically accurate and contextually relevant.

[0075] The system fuzzy LIONN system may provide personalized recommendations or insights, and continually adapt to new data and evolving scenarios. For example, mechanisms may be implemented for continuous learning and adaptation, thereby enhancing the accuracy and relevance of the fuzzy LIONN

system over time. The integration of large language models, knowledge graphs, and fuzzy logic in fuzzy LIONN systems may be a finely tuned process, ensuring efficient data processing, analysis, and interpretation. Each step in the implementation may be configured to build upon the previous step, culminating in a system capable of delivering personalized, adaptive, and insightful content for complex decision-making scenarios.

[0076] Fuzzy LIONN system architectures may include a multi-layered structure configured to provide efficient handling of data, seamless integration of different AI components, robust processing for analysis and interpretation, and effective interaction with end-users. The multi-layered structures may include a data layer, an integration layer, a processing layer, and an application layer.

[0077] The data layer may provide a foundation for the fuzzy LIONN system, and may be tasked with management of both raw and processed data. To this end, the data layer may handle the storage, retrieval, and management of data, so that the fuzzy LIONN system has continuous access to the data necessary for its operation. The data layer may be configured to accommodate large volumes of diverse data types, ranging from structured data in databases to unstructured data such as text and multimedia. The data layer may be configured to emphasize scalability and security, so that as the fuzzy LIONN system grows and evolves, data integrity and accessibility are maintained.

Agentic AI System

[0078] Precision education represents a shift in medical education that draws parallels to the precision medicine approach of tailoring treatment to each patient's unique characteristics. Precision education personalizes learning by addressing the distinct needs, paces, and knowledge gaps of each student through a focus on data-driven methods. These data-driven methods may include transforming assessment data into actionable feedback with the objective of improving learning efficiency and student outcomes.

[0079] The agentic AI systems disclosed herein provide precision education by combining the adaptability and nuanced understanding of fuzzy logic with the generative capabilities of large language models to provide a flexible and responsive educational environment. Shapley values may also be used to quantify the contribution of each input element of a prompt, enabling a deeper insight into the learning process and ensuring that the agentic AI system's adaptability is grounded

in empirical evidence. The agentic AI system may be configured to dynamically adapt and fulfill the needs of individual students with precision and efficacy by providing the students with deliberate repetitive practice including immediate and actionable feedback that enables real-time realignment of individual education objectives.

[0080] The agentic AI systems may use generative AI to bridge knowledge gaps, and may provide precision education by harnessing the broad informational and processing capabilities of large language models to create adaptive, personalized learning experiences for medical students. These learning experiences may include simulated clinical scenarios that allow students to engage in a conversational manner with an AI-generated patient as part of their pre-clinical curriculum. This allows students to engage in deliberate practice of communication, application of knowledge, and clinical reasoning with each scenario varying in context to enhance learning. Embodiments of the agentic AI system may use large language models to provide reliable and valuable feedback to students, including feedback in domain-specific arenas. The immediate structured feedback provided to students based on their simulated clinical encounters may help students identify areas of weakness that need improvement.

[0081] A large language model is generally only as powerful as the information on which it has been trained. In rapidly evolving domains, a general-purpose large language model that has not been trained using domain-specific training data may provide less reliable content than a large language model which has been trained using domain-specific training data. A large language model may also reflect biases in the training data, and any such biases may be perpetuated through content generated by the model. Large language model bias may pose particular issues in educational systems by perpetuating unreasoned judgments, such as decisions based on race. A large language model may also be susceptible to “hallucinations”. Hallucinations is the term used for generated content that is factually incorrect, nonsensical, or disconnected from the original prompt provided to the model. The frequency of hallucinations has been found to increase with increasing task complexity. Thus, large language models used in precision education systems may be particularly susceptible to this problem. Because large language models utilize advanced, probabilistic algorithms far beyond simple binary logic, they may generate responses that appear reasonable but are not reliably verified or otherwise

explained. Coupled with this, the complex nature of large language models typically results in there being little or no transparency in the underlying decision-making process that resulted in the generation of content.

[0082] The lack of transparency and reliability of content generated by large language models is a potential issue for their use in education, where transparency in evaluation criteria and reasoning has been shown to improve understanding and learning outcomes. For many students, comprehending the rationale underlying a correct answer enables them to understand the ‘why’ that facilitates the development of critical thinking and the ability to apply knowledge in practical situations.

[0083] Additional potential issues with large language models may include the emergence of superintelligence. Such advancements could result in AI systems with capabilities far exceeding those of human intelligence, bringing both unprecedented benefits and complex risks. The ability of large language models to generate certain types of content (e.g., highly complex code) raises questions about the ability of humans to evaluate and trust the safety and objectives of advanced systems employing large language models. To address this issue, embodiments of the disclosed system may use one or more weak AI systems (e.g., a weak large language model) to train, supervise, and refine content generated by one or more strong AI systems (e.g., a strong large language model) used to correspond with students.

[0084] As used herein, the term “weak AI system” refers to a system having relatively narrow boundaries and an inability to modify itself based on external information. In contrast, the term “strong AI system” refers to an AI system having the ability to modify or otherwise update itself (i.e., “learn”) based on input from external sources, such as students or trainers. The stability of a weak large language model may enable it to manage a strong large language model in a manner that prevents the strong large language model from rapidly developing advanced capabilities (e.g., artificial general intelligence) or providing inappropriate content to students. The use of weaker large language models to manage stronger large language models may reduce the need for humans to train, supervise, and refine the strong large language models. The use of weak large language models to monitor strong large language models may also improve interpretability, refine alignment, and provide scalable oversight for accuracy assurance and resistance to misalignment of the strong large language model being monitored. Moreover, these

advantages may be provided with minimal impact on the overall capability of the strong large language model.

[0085] As used herein, the term “agentic AI system” refers to a system that includes one or more AI agents which are used to achieve complex tasks with limited human supervision. The use of multiple AI agents may include the use of one or more of weak-to-strong generalization and scalable oversight. The resulting agentic AI system may be capable of interpreting vast amounts of data to create a personalized and adaptive learning environment, taking part in every step of a precision education concept by tracking a student’s progress in real-time, adjusting to their needs, and offering immediate feedback. The learning experience may thereby be tailored to both the strengths of each individual student as well as any areas needing further development.

[0086] Fuzzy logic may be used to improve the learning experience by providing explainability and transparency to the decision-making process of the agentic AI system. Fuzzy logic is rooted in fuzzy set theory, and transcends the binary confines of conventional logic systems by embracing multi-valued reasoning. Fuzzy logic embodies the concept of partial truth, allowing for a more nuanced representation of the real world. This feature can make fuzzy logic well-suited to tackle the ambiguities and complexities inherent in human reasoning and decision-making processes. By integrating fuzzy logic with large language models, the agentic AI system may be configured to interpret the subtleties and variability inherent in educating humans. The integration of fuzzy logic and large language models may increase the reliability of AI-generated content as compared to AI systems using large language models without fuzzy logic. The use of fuzzy logic may also increase the probability that information and feedback presented to students is both contextually relevant and thoroughly vetted for accuracy and alignment with domain-specific expert opinions. To this end, fuzzy logic may be used filter the content generated by large language models to reduce the occurrence of hallucinations and other errors. Fuzzy logic may reduce the occurrence of hallucinations and other errors by providing an interpretive layer between the large language model and students. The fuzzy logic may then assess and verify the content generated by the large language model before the content is provided to the student.

[0087] Shapley values are derived from cooperative game theory, and may be used to provide a systematic way to assign a contribution score to the input elements

of prompts provided to a large language model. By calculating the marginal contributions of individual input elements, Shapley values may provide a quantifiable measure of the impact and relevance of individual input elements on the content generated by the large language model in response to receiving the prompt.

Shapley values may thereby be used in the agentic AI system to improve the interpretability and explainability of content generated by the large language model.

[0088] Integrating Shapley values and fuzzy logic with large language model-based AI systems may enable embodiments of the agentic AI system to tackle some of the challenges associated with the evolving capabilities of large language models. A combination of Shapley values and fuzzy logic may be used to provide the agentic AI system with reliable oversight of how the large language model interprets input received from students and feedback provided to students. Fuzzy logic may be particularly applicable to the processing of subjective assessment criteria, and specifically to assessments generated by human trainers. Shapley values may be used to complement the fuzzy logic by providing clear, quantitative insights into the decision-making processes of the large language model. This dual approach may improve the agentic AI system's ability to interpret content, enabling the system to discern the weight and relevance of various educational inputs and reducing the need for human oversight. Embodiments of the agentic AI system may thereby capitalize on the strengths of both fuzzy logic and Shapley values to refine the reliability and clarity of large language models in interpreting complex cognitive processes used to generate workplace-based assessments (such as decision-making, reasoning, and communication skills) that avoid the subjectivity of human judgment.

[0089] The agentic AI system may enhance precision education by autonomously performing diverse tasks with limited supervision, and may include features that enforce alignment strategies. Enforcement of alignment strategies may ensure the content provided to students is aligned with human values and educational standards, even if the underlying AI advances towards artificial general intelligence (AGI) and artificial superintelligence (ASI). Challenges proximal to the need for alignment may include overcoming weak-to-strong generalization for reliable autonomy, ensuring transparency and interpretability for trust and reliability, scaling to AI-dominant oversight models for sustainability, and securing systems against adversarial threats. Alignment strategies may also be configured to safeguard the

integrity and efficacy of the agentic AI system and maintain alignment of the system with educational goals and ethical standards as the underlying AI technology evolves.

[0090] FIG. 2 depicts an exemplary agentic AI system 200 in accordance with an embodiment of the invention. The agentic AI system 200 includes a user interface 202, one or more educational modules 204 (e.g., four educational modules), an input layer 206, a content generating large language model 208, an oversight layer 210 having one or more agents 212-214 (e.g., three agents), an aggregation layer 216, and an output layer 218. In operation, a system user 220 (e.g., a student) may select an educational module 204 representing an area they wish to study. In response to the user 220 selecting the educational module 204, the input layer 206 may provide a prompt to the large language model 208 indicative of the selection. The prompt may include content (e.g., text) that prompts the large language model 208 to generate a response in accordance with the selected educational module 204. For example, the prompt may cause the large language model 208 to generate a response including content that describes an educational scenario in accordance with the content included in the prompt, e.g., a clinical scenario involving a patient.

[0091] As part of the process of generating a response, the large language model 208 may partition the prompt into a plurality of input elements. The input elements may be referred to as “tokens”, and the process of partitioning the content of the prompt may be referred to as “tokenization”. Each token may comprise a portion of the prompt (e.g., a sequence of one or more text characters) that is converted into a numeric value for processing by the large language model 208. The large language model 208 may apply a statistical analysis to the tokens that generates a response to the prompt. For example, certain numbers of tokens may be grouped together into strings, sometimes referred to as “ n -grams”, where n represents the number of tokens in the n -gram. These n -grams may be subjected to one or more statistical analyses that generate content. The statistical analyses may be performed using one or more neural networks that have been trained using a large amount of training data. The generated content may be assembled into a response to the prompt, e.g., a string of text characters, n -grams, words, phrases, sentences, etc. The resulting response may represent what the large language model 208 has determined would be the most likely response based on statistics extracted from the training data. Language models may be classified based on the longest n -gram they can process,

with large language models generally referring to language models that can process n -grams with values of n greater than three, e.g., $n \geq 4$.

[0092] The “raw content” generated by the large language model 208 may be processed by the oversight layer 210 to generate “filtered content”. If the raw content is not flagged by one or more of the agents 212-214 as being inappropriate, the filtered content may be provided to the output layer 218 for display by the user interface 202. In cases where the raw content is closely aligned with approved standards, the filtered content may be essentially unchanged from the raw content. In other cases, the raw content may be modified by the agents 212-214 so that the filtered content is consistent with approved terminology or standards. Each agent 212-214 may interpret the raw content according to how the agent 212-214 is configured, and generate filtered content (or flag content for human review) based on this interpretation. The oversight layer 210 may thereby review and (if necessary) modify the raw content for accuracy and reliability. The filtered content may then be consolidated and verified for educational content by the aggregation layer 216 before being provided to the output layer 218 for display to the user 220.

[0093] Input received from the user 220 (e.g., questions, proposed actions, etc.) in response to the scenario presented by the filtered content may be provided to the large language model 208 via the input layer 206. The large language model 208 may then generate new raw content (e.g., module content, feedback, or assessments) based on the user input received via the input layer 206 and the filtered content displayed to the user 220. This new filtered content may be consistent with both the previous filtered content and the input received from the user 220. The agentic AI system 200 may thereby step the user 220 through the scenario in stages by repeatedly receiving input from the user 220 and updating the scenario in accordance with the input using the large language model 208.

[0094] The oversight layer 210 of agentic AI system 200 may flag portions of the raw content that have been interpreted as being inaccurate, unreliable, or otherwise inappropriate. This flagged content may be forwarded to a domain agent (not shown) for review by a human domain expert 222, who may in turn provide iterative feedback and model tuning 224. This feedback and tuning may be based on decisions by the domain expert 222 about how to adjust one or more of the input layers 206, large language model 208, and oversight layer 210 to avoid generation of inappropriate content or false flagging of appropriate content under similar

circumstances in the future. This human review of content may enable the agentic AI system 200 to become more finely tuned over time. Advantageously, the oversight layer 210 may both reduce the need for human intervention and be tuned by feedback from the domain expert 222. The interactive feedback provided by these feedback loops may, over time, continue to improve the quality of the raw content generated by the large language model 208 and reduce the need for human review.

[0095] The above-described filtered content generation process may capitalize on the strongest aspects of explainable AI (e.g., fuzzy logic and Shapley values), and integrate the explainable AI into a plurality of agents 212-214 (e.g., a supervision agent 212, an interpretability agent 213, and an oversight agent 214). The supervision agent 212 may tokenize the raw content generated by the large language model 208 and provide the tokens to a weak large language model. The weak large language model may replace any suspect or prohibited terminology in the raw content as well as flag the raw content if any inappropriate content is detected. The interpretability agent 213 may tokenize the raw content and apply a set of fuzzy logic rules to the resulting raw content tokens to generate filtered content, identify raw content to flag, or both generate and flag content. The oversight agent 214 may determine Shapley values for one or more of the prompt tokens generated by the large language model 208. These Shapley values may provide an indication of the impact of the prompt tokens on the raw content.

[0096] The agents 212-214 may work cooperatively to process raw content generated by the large language model 208, and contribute to an iterative process that refines the performance of the large language model 208 by flagging inappropriate content for review by the domain expert 222. This iterative feedback and model tuning 224 may enable fine-tuning of the agentic AI system 200 while minimizing the need for constant human oversight. Advantageously, the architecture of the depicted agentic AI system 200 may facilitate alignment at scale as the capabilities of the agentic AI system 200 increase, ensuring the content provided to the user 220 remains consistent with educational objectives.

[0097] FIG. 3 depicts a flowchart of an exemplary process 300 that may be implemented using the agentic AI system 200. In block 302, the process 300 may receive an indication from the user interface 202 that the user 220 has selected an educational module 204. In response, the process 300 may proceed to block 304

and generate a clinical scenario in accordance with the selected educational module 204. The clinical scenario may be generated by the large language model 208 based on a prompt received via the input layer 206, or the clinical scenario may be selected from a plurality of previously defined clinical scenarios. In cases where the clinical scenario is generated on demand by the large language model 208, the prompt may include keywords, a fact pattern, or other content that prompts the large language model 208 to generate a clinical scenario consistent with the area of study covered by the educational module 204 in question. The clinical scenario generated by the large language model 208 may describe, for example, a patient and a set of symptoms presented by the patient, or some other suitable fact pattern.

[0098] If the clinical scenario is previously defined rather than generated on demand, the process 300 may proceed directly to block 306 and display the scenario to the user 220. If the clinical scenario is generated by the large language model 208, the process 300 may instead proceed to block 308 and determine if the clinical scenario is appropriate. The determination process may include providing the raw content generated by the large language model 208 to each of the agents 212-214. Each agent 212-214 may then generate a response based on the raw content indicating whether the clinical scenario is appropriate or inappropriate. The process 300 may determine, for example, that the clinical scenario is appropriate if no agents generate an inappropriate verdict, and inappropriate if one or more agents 212-214 determines the clinical scenario is inappropriate. In an alternative embodiment, the process 300 may merely require a majority of the agents 212-214 to indicate the raw content is appropriate rather than a unanimous verdict.

[0099] If the clinical scenario is inappropriate ("NO" branch of decision block 308), the process 300 may proceed to block 310, flag the clinical scenario for review by a human domain expert 222, and return to block 304 to generate a new clinical scenario. If the clinical scenario is appropriate ("YES" branch of decision block 306), the process 300 may proceed to block 308 and display the clinical scenario to the user 220.

[00100] In block 312, the process 300 may receive input from the user 220 in the form of a response to the clinical scenario. This input may include, for example, questions for a patient described by the clinical scenario, physical examination steps, tests to run on the patient, or any other suitable actions that might be taken by a physician presented with the clinical scenario in question.

[00101] In response to receiving the input from the user 220, the process 300 may proceed to block 314 and update the scenario. The scenario may be updated, for example, by providing the current scenario and input from the user 220 to the large language model 208, thereby prompting the large language model 208 to generate new raw content based thereon. The process 300 may then proceed to block 315 and process the new raw content through the oversight layer 210 as described above. If the clinical scenario is inappropriate (“NO” branch of decision block 315), the process 300 may proceed to block 316, flag the clinical scenario for review by a human domain expert 222, and return to block 314 to generate a new updated clinical scenario. If the new raw content is determined to be appropriate (“YES” branch of decision block 315), the process may proceed to block 320 and display the updated scenario to the user 220. The updated scenario may have an updated fact pattern corresponding to a likely outcome determined by the large language model 208 based on the previous scenario fact pattern and actions taken by the user 220.

[00102] In block 322, the process 300 may determine if the scenario has been resolved, e.g., whether the user 220 has correctly diagnosed the patient and prescribed an appropriate course of treatment. If the clinical scenario has not been resolved (“NO” branch of decision block 322), the process 300 may proceed to block 312 and await additional input from the user 220. The steps of receiving user input 312, updating the scenario 314-316, and displaying the updated scenario 320 to the user 220 may be iterated a plurality of times until either the scenario is resolved (e.g., the patient is properly diagnosed and prescribed an appropriate treatment), or the user 220 decides to end the scenario. The process 300 may also allow the user 220 to pause the training session, and return to it at a later time.

[00103] If the scenario has been resolved (“YES” branch of decision block 322), the process 300 may proceed to block 324 and generate a report. The report may include a session summary or transcript, feedback on how well the user 220 performed at one or more steps of the training session, suggestions on how the user 220 could have handled certain aspects of the training session differently, or any other suitable feedback. Feedback may be generated during the session (e.g., by one or both of the large language model 208 and oversight layer 210) and provided to the user 220 at the end of the session. For example, for each updated scenario, the large language model 208 may be prompted to generate a model response by the user 220. This model response could then be added to a training record for

comparison to the user's actual response as part of the above evaluation of the user 220.

[00104] Reinforcement learning from human feedback is a technique that may be used to train a large language model. Human feedback may be used to shape a large language model by reinforcing human-endorsed behaviors and discouraging undesirable behaviors. This type of feedback may guide large language models from human-led (strong) to independent (weak) generalization. However, reinforcement learning from human feedback is typically not scalable. This lack of scalability makes adapting large language models in evolving fields (such as medical education) difficult due to the breadth and continual evolution of the field exceeding what can be directly evaluated by humans.

[00105] By way of example, consider the creation of clinical scenarios from start to finish. Each scenario may involve thousands of variables and outcomes, reflecting the intricate nature of real-world medical cases. One challenge is thus making sure the clinical scenarios are all aligned with educational objectives, accurately depict medical situations, and adhere to safety protocols without human domain experts reviewing every detail of each scenario line by line.

[00106] By integrating fuzzy logic into the interpretability agent 213 of oversight layer 210, the agentic AI system 200 may refine data alignment before fine-tuning advanced models. The interpretability agent 213 may clarify inconsistencies and reduce ambiguity, thus improving the quality of supervision for advanced models so that the models surpass their training limits. This feature of the agentic AI system 200 may compensate for the effects of low-quality training data and expand the capabilities of the large language model 208 beyond those provided by its initial training.

[00107] Fuzzy logic may be specifically suited for the above issues due to its ability to model and interpret the uncertainty and vagueness that typify human observations and decision-making processes. Unlike traditional binary logic systems, which rigidly classify inputs as true or false, fuzzy logic accommodates degrees of truth. This flexibility may be particularly advantageous in medical education, where the assessment of competencies like clinical reasoning and decision-making involves substantial ambiguity, idiosyncrasies, and subjective judgment. The fuzzy logic of interpretability agent 213 may enable the agentic AI system 200 to interpret and navigate the inherent imprecision of human feedback more accurately than

conventional AI systems, resulting in a more accurate and nuanced understanding of the desired outcomes.

[00108] The agentic AI system 200 may use a multi-pronged strategy for employing fuzzy logic to enhance system proficiency under weak supervision conditions. One prong may employ fuzzy logic to model the gradations of correctness and relevance of student responses. This modelling of gradations may enable a more sophisticated evaluation of student work product, and acknowledges that in the realm of clinical reasoning, answers are not strictly right or wrong, but rather often lie on a spectrum. Another prong may use fuzzy logic to enable the development of AI-driven assessment tools that provide tailored feedback to students. This feedback may identify both areas of deficiency and provide guidance to the student on how to approach learning tasks more effectively. Fuzzy logic may thereby facilitate a dynamic learning environment in which the agentic AI system 200 adjusts educational content in real time, based on the evolving needs and competencies of each student, thereby enabling deliberate practice in medical education.

[00109] Interpretability of large language model content may be advantageous in educational contexts, such as medical training, because learning is typically improved when educators and students trust and comprehend the logic behind AI-generated suggestions. This trust may be engendered, for example, by educators and students being able to see for themselves that the logic behind the AI-generated suggestions is both precise and ethically responsible. The inherent complexity of large language models often obfuscates the reasoning behind their generated content, posing a challenge to users who would benefit from trusting these systems to make accurate and ethically sound decisions.

[00110] The use of multiple agents 212-214 may enhance the performance of the large language model 208 by providing a diversity of interpretations and analytic processes. In an embodiment of the invention, the agents 212-214 may be configured to emphasize the critical evaluation of content reliability and quality rather than focus solely on performance improvement. To this end, one or more agents 212-214 of oversight layer 210 may analyze the content generated by the large language model 208 for accuracy, robustness, and ethics using fuzzy logic to generate a confidence rating. This collective review of the content may improve content quality. The agents 212-214 may thereby function as mediators, making the

analysis of student responses provided by the large language model 208 clear and understandable for educators and students, and providing a layer of interpretability and trust unattainable by conventional large language models.

[00111] Shapley values may also increase the interpretability of the content generated by the large language model 208 by quantifying the impact of individual input elements on generated content. This quantification may provide insights into what drives decisions made by the large language model 208. Using Shapley values to interpret the raw content generated by the large language model 208 may improve the interpretability of the filtered content in multiple ways. For example, Shapley values may facilitate refinement of student prompts by determining the influence of different input elements on the output of the large language model 208, and directing the large language model 208 toward input elements with higher educational impact. The use of Shapley values may facilitate improvements in the large language model 208 so that it generates content with increased relevance and that is conducive to deeper learning.

[00112] The use of Shapley values may also enable adaptive prompting. Adaptive prompting may in turn enable the agentic AI system 200 to modify prompts in real-time based on student responses. This dynamic approach may address individual learning needs and clarify areas of confusion, personalizing the learning experience. Shapley values may provide educators with deeper insights into the development of complex competencies, such as clinical decision-making. By identifying key factors that influence these processes, educators may better understand and teach critical skills. Shapley values may also inform the creation of personalized learning paths by the agentic AI system 200. These personalized learning paths may highlight specific topics or concepts on which a student needs to concentrate. This tailored approach may promote alignment of educational content with each student's unique learning experience, thereby providing a more effective and individualized education.

[00113] Scalable oversight may complement weak supervision by adding an additional layer of validation to the filtering of the content generated by the large language model 208. Weak supervision may be used to align the large language model 208 via fine-tuning, while scalable oversight may employ an external agent to check content to ensure it meets specific goals or tasks. Scalable oversight may be implemented by breaking down tasks into simpler subtasks or by application of the principle that evaluating content is typically simpler than creating content. For

example, assessing student feedback on nuanced and subjective topics may surpass the capabilities of the agents 212-214, even when guided by clear criteria, to accurately evaluate the output of the large language model 208.

[00114] Integration of fuzzy logic may address the above concept of scalable oversight as it applies to the complexity of medical education and aligning large language model generated content with human standards. Embodiments of the agentic AI system 200 may include a feedback loop where an interpretability agent 213 performs initial assessments of raw content and flags content that is misaligned. Flagged content may be reviewed by a human domain expert 222, whose feedback may be used to refine the parameters of the interpretability agent 213. This scalable oversight may be implemented using a multifaceted approach that includes initial content assessment, human domain expert review and feedback, and iterative refinement.

[00115] Initial content assessment may be implemented by using the interpretability agent 213 to evaluate the raw content generated by the large language model 208. The interpretability agent 213 may assign truth values to the content based on criteria like accuracy, relevance, and educational objectives. These criteria may be initially informed by human assessments that are used to train the interpretability agent 213. Content flagged by fuzzy logic as uncertain or irrelevant may be reviewed by human experts, whose feedback may be used to refine the fuzzy logic rules or tune the large language model 208 to improve the accuracy and reliability of the agentic AI system 200. Expert human feedback may be used to iteratively update the fuzzy logic evaluation criteria so that the agentic AI system 200 adapts to changing educational standards and content complexities.

[00116] The interpretability agent 213 may facilitate scalable oversight by automating the initial stages of content evaluation. This automation may enable an increased volume of raw content generated by the large language model 208 to be evaluated by effectively filtering out and flagging content for human review while allowing validated content to be used without human review. Fuzzy logic may thereby enable scalable oversight and maintain content quality and relevance in the agentic AI system 200. As the agentic AI system 200 is expanded across various subjects and complexities, the interpretability agent 213 may provide a systematic and scalable solution for content evaluation that supports personalized learning and maintains high educational standards for medical training.

Methods of Evaluating the Agentic AI System

[00117] Evaluation of large language models include a human-large language model collaboration that combines human insight with AI efficiency. Human-large language model collaboration may enable swift data assessment, beyond human capacity alone, while ensuring human oversight. This method of evaluation may capitalize on human expertise for large language model alignment, and may be advantageous in educational environments where subjective judgments and consensus are standard for evaluating competencies. Additionally, by balancing human subjectivity with AI objectivity, incorporating collaborative large language models may reduce potential biases in evaluations of student work by the agentic AI system.

[00118] One type of human-large language model collaboration that may be used to evaluate the agentic AI system 200 is known as CoEVAL, which is described in detail by *Collaborative Evaluation: Exploring the Synergy of Large Language Models and Humans for Open-ended Generation Evaluation*, Qintong Li et al., arXiv:2310.19740 (30 Oct 2023). CoEVAL blends large language model ideation with human oversight to generate refining assessments. This evaluation may be tailored for tasks with specific criteria like factual accuracy, clarity, logical coherence, input relevance, and tone. CoEVAL may complement fuzzy logic's goal of enhancing explainability and oversight in the agentic AI system, and demonstrate consistency in evaluation outputs and alignment when compared to human evaluators acting alone.

[00119] Referring now to FIG. 4, the evaluation process 400 for the agentic AI system 200 may be implemented in multiple steps. Initially, the outputs of the agentic AI system 200 may be compared with human assessments to gauge the system's current alignment with human judgment standards. This step may set a baseline for measuring the effects of integrating fuzzy logic on aspects like weak-to-strong generalization and improving interpretability.

[00120] Scalable oversight refers to the process of assessing an agent designed to excel in tasks beyond the capabilities of its overseeing agent. Additional performance metrics may include scalable oversight, which may be evaluated using a tailored approach due to the unique challenge of evaluating scalable oversight. The use of scalable oversight as a system evaluation strategy is described in more detail below.

[00121] Integrating fuzzy logic may enhance weak-to-strong generalization, enabling a more advanced model to perform its tasks more efficiently with less or more generalized training data, thanks to the increased explainability that fuzzy logic offers. To evaluate the effectiveness of the interpretability agent 213, the agentic AI system 200 may be engaged with an array of prompts, both with and without the interpretability agent 213 being activated, and monitoring the responses of the large language model 208 through the oversight layer 210. This method may allow evaluation of the ability of the agentic AI system 200 in weak-to-strong generalization, with the objective being performance that aligns or surpasses the baseline standards set by human judgment. An enhancement in accuracy and interpretability of the responses of the large language model 208 to the prompts may be quantified, akin to models trained on expansive datasets. One metric indicative of an improvement provided by the interpretability agent 213 may be improvement in the quality and relevance of the responses generated by the large language model 208. This type of improvement may demonstrate the effectiveness of fuzzy logic in refining the output of the large language model 208 without extensive training data.

[00122] Integrating the CoEVAL approach into evaluation process 400 may enhance the evaluation methodology by incorporating human-large language model collaboration in a way specifically designed to tackle the challenges of interpreting content generated by large language models. This approach may effectively address the subjective nature of interpretability, which is a measure of appropriateness as judged by humans, and directly addresses limitations posed by the vast amount of data and the finite capacity of human evaluative efforts to refine the interpretability of the agentic AI system 200 and ensure its technical accuracy.

[00123] The evaluation process 400 may begin by establishing a consensus among human experts on the core criteria for interpretability. This consensus may include metrics such as accuracy, understandability, and efficiency, while also permitting the addition of criteria to augment the educational value of content generated by the agentic AI system 200. Human experts may also identify additional factors to enhance the educational impact of large language model generated explanations. Shapley values may be used to assign weights to these interpretability criteria, quantifying each criterion's contribution to overall interpretability. This approach may provide a balanced and mathematically grounded framework for

evaluating the content generated by the agentic AI system 200, and reflect the complex expectations inherent in educational applications of AI.

[00124] The role of CoEVAL in the evaluation process may include leveraging large language models for the initial ideation of evaluation criteria, subsequently refined through meticulous human scrutiny. This hybrid evaluation mechanism may facilitate rapid data assessment beyond what would be feasible through human efforts alone, while maintaining human insights to align large language model outputs with nuanced educational standards.

[00125] Benchmarking the performance of the agentic AI system 200 against the predefined criteria, both with and without the integration of fuzzy logic and under the guidance of the CoEVAL framework, may provide an understanding of fuzzy logic's impact on interpretability. This evaluation may also explore the efficacy of applying fuzzy logic at various stages of data handling—namely, pre-processing input data and post-processing large language model output data. This structured approach may systematically assess how fuzzy logic integration influences the interpretability of the agentic AI system 200, providing valuable insights into optimizing educational content delivery.

[00126] The evaluation of scalable oversight presents a challenge, especially when applied to subjective areas like medical education. One approach is to use a “sandwiching” technique, which compares the performance of non-experts, both unassisted and assisted by AI, against that of domain experts. This approach assesses the ability of the AI to elevate non-expert task performance to the level of experts, providing a measure of the AI agent's alignment effectiveness.

[00127] A type of sandwiching approach may be used to assess the impact of integrating fuzzy logic on the performance of the agents 212-214. Task completions may be conducted across various medical education domains—such as medical knowledge and clinical reasoning. These tasks may be performed by non-experts unassisted, non-experts assisted by the oversight layer 210 without the interpretability agent 213, and non-experts assisted by the oversight layer 210 with the interpretability agent 213. The results may then be compared to the performance of domain experts in medical education.

[00128] The above comparative analysis may provide insights into the ability of the oversight layer 210 to improve non-expert performance towards expert-level competence, thereby providing a quantitative metric for performance improvement.

A performance gap recovery value of zero may indicate no improvement over unassisted non-experts, while a performance gap recovery value of one may signify alignment with domain-expert performance levels. The performance gap recovery value may thereby quantify the potential efficacy of fuzzy logic integration in enhancing the supervisory capabilities of the agentic AI system 200 in the context of medical education.

Conclusion

[00129] The integration of fuzzy logic and Shapley values into an agentic AI system may improve medical education and assessment by facilitating personalized, timely feedback to students and leveraging advanced data integration for a comprehensive assessment of student capabilities. An initial step in assessing the impact of this innovation may include evaluating its educational value from the students' perspective, analyzing objective performance metrics pre- and post-implementation, and correlating the outcomes from the agentic AI system with conventional assessment methods. However, these metrics alone may not fully capture the efficacy and potential improvements in medical education and assessment.

[00130] One challenge in evaluating the performance of the agentic AI system 200 lies in the lack of objective standards for comparison. Beyond standardized tests, the assessment of medical students mostly involves a group consensus by domain experts after triangulating many subjective data points. The use of fuzzy logic and AI in this context introduces the capability to process and analyze more data than traditional methods across multiple contexts. This may lead to a richer and more reliable understanding of student capabilities and enhanced performance.

[00131] Assessing learning outcomes presents its own set of challenges due to the subjective and complex nature of medical training environments. The synergy between fuzzy logic and the agentic AI system 200 may provide a unique opportunity to track student performance longitudinally and potentially create learning curves, offering a nuanced view of competency achievement over time. This longitudinal monitoring may facilitate a more targeted and individualized educational approach, with the duration to competency thresholds serving as one potential proxy for learning outcomes.

[00132] Measuring outcomes in patient care may pose a more formidable challenge. Patient care may represent the ultimate metric for evaluating the real-

world impact of the agentic AI system 200 on medical education. Potential gains made with the agentic AI system 200 may make medical students experience in the clinical environment more efficient, shortening curricular components and removing certain one-size-fits all elements that are no longer necessary. Additionally, the individualized approach that the agentic AI system 200 provides to optimize the clinical reasoning of students has the potential to minimize cognitive errors, thereby reducing a significant source of morbidity and mortality in healthcare.

[00133] Finally, the agentic AI system 200 may optimize resource utilization in the remediation of medical students. Educational resources are typically limited, but the individualized and precise nature of the agentic AI system 200 may enable educators to consolidate efforts around students that need the most help. The agentic AI system 200 provides an educational environment that correlates nuanced aspects of patient care—often overlooked in traditional education settings—with student performance in clinical environments, offering a new avenue for research. This research could explore how clinical environment data can be integrated into feedback for students and physicians, promoting behaviors aligned with enhanced patient outcomes.

[00134] Referring now to FIG. 5, embodiments of the invention described above, or portions thereof, may be implemented using one or more computer devices or systems, such as exemplary computer 500. The computer 500 may include a processor 502, a memory 504, an input/output (I/O) interface 506, and a Human Machine Interface (HMI) 508. The computer 500 may also be operatively coupled to one or more external resources 510 via a network 512 or I/O interface 506. External resources may include, but are not limited to, servers, databases, mass storage devices, peripheral devices, cloud-based network services, or any other resource that may be used by the computer 500.

[00135] The processor 502 may include one or more devices selected from microprocessors, micro-controllers, digital signal processors, microcomputers, central processing units, field programmable gate arrays, programmable logic devices, state machines, logic circuits, analog circuits, digital circuits, or any other devices that manipulate signals (analog or digital) based on operational instructions stored in memory 504. Memory 504 may include a single memory device or a plurality of memory devices including, but not limited to, read-only memory (ROM), random access memory (RAM), volatile memory, non-volatile memory, static random

access memory (SRAM), dynamic random access memory (DRAM), flash memory, cache memory, or data storage devices such as a hard drive, optical drive, tape drive, volatile or non-volatile solid state device, or any other device capable of storing data.

[00136] The processor 502 may operate under the control of an operating system 514 that resides in memory 504. The operating system 514 may manage computer resources so that computer program code embodied as one or more computer software applications, such as an application 516 residing in memory 504, may have instructions executed by the processor 502. In an alternative embodiment, the processor 502 may execute the application 516 directly, in which case the operating system 514 may be omitted. One or more data structures 518 may also reside in memory 504, and may be used by the processor 502, operating system 514, or application 516 to store or manipulate data.

[00137] The I/O interface 506 may provide a machine interface that operatively couples the processor 502 to other devices and systems, such as the external resource 510 or the network 512. The application 516 may thereby work cooperatively with the external resource 510 or network 512 by communicating via the I/O interface 506 to provide the various features, functions, applications, processes, or modules comprising embodiments of the invention. The application 516 may also have program code that is executed by one or more external resources 510, or otherwise rely on functions or signals provided by other system or network components external to the computer 500. Indeed, given the nearly endless hardware and software configurations possible, persons having ordinary skill in the art will understand that embodiments of the invention may include applications that are located externally to the computer 500, distributed among multiple computers or other external resources 510, or provided by computing resources (hardware and software) that are provided as a service over the network 512, such as a cloud computing service.

[00138] The HMI 508 may be operatively coupled to the processor 502 of computer 500 to allow a user to interact directly with the computer 500. The HMI 508 may include video or alphanumeric displays, a touch screen, a speaker, and any other suitable audio and visual indicators capable of providing data to the user. The HMI 508 may also include input devices and controls such as an alphanumeric keyboard, a pointing device, keypads, pushbuttons, control knobs, microphones,

etc., capable of accepting commands or input from the user and transmitting the entered input to the processor 502.

[00139] A database 520 may reside in memory 504, and may be used to collect and organize data used by the various systems and modules described herein. The database 520 may include data and supporting data structures that store and organize the data. In particular, the database 520 may be arranged with any database organization or structure including, but not limited to, a relational database, a hierarchical database, a network database, or combinations thereof. A database management system in the form of a computer software application executing as instructions on the processor 502 may be used to access the information or data stored in records of the database 520 in response to a query, which may be dynamically determined and executed by the operating system 514, other applications 516, or one or more modules.

[00140] In general, the routines executed to implement the embodiments of the invention, whether implemented as part of an operating system or a specific application, component, program, object, module or sequence of instructions, or a subset thereof, may be referred to herein as “program code.” Program code typically comprises computer-readable instructions that are resident at various times in various memory and storage devices in a computer and that, when read and executed by one or more processors in a computer, cause that computer to perform the operations necessary to execute operations or elements embodying the various aspects of the embodiments of the invention. Computer-readable program instructions for carrying out operations of the embodiments of the invention may be, for example, assembly language, source code, or object code written in any combination of one or more programming languages.

[00141] Various program code described herein may be identified based upon the application within which it is implemented in specific embodiments of the invention. However, it should be appreciated that any particular program nomenclature which follows is used merely for convenience, and thus the invention should not be limited to use solely in any specific application identified or implied by such nomenclature. Furthermore, given the generally endless number of manners in which computer programs may be organized into routines, procedures, methods, modules, objects, and the like, as well as the various manners in which program functionality may be allocated among various software layers that are resident within a typical computer

(e.g., operating systems, libraries, API's, applications, applets, etc.), it should be appreciated that the embodiments of the invention are not limited to the specific organization and allocation of program functionality described herein.

[00142] The program code embodied in any of the applications/modules described herein is capable of being individually or collectively distributed as a computer program product in a variety of different forms. In particular, the program code may be distributed using a computer-readable storage medium having computer-readable program instructions thereon for causing a processor to carry out aspects of the embodiments of the invention.

[00143] Computer-readable storage media, which is inherently non-transitory, may include volatile and non-volatile, and removable and non-removable tangible media implemented in any method or technology for storage of data, such as computer-readable instructions, data structures, program modules, or other data. Computer-readable storage media may further include RAM, ROM, erasable programmable read-only memory (EPROM), electrically erasable programmable read-only memory (EEPROM), flash memory or other solid state memory technology, portable compact disc read-only memory (CD-ROM), or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium that can be used to store data and which can be read by a computer. A computer-readable storage medium should not be construed as transitory signals per se (e.g., radio waves or other propagating electromagnetic waves, electromagnetic waves propagating through a transmission media such as a waveguide, or electrical signals transmitted through a wire). Computer-readable program instructions may be downloaded to a computer, another type of programmable data processing apparatus, or another device from a computer-readable storage medium or to an external computer or external storage device via a network.

[00144] Computer-readable program instructions stored in a computer-readable medium may be used to direct a computer, other types of programmable data processing apparatuses, or other devices to function in a particular manner, such that the instructions stored in the computer-readable medium produce an article of manufacture including instructions that implement the functions, acts, or operations specified in the text of the specification, the flowcharts, sequence diagrams, or block diagrams. The computer program instructions may be provided to one or more

processors of a general purpose computer, a special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the one or more processors, cause a series of computations to be performed to implement the functions, acts, or operations specified in the text of the specification, flowcharts, sequence diagrams, or block diagrams.

[00145] The flowcharts and block diagrams depicted in the figures illustrate the architecture, functionality, or operation of possible implementations of systems, methods, or computer program products according to various embodiments of the invention. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function or functions.

[00146] In certain alternative embodiments, the functions, acts, or operations specified in the text of the specification, the flowcharts, sequence diagrams, or block diagrams may be re-ordered, processed serially, or processed concurrently consistent with embodiments of the invention. Moreover, any of the flowcharts, sequence diagrams, or block diagrams may include more or fewer blocks than those illustrated consistent with embodiments of the invention. It should also be understood that each block of the block diagrams or flowcharts, or any combination of blocks in the block diagrams or flowcharts, may be implemented by a special purpose hardware-based system configured to perform the specified functions or acts, or carried out by a combination of special purpose hardware and computer instructions.

[00147] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the embodiments of the invention. As used herein, the singular forms “a”, “an” and “the” are intended to include both the singular and plural forms, and the terms “and” and “or” are each intended to include both alternative and conjunctive combinations, unless the context clearly indicates otherwise. It will be further understood that the terms “comprises” or “comprising,” when used in this specification, specify the presence of stated features, integers, actions, steps, operations, elements, or components, but do not preclude the presence or addition of one or more other features, integers, actions, steps, operations, elements, components, or groups thereof. Furthermore, to the

extent that the terms “includes”, “having”, “has”, “with”, “comprised of”, or variants thereof are used in either the detailed description or the claims, such terms are intended to be inclusive in a manner similar to the term “comprising”.

[00148] While all the invention has been illustrated by a description of various embodiments, and while these embodiments have been described in considerable detail, it is not the intention of the Applicant to restrict or in any way limit the scope of the appended claims to such detail. Additional advantages and modifications will readily appear to those skilled in the art. The invention in its broader aspects is therefore not limited to the specific details, representative apparatus and method, and illustrative examples shown and described. Accordingly, departures may be made from such details without departing from the spirit or scope of the Applicant's general inventive concept.

CLAIMS

What is claimed is:

1. A method for providing an educational environment using an agentic artificial intelligence system, comprising:
 - receiving a first prompt;
 - generating a first response based on the first prompt using a first large language model, the first response including first raw content that defines a first clinical scenario consistent with the first prompt;
 - interpreting the first raw content by using fuzzy logic to generate at least a first portion of a first filtered content based on the first raw content; and
 - displaying the first filtered content to a system user.
2. The method of claim 1, further comprising:
 - receiving an input from the system user responsive to the first filtered content;
 - generating a second prompt based on the input from the system user;
 - providing the second prompt to the first large language model;
 - generating a second response to the second prompt using the first large language model, the second response including second raw content that defines a second clinical scenario consistent with both the first clinical scenario and the second prompt;
 - interpreting the second raw content by using the fuzzy logic to generate a second filtered content based on the second raw content; and
 - displaying the second filtered content to the system user.
3. The method of claim 1, wherein generating the at least the first portion of the first filtered content includes:
 - tokenizing the first raw content to generate a plurality of raw content tokens;
 - and
 - applying a set of fuzzy logic rules to the raw content tokens to generate the at least the first portion of the first filtered content.

4. The method of claim 1, wherein generating the first response includes tokenizing the prompt to generate a plurality of prompt tokens, and further comprising:
 - determining a Shapley value for each prompt token that quantifies a contribution of the prompt token to the raw content.
5. The method of claim 1, further comprising:
 - forwarding the first raw content to a domain agent for human review in response to the fuzzy logic interpreting the first raw content as including at least some uncertain content or some irrelevant content.
6. The method of claim 1, wherein generating the first filtered content further comprises:
 - using a second large language model to interpret the first raw content and generate at least a second portion of the first filtered content.
7. The method of claim 6, wherein the second large language model interprets the first raw content by:
 - tokenizing the first raw content into a plurality of raw content tokens; and
 - generating the at least the second portion of the first filtered content by statistically analyzing the raw content tokens.
8. The method of claim 7, wherein each of the first large language model and the second large language model are configured to analyze strings of tokens, and a maximum length of the strings of tokens analyzed by the first large language model is larger than the maximum length of the strings of tokens analyzed by the second large language model.
9. The method of claim 7, wherein the first large language model is a strong large language model, and the second large language model is a weak large language model.
10. The method of claim 6, further comprising:
 - aggregating the first and second portions of the first filtered content.

11. An agentic artificial intelligence system for providing an educational environment, comprising:
- one or more processors; and
 - a memory coupled to the one or more processors and including program code that, when executed by the one or more processors, causes the system to:
 - receive a first prompt;
 - generate a first response based on the first prompt using a first large language model, the first response including first raw content that defines a first clinical scenario consistent with the first prompt;
 - use an interpretability agent including fuzzy logic that interprets the first raw content to generate at least a first portion of a first filtered content based on an interpretation of the first raw content by the fuzzy logic; and
 - display the first filtered content to a system user.
12. The system of claim 11, wherein the program code further causes the system to:
- receive an input from the system user responsive to the first filtered content;
 - generate a second prompt based on the input from the system user;
 - provide the second prompt to the first large language model;
 - generate a second response to the second prompt using the first large language model, the second response including second raw content that defines a second clinical scenario consistent with both the first clinical scenario and the second prompt;
 - use the interpretability agent to generate a second filtered content based on the interpretation of the second raw content by the fuzzy logic; and
 - display the second filtered content to the system user.

13. The system of claim 11, wherein the interpretability agent interprets the first raw content and generates the at least the first portion of the first filtered content by:
- tokenizing the first raw content to generate a plurality of raw content tokens;
 - and
 - applying a set of fuzzy logic rules to the raw content tokens to generate the at least the first portion of the first filtered content.
14. The system of claim 11, wherein generating the first response includes tokenizing the prompt to generate a plurality of prompt tokens, and the program code further causes the system to:
- use an oversight agent to determine a Shapley value for each prompt token, the Shapley value quantifying a contribution of the prompt token to the raw content.
15. The system of claim 11, wherein the program code further causes the system to:
- forward the first raw content to a domain agent for human review in response to the interpretability agent interpreting the first raw content as including at least some uncertain content or some irrelevant content.
16. The system of claim 11, wherein generating the first filtered content further comprises:
- using a supervision agent to interpret the first raw content and generate at least a second portion of the first filtered content.
17. The system of claim 16, wherein the supervision agent interprets the first raw content by:
- tokenizing the first raw content into a plurality of raw content tokens; and
 - providing the raw content tokens to a second large language model that generates the at least the second portion of the first filtered content.

18. The system of claim 17, wherein each of the first large language model and the second large language model are configured to analyze strings of tokens, and a maximum length of the strings of tokens analyzed by the first large language model is larger than the maximum length of the strings of tokens analyzed by the second large language model.

19. The system of claim 17, wherein the first large language model is a strong large language model, and the second large language model is a weak large language model.

20. A computer program product for providing an educational environment using an agentic artificial intelligence system, comprising:

- a non-transitory computer-readable storage medium; and
- program code stored on the non-transitory computer-readable storage medium that, when executed by one or more processors, causes the one or more processors to:
 - receive a first prompt;
 - generate a first response based on the first prompt using a first large language model, the first response including first raw content that defines a first clinical scenario consistent with the first prompt;
 - use an interpretability agent including fuzzy logic that interprets the first raw content to generate at least a first portion of a first filtered content based on an interpretation of the first raw content by the fuzzy logic; and
 - display the first filtered content to a system user.

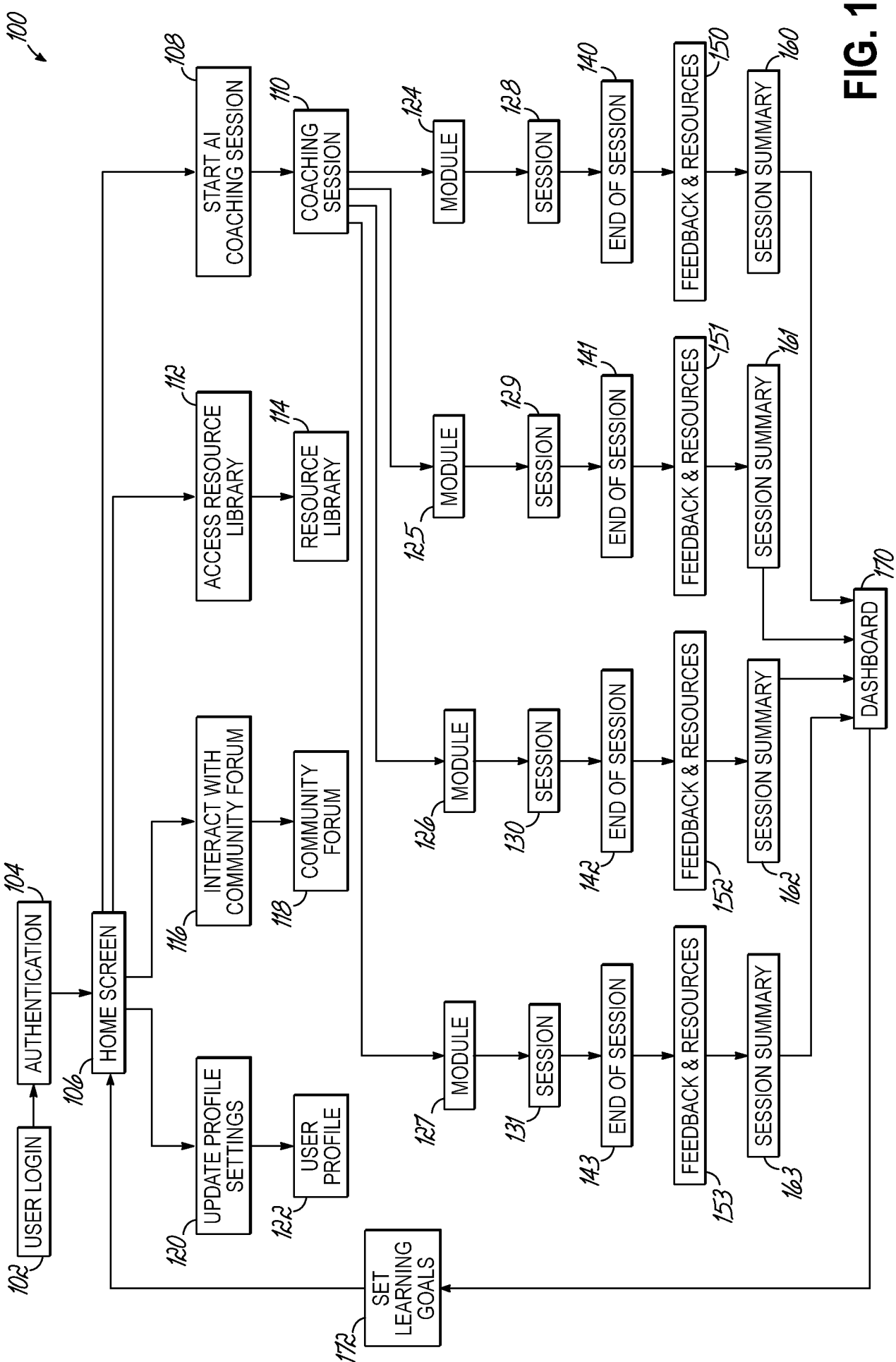


FIG. 1

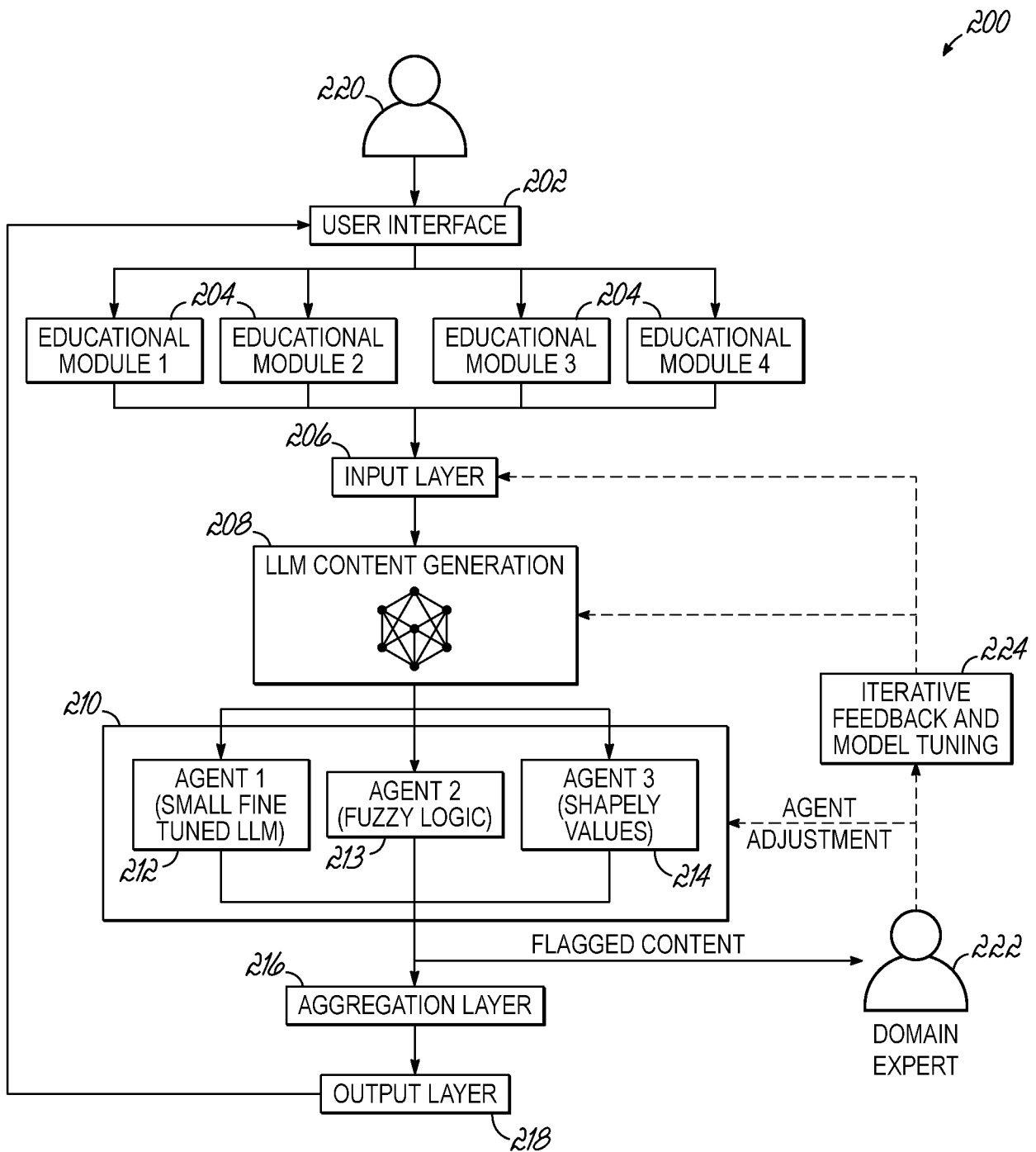


FIG. 2

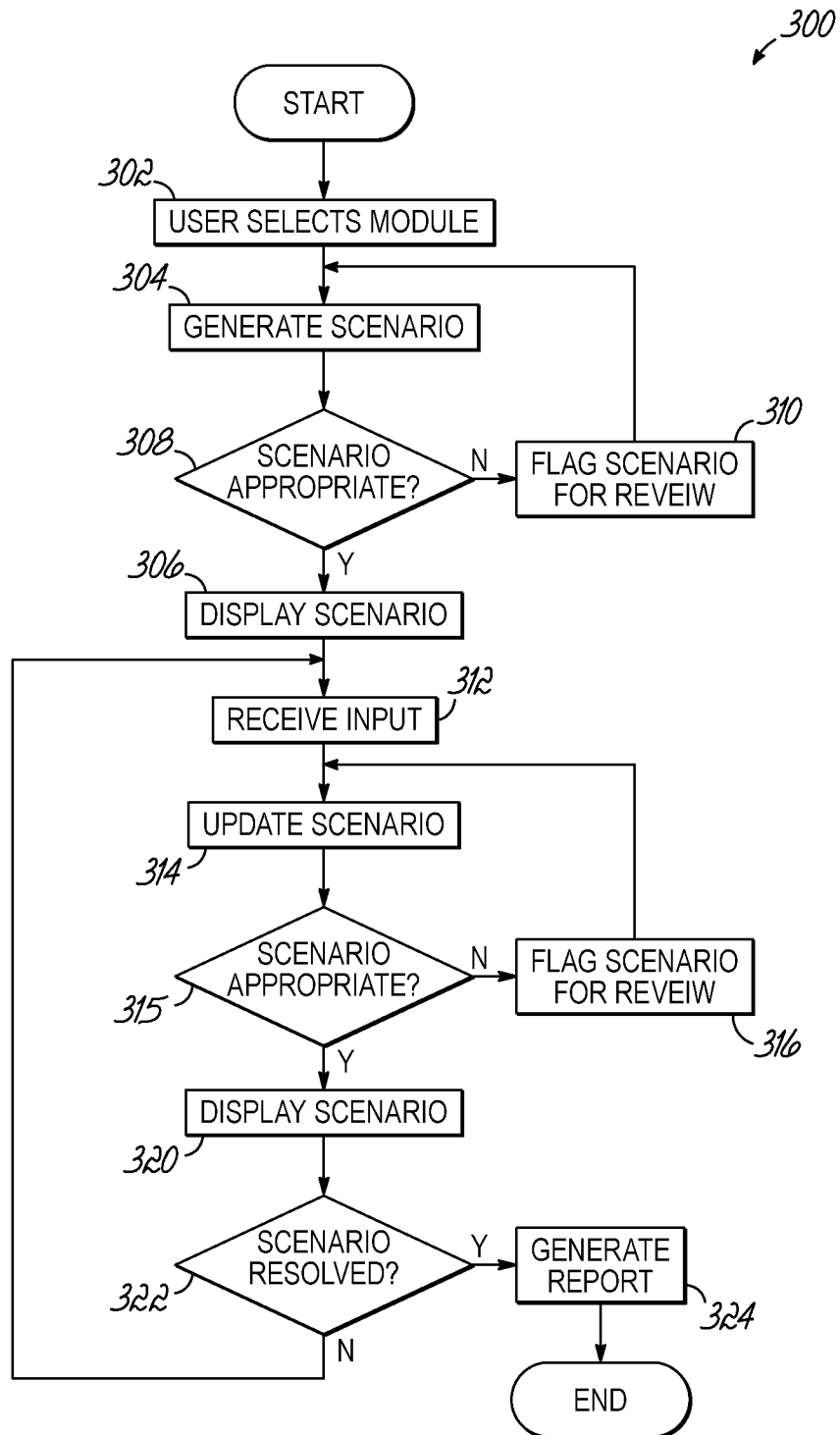


FIG. 3

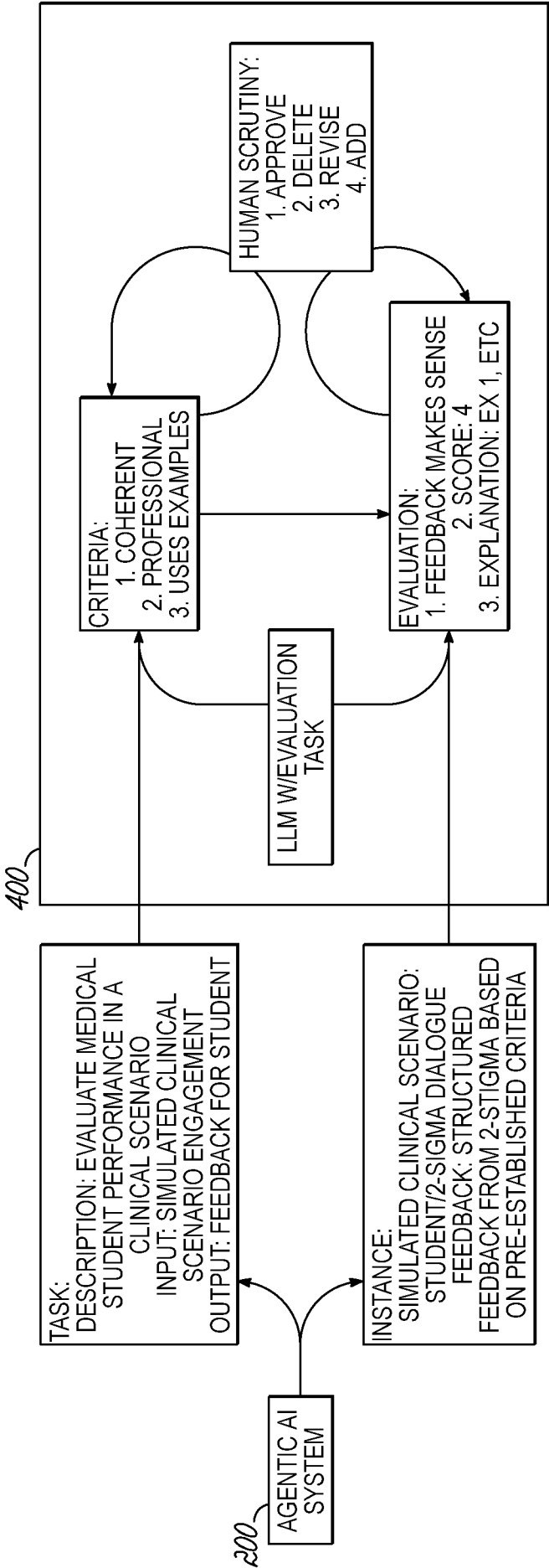


FIG. 4

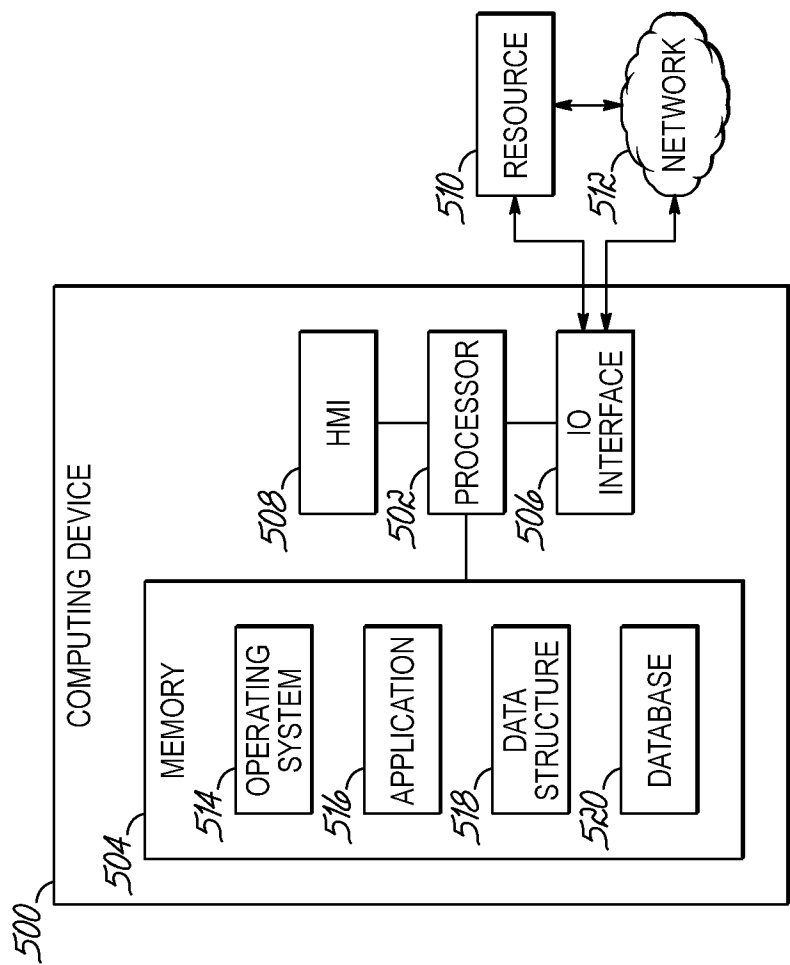


FIG. 5