US 20250259754A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2025/0259754 A1**

Hamilton et al. (43) **Pub. Date: Aug. 14, 2025**

(54) **ARTIFICIALLY INTELLIGENT DIALOG EVALUATION SYSTEM AND ASSOCIATED METHODS**

(71) Applicant: **ARIZONA BOARD OF REGENTS ON BEHALF OF THE UNIVERSITY OF ARIZONA**, Tucson, AZ (US)

(72) Inventors: **Allan Hamilton**, Tucson, AZ (US); **Kyle McLemore**, Tucson, AZ (US)

(21) Appl. No.: **19/051,667**

(22) Filed: **Feb. 12, 2025**

Related U.S. Application Data

(60) Provisional application No. 63/552,516, filed on Feb. 12, 2024.

Publication Classification

(51) **Int. Cl.**

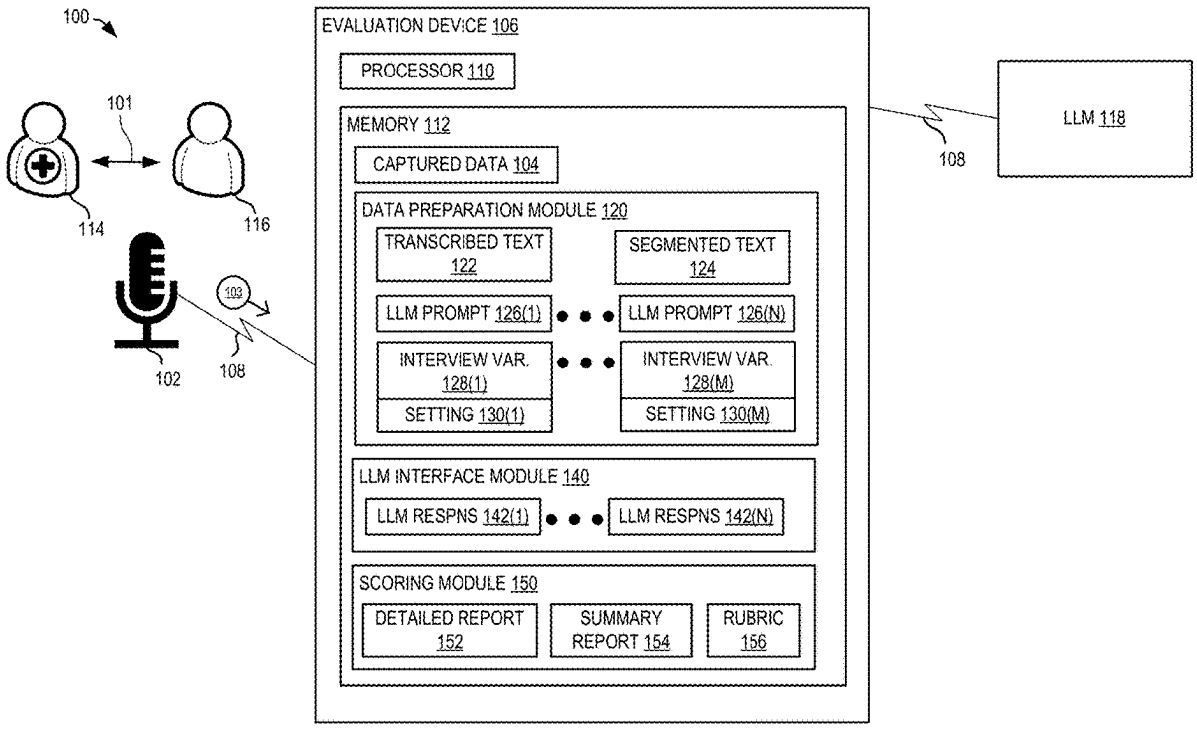
G16H 80/00	(2018.01)
G06F 40/40	(2020.01)
G10L 17/02	(2013.01)
G16H 10/60	(2018.01)

(52) **U.S. Cl.**

CPC ***G16H 80/00*** (2018.01); ***G06F 40/40*** (2020.01); ***G10L 17/02*** (2013.01); ***G16H 10/60*** (2018.01)

(57) **ABSTRACT**

An evaluation device receives captured data of a medical history evaluation interview between a patient and a medical provider. The evaluation device transcribes audio of the captured data into transcribed text and segments the transcribed text into segmented lines according to a speaker of the transcribed text within the audio. The evaluation device generates a plurality of prompts, each prompt corresponding to one of a plurality of interview analysis variables, to control a large language model (LLM) to analyze each of the segmented lines in context of the transcribed text. The evaluation device transmits the plurality of prompts to the LLM and receives LLM responses from the LLM for each of the plurality of prompts. The evaluation device analyzes the LLM responses with respect to a scoring rubric and generates a detail report defining performance of the medical provider during interaction between the patient and the medical provider.



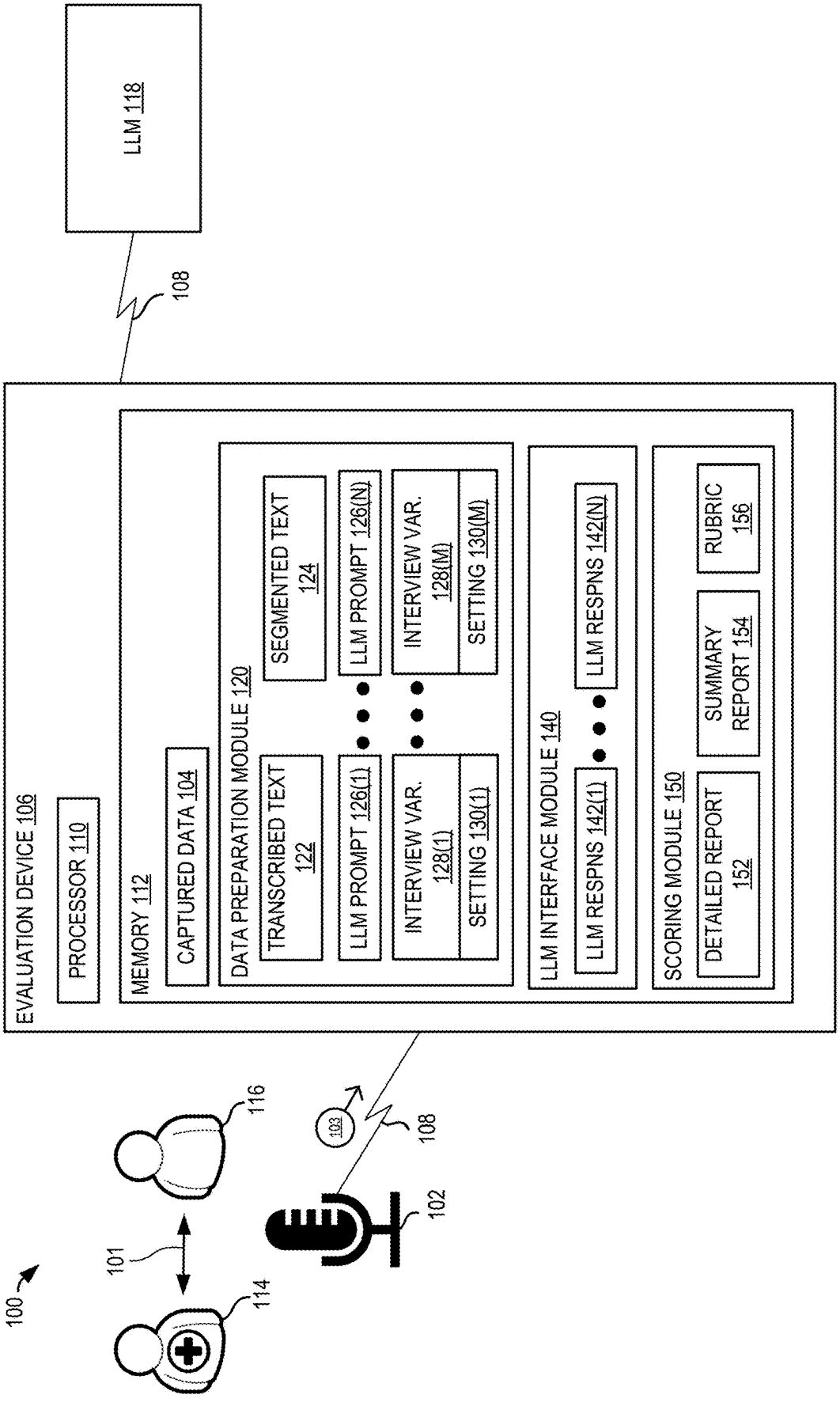

LLM 118
108
EVALUATION DEVICE 106
PROCESSOR 110
MEMORY 112
CAPTURED DATA 104
DATA PREPARATION MODULE 120
TRANSCRIBED TEXT 122
SEGMENTED TEXT 124
LLM PROMPT 126(1)
LLM PROMPT 126(N)
INTERVIEW VAR. 128(1)
INTERVIEW VAR. 128(M)
SETTING 130(1)
SETTING 130(M)
LLM INTERFACE MODULE 140
LLM RESPNS 142(1)
LLM RESPNS 142(N)
SCORING MODULE 150
DETAILED REPORT 152
SUMMARY REPORT 154
RUBRIC 156
100
101
114
116
103
102
108

FIG. 1

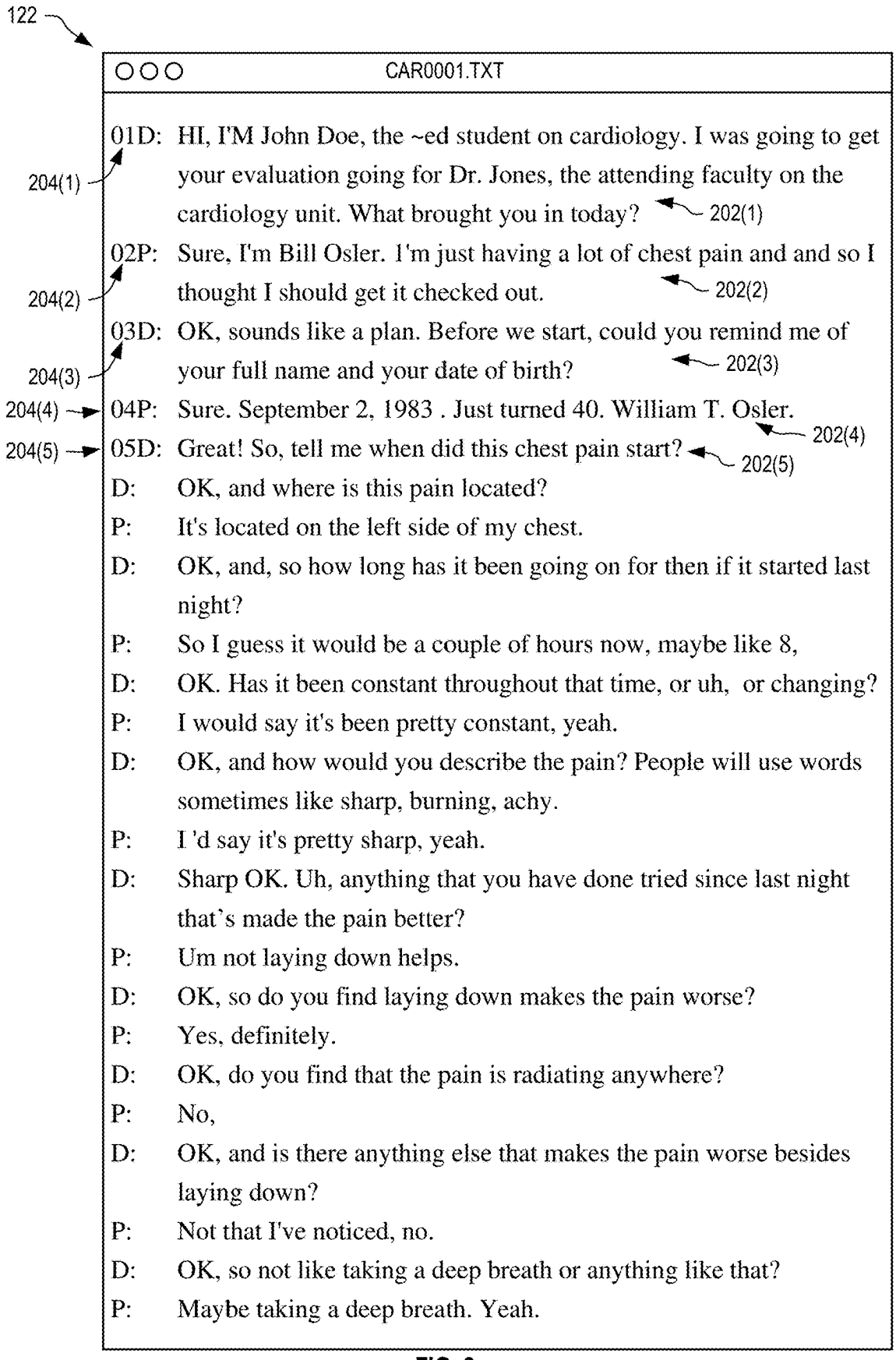
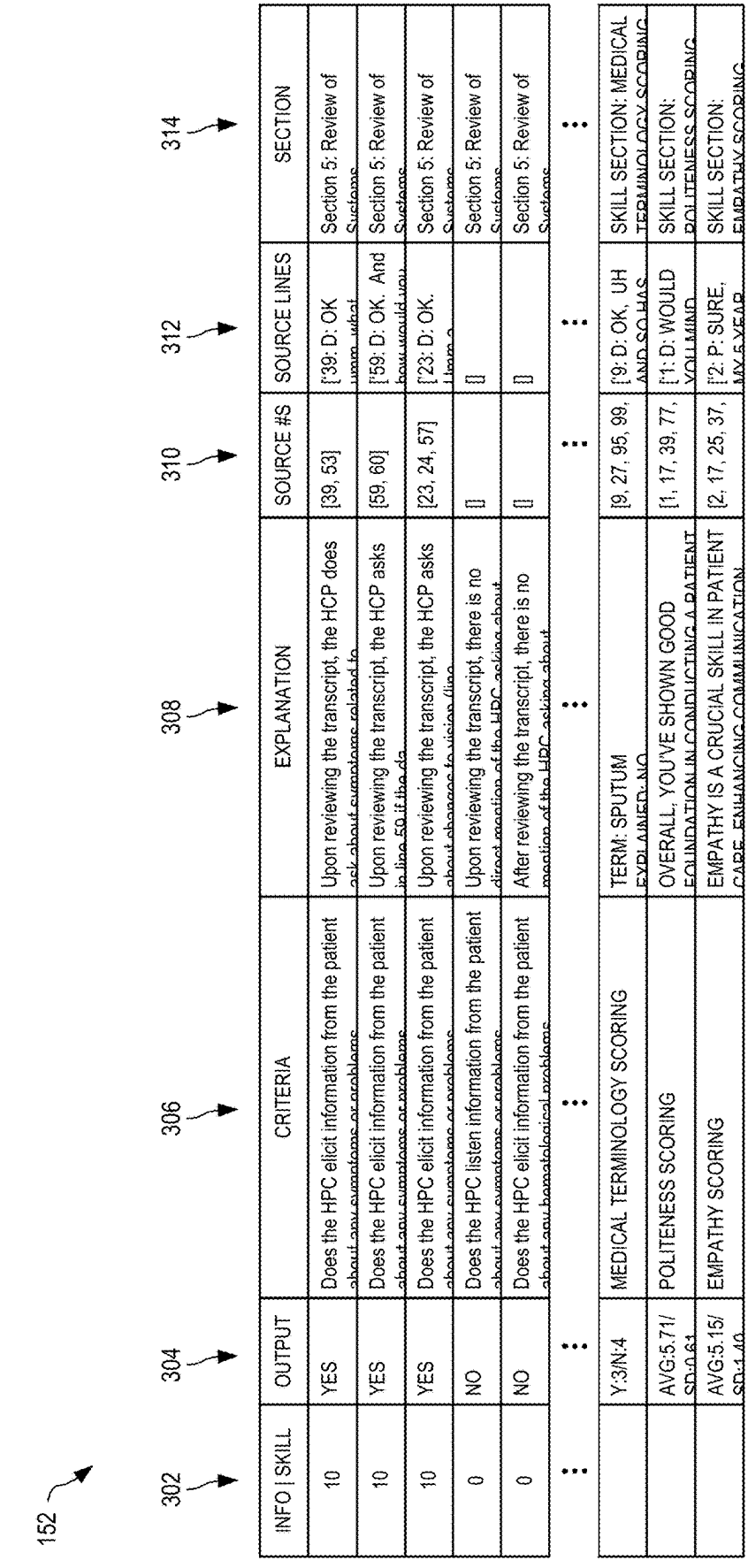

122
CAR0001.TXT
01D: HI, I'M John Doe, the ~ed student on cardiology. I was going to get your evaluation going for Dr. Jones, the attending faculty on the cardiology unit. What brought you in today?
204(1)
202(1)
02P: Sure, I'm Bill Osler. 1'm just having a lot of chest pain and and so I thought I should get it checked out.
204(2)
202(2)
03D: OK, sounds like a plan. Before we start, could you remind me of your full name and your date of birth?
204(3)
202(3)
04P: Sure. September 2, 1983 . Just turned 40. William T. Osler.
204(4)
202(4)
05D: Great! So, tell me when did this chest pain start?
204(5)
202(5)
D: OK, and where is this pain located?
P: It's located on the left side of my chest.
D: OK, and, so how long has it been going on for then if it started last night?
P: So I guess it would be a couple of hours now, maybe like 8,
D: OK. Has it been constant throughout that time, or uh, or changing?
P: I would say it's been pretty constant, yeah.
D: OK, and how would you describe the pain? People will use words sometimes like sharp, burning, achy.
P: I 'd say it's pretty sharp, yeah.
D: Sharp OK. Uh, anything that you have done tried since last night that's made the pain better?
P: Um not laying down helps.
D: OK, so do you find laying down makes the pain worse?
P: Yes, definitely.
D: OK, do you find that the pain is radiating anywhere?
P: No,
D: OK, and is there anything else that makes the pain worse besides laying down?
P: Not that I've noticed, no.
D: OK, so not like taking a deep breath or anything like that?
P: Maybe taking a deep breath. Yeah.

FIG. 2



INFO	SKILL	OUTPUT	CRITERIA	EXPLANATION	SOURCE #S	SOURCE LINES	SECTION
	10	YES	Does the HPC elicit information from the patient about any symptoms or problems	Upon reviewing the transcript, the HCP does ask about symptoms related to	[39, 53]	['39: D: OK umm what	Section 5: Review of Systems
	10	YES	Does the HPC elicit information from the patient about any symptoms or problems	Upon reviewing the transcript, the HCP asks in line 59 if the ds	[59, 60]	['59: D: OK. And how would you	Section 5: Review of Systems
	10	YES	Does the HPC elicit information from the patient about any symptoms or problems	Upon reviewing the transcript, the HCP asks about changes to vision (line	[23, 24, 57]	['23: D: OK. Hmm o	Section 5: Review of Systems
	0	NO	Does the HPC listen information from the patient about any symptoms or problems	Upon reviewing the transcript, there is no direct mention of the HPC asking about	[]	[]	Section 5: Review of Systems
	0	NO	Does the HPC elicit information from the patient about any hematological problems	After reviewing the transcript, there is no mention of the HPC asking about	[]	[]	Section 5: Review of Systems
...			...	...	...	...	...
		Y:3/N:4	MEDICAL TERMINOLOGY SCORING	TERM: SPUTUM EXPLAINED: NO	[9, 27, 95, 99,	['9: D: OK, UH AND SO HAS	SKILL SECTION: MEDICAL TERMINOLOGY SCORING
		AVG:5.71/ SD:0.61	POLITENESS SCORING	OVERALL, YOU'VE SHOWN GOOD FOUNDATION IN CONDUCTING A PATIENT	[1, 17, 39, 77,	['1: D: WOULD YOU MIND	SKILL SECTION: POLITENESS SCORING
		AVG:5.15/ SD:1.40	EMPATHY SCORING	EMPATHY IS A CRUCIAL SKILL IN PATIENT CARE. ENHANCING COMMUNICATION	[2, 17, 25, 37,	['2: P: SURE, MY 5 YEAR	SKILL SECTION: EMPATHY SCORING

FIG. 3

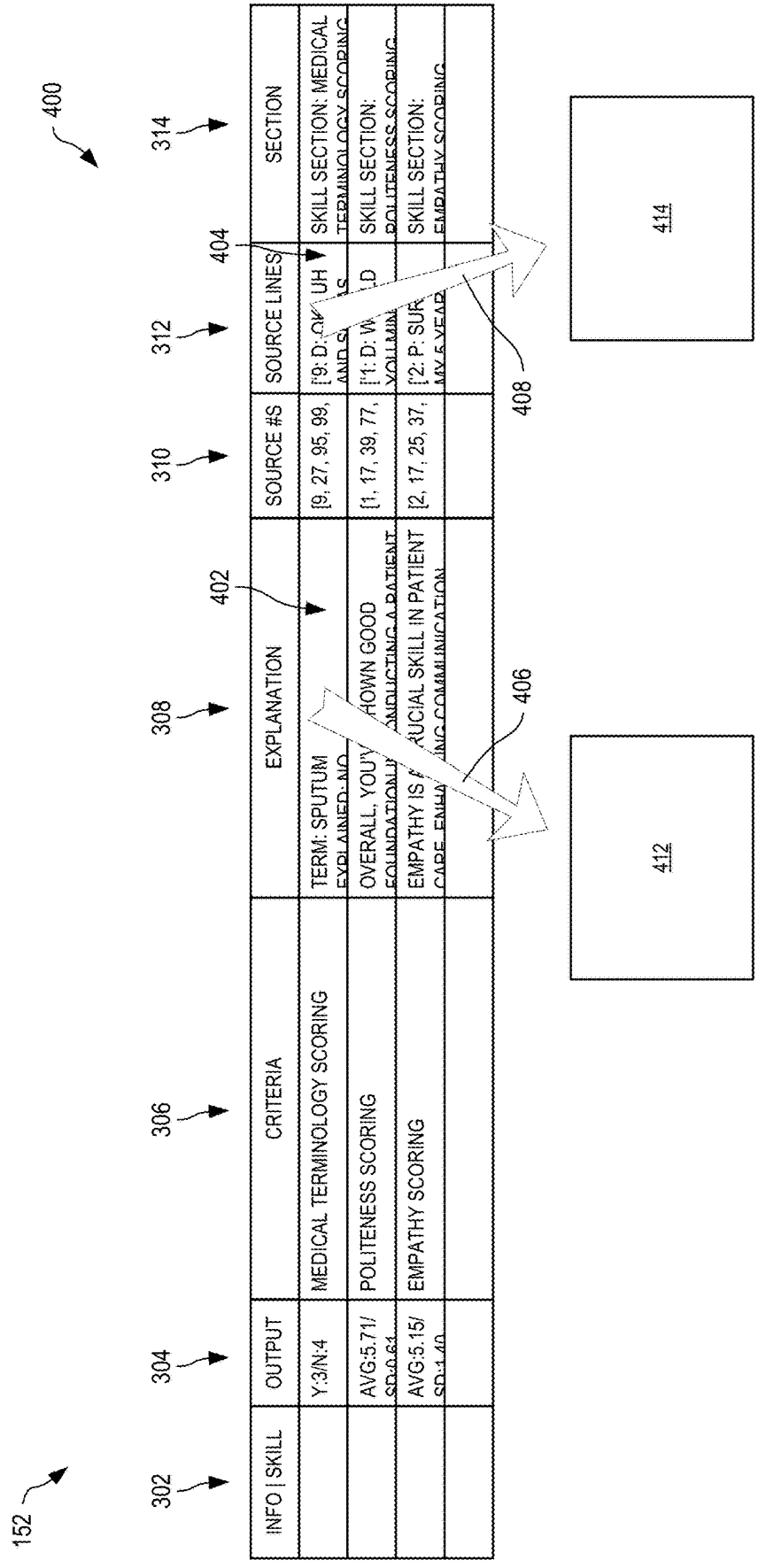

400
152
302
304
306
308
310
312
314
402
404
406
408
412
414
INFO | SKILL
OUTPUT
CRITERIA
EXPLANATION
SOURCE #S
SOURCE LINES
SECTION
MEDICAL TERMINOLOGY SCORING
POLITENESS SCORING
EMPATHY SCORING
SKILL SECTION: MEDICAL
SKILL SECTION:
SKILL SECTION:

FIG. 4A

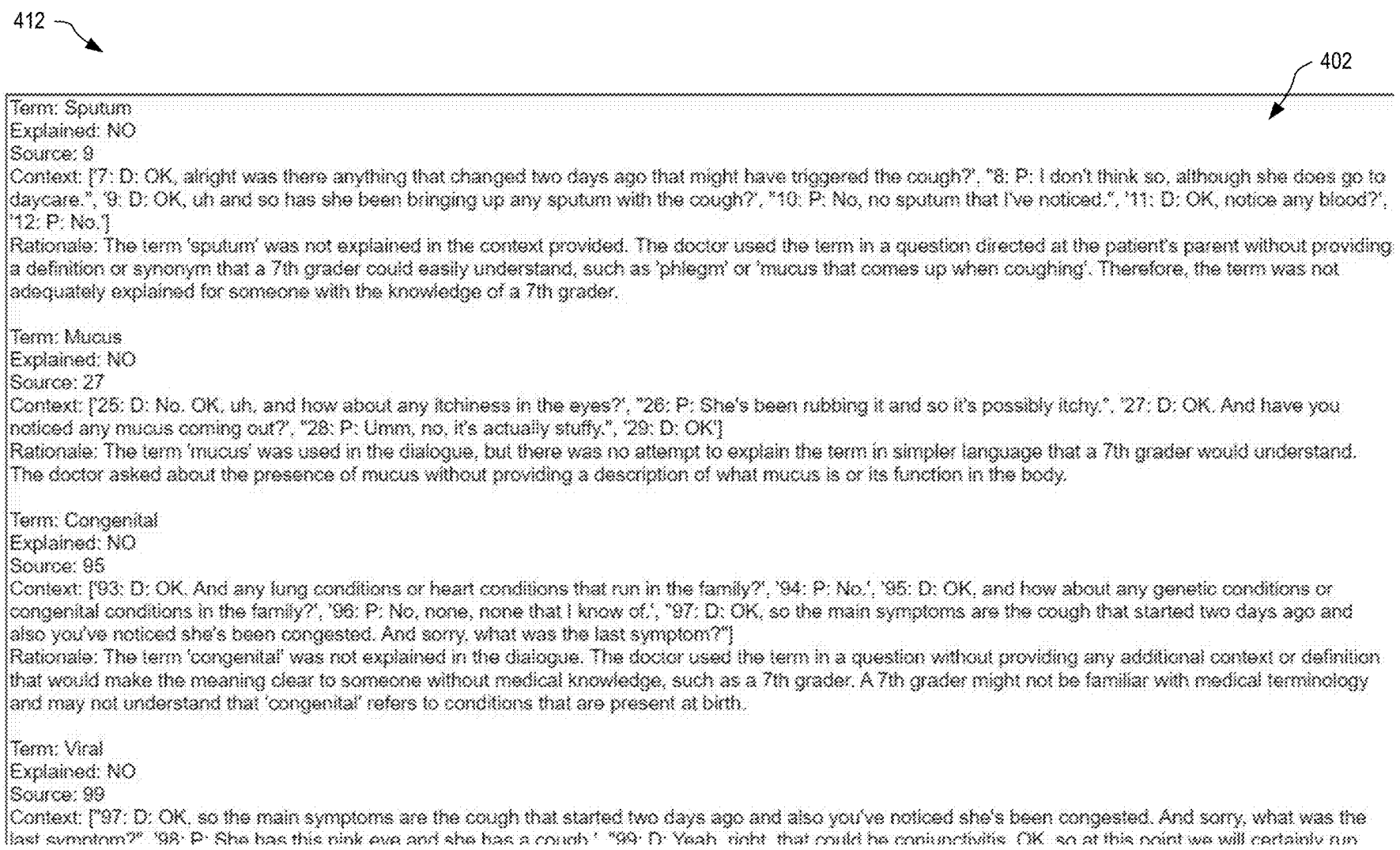

412
402
Term: Sputum
Explained: NO
Source: 9
Context: ['7: D: OK, alright was there anything that changed two days ago that might have triggered the cough?', "8: P: I don't think so, although she does go to daycare.", '9: D: OK, uh and so has she been bringing up any sputum with the cough?', "10: P: No, no sputum that I've noticed.", '11: D: OK, notice any blood?', '12: P: No.']
Rationale: The term 'sputum' was not explained in the context provided. The doctor used the term in a question directed at the patient's parent without providing a definition or synonym that a 7th grader could easily understand, such as 'phlegm' or 'mucus that comes up when coughing'. Therefore, the term was not adequately explained for someone with the knowledge of a 7th grader.
Term: Mucus
Explained: NO
Source: 27
Context: ['25: D: No. OK, uh, and how about any itchiness in the eyes?', "26: P: She's been rubbing it and so it's possibly itchy.", '27: D: OK. And have you noticed any mucus coming out?', "28: P: Umm, no, it's actually stuffy.", '29: D: OK']
Rationale: The term 'mucus' was used in the dialogue, but there was no attempt to explain the term in simpler language that a 7th grader would understand. The doctor asked about the presence of mucus without providing a description of what mucus is or its function in the body.
Term: Congenital
Explained: NO
Source: 95
Context: ['93: D: OK. And any lung conditions or heart conditions that run in the family?', '94: P: No.', '95: D: OK, and how about any genetic conditions or congenital conditions in the family?', '96: P: No, none, none that I know of.', '97: D: OK, so the main symptoms are the cough that started two days ago and also you've noticed she's been congested. And sorry, what was the last symptom?"]
Rationale: The term 'congenital' was not explained in the dialogue. The doctor used the term in a question without providing any additional context or definition that would make the meaning clear to someone without medical knowledge, such as a 7th grader. A 7th grader might not be familiar with medical terminology and may not understand that 'congenital' refers to conditions that are present at birth.
Term: Viral
Explained: NO
Source: 99
Context: ["97: D: OK, so the main symptoms are the cough that started two days ago and also you've noticed she's been congested. And sorry, what was the last symptom?", '98: P: She has this pink eye and she has a cough.', "99: D: Yeah, right, that could be conjunctivitis. OK, so at this point we will certainly run

FIG. 4B

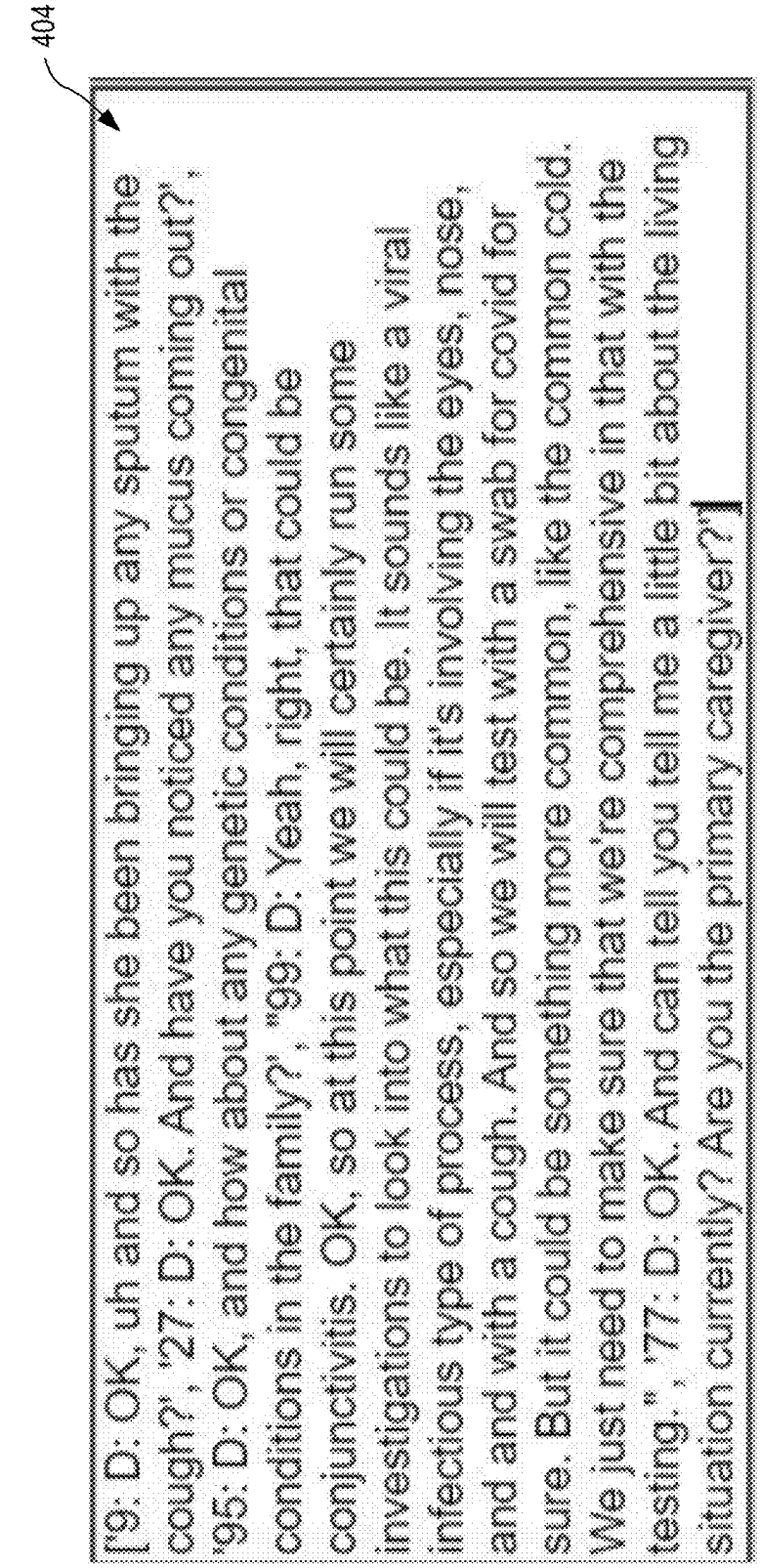

404
414
['9: D: OK, uh and so has she been bringing up any sputum with the cough?', '27: D: OK. And have you noticed any mucus coming out?', '95: D: OK, and how about any genetic conditions or congenital conditions in the family?', '99: D: Yeah, right, that could be conjunctivitis. OK, so at this point we will certainly run some investigations to look into what this could be. It sounds like a viral infectious type of process, especially if it's involving the eyes, nose, and and with a cough. And so we will test with a swab for covid for sure. But it could be something more common, like the common cold. We just need to make sure that we're comprehensive in that with the testing.', '77: D: OK. And can tell you tell me a little bit about the living situation currently? Are you the primary caregiver?']

FIG. 4C

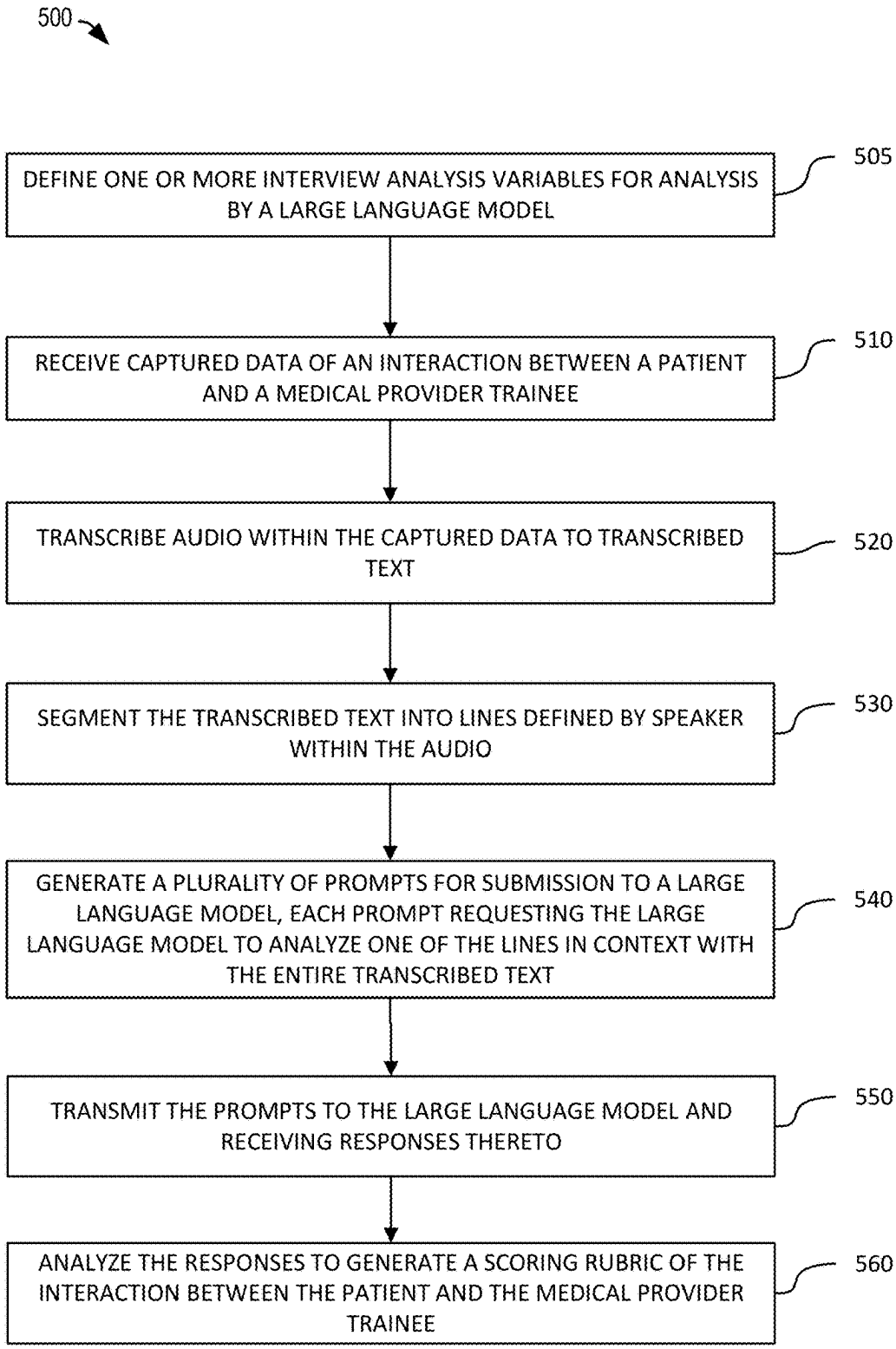

500
DEFINE ONE OR MORE INTERVIEW ANALYSIS VARIABLES FOR ANALYSIS BY A LARGE LANGUAGE MODEL
505
RECEIVE CAPTURED DATA OF AN INTERACTION BETWEEN A PATIENT AND A MEDICAL PROVIDER TRAINEE
510
TRANSCRIBE AUDIO WITHIN THE CAPTURED DATA TO TRANSCRIBED TEXT
520
SEGMENT THE TRANSCRIBED TEXT INTO LINES DEFINED BY SPEAKER WITHIN THE AUDIO
530
GENERATE A PLURALITY OF PROMPTS FOR SUBMISSION TO A LARGE LANGUAGE MODEL, EACH PROMPT REQUESTING THE LARGE LANGUAGE MODEL TO ANALYZE ONE OF THE LINES IN CONTEXT WITH THE ENTIRE TRANSCRIBED TEXT
540
TRANSMIT THE PROMPTS TO THE LARGE LANGUAGE MODEL AND RECEIVING RESPONSES THERETO
550
ANALYZE THE RESPONSES TO GENERATE A SCORING RUBRIC OF THE INTERACTION BETWEEN THE PATIENT AND THE MEDICAL PROVIDER TRAINEE
560

FIG. 5

154

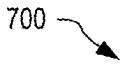
MEDICAL INTERVIEW FORMAT & STRUCTURE:
The interview was evaluated with respect to the format and structure of the medical interview.
1. Healthcare Provider Identification
The health care provider provided [**complete**] identification. This [**included**] [**identification of their appropriate professional rank or title**], their [**first**] [**and**] [**their last name**], and [**notification to the patient of the objective of their visit**].
Healthcare Provider ID: [40;40] points
2. Patient-Subject-Client Identification
The health care provider [**appropriately**] identified the patient by their [**first**][[**and last name**] and their [**date of birth**]. [**No other identification was requested**]. The patient was [**an adult**], and during the course of the interview, the health care provider [**appropriately**] referred to the patient with the appropriate salutation and their last name.
During the course of the interview (excluding introduction and concluding remarks) the health
care provider referred to the patient by name on [**6**] occasions.
Patient/Client Subject Identification: [30/40] points
Comfort Check [0/10] points
Proper Mode of Addressing Patient: [10 + 10 + (6 x 10 points): 80/80] points

FIG. 6A

154

SUMMARY: In summary, [**John Doe, Class 2028**] demonstrated [**good**] command of the elements of the medical history interview at this
stage of his medical training. There were [**no**] areas of unacceptable performance1. Areas of improvement should on items like [el**iciting more information about (1) allergic reaction to penicillin and morphine, (2) vaccinations and immunizations, (3) exotic or recent travel, (4) trying to avoid excessive use of medical or technical jargon without explanation for the patient in layman's terms, (5) doing a more thorough review of systems, and 5) providing a concluding statement or remark at the end of the interview.**].
The overall score for the medical history interview was [**420 out of 560 (83%)**]
A full scoring off all areas of the evaluation is available for review.
The following summary was generated by an artificial intelligence-enhanced evaluation algorithm. Please feel free to contact your faculty instructor should there be any questions or concerns about the way this was scored by the computer.
This evaluation of a medical history was reviewed by [**print faculty members name and title**] and review is attested by the signature below:
_____. _____ :
_____
[**Print faculty name and title**] Signature [**Date**]
1 Items of unacceptable performance: failure to identify oneself as healthcare provider, failure to properly identify patient, failure to document allergy(ies), failure to record medications patient is taking.

FIG. 6B



This report relates to an AI-derived evaluation of a medical history interview conducted by healthcare provider (HCP) HCP_NAME (HCP_YEAR). They interviewed patient ID PATIENT_ID on DATE_TODAY as part of their primary care medicine clerkship. The human supervisor (HS) or faculty member was HUMAN_SUPERVISOR.

The overall score for the medical history interview was POINTS_AWARDED out of POINTS_TOTAL (POINTS_PERCENTAGE). In summary, HCP_NAME, HCP_YEAR demonstrated ADJECTIVE_SCORE proficiency (PERCENTILE_SCORE percentile) of the elements of the medical history interview at this stage of their medical training. There were UNACCEPTABLE_PERFORMANCE_AREAS[1]. Regardless, areas of improvement are listed on the next page.

A full scoring of all areas of the evaluation is available for review. The following summary was generated by an artificial intelligence-enhanced evaluation algorithm. Please feel free to contact the assigned faculty instructor should there be any questions or concerns about the way this was scored by the computer.

This evaluation of a medical history was reviewed by HUMAN_SUPERVISOR and the review is formally attested by the signature below:

Signature

FIG. 7

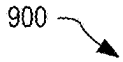
800

Information Section	Correct	Incorrect	Total	Percentage
Introduction	9	3	12	75.0%
Present Illness	15	2	17	88.24%
Past Medical History	8	0	8	100.0%
Family and Social History	15	6	21	71.43%
Review of Systems	10	5	15	66.67%
Obstetrical/Gynecological and Reproductive History	3	6	9	33.33%
Conclusion & Summarizing	4	0	4	100.0%
Sub Totals	**64**	**22**	**86**	**74.42%**

Skill Section	Correct	Incorrect	Total	Percentage
Medical Terminology Scoring	1.82	Y:2/N:9	10	18.18%
Politeness Scoring	5.13	SD:1.26	10	51.3%
Empathy Scoring	5.91	SD:0.93	10	59.1%
Sub Totals	**12.86**	**17.14**	**30**	**42.86%**

Total Score	58.64%

FIG. 8



SECTION 6: OBSTETRICAL/GYNECOLOGICAL AND REPRODUCTIVE HISTORY

STRENGTHS:
The HCP successfully elicited information about menstrual history and problems or concerns relating to reproduction from female patients of reproductive age. The HCP also ensured to ask about regular gynecological visits and follow-up, which is crucial for maintaining reproductive health.

WEAKNESSES:
The HCP did not ask about the patient's gender identity, which is important for understanding the patient's health needs and providing personalized care. There was also a lack of inquiry into the patient's sexual history, level of activity, and relations with multiple individuals, which are essential for assessing risk factors for sexually transmitted diseases (STDs) and providing appropriate counseling. The HCP failed to elicit a history of STDs or concerns about birth control and protection from STDs. For female patients with children, the HCP did not ask about prenatal and birth history, which can provide valuable insights into current or future pregnancy-related health issues. Additionally, the HCP did not inquire about fertility concerns for females of reproductive age or about sexual or erectile dysfunction in male patients above the age of 35.

OVERALL:
To improve the medical history-taking interview in the Obstetrical/Gynecological and Reproductive History section, the HCP should begin by asking about the patient's gender identity to ensure that the conversation is respectful and relevant to the patient's experience. It is important to create a safe and nonjudgmental environment to discuss sexual history, including the number of partners and types of sexual relations, as this can impact the patient's risk for STDs and inform preventive health strategies. The HCP should also actively inquire about any history of STDs and discuss birth control options and STD prevention methods. For female patients with children, obtaining a detailed prenatal and birth history is crucial for identifying any complications or issues that may affect future pregnancies or overall health. When interviewing female patients of reproductive age, the HCP should ask about fertility concerns to address any potential issues early on. For male patients over 35, discussing sexual function is important as it can be an indicator of other health conditions. By incorporating these elements into the interview, the HCP can provide comprehensive care that addresses all aspects of the patient's reproductive and sexual health.

FIG. 9

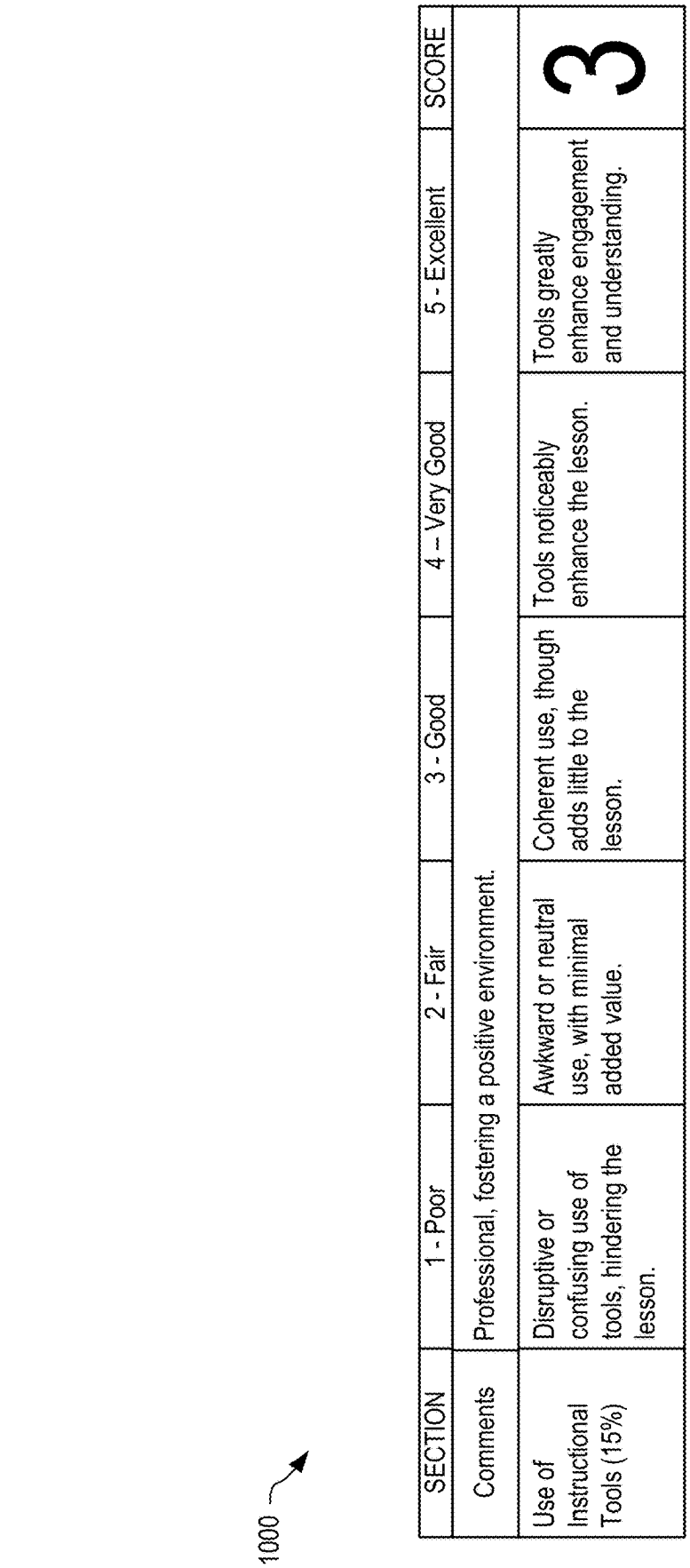

1000
SECTION
1 - Poor
2 - Fair
3 - Good
4 – Very Good
5 - Excellent
SCORE
Comments
Professional, fostering a positive environment.
Use of Instructional Tools (15%)
Disruptive or confusing use of tools, hindering the lesson.
Awkward or neutral use, with minimal added value.
Coherent use, though adds little to the lesson.
Tools noticeably enhance the lesson.
Tools greatly enhance engagement and understanding.
3

FIG. 10

1100

[Speaker 2 - Debriefer - (instructor name: Deana)] Well thank you all for being a part of this simulation. You know, I think we can all agree that. Um, Sometimes when we have a patient, who is acutely ill, it's a little chaotic at times. Um, I'd like to First go around the room and introduce yourself and your role that you've played in the simulation.

[Speaker 2] Here,

[Speaker 3] My name is Lucas. I was the resident during the simulation.

[Speaker 4] And,

[Speaker 3] Uh, my

[Speaker 5] Name is Alan. I was the physician

[Speaker 6] My name is Kyle. I was the RN

[Speaker 1] I'm Merrill. I was a triage RN

[Speaker 4] I'm Dave. Pct.

[Speaker 7] That was Katie, I was a nursing assistant.

[Speaker 1] Okay.

[Speaker 2] Okay. So, um You know, I'd like to kind of like start, anybody can start where, it's just kind of like a open dialogue as a group and and If I have a volunteer, how do you think the communication went with your team during the simulation? Anyone can start?

[Speaker 2] Well, how about we start with the person who first saw the patient like brought him in the hallway? Uh, that would be that triage are in, right?

[Speaker 1] Yes, what was the question again?

[Speaker 2] How do you feel the communication was with your team?

[Speaker 1] Um, I thought it was a little bit over. All over the place in the beginning. But when we started the code, it got a little better. I mean, we got the patient

[Speaker 2] Back. So you're you felt it was good because you had a good outcome. Yes.

[Speaker 1] Okay.

[Speaker 2] Uh, I think that the next person with that patient was, um, Our paramedic nursing assistant. Tara, how do you feel? It was It was overwhelming.

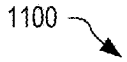
[Speaker 7] I think that it was hard getting his clothes off and stuff. And so I wasn't really Not used to working with a mannequin and so that was kind of weird for me.

[Speaker 2] Yeah, so do you feel that? Um, Do you feel like a stronger pre-brief and what the mannequin can't do? What you can do? Would have been beneficial. I

[Speaker 7] Don't know, I think it would have helped to know like if we really were supposed to take his clothes off and stuff or if I was just supposed to pretend, I didn't really know like how real it was supposed to be.

[Speaker 2] Okay, that's fair. Um, anybody else want to comment on that?

FIG. 11A



[Speaker 3] I

[Speaker 5] Thought the team did a pretty good job and communicated pretty well. I mean Think, you know, every Med got delivered and I think people were checking in people were really helping out. I was pretty, I thought the team worked pretty smoothly.

[Speaker 2] That's yeah, I think um for the most part they had some strong communication, they give some closed-loop communication And um, I from what I observed.

[Speaker 2] The staff. Was communicating, well, but I think at times they didn't know who they were communicating to. So, You know, Dave you came in um, as a respiratory therapist later on in the simulation, how do you feel when you came into the room?

[Speaker 4] Well, it came a little late.

[Speaker 2] And

[Speaker 4] So I of course started to back valve Mass by that point, the patient was coding already. I wasn't quite sure.

[Speaker 4] Would all I should should have done. Um, Maybe that's because I wasn't clear of the objectives. Because I thought maybe we should intubate. Actually intubate, like we would in real life. Um, But yeah, I Was sort of at at where I was supposed to be in Bag valve mask, and sort of took the Took the, the doctor's lead.

[Speaker 2] I think that um, your role was pretty critical in the simulation base that you were doing Airway and could check for pulses and Um, how is your communication with the provider and ensuring that, you know, Respirations were given and that, you know, we had pulses and that everyone was kind of Um, on task. Did you feel that I felt like

[Speaker 4] For the most part? Uh, we were doing what we were supposed to do and it was communicated. I think we just could have been a little smoother. And Tighter.

[Speaker 2] Yeah, Kyle, what do you think? I I I kind of noticed that you were

[Speaker 2] Kind of jolly. Yeah I was

[Speaker 6] Laughing a lot. I don't think we were taking it as seriously as we could have. I don't think we have all had a lot of experience with mannequin based training so this was I thought it was kind of fun and funny.

[Speaker 2] I did appreciate that. You verified what the next med should be in the next process. Coding, the patient. I think that was really great on your part in order to Follow the algorithm.

[Speaker 3] Yeah, yeah, I

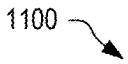
[Speaker 5] Did too. I thought that was very helpful and somebody reminded me about getting the second Epi and getting electrodes on. So, I was very much appreciated getting that help.

[Speaker 2] Yeah, how do we feel about our compressions? Um, Tara, I think you started some compressions, didn't you?

[Speaker 7] I didn't start a no.

[Speaker 1] I started. Yeah, I started I was on the show.

FIG. 11B



[Speaker 2] And then and then someone took over for you.

[Speaker 1] Yeah, what's that? Well, we probably could have done a better job communicating when we were supposed to switch. Because after the fact I realized there was quite a bit of time that we were off the chest and we could have Minimize that time when we were trading? Now,

[Speaker 2] Lucas being a resident. Probably haven't seen as many codes. How do you feel that this code went in your opinion? Um,

[Speaker 3] I think as a code it went pretty well. Going back to what Dave said. As a simulation though, it was kind of hard to follow what had been given and what had been done already, they kind of went through this saying, oh yeah, we gave this. We gave this, but looking at the, the mannequin and the actual room, we I didn't see any medication actually hanging or anything being done, physically, to the mannequin. So, it's hard to keep track of that as someone not as familiar with these kind of situations.

[Speaker 2] Okay.

[Speaker 3] Yeah, I think that,

[Speaker 2] Um, Something that we can practice is closed communication and that's verifying what bed, your the order that was given by The Physician

[Speaker 2] Alan and giving that designation to who should be giving it and when and then that person closing that Loop would have been probably a little beneficial in the simulation sort

[Speaker 5] Of got a little confused. When I started looking at the airway while they were checking the pulse and had to stop in the middle of that and it was sort of out of step that was my fault. Yeah, so I think that could have been smoother and

[Speaker 3] The shock was given, and then they checked the pulse. Yeah, instead. Yeah, I agree.

[Speaker 2] Well I can I think we can all agree that you know having these types of simulations and practicing them is a safe way to kind of hone those skills. And so overall, I think you did a great job. I think there's definitely some opportunities to Hone in on our closed loop, communication and roles and just ensuring that someone's has the algorithm. And we're following the algorithm when we're doing CPR and ACLS. Um, does anybody else want to share any comments or feelings that they have about the simulation?

[Speaker 5] I've been trying it a couple more times would help.

[Speaker 1] Okay. And

[Speaker 5] I think it would, I mean, you know, I think if we go back and practice it some more I think we'd be a little bit smoother on the handoff and when we're checking for a pulse and when we're giving CPR and that sort of thing,

FIG. 11C

1200

152

Evaluators:	AIDSET	Instructor:	Deana
Date:	10/18/2024	Course/Section:	Simulation Training
Location:	ASTEC	Class Size:	
Context:	Simulation with healthcare roles		

1250

SECTION	1 - Poor	2 - Fair	3 - Good	4 – Very Good	5 - Excellent	SCORE
Structure of Lesson (10%)	Significant confusion, unresolved issues, or off-topic digressions.	Minor confusion or off-topic elements; lacks clarity.	Clear and Logical structure, but lacking nuance.	Key ideas clear with good flow and transitions.	Nuanced and engaging structure, anticipating student questions.	4
Comments	Clear structure, but more depth needed.					
Style of Delivery (15%)	Lifeless, distracted, or indifferent presentation.	Low energy or inconsistent delivery.	Clear delivery but lacking enthusiasm.	Enthusiastic keeping students attention.	Captivating and passionate delivery, fully engaging students.	3
Comments	Clear delivery, but lacked enthusiasm.					
Professional Demeanor (10%)	Unprofessional or lacking creadibility.	Some issues affecting classroom management or engagement.	Acceptable, but lacks a strong professional presence.	Committed and professional, supporting a positive classroom environment.	Role model quality, elevating student engagement and professionalism.	4
Comments	Professional, fostering a positive environment.					

FIG. 12A

1250

SECTION	1 - Poor	2 - Fair	3 - Good	4 – Very Good	5 - Excellent	SCORE
Use of Functional Tools (15%)	Disruptive of confusing use of tools, hindering the lesson.	Awkward or neutral use, with minimal added value.	Coherent use, though adds little to the lesson.	Tools noticeably enhance the lesson.	Tools greatly enhance engagement and understanding.	3
Comments	Tools used well, more integration needed.					
Student Engagement (20%)	No participation; students appear disengaged.	Minimal participation; students unsure about contributing.	Sporadic participation; students attentive but not fully engaged.	Comfortable and active participation by students.	Highly engaged students, actively shaping the lesson.	4
Comments	Active participation, enhance algorithm adherence.					
Attentiveness to Students' Active Learning (20%)	No adaption to students' comprehension or engagement.	Limited adaption; basic awareness of student needs.	Encourages thinking but lacks depth in engagement.	Encourages active exploration; adapts pedagogy based on feedback.	Masterfully adapts pedagogy; students actively explore and enrich the lesson.	4
Comments	Adapted to feedback, more exploration needed.					
Cumulative Impressions (10%)	Overall lesson was ineffective, and failed to meet learning objectives.	Lesson had some strengths but was inconsistent and partially met objectives	Solid lesson with clear objectives met, though lacking some elements.	Strong lesson with clear objectives, effective delivery, and engagement.	Highly effective in achieving objectives, engaging and memorable.	4
Comments	Strong lesson, effective yet improve engagement.					

FIG. 12BA

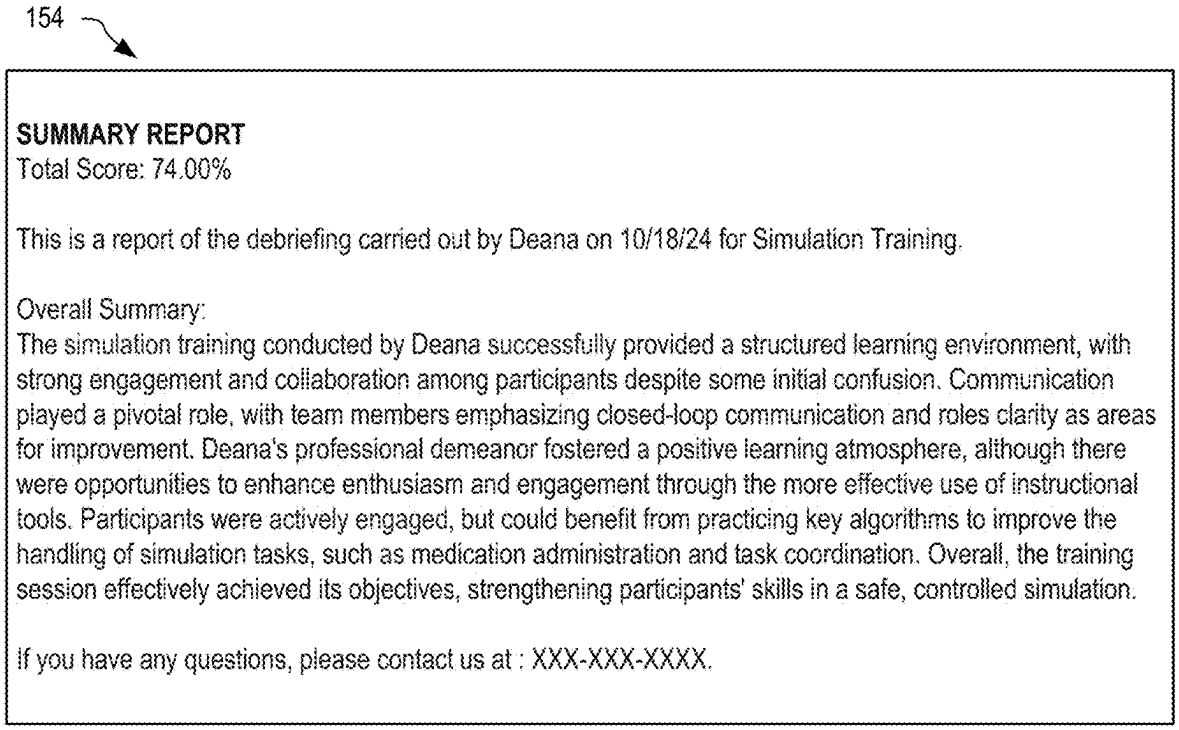

154
SUMMARY REPORT
Total Score: 74.00%
This is a report of the debriefing carried out by Deana on 10/18/24 for Simulation Training.
Overall Summary:
The simulation training conducted by Deana successfully provided a structured learning environment, with strong engagement and collaboration among participants despite some initial confusion. Communication played a pivotal role, with team members emphasizing closed-loop communication and roles clarity as areas for improvement. Deana's professional demeanor fostered a positive learning atmosphere, although there were opportunities to enhance enthusiasm and engagement through the more effective use of instructional tools. Participants were actively engaged, but could benefit from practicing key algorithms to improve the handling of simulation tasks, such as medication administration and task coordination. Overall, the training session effectively achieved its objectives, strengthening participants' skills in a safe, controlled simulation.
If you have any questions, please contact us at : XXX-XXX-XXXX.

FIG. 13

ARTIFICIALLY INTELLIGENT DIALOG EVALUATION SYSTEM AND ASSOCIATED METHODS

RELATED APPLICATION

[0001] This application claims priority to U.S. Patent Application No. 63/552,516, titled "System and Method for Artificially Intelligent Medical History Evaluation Instrument," and filed Feb. 12, 2024, which is incorporated herein by reference in its entirety.

BACKGROUND

[0002] Taking a complete and thorough medical history (MHx) during an interview between a healthcare professional and a patient is a crucial skill for the healthcare professional. The interview process involves various elements, including documenting the interview participants, chief complaint (CC), history of present illness (HPI), past medical history (PMHx), and reviewing organ systems (ROS). Beyond documentation, it's essential to establish a trusting relationship between the patient (Pt) and healthcare provider (HCP).

[0003] Training in MHx capture requires developing clear communication skills, and avoiding narrow medical vocabulary, technical jargon, and slang. The HCP should ask questions clearly, respectfully, and empathetically, demonstrating active listening to a patients' responses. This approach ensures comprehensive understanding and accurate recording of medical history, facilitating effective diagnosis and treatment. Moreover, empathetic communication builds rapport, enhancing patient comfort and cooperation. The integration of technical knowledge and interpersonal skills in MHx capture is pivotal in healthcare, directly impacting patient outcomes and healthcare quality.

[0004] Large language models (LLMs) have become increasingly popular and utilized in various prompt/response applications such as search engines, customer service, etc. LLMs often include generative AI models trained based on extensive datasets built on historical or bulk general information. To interact with an LLM, a prompt is given to the LLM, and the LLM provides a response back.

SUMMARY

[0005] One aspect of the present embodiments includes the realization that current training schema for medical history (MHx) evaluation interviews are limited, in that each trainee, whether a student or current medical professional, requires a trainer (e.g., a person to assess the performance of the trainee) in the room during the medical interview. However, an additional person in the room with the patient raises additional HIPPA issues that require resolution prior to the interview. Moreover, there is a limited number of trainers available, and therefore not every interview undergoes an evaluation process, because there are not enough trainers/professors available to review every medical interview. Where a trainer reviews a transcript or recording of the trainee's interview with the patient, the trainer is limited to information in the transcript, may exhibit an inherent bias, and may be unaware of cultural norms associated with the interviewee.

[0006] The present embodiments solve these problems by implementing an artificially intelligent (AI) medical history evaluation system and associated methods to evaluate multiple social aspects of speech during the medical history evaluation interview, including, but not limited to, politeness, empathy, and the use of jargon or specialized vocabulary. The medical history evaluation interview is automatically transcribed and excerpts from the transcript are analyzed by the AI dialog evaluation system, which provides suggestions on how the trainee may improve their performance in these areas. Many new virtual reality (VR) and avatar-based patient interviewing and physical examination platforms, offer immediate speech-to-text transformation, and provide an output that is suitable for evaluation by an AI-based system.

[0007] In certain embodiments, the techniques described herein relate to a method for artificially intelligent medical history interview evaluation, including: receiving captured data of a medical history evaluation interview between a patient and a medical provider; transcribing audio of the captured data into transcribed text; segmenting the transcribed text into segmented lines according to a speaker of the transcribed text within the audio; generating a plurality of prompts, each prompt corresponding to one of a plurality of interview analysis variables, to control a large language model (LLM) to analyze each of the segmented lines in context of the transcribed text; transmitting the plurality of prompts to the LLM; receiving LLM responses from the LLM for each of the plurality of prompts; analyzing the LLM responses with respect to a scoring rubric; and generating a detail report defining performance of the medical provider during interaction between the patient and the medical provider.

[0008] In certain embodiments, the techniques described herein relate to a system for artificially intelligent medical history evaluation, including: a capture device configured to capture information an medical history evaluation interview between a patient and a medical provider; an evaluation device having at least one processor and memory storing non-transitory executable instructions that, when executed by the processor operate to control the evaluation device to: receive, from the capture device, captured data of the medical history evaluation interview; transcribe audio of the captured data into transcribed text; segment the transcribed text into segmented lines according to a speaker of the transcribed text within the audio; generate a plurality of prompts, each prompt corresponding to one of a plurality of interview analysis variables, to control a large language model (LLM) to analyze each of the segmented lines in context with the transcribed text; transmit the plurality of prompts to the LLM; receive LLM responses from the LLM for each of the plurality of responses; analyze the LLM responses with respect to a scoring rubric; and generate a detail report defining performance of the medical provider during interaction between the patient and the medical provider.

[0009] In certain embodiments, the techniques described herein relate to a method for artificially intelligent debriefing dialog evaluation, including: receiving captured data of a debriefing dialog for a scenario-based medical simulation; transcribing audio of the captured data into transcribed text; segmenting the transcribed text into segmented lines according to a speaker of the transcribed text within the audio; generating a plurality of prompts, each prompt corresponding to one of a plurality of interview analysis variables, to control a large language model (LLM) to analyze each of the segmented lines in context of the transcribed text; transmit-

ting the plurality of prompts to the LLM; receiving LLM responses from the LLM for each of the plurality of prompts; analyzing the LLM responses with respect to a scoring rubric; and generating a detail report defining performance of a trainee demonstrating scenario-based medical simulation based on the debriefing dialog.

BRIEF DESCRIPTION OF THE FIGURES

[0010] FIG. **1** is a schematic diagram showing one example artificial intelligence (AI) dialog evaluation system for evaluating a medical history interview, in embodiments.

[0011] FIG. **2** shows example processing of transcribed text of FIG. **1** into segmented text, in embodiments.

[0012] FIG. **3** is a schematic diagram illustrating one example detailed report generated by the evaluation device of FIG. **1**, in embodiments.

[0013] FIG. **4**A shows a portion of the detailed report of FIG. **3**, illustrating selections of fields to cause pop-out of an additional detail windows, in embodiments.

[0014] FIG. **4**B shows example content of a first additional detail window of FIG. **4**A, in embodiments.

[0015] FIG. **4**C shows example content of a second additional detail window of FIG. **4**A, in embodiments.

[0016] FIG. **5** is a flowchart illustrating one example method for artificially intelligent dialog evaluation, in embodiments.

[0017] FIGS. **6**A and **6**B show an example summary report generated by the evaluation device of FIG. **1**, in embodiments.

[0018] FIG. **7** shows one example supervisor report template used by the evaluation device of FIG. **1** to generate a supervisor summary of the medical history interview, in embodiments.

[0019] FIG. **8** shows one example scoring table generated by the scoring module of FIG. **1**, in embodiments.

[0020] FIG. **9** shows one example summary section generated by the evaluation device of FIG. **1**, in embodiments.

[0021] FIG. **10** shows one example section of summary report and/or scoring rubric corresponding to a first example category, in embodiments.

[0022] FIGS. **11**A, **11**B, and **11**C show one example debriefing dialog between a group of participants in a scenario-based medical simulation as captured by capture device and transcribed into transcribed text.

[0023] FIGS. **12**A and **12**B show a detailed report with a heading block identifying context details of debriefing dialog and example result sections that define performance of a trainee of the scenario-based medical simulation as determined from a debriefing dialog evaluated against scoring rubric, in embodiments.

[0024] FIG. **13** shows one example summary report section generated by the LLM and/or the scoring module of the system of FIG. **1** based on analysis of the debriefing dialog, in embodiments.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[0025] It should be noted that the described system and methods do not provide medical analysis of a captured medical history. Rather, the systems and methods disclosed herein analyze a trainee's techniques for capturing the medical history of a patient during an interview with the patient.

[0026] A large library of anonymous medical interviews has been accumulated for machine learning. A scoring template, or scoring rubric, and an interactive user interface for individual patient interviews provides each student/trainee with detailed scoring and coaching after the interview is complete. Certain embodiments provide additional real-time feedback, including one or more of audio feedback, visual feedback, and haptic feedback to the interviewee during the interview. The AI dialog evaluation system saves thousands of faculty and instructor hours annually by eliminating the need for manual interview scoring. The AI dialog evaluation system is capable of monitoring multiple trainees simultaneously. The AI dialog evaluation system interacts with large language models (LLMs) in specific ways to decrease the time required to provide feedback, and to reduce/eliminate inter-and intra-evaluator variability and bias.

[0027] FIG. **1** is a schematic diagram showing one example AI dialog evaluation system **100** for evaluating a medical history interview **101**, in embodiments. System **100** includes a capture device **102** that captures audio, video, and/or other audio/visual information, shown as information **103**, of medical history evaluation interview **101** between a medical provider **114** and a patient **116**. System **100** also includes an evaluation device **106** that includes a processor **110** and a memory **112**. Evaluation device **106** may be implemented as one or more of an embedded device, a stand-alone computer, a server, and a cloud service.

[0028] Capture device **102** may include any one or both of a microphone and a camera. Capture device **102** may be a stand-alone device, or may be integrated into another device, such as any one or more of a laptop computer, a desktop computer, a smart phone, a tablet, a medical device, and other such devices. In certain embodiments, capture device **102** hosts an application or web-browser that assists in capture of information **103** of interview **101**. Capture device **102** sends information **103** to evaluation device **106** where it is stored as captured data **104** (e.g., raw or unmodified digital data corresponding to information **103**) in memory **112**. In certain embodiments, evaluation device **106** is implemented with capture device **102**. Where evaluation device **106** is separate from capture device **102**, both capture device **102** and evaluation device **106** may communicate (wirelessly and/or wired) via a network **108**. Network **108** is implemented using one or more of Ethernet, Wi-Fi, Cellular, Bluetooth, ANT+, and any other similar wired or wireless connectivity protocol.

[0029] Evaluation device **106** includes data, software, and firmware, stored within memory **112**, that implements functionality of evaluation device **106** as described herein. Processor **110** may be any type of circuit or integrated circuit capable of performing logic, arithmetic, control, and input/output operations. For example, processor **110** may include one or more of a microprocessor with one or more central processing unit (CPU) cores, a graphics processing unit (GPU), a digital signal processor (DSP), a field-programmable gate array (FPGA), a system-on-chip (SoC), a microcontroller unit (MCU), and an application-specific integrated circuit (ASIC). Processor **110** may also include a memory controller, bus controller, and other components that manage data flow between processor **110**, memory **112**, and other components connected to processor **110** and evaluation device **106**. For example, memory **112** stores non-transitory computer-readable instructions that, when

executed by processor **110**, control processor **110** to implement the described functionality of evaluation device **106**. It should be appreciated that none, all, or some but not all components and/or functionality of evaluation device **106** may be implemented using a cloud computing service, such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud, Hostwinds, Cloudways, Hostinger, and the like.

[0030] In one example of operation, capture device **102** is positioned within a patient room, or a training facility, and positioned to monitor interaction between medical provider **114** and patient **116**. Patient **116** may be an actual patient, or may be a training patient. In certain embodiments, capture device **102** is a component of a virtual training device, such as a virtual reality (VR) device, where patient **116** is a virtual patient. Examples of VR devices include, but are not limited to, the PCS Spark Virtual Patient System (by Team PCS North America LLC), SimX VR System (by SimX, Inc.), Osso VR Systems (by Osso VR, Inc.), Apple Vision Pro implementing a virtual patient software program, etc.

Data Preparation for Submission to LLM

[0031] Evaluation device **106** communicates (e.g., via network **108**) with at least one LLM **118**, which is controlled by evaluation device **106** to analyze captured data **104** to provide interview analysis of the monitored interaction between medical provider **114** and patient **116**. In certain embodiments, LLM **118** is a locally-installed/managed LLM within evaluation device **106**. In other embodiments, LLM **118** is remote from evaluation device **106** and evaluation device **106** communicates with LLM **118** via network **108**, whereby evaluation device **106** transmits prompts to, and receives responses from LLM **118** via API calls.

[0032] To use LLM **118** efficiently, evaluation device **106** conditions the data used in the interactions with LLM **118**. Simply transmitting captured data **104** without preparing captured data **104** would result in inefficient use of LLM **118** and would prevent real-time analysis and feedback of the monitored interaction between medical provider **114** and patient **116**. Accordingly, evaluation device **106** includes a data preparation module **120** that performs a transcription of captured data **104** into a transcribed text **122**. Data preparation module **120** is software (e.g., computer-readable instructions executable by processor **110**) implementing one or more algorithms that control processor **110** to process **104** prior to interaction with LLM **118**.

[0033] Data preparation module **120** may implement various voice-to-text transcription protocols, known in the art, to transcribe information **103** of captured data **104** into transcribed text **122**. Data preparation module **120** then segments transcribed text **122** into a segmented text **124** that associated each line of transcribed text **122** with a given speaker.

[0034] FIG. **2** shows example processing of transcribed text **122** into segmented text **124**, in embodiments. In one example of operation, to generate segmented text **124**, data preparation module **120** identifies each line, or set of adjacent lines, that is/are associated with a given speaker (e.g., a first speaker is medical provider **114** and a second speaker is patient **116**). Each line/set of lines may be assigned a reference number **204** that may be used to identify the line/set of lines subsequently. The example of FIG. **2** shows a first five segmented lines **202**(**1**)-**202**(**5**), that are lines of transcribed text **122** processed into segmented text **124**, whereas the remaining lines of transcribed text **122** are not fully segmented. The reference numerals **204**(**1**)-(**5**) of segmented lines **202**(**1**)-(**5**) each include one of a "P" to identify patient **116** as speaker, and "D" to identify medical provider **114** (e.g., doctor) as a speaker. Although shown as a complete set of transcribed lines, processing of captured data **104** into transcribed text **122** and segmented text **124** may occur in real-time as each speaker is talking. That is, processing of captured data **104** does not wait until medical history interview **101** is completed.

[0035] Evaluation device **106** is configured with a set of interview analysis variables **128**, each with a corresponding settings module **130**, that define metrics for analyzing medical history interview **101**, and in particular for analyzing segmented text **124**. For segmented lines **202**, data preparation module **120** generates one or more prompts **126**(**1**)-**126**(N) for use with LLM **118**, where each prompt **126** represents a query that causes LLM **118** to analyze medical history interview **101** according to one or more interview analysis variables **128**. Prompts **126** are generated to cause LLM **118** to analyze each segmented line **202** of segmented text **124** in the context of interview **101**. For example, a first prompt **126**(**1**) may query segmented line **202**(**1**) "01D" against one or more interview analysis variables **128**. A second prompt **126**(**2**) may query line "02P" against one or more of interview analysis variables **128**. Data preparation module **120** generates prompts **126** such that each of segmented lines **202** of segmented text **124** are associated with at least one prompt **126**.

Interview Analysis Variables

[0036] Interview analysis variables **128** may be configured by an administrator or may be generated at least in part by an automated process. In one example, interview analysis variables **128** are generated by controlling an LLM to identify appropriate metrics for analyzing a medical history interview by analyzing one or more of textbooks, well-established medical sources, and so on. In one example, Bate's Guide to Physical Examination and Medical History, Robert Wood Johnson Foundation (2017): *Patient Centered Medicine: Guidebook for History Taking and Physical Examination*, the American College of Physicians syllabi, WHO, and LCME were analyzed to determine interview analysis variables **128**(**1**)-**128**(M). For example, each interview analysis variable **128** may defining analysis parameters for LLM **118** based on one or more of: (1) Healthcare Provider identification (ID); (2) Patient/Client Subject Identification; (3) Comfort Check; (4) Proper Mode of Addressing Patient; (5) Chief Complaint; (6) Health Provider Identification (HPI); (7) Past Medical History (PMHx); (8) Family and Social History; (9) Familial, Genetic History and Travel History; (10) Review of Systems (ROS); (11) Rephrasing; (12) Reformulation; (13) Empathetic expression; (14) Politeness score; (15) Medical or technical jargon; and/or (16) Wrapping up & Conclusion.

[0037] Each variable of interview analysis variables **128** may be further configurable through an associated settings module **130**. Settings module **130** is, for example, a python code module, or the like, that includes a definition of constraints for the associated interview analysis variable **128**. As an example, where interview analysis variable **128** corresponds to "medical or technical jargon," the associated settings module **130** may include a list of medical or technical terms, or a list of approved resources to use when analyzing a given term or set of terms of segmented text **124**.

Each interview analysis variable **128** may include the corresponding settings module **130**. Alternatively, settings module **130** may be transmitted to LLM **118** and used by LLM **118** for configure to set a context for analyzing segmented text **124**, or certain segmented lines **202**, for each prompt **126** received by the LLM **118**.

[0038] Advantageously, since each interview analysis variables **128** has a corresponding settings module **130**, each setting module **130** is easier to manipulate and change, making changing of a given setting efficient. For example, since cultural norms may vary by country, state, region, etc., settings module **130** corresponding to interview analysis variables **128** for empathy, or proper mode of addressing the patient, or any other variable affected by cultural norms, may be individually configured by adjusting associated settings module **130** accordingly. Thus, system **100** is easily adapted for use in different regions, and/or for different characteristic (e.g., gender, sexual preference, race, ethnic identity, etc.) of one patient versus another patient.

LLM Interface Module

[0039] Evaluation device **106** may also include an LLM interface module **140**. LLM interface module **140** may be implemented as software (e.g., non-transitory computer readable instructions executable by processor **110**) that controls processor **110** to implement the following functionality.

[0040] LLM interface module **140** queues each prompt **126**, received from data preparation module **120**, for example, for transmittal to LLM **118**. In certain embodiments, each prompt **126** is submitted sequentially to LLM **118**. In other embodiments, each prompt **126** is transmitted to an individual instance of LLM **118**, such that prompts **126** are transmitted in parallel. The number of individual instances of LLM **118** may be configured based on the types of prompt **126**, the types of interview analysis variables **128**, etc. Advantageously, the use of multiple instances of LLM **118** and parallel transmission of prompts **126** enables faster interfacing with the LLM **118** such that results and analysis of the medical interview are provided in an expedited manner. In certain embodiments, evaluation device **106** uses sixteen instances of LLM **118** such that the sixteen different interview analysis variables **128**, discussed above, are processed simultaneously, such that segmented text **124** of medical provider **114** is sampled as many as **36,000** times for transcribed text **122** of one-hundred lines. Sequential query/response processing by LLM **118** may take twenty minutes or more; however, by grouping prompts **126** and using multiple instances of LLM **118** to query the prompts in parallel, the query/response processing time may be reduced to less than one minute.

[0041] LLM interface module **140** receives LLM responses **142(1)**-**142**(N) generated by LLM **118** for each prompt **126(1)**-**126**(N), and stores each LLM response **142** in memory **112**. Each LLM response **142** includes an identification of interview analysis variable **128** for which it was generated. For example, LLM response **142(1)** may identify any of interview analysis variables **128** based on the one or more interview analysis variable defined by prompt **126(1)**. Where prompt **126** identifies more than one interview analysis variable **128**, LLM **118** generates more than one LLM response. Thus, there may be more LLM responses **142** than prompts **126**, since each LLM response **142** is associated with one interview analysis variable **128**.

[0042] Certain LLM responses **142** may be "binary" (e.g., "yes" or "no") for a given interview analysis variable **128**. For example, where interview analysis variable **128** relates to whether or not patient **116** is properly identified, the corresponding LLM response **142** includes "yes" when LLM **118** determines that patient **116** was properly identified and "no" when patient **116** was not properly identified. Certain LLM responses **142** that are binary may include additional information such as indicating a portion of transcribed text **122** that satisfied (or did not satisfy) the corresponding interview analysis variables **128**. Continuing the above example, LLM response **142** may include reference number **204(2)** to indicate that segmented line **202(2)** included a proper identification of patient **116**.

[0043] Certain LLM responses **142** may be a scaled responses. Scaled responses may provide a sliding score for a given interview analysis variable **128**. As an example, the "Empathy" and/or "Politeness" variable may include a sliding scale indicative of a perceived level of empathy and/or politeness. To generate a scaled response, LLM **118**, or LLM interface module **140**, may generate a series of prompts **126** for the corresponding interview analysis variables **128**, as follows. A first scaled-response prompt **126** is generated to include at least one segmented line **202** for analysis. A second scaled-response prompt **126** is associated with the first scaled-response prompt to query LLM **118** to provide a first example text scoring at a low end of the scale. A third scaled-response prompt **126** associated with the first and second scaled-response prompts queries LLM **118** to provide a second example text scoring at a high end of the scale. A fourth scaled-response prompt **126** associated with the first, second, and third scaled-response prompts queries LLM **118** to provide a rating of the segmented line, the first example text, and the second example text on the scale. In certain embodiments, the rating is generated by LLM **118** using a cosine similarity function. The second and third scaled responses may be configured according to cultural norms associated with patient **116** as well, by utilizing associated settings module **130** to alter the requirements of LLM **118** to meet those cultural norms.

Scoring Module

[0044] Evaluation device **106** may also include a scoring module **150**. Scoring module **150** may be implemented as software (e.g., non-transitory computer readable instructions executable by processor **110**) that controls processor **110** to implement the following functionality. Scoring module **150** receives LLM responses **142** stored in memory **112** by LLM interface module **140** and configures them into a detailed report **152** based on a scoring rubric **156**. Scoring rubric **156** corresponds to interview analysis variables **128**, for example, and may provide a template for generating detailed report **152**. In certain embodiments, detailed report **152** is an interactive scoring rubric **156** that includes information from LLM responses **142**.

[0045] FIG. **3** is a schematic diagram illustrating one example detailed report **152** generated by evaluation device **106** of FIG. **1**, in embodiments. Detailed report **152** may be an interactive rubric where certain fields may be selected, via interaction with detailed report **152** on a computing device, to view further detail on how and/or why LLM response **142** associated with that field was generated. Detailed report **152** is formed with a plurality of columns including an info/skill column **302**, an output column **304**, a

criteria column **306**, an explanation column **308**, a source number column **310**, a source lines column **312**, and a section column **314**. Each row of detailed report **152** corresponds to one defined criteria of criteria column **306**.

[0046] FIGS. **4**A, **4**B, and **4**C are schematic diagrams illustrating expansion of two fields **402** and **404** of detailed report **152** in response to selection by a user, to show additional detail, in embodiments. FIG. **4**A shows a portion **400** of detailed report **152** where the user has selected field **402**, causing pop-out **406** of an additional detail window **412** and selection of field **404**, causing pop-out **408** of an additional detail window **414**. FIG. **4**B shows example content of additional detail window **412** of FIG. **4**A, illustrating details of terms that were not visible on portion **400** of LLM response **142**, in embodiments. FIG. **4**C shows example content of additional detail window **414**, illustrating many lines of segmented text **124** that are relevant to the corresponding line of portion **400** of detailed report **152**, in embodiments. Advantageously, by providing interactive detailed report **152**, a user (e.g., medical provider **114**) may see details of selected fields that would otherwise not be displayed.

[0047] In this example, field **402** corresponds to a detailed explanation of a medical terminology scoring criteria and field **404** corresponds to a corresponding portion of segmented text **124**. Additional detail window **412** displays the detailed explanation corresponding to field **402** and, in this example, provides rationale of why the "Sputum" term use in the segmented text **124** indicated a poor score, notably that "the doctor used the term in a question directed at the patient's parent without providing a definition or synonym that a 7^{th} grader could easily understand, such as 'phelgm' or 'mucus that comes up when coughing'. Therefore the term was not adequately explained for someone with the knowledge of a 7^{th} grader." Particularly, LLM **118** has identified the term "Sputum" as medical jargon, and as a likely term that is unknown to the patient. For example, the associated settings module **130** corresponding to interview analysis variables **128** for the medical terminology scoring defines that LLM **118** compare terminology to terms known to a person with the knowledge of a 7^{th} grader. Associated settings module **130** may be configured with parameters other than the knowledge of a 7^{th} without departing from scope hereof. The user selects field **404** to view a portion of segmented text **124** associated with the given term/source for that criteria. Fields **402** and **404** are used for examples, and other fields of **152** may be similarly selected to show corresponding additional detail windows.

[0048] Output column **304**, which may also be referred to as a score column, may display either binary or scaled values as discussed above for each variable (e.g., interview analysis variables **128**). The criteria includes a definition of required criteria associated with each interview analysis variables **128**. Criteria column **306** may display information identified by LLM **118** analyzing the context of transcribed text **122** (e.g., the entire transcribed text as currently determined), particularly where each prompt **126** includes the specific line within segmented text **124** to be analyzed, as well as the remainder of the transcribed text **122**. Explanation column **308** may display a description provided by LLM **118** within LLM response **142** of how and why the given criteria/interview analysis variables **128** was analyzed and why the displayed score in output column **304** was provided. Source number column **310** and source lines column **312** definition the location of the processed text within segmented text **124** and/or transcribed text **122** that led to the results displayed in the fields of criteria column **306** and explanation column **308**.

[0049] Scoring module **150** may further generate a summary report **154** (see FIG. **1**). Appendix B, attached hereto, provides an example summary report **154**. It includes a scoring section including numerical scores for one or more of the interview analysis variables **128**, as well as a summary section. The summary section may be a plain-language text generated via a query/response process with LLM **118**. The summary section may further be generated using a standard format where LLM **118** and/or scoring module **150** are used to complete specific fields within the standard format. FIGS. **6**A and **6**B show an example summary report **154** generated by evaluation device **106** of FIG. **1** with a summary section where the language in brackets [] is inserted using LLM **118** and/or scoring module **150**.

[0050] As discussed above, evaluation device **106** may monitor the interaction between medical provider **114** and patient **116** in real time. The use of LLM **118** may additionally occur in real time as the interview is occurring. If the generated detailed report **152** and or summary report **154** indicates a poor score, the evaluation device **106** may interact with a device worn by medical provider **114**, such as a watch or smartphone, and provide real-time feedback to the medical provider **114** during the interaction between medical provider **114** and patient **116**. Additionally or alternatively, where the score generated by scoring module **150** is below a predefined threshold, the summary report **154** may include a further detailed section that includes description of how the medical provider **114** could have altered the interaction between medical provider **114** and patient **116**.

[0051] FIG. **5** is a flowchart illustrating one example method **500** for artificially intelligent dialog evaluation of a medical history interview. Method **500** may be implemented using one or more components of system **100**, including but not limited to evaluation device **106** and LLM **118**. Method **500** includes one or more of the following blocks.

[0052] Block **505** is optional in that block **505** may be performed external to method **500**. In block **505**, method **500** defines one or more interview analysis variables to be analyzed by the LLM. In one example of block **505**, for one or more of interview analysis variables **128**, one or more associated settings modules **130** are defined, as discussed above, and transmitted by evaluation device **106** to LLM **118** to configure the LLM for analysis of each prompt **126**.

[0053] In block **510**, method **500** receives captured data of an interaction between a patient and a medical provider trainee. In one example of block **510**, evaluation device **106** receives captured data **104** from capture device **102** and stores captured data **104** within memory **112**.

[0054] In block **520**, method **500** transcribes audio within the captured data to transcribed text. In one example of block **520**, data preparation module **120** generates transcribed text **122**, which is stored within memory **112**.

[0055] In block **530**, method **500** segments the transcribed text into lines defined by speaker within the audio. In one example of block **530**, data preparation module **120** segments transcribed text **122** in to segmented text **124**, which is stored within memory **112**.

[0056] In block **540**, method **500** generates a plurality of prompts for submission to an LLM, each prompt requesting the LLM to analyze one of the lines in context with the entire

transcribed text. In one example of block **540**, data preparation module **120** generates one or more prompts **126**, each based on one or more of interview analysis variables **128**.

[0057] In at least one embodiment, generating a plurality of prompts **126** includes generating a prompt for a binary analysis of one or more interview analysis variables **128**.

[0058] In at least one embodiment, generating a plurality of prompts **126** includes generating a prompt for a scaled analysis of one or more interview analysis variables **128**. In at least one embodiment, generating a prompt **126** for a scaled analysis includes: generating a first scaled-response prompt **126** to include a corresponding segmented text **124** to be analyzed; generating a second scaled-response prompt **126**, associated with the first scaled-response prompt, querying LLM **118** to provide an example of a low-scoring end of the scale; generating a third scaled-response prompt **126**, associated with the first and second scaled-response prompts, querying LLM **118** to provide an example of a high-scoring end of the scale; and generating a fourth scaled-response prompt **126** associated with the first, second, and third scaled-response prompts, querying LLM **118** to provide a rating of the provided segmented text **124** on a given scale. The rating may include a cosine similarity analysis.

[0059] In block **550**, method **500** transmits the prompts to the LLM and receiving responses thereto. In one example of block **550**, evaluation device **106** transmits prompt **126** to LLM **118**, receives LLM responses **142** from LLM **118**, and stores each LLM response **142** within memory **112**. Transmitting the prompts **126** to LLM **118** may include transmitting one or more of prompts **126** in parallel to individual instances of LLM **118**, thereby analyzing the prompts concurrently to receive LLM responses **142** quicker, as compared to sending prompts **126** sequentially to LLM **118**.

[0060] In block **560**, method **500** analyzes the responses to generate a scoring rubric of the interaction between the patient and the medical provider trainee. In one example of block **560**, scoring module **150** generates one or more of detailed report **152** and summary report **154** as discussed above. Evaluation device **106** may output the detailed report **152** and/or summary report **154** to an external device for evaluation by medical provider **114**.

[0061] FIG. **7** shows one example supervisor report template **700** that is used by evaluation device **106** of FIG. **1** to generate a supervisor summary of the medical history interview **101**, in embodiments. Capital letters represent fields that are automatically completed by scoring module **150** based on LLM response **142** and/or detailed report **152**. As appreciated, system **100** may be configured to generate many types of report based on information in detailed report **152**.

[0062] FIG. **8** shows one example scoring table **800** generated by scoring module **150** of FIG. **1**, in embodiments. Scoring table **800** provides an overview of the evaluated performance of medical provider **114** during medical history interview **101**. Scoring table **800** may form part of summary report **154**.

[0063] FIG. **9** shows one example summary section **900** generated by evaluation device **106** of FIG. **1**, in embodiments. In this example, summary section **900** relates to a section titled "Obstetrical/Gynecological And Reproductive History." Evaluation device **106** may generate summaries of other sections of scoring rubric **156**. These summaries provide feedback on strengths, weaknesses, and overall performance of **114** during medical history interview **101**.

[0064] System **100** is versatile and adaptable to evaluate different human interactions and provide feedback and reports on performance of a trainee. The following example illustrates this versatility by adapting system **100** for evaluating debriefing techniques for scenario-based medical simulations (SBMS).

Artificially Intelligent Comprehensive Debriefing Assessment Tool

[0065] Traditionally, multiple inventories and batteries have been used to assess how educators and facilitators conduct debriefing sessions with healthcare students after simulation scenarios. These tools, designed to "train the trainer," typically involve another senior educator or facilitator observing the SBMS debriefing in real-time or through a delayed video recording, using inventory or paper-based scoring sheets. However, this process is both time-consuming and subject to evaluator bias.

[0066] The process of debriefing involves reviewing a conversation between two or more people as they evaluate and review a simulated event or activity in which participants explore, analyze, and synthesize their actions, their thought processes, their emotional states, and other information to improve performance in real-life situations. High participant engagement is a hallmark of strong debriefings because it leads to deeper levels of learning and increases the likelihood of transfer to the clinical setting.

[0067] Existing tools evaluate the strategies and techniques used to conduct debriefings by examining concrete behaviors, based on the theory that people learn and apply information when they have an experiential context in which they can use it. However, in almost all situations, debriefing assessment is carried out for the benefit of the person leading the debriefing, usually a senior facilitator or faculty member.

[0068] To address the limitations of the prior art, system **100** is adapted to evaluate the transcript of the debriefing session by combining best practices of validated debriefing tools and thereby create a superior amalgam of attributes. Accordingly, system **100** is configured with specific criteria and algorithms to provide a comprehensive assessment of debriefing conversations. System **100** evaluates debriefing transcripts and assigns scores across elements considered critical to debriefing technique and outcome. These elements align with the key objectives of SBMS debriefing, including education and skill development, patient safety and quality improvement, professional development, research and development, and health system integration. The use of AI-powered analysis by system **100** offers several advantages over traditional methods. System **100** provides a highly individualized evaluation with coaching suggestions, custom-tailored to the particular debriefing and facilitator. Moreover, system **100** delivers feedback directly to the trainee (e.g., the individual being evaluated), eliminating the need for assessment by another person. Advantageously, this approach not only saves time but also makes the critique and evaluation free from individual bias and ensures privacy in conveying the results to the trainee. By leveraging AI technology, system **100** addresses the limitations of conventional assessment methods while providing a more efficient, consistent, and objective evaluation of debriefing sessions. Accordingly, system **100** significantly enhances the quality of debriefing in medical simulation education, ultimately

contributing to improved learning outcomes for trainees and better patient care. System **100** is adapted to apply specific categories and schema for evaluation of debriefing performance without bias. However, it's important to note that while system **100** represents a promising advancement, human expertise and judgment still play a crucial role in medical education.

[0069] In one example, interview analysis variables **128** and associated settings module **130** are adapted to allow data preparation module **120** and LLM interface module **140** to control LLM **118** to analyze a debriefing dialog based on the following categories: (1) Identification section for the session; (2) Structure of lesson (10% of composite scoring); (3) Style of delivery (15% of composite score); (4) Professional demeanor (10% of composite score); (5) Use of instructional tools (15% of composite score); (6) Student engagement (20% of composite score); (7) The attentiveness of students to active learning (20% of composite score); and (8) Cumulative Impressions (10% of composite score). As appreciated, interview analysis variables **128** and associated settings module **130** may be configured with more or fewer categories without departing from the scope hereof. System **100** is adapted to simplify scoring into range of 1-5, where 1 represents poor, 2 represent fair, 3 represents good, 4 represents very good and 5 represents excellent. For each of the above categories, LLM **118** (e.g., implemented using ChatGPT4) is trained using specific criteria and examples of how to assign a score to a training debriefing sessions.

[0070] FIG. **10** shows one example section **1000** of summary report **154** and/or scoring rubric **156** corresponding to a first example category. FIGS. **11**A, **11**B, and **11**C show one example debriefing dialog **1100** between a group of participants in a SBMS as captured by capture device **102** and transcribed into transcribed text **122**. In this embodiment of system **100**, LLM **118** is trained to analyze debriefing dialog **1100** against scoring rubric **156** which is defined based on the categories listed above. Since system **100** is adaptable to use composite sections, it may also be adapted to incorporate any new useful evaluations that appear in the public domain. Advantageously, system **100** may evolve as assessment methodologies are developed and disclosed in the literature. One significant advantage of system **100**, because of its AI driven assessment, is that it is free of inter-rater variability and free of individual bias.

[0071] FIGS. **12**A and **12**B show one example detailed report **152** generated by scoring module **150** using scoring rubric **156** and based on LLM responses **142** resulting from evaluation of transcribed text **122** by LLM **118** in response to generated prompts **126**, interview analysis variables **128**, and corresponding settings modules **130**. System **100** is adapted to implement a composite of four debriefing assessment or evaluation tools and may be referred to as a "Composite Debriefing Assessment Tool" (CDAT). FIG. **12**A shows detailed report **152** with a heading block **1200** identifying context details of debriefing dialog **100** and examples result sections **1250** that define performance results determined from debriefing dialog **1100** against scoring rubric **156**, for example.

[0072] FIG. **13** shows one example summary report **154** section generated by LLM **118** and/or scoring module **150** based on analysis of debriefing dialog **1100** based on prompt **126**, interview analysis variables **128**, associated settings module **130**, and scoring rubric **156**.

[0073] Advantageously, system **100** generates detailed report **152** and summary report **154** without bias. Accordingly, system **100** provides an improvement over the art by removing bias that is unavoidably subconsciously imparted by a human evaluator.

[0074] Changes may be made in the above methods and systems without departing from the scope hereof. It should thus be noted that the matter contained in the above description or shown in the accompanying drawings should be interpreted as illustrative and not in a limiting sense. The following claims are intended to cover all generic and specific features described herein, as well as all statements of the scope of the present method and system, which, as a matter of language, might be said to fall therebetween.

What is claimed is:

1. A method for artificially intelligent medical history interview evaluation, comprising:

receiving captured data of a medical history evaluation interview between a patient and a medical provider;

transcribing audio of the captured data into transcribed text;

segmenting the transcribed text into segmented lines according to a speaker of the transcribed text within the audio;

generating a plurality of prompts, each prompt corresponding to one of a plurality of interview analysis variables, to control a large language model (LLM) to analyze each of the segmented lines in context of the transcribed text;

transmitting the plurality of prompts to the LLM;

receiving LLM responses from the LLM for each of the plurality of prompts;

analyzing the LLM responses with respect to a scoring rubric; and

generating a detail report defining performance of the medical provider during interaction between the patient and the medical provider.

2. The method of claim **1**, the plurality of interview analysis variables each defining analysis parameters for the LLM.

3. The method of claim **1**, further comprising transmitting one or more settings, respectively defined for each of the plurality of interview analysis variables, to the LLM to configure the LLM to analyze the segmented line in context with the transcribed text.

4. The method of claim **1**, wherein transmitting the prompts to the LLM includes transmitting at least two of the prompts in parallel to at least two different instances of the LLM.

5. The method of claim **1**, the generating a plurality of prompts comprises generating a prompt for a binary analysis of one or more of the plurality of interview analysis variables.

6. The method of claim **1**, the generating a plurality of prompts comprises generating a prompt for a scaled analysis of one or more of the plurality of interview analysis variables.

7. The method of claim **6**, the generating a prompt for a scaled analysis comprises:

generating a first scaled-response prompt including the segmented line;

generating a second scaled-response prompt associated with the first scaled-response prompt querying the LLM to provide an example of a low-scoring end of a scale;

generating a third scaled-response prompt associated with the first and second scaled-response prompts querying the LLM to provide an example of a high-scoring end of the scale; and

generating a fourth scaled-response prompt associated with the first, second, and third scaled-response prompts querying the LLM to provide a rating on a given scale.

8. The method of claim **7**, wherein the rating is a cosine similarity analysis between LLM responses associated with the first, second, and third scaled-response prompts.

9. A system for artificially intelligent medical history evaluation, comprising:

a capture device configured to capture information an medical history evaluation interview between a patient and a medical provider;

an evaluation device having at least one processor and memory storing non-transitory executable instructions that, when executed by the processor operate to control the evaluation device to:

receive, from the capture device, captured data of the medical history evaluation interview;

transcribe audio of the captured data into transcribed text;

segment the transcribed text into segmented lines according to a speaker of the transcribed text within the audio;

generate a plurality of prompts, each prompt corresponding to one of a plurality of interview analysis variables, to control a large language model (LLM) to analyze each of the segmented lines in context with the transcribed text;

transmit the plurality of prompts to the LLM;

receive LLM responses from the LLM for each of the plurality of responses;

analyze the LLM responses with respect to a scoring rubric; and

generate a detail report defining performance of the medical provider during interaction between the patient and the medical provider.

10. The system of claim **9**, wherein the patient is a virtual patient, and the capture device is a component of a virtual training device.

11. The system of claim **9**, wherein the LLM operating on a processing device, or a group of processing devices, of the evaluation device.

12. The system of claim **9**, wherein the LLM is executed remotely from the evaluation device.

13. The system of claim **9**, the plurality of interview analysis variables each defining analysis parameters for the LLM.

14. The system of claim **9**, the non-transitory executable instructions further comprising non-transitory executable instructions that, when executed by the processor operate to control the evaluation device to transmit one or more settings, respectively defined for each of the plurality of interview analysis variables, to the LLM to configure the LLM to analyze the segmented line in context with the transcribed text.

15. The system of claim **9**, the non-transitory executable instructions further comprising non-transitory executable instructions that, when executed by the processor operate to control the evaluation device to transmit at least two of the prompts in parallel to at least two different instances of the LLM.

16. The system of claim **9**, the non-transitory executable instructions further comprising non-transitory executable instructions that, when executed by the processor operate to control the evaluation device to generate a prompt for a binary analysis of one or more of the plurality of interview analysis variables.

17. The system of claim **9**, the non-transitory executable instructions further comprising non-transitory executable instructions that, when executed by the processor operate to control the evaluation device to generate a prompt for a scaled analysis of one or more of the plurality of interview analysis variables.

18. The system of claim **17**, the non-transitory executable instructions further comprising non-transitory executable instructions that, when executed by the processor operate to control the evaluation device to:

generate a first scaled-response prompt including the segmented line;

generate a second scaled-response prompt associated with the first scaled-response prompt querying the LLM to provide an example of a low-scoring end of the scale;

generate a third scaled-response prompt associated with the first and second scaled-response prompts querying the LLM to provide an example of a high-scoring end of the scale; and

generate a fourth scaled-response prompt associated with the first, second, and third scaled-response prompts querying the LLM to provide a rating on a given scale.

19. The system of claim **18**, wherein the rating is a cosine similarity analysis between LLM responses associated with the first, second, and third scaled-response prompts.

20. A method for artificially intelligent debriefing dialog evaluation, comprising:

receiving captured data of a debriefing dialog for a scenario-based medical simulation;

transcribing audio of the captured data into transcribed text;

segmenting the transcribed text into segmented lines according to a speaker of the transcribed text within the audio;

generating a plurality of prompts, each prompt corresponding to one of a plurality of interview analysis variables, to control a large language model (LLM) to analyze each of the segmented lines in context of the transcribed text;

transmitting the plurality of prompts to the LLM;

receiving LLM responses from the LLM for each of the plurality of prompts;

analyzing the LLM responses with respect to a scoring rubric; and

generating a detail report defining performance of a trainee demonstrating scenario-based medical simulation based on the debriefing dialog.

* * * * *