

Empowering Private Tutoring by Chaining Large Language Models

Anonymous ACL submission

Abstract

Artificial intelligence has been applied in various aspects of online education to facilitate teaching and learning. However, few approaches have been made towards a complete AI-powered tutoring system. In this work, we explore the development of a full-fledged intelligent tutoring system based on large language models (LLMs). The proposed system CHATTUTOR, powered by state-of-the-art LLMs, is equipped with automatic course planning and adjusting, informative instruction, and adaptive quiz offering and evaluation. CHATTUTOR is decomposed into three inter-connected core processes—*interaction*, *reflection*, and *reaction*. Each process is implemented by chaining LLM-powered tools along with dynamically updated memory modules. To demonstrate the mechanism of each working module and the benefits of structured memory control and adaptive reflection, we conduct a wide range of analysis based on statistical results and user study. The analysis shows the designed processes boost system consistency and stability under long-term interaction and intentional disruptions, with up to 5% and 20% increase in performance respectively. Meanwhile, we also compare the system with scripts from real-world online learning platform and discuss the potential issues unique to LLM-based systems.

1 Introduction

Online education, along with artificial intelligence (AI) technology, brought the aspiration of personalized tutoring within reach (Bloom, 1984). AI has been used to assist education in multiple aspects, ranging from adaptive content recommendation (Costello and Mundy, 2009), automatic performance evaluation (McDonald et al., 2013; Grivokostopoulou et al., 2017), to personalized instruction and dynamic feedback (Bhutoria, 2022; Tang et al., 2020; Gordon et al., 2016; Grawemeyer et al., 2016; Dzikovska et al., 2014). Although a

few early approaches have been made towards a dialogue-based intelligent tutoring system (ITS) (Graesser et al., 2001; Rus et al., 2014), most of them are domain-specific and focus primarily on guiding the users to solve a pre-defined problem. Nevertheless, a more ultimate exploration lies in the pursuit of a full-fledged AI-driven tutoring system with greater flexibility and generalizability that teaches in a systematic and consistent manner on a much broader range of knowledge.

While previous works often employ diverse techniques jointly, including learner style classification (Nihad et al., 2017), data mining (Echeverria et al., 2015), Bayesian learning (Grawemeyer et al., 2016), etc, the recent emergence of large language models (LLMs) (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023; Team et al., 2023), like ChatGPT (OpenAI, 2022), has broadened our imagination on new designs of intelligent tutoring systems. LLMs impressed people firstly with the ability to generate and transform information following human instructions, then with the potential in task planning and tool usage. Given their extraordinary ability in providing detailed and accessible content and summarizing information, LLMs become great sources for consulting a wide range of knowledge. Additionally, LLMs have shown remarkable capabilities in engaging in multi-turn dialogues, which makes them well-suited for tutoring applications where there is a back-and-forth exchange with a student. Although challenges like hallucination remain (Bang et al., 2023), LLMs can adapt to expertise in specific domains and pedagogical strategies with further fine-tuning.

In this work, we explore the potential of employing generative large language models to build a full-fledged dialogue-based personalized tutoring system. One specialty about an ITS compared to other LLM-powered agents is that, education is a long-term cooperative process accomplished by AI and human users jointly. A well-designed

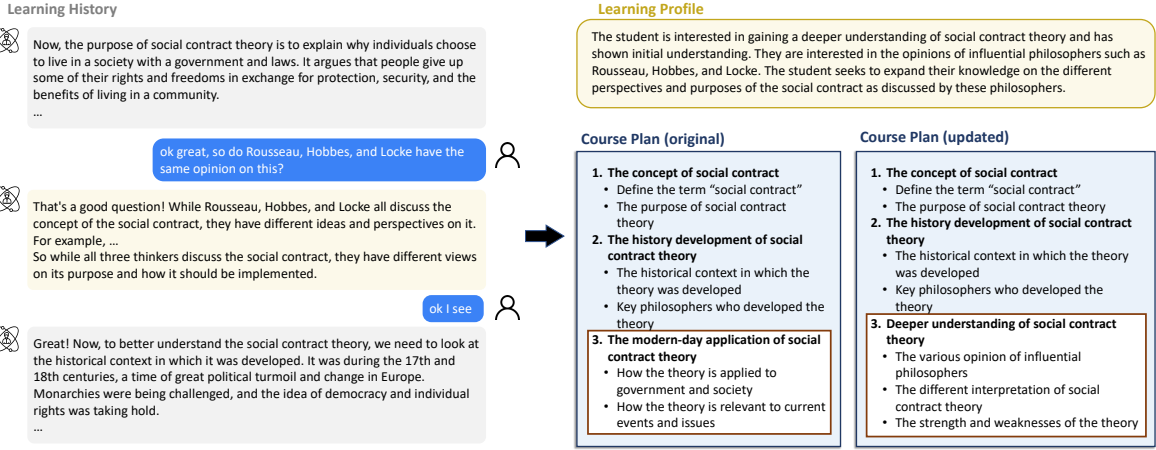


Figure 1: An example of the learning progress. The left side is the user interface directly controlled by the interaction process. The right side is the backend memory changes brought by reflection and reaction processes.

tutoring system should correctly infer the human user’s mental state to achieve adaptive teaching, and meanwhile, the user should be informed about the learning progress in order to cooperate more effectively. Therefore, the system faces some unique challenges in how to maintain an explainable and consistent control over the learning progress and how to attain an effective way of capturing human users’ behavior and responses correspondingly.

To solve the problem, CHATTUTOR has a modularized design, encompassing three core processes—*interaction*, *reflection*, and *reaction*, each further composed of chained LLM-powered tools to perform atomic tasks. The processes are connected to each other through various memory modules, which store the essential data describing the overall progress and support update and retrieval. Such design enables **structured memory control** and **adaptive reflection** on status quo. CHATTUTOR carries out every stage in education systematically and dynamically, including instructing, question answering, exercise offering and evaluating. Note that the system is for general purpose of learning for everyone who can read and type. It is currently not designed for specific target audience since we focus more on exploring the potential of LLMs in education process in general.

Evaluation of the proposed system is conducted by analyzing statistics collected from learning logs and subjective human feedback. Results show that CHATTUTOR can satisfactorily handle various educational activities, including adaptive course plan design, consistent instructing, impromptu question answering, etc. Meanwhile ablation study demonstrates the advantage in performance stability and

consistency over long-term interaction and disruption situations. It further supports the plausibility of the design principle.

2 Overview of CHATTUTOR

CHATTUTOR is essentially a dialogue-based tutoring system that aims to help learners acquire knowledge on one given topic systematically. As shown in Figure 1, the whole learning process is carried out in natural language conversations, with time-to-time backend reflection and reaction to update the memories. This section gives a general picture of the system’s workflow. We start with explaining the design principles by introducing three underlying processes within the system. Then, we briefly go over the components employed to realize each process. Finally, we provide a complete introduction to how each process and components work together. Note that the proposed system mainly explores personalized tutoring and adaptive system reflection with chained LLMs, while we do acknowledge the potential of fallacious and biased output due to inherent flaws in LLMs.

2.1 Design Principles

The system design demonstrates the breakdown into three core system processes: *Interaction*, *Reflection*, and *Reaction*. They each has a modularized implementation and is connected to one another to form an execution loop that empowers the whole education process.

Interaction. The interactive dialogue between the system and the user is the media for tutoring and learning, and therefore is the major process of the designed system. LLMs like ChatGPT can inter-

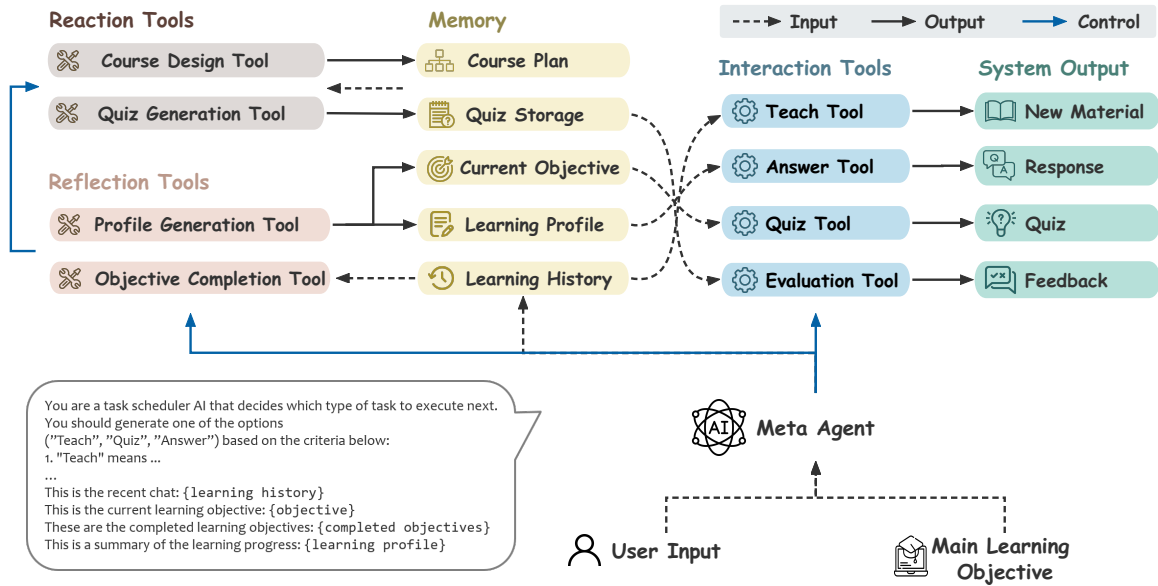


Figure 2: An overview of the system’s modularized implementation and execution in a single round of conversation.

act with users in a responsive and robust way in daily chit-chat. However, tasking them with long-term purposeful interaction is still tricky given the restriction on context length. As for educational purposes, it is especially important to keep the interaction on track, meanwhile ensuring its accessibility and informativeness.

Reflection. To facilitate interaction, we devise a reflection process to generate high-level insights on the learning progress, which serves as global information (Park et al., 2023). Meanwhile, it is expected to help adjust system response dynamically based on user preference and behavior to achieve personalized tutoring.

Reaction. Along with reflection, reaction refers to the automatically triggered system behavior afterward, including adjustment of course plan and quiz generation. It differs from the interaction process in that *interaction* is always triggered directly by a new round of response, while *reaction* is performed at the backend from time to time, subsequent to the reflection process.

2.2 Components

Tools. Under the principled design, each process is embodied by a set of tasks performed either sequentially or in parallel. For instance, there are diverse ways of engaging with the student, such as providing instructions, addressing questions, administering quizzes, and offering feedback. This variation in approaches complicates the development of a single unified solution. We therefore devise separate modules for each specific task to ensure

performance. We term those modules as “tools”, and that each tool is a task-specific prompted LLM responsible for generating system output or updating memories, as shown in Figure 2. For example, *interaction* is broken down into four types of response in terms of education function, each hosted by one well-prompted tool. At each round, only one tool is used to generate the response.

Memories. Apart from tools, data storage is required to host information generated by reflection and reaction processes, while also supporting querying and retrieving. We propose four types of memories to record the progress and current status of learning, each stored in distinct data format and supports different ways of querying and updating. Another critical feature of the memories is that they serve as a linkage between different sets of tools to pass on information to control tool output. The detailed description of each tool and memory can be found in the next section.

Meta Agent. Above all three processes, we introduce meta agent, the single access of the control flow. It is powered by LLM and prompted to decide what specific tasks to execute next. See Figure 2 for an example prompt for controlling the interaction process. The template contains helpful information retrieved from the memory and asks for an output deciding the type of interaction process. In our implementation, the meta agent only controls the interaction tools, while we set a fixed time interval for the execution of reflection process.

2.3 Overview of Control Flow

Above all, all designs serve for the ultimate goal of better interaction with the users. The system reflects from time to time to update cognition on the overall progress, and in turn refines the interaction production with new insights. At the frontend, the user first inputs what to learn with desired difficulty level. Then the system automatically calls the course design tool to generate the initial course plan, and starts the conversation accordingly. Upon receiving a new round of user input, the meta agent decides which interaction tool to use and the tool executes the task correspondingly to generate a new response with queried information from memories. At the backend, the reflection tools are triggered to reflect on the status quo and update the learning profile and current objective, after which the reaction tools will be triggered immediately to generate new quiz questions and update the course plan.

As shown in Figure 2, the right side represents the interaction process that is presented on the user interface, while the left side demonstrates the backend processes that are responsible for generating and updating memory modules. Practically, throughout each dialogue session, the reflection and reaction processes run alternatively at the backend, where the output result is periodically utilized by the interaction process to produce the final response to the user in each round. Table 1 presents detailed usage of each tool in the three processes, including the input and output memory content, and the condition for tool execution. The learning proceeds until all objectives in the course plan have been completed.

3 Structured Control and Adaptive Reflection

What sets ITS apart from other LLM-powered agents is the cooperative feature between human user and the system, which contributes to the three-process design and the separation of memory and tools. Similarly, two crucial mechanisms throughout the system implementation ensure both the communicative and adaptive feature of the system. In this section, we present the rationale for the mechanisms-*structure memory control* and *adaptive reflection and reaction* and the detailed application of each within the system.

3.1 Structured Memory Control

The interactive and cooperative feature of a tutoring system calls for the need to communicate with the

users effectively about current and future progress. Meanwhile, it is also important to keep the system itself aware of the progress to ensure better stability. We therefore design various memory modules in different storage format and function to support the mutual communication.

Course Plan. The course plan is stored in a tree structure, with each node representing an atomic topic in the course, and its child nodes representing the sub-topics. The course is expected to be taught and learnt in depth-first traverse order. Current objective is a pointer pointing to the next uncompleted objective node in the tree to denote current progress. Such structure allows for presentation to the users, informing them of the overall status of learning, while enabling mechanistic operation by the system. Specifically, the course design tool is used at the beginning of the learning to generate the initial course plan based on user’s desired topic and difficulty level. In each new round of conversation, objective completion tool is called to update status of the current objective based on the recent and relevant learning history. Then, the course design tool is asked to update the current course plan while maintaining the completed objectives.

Learning History. As for learning histories, the recent history is stored as plain text that can be directly fed into the LLM, whereas the relevant history is stored along with their embedding and queried with cosine similarity with embeddings of current objective upon usage. The detailed mechanism can be seen in Figure 3. Meanwhile, the benefit of explicitly collecting completed objectives also extends to more effective quiz offerings, which will be detailed in the next section.

Quiz Storage. A crucial function of an ITS system is to offer adaptive quiz that helps the learner review and master what has been learned. In CHAT-TUTOR, the LLM is instructed to generate quiz questions based on learning materials and formatted as a structured json string. The questions are stored corresponding to each learning objective with explicit status marking, and extracted in order whenever the quiz tool is called. The quiz questions will keep appearing in the next quiz batch until it is answered correctly by the learner.

3.2 Adaptive Reflection and Reaction

Reflection and reaction processes at the backend are closely bound to each other in CHAT-TUTOR, whereas reflection process generates high-level insight about the learning progress, reaction pro-

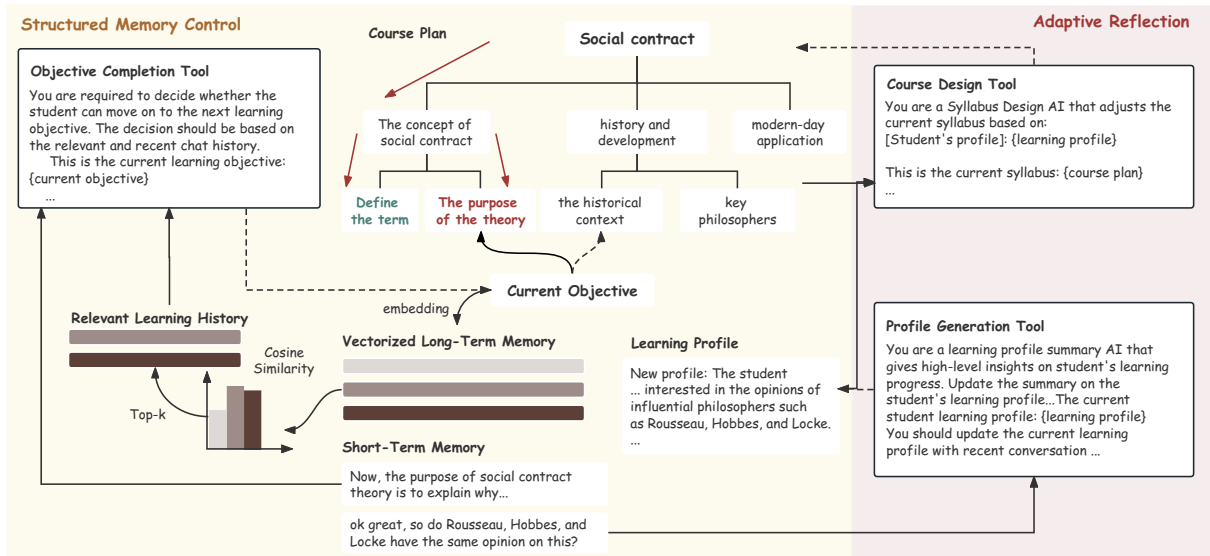


Figure 3: A detailed illustration of how course plan is stored and manipulated structurally and how reflection process helps customize the reaction followed.

Process	Tool Name	Execution Condition	Input	Output/Update
Interaction	Teach Tool	Meta agent	Learning history, Current objective, Learning profile	System output
	Answer Tool	Meta agent	Learning history	System output
	Quiz Tool	Meta agent	Quiz storage, Learning profile	System output
	Evaluation Tool	Quiz	Learning history	System output
Reflection	Profile Generation	Each round	Learning history, Learning profile	Memory: Learning profile
	Objective Completion	Each round	Learning history, Current objective	Memory: Current objective
Reaction	Course Design	Profile generation	Course plan, Learning profile	Memory: Course plan
	Quiz Generation	Objective completion	Learning history, Current objective	Memory: Quiz storage

Table 1: A summary of the detailed tool usage. “Input” means the memories are part of the prompt.

315 cess updates the structured memory modules based on the insights. The sequential and dependent
 316 design could more accurately infer about the status and enhance the stability and adaptiveness of the
 317 system behavior.
 318
 319

320 **Learning Progress Control.** One core function of reflection process is to control the learning progress
 321 by determining when to move on to next learning objective. Objective completion tool is prompted to
 322 judge whether the current objective has been completed based on queried learning histories based
 323 on text embedding similarity. Whenever the current objective is considered completed, the status
 324 of the course plan and the pointer will automatically be updated. Meanwhile, the reaction process
 325 “quiz generation” will be triggered as well. It is prompted to generate several representative quiz
 326 questions for the completed current objective, with relevant queried learning history provided, which
 327 ensures the stability and relevance of the generated quiz question. The questions are stored in the mem-
 328 ory until the meta agent decides it is time for a
 329
 330
 331
 332
 333
 334
 335
 336

337 quiz, where the corresponding quiz questions are retrieved from the storage for the completed objec-
 338 tives and further filtered and organized by the quiz tool in the interaction process to present to the user.
 339
 340

341 **Profile and Course Plan Update.** Apart from reflecting on objective status, an important compo-
 342 nent is user’s learning profile. Learning profile summarizes what the user has learned and gives
 343 high-level insight on the user’s preference based on conversation history. Though not directly presented
 344 to users, it is crucial to the stability of system’s memory update and overall understanding of the
 345 learning process. It is especially useful as part of the input to course design tool to provide direction
 346 for course plan adjustment.
 347
 348
 349
 350
 351

352 At each round of conversation, the system automatically reflects on the recent dialogue and up-
 353 dates learning profile with profile generation tool. The tool is a prompted LLM that takes recent di-
 354 alogues and current profile summary as input and outputs a new version of learning profile, summariz-
 355 ing the learned knowledge, the user’s reaction and
 356
 357
 358

359 preference mainly. Then it is fed into the course
360 design tool for a new version of course plan gener-
361 ation. Figure 3 provides an example of profile gener-
362 ation tool generating high-level insight of “*the*
363 *student seeks to expand their knowledge on the*
364 *different perspectives and purposes of the social*
365 *contract as discussed by the philosophers.*” after
366 the user asks a follow-up question about different
367 philosophers’ opinion. This further leads to an
368 updated course plan that enhances deeper under-
369 standing of the theory.

370 4 Experiments

371 To demonstrate and analyze the features of our
372 tutoring system, the experiments are conducted in
373 two folds. We invite a number of users to learn a se-
374 ries of pre-defined topics using the system. During
375 interaction, we collect critical statistics and record
376 the conversation for future analysis. After learning
377 completes, the users are required to answer a ques-
378 tionnaire to rate their experience with the system
379 from multiple perspectives. We also develop abla-
380 tion systems to better understand the effect of each
381 process and module.

382 4.1 Experimental Design

383 **System Setup.** In addition to the main system,
384 we implement two ablation systems with only par-
385 tial functions. Specifically, we have one system
386 without reflection process and another with both
387 reflection and reaction processes removed. The de-
388 tailed description of the two ablation systems can
389 be found in Appendix B.1. We collect 80 different
390 topics as the main learning objectives for evaluation
391 and randomly assign them with different difficulty
392 levels. The topics cover a wide range of academic
393 subjects and some daily life phenomena, varying
394 in granularity and language format. In evaluation,
395 each topic is learned independently with three sys-
396 tems, making up altogether 240 courses. Further
397 details about topic collection and tool prompt de-
398 sign and used can be found in Appendix B.2.

399 **Participants.** We invite 13 average adult users who
400 are proficient in English to participate in learning.
401 Every single course is randomly assigned to one
402 user, while we make sure that each participant does
403 not get repeated course topics.

404 **Statistical Analysis.** We collect various statistics
405 for analysis, including (1) **Complexity of course**
406 **plan** reflects the ability to design adaptive course
407 plan; (2) **Average length of system response** and

408 **Average number of objectives per response** are
409 indicators of instruction informativeness; (3) **Fre-**
410 **quency of course plan update** shows the reflective
411 feature of the system; and (4) **Frequency of in-**
412 **course quiz** explores the pattern of quiz offerings.
413 **Survey Design.** After completing the course, the
414 learner is required to answer a survey composed of
415 9 questions targeting different aspects of the sys-
416 tem. Each question is a statement to be rated on a
417 1~5 scale, where higher scores mean better qual-
418 ities. Table 8 in appendix presents the statements
419 by category.

420 **Stability Analysis.** To further demonstrate the ben-
421 efit of our system design, intentional disruption to
422 the learning process is conducted to test the stabil-
423 ity of the system. We take 15 most difficult topics
424 (difficulty 4~5) in our list and for each learning
425 process apply 3 consecutive rounds of attacks with
426 ChatGPT generating a random question. The sys-
427 tem is expected to answer robustly to the question
428 and resume the original learning course after the
429 disruption. we manually annotate the quality of
430 resumed learning after disruption and the quality
431 of response given to attack questions. The details
432 for annotation can be found in Appendix B.5.

433 4.2 Results

434 **Statistical Results.** Figure 4 presents the statisti-
435 cal characteristics related to teach tool, including
436 average length of output and the average number
437 of objectives covered in each generation. Overall
438 all systems can generate tailored output according
439 to difficulty level. Higher difficulty comes along
440 with longer and more informative output. It means
441 the teach tool is successfully aware of the dynamic
442 prompting controlled by difficulty. What is worth
443 noting is that the main system generates signifi-
444 cantly longer output with the smallest variation. It
445 demonstrates that the main system is able to con-
446 sistently generate informative content on the given
447 topic, which is further testified by the number of ob-
448 jectives covered in each output. This phenomenon
449 shows the benefits of structured memory control,
450 where the objective completion tool reflects on and
451 updates the current objective so that the teach tool
452 is prompted to give new materials on one specific
453 atomic objective.

454 Figure 5 showcases the behavior of course de-
455 sign tool. Note that all systems start with the same
456 initial course plan. The system without reflec-
457 tion relies on learning history in course plan up-
458 dates, while the main system takes advantage of

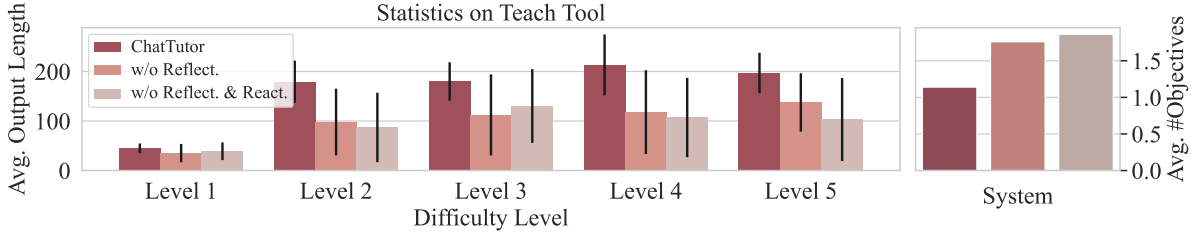


Figure 4: Average output length (calculated by the number of words) and the number of objectives covered in each output for different systems. Average number of objectives are manually annotated with 50 randomly sampled response from each system.

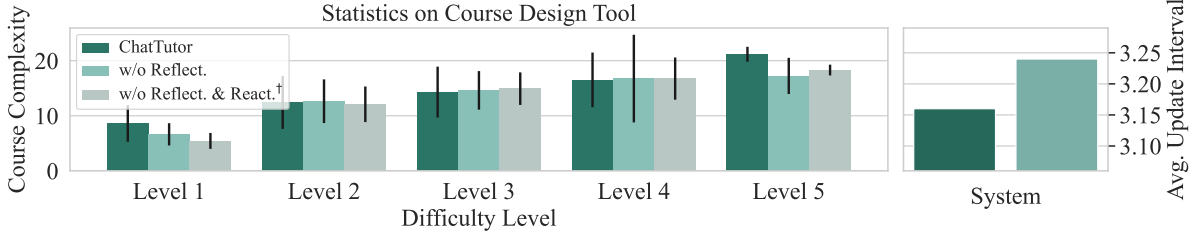


Figure 5: Average course plan complexity (calculated by the number of objectives) and update interval (calculated by the number of conversation rounds in between) by course design tool for different systems. † means this is the baseline statistics as the system without reflection or reaction processes has a fixed course plan throughout learning.

the learning profile. Similarly, the tool can generate difficulty-aware course plans, seen in a positive correlation between complexity and difficulty. It is also observed that the future updates in course plan tend to do more addition than deletion, making the course plan more complex than the initial state. Note the main system shows more stability in complexity control compared to ablation system without intermediate learning profile generation step, demonstrating the effectiveness of reflection process stabilizing the reaction process. In terms of frequency of updates, we observe that the main system seems more prone to making updates, meaning that it is easier to detect helpful information in a compressed learning profile compared to verbose learning history.

As for in-course quizzes, the average interval between quizzes for the main system and ablation system are 6.91 and 10.20, respectively. Without the backend processes, the meta agent has little information available to suggest quiz generation. We also calculate the average number of objectives covered in each quiz for the main system, the result is reasonably 3.96 with good stability.

Survey Results. Table 2 and Table 3 present results on survey questions. Overall, with powerful ChatGPT, all systems demonstrate promising usability and quality. The designed reflection and reaction processes offer advantages in complex learning set-

tings. For intricate course plan updates, the main system achieves better coherence and compatibility with difficulty levels. Through reflection and memory querying, the system maintains better control with fewer repetitive instructions during longer conversations. While all three systems seem to generate highly relevant quiz questions, the main system generates much more fine-grained questions querying about detail information in the learning material, while the questions from ablation systems are largely answerable based on the course plan (see Appendix C). This, together with better evaluation accuracy, showcases the benefit of online quiz storage. However, it should also be noted that although the main system performs satisfactorily on QA consistency, the timeliness in response is compromised by backend processes. It indicates that prompts with global information may interfere with the LLM’s ability to focus locally and generate timely and coherent response.

Stability Analysis. As shown in Table 4, thanks to the reflection and reaction processes, CHATTUTOR has the overall best performance in terms of stability and robustness, with up to 20% increase compared to ablation systems. Whereas there is a clear trade-off between the repetition and omission of course material in the two ablation systems, while both signify instability. The fact that CHATTUTOR tends to ignore user’s irrelevant questions more of-

System	Course Plan			Instruction		Question Answering		Quiz	
	Relevance	Coherence	Compatibility	No Repetition	Accessibility	Timeliness	Consistency	Relevance	Judgment
CHAT TUTOR	4.72	4.51	4.65	4.29	4.77*	4.41	4.82	4.88	4.24
w/o Reflect.	4.71	4.62	4.77	4.42	4.66	4.64	4.85	4.75	4.65
w/o Reflect. & React.	4.97 [†]	4.77 [†]	4.78 [†]	4.38	4.77	4.75	4.95	4.86	4.36

Table 2: Survey results for learning courses at difficulty level 1~3. [†] means the score evaluates the initial course plan only, as no changes in course plan happen throughout the learning process. It could be viewed as the static quality evaluation of course plan generated from scratch. * means p-value < 0.1 using t-test.

System	Course Plan			Instruction		Question Answering		Quiz	
	Relevance	Coherence	Compatibility	No Repetition	Accessibility	Timeliness	Consistency	Relevance	Judgment
CHAT TUTOR	4.87*	4.87**	4.73	4.27	4.53	3.67	5.00	4.87	4.20
w/o Reflect.	4.67	4.60	4.67	3.87	4.87	4.80	4.67	4.40	4.13
w/o Reflect. & React.	4.93 [†]	4.73 [†]	4.80 [†]	4.07	5.00	4.33	4.93	4.93	4.13

Table 3: Survey results for learning courses at difficulty level 4~5. [†] means the score evaluates the initial course plan only, as no changes in course plan happen throughout the learning process. It could be viewed as the static quality evaluation of course plan generated from scratch. * and ** means p-value < 0.1 and < 0.05 using t-test.

System	Repeat	Omit	Response	Overall
CHAT TUTOR	0.50	0.93	0.60	0.68
w/o Reflect.	0.60	0.47	0.40	0.49
w/o Reflect. & React.	0.33	0.67	0.67	0.56

Table 4: Results of stability evaluation. “Repeat” means the repetition of learning materials after disruption and “omit” means some designed topics are skipped during learning. Higher score indicates better quality.

ten also echoes finding in user study and highlights the reconciling effect between robustness and stability and controlability.

4.3 Case Study

In this section, we further demonstrate how CHAT TUTOR behaves with a detailed case study. Specifically, we compare the learning script from CHAT TUTOR and the original script, and discuss the effect of adaptive course design and quiz offering. Meanwhile, we also discuss the potential issues from LLMs like hallucination. To compare with real-world education scenario, we adopt the machine learning course on Coursera platform. In terms of course plan design, it can be seen that CHAT TUTOR can satisfactorily cover the major topics of the course, meanwhile also maintaining a logical dependence between crucial concepts. On the other hand, course plan produced by CHAT TUTOR is more coarse-grained than a human-devised syllabus. It should be noted that while CHAT TUTOR tends to propose a wide range of concepts, real world teaching pays more attention to technical problems, including how to solve a specific machine learning problem and what practical tricks are commonly used. This down-side, of course could be compensated by the adaptiveness and timely

responsiveness of CHAT TUTOR, where users can motivate more in-depth discussion with impromptu questions.

While CHAT TUTOR largely provides accurate information on “machine learning”, in another case featuring “volcanoes”, we find that the system stumbles when users ask for more details. For example, the system makes typical hallucination due to the knowledge cutoff of training data, and also confuses two towns in Hawaii in a specific eruption. When being pushed to provide more information, it hallucinates the eruption duration as well. Problems like this could be mitigated with retrieval-augmented generation technique given a relevant knowledge base. Details about our case study can be found in Appendix D.

5 Conclusion

This work is a pioneering exploration of an LLM-powered intelligent tutoring system, with an emphasis on the possibility of employing LLMs to complete complex and dynamic tasks based on natural language conversation with users. The proposed system, CHAT TUTOR, can satisfactorily complete the core functions of an intelligent tutoring system. As ablation study shows, the three-process system design provides unique benefit in ensuring the stability and consistency of the system behavior, meanwhile maintaining flexibility and adaptiveness. This work demonstrates the promising application of a pure LLM-based purposeful interaction system based on reflection and memory mechanism, which might inspire future efforts in employing LLMs to build increasingly interactive and reflective systems.

577 Limitations

578 CHATTUTOR is among the very first approaches to
579 realize an ITS supported by LLMs only. It should
580 be acknowledged that there are some limitations in
581 the system design and the evaluation methods.

582 Firstly, the system itself inherits the limitations
583 from LLMs. Problems like hallucination and bi-
584 ased generation should be handled before put into
585 real use. The system could be improved in the
586 future in many ways, including but not limited to
587 incorporating retrieval techniques to enhance
588 more faithful content generation, further tuning on
589 existing learning scripts to adapt to pedagogical
590 language, and better modeling of student’s mental
591 state during learning with insight from educational
592 theories.

593 Secondly, we position CHATTUTOR as a com-
594 mon knowledge tutor for all people who are starters
595 in one domain of knowledge. The case study results
596 also suggest the fact that CHATTUTOR is far from
597 a domain expert and therefore the use cases will
598 be further limited. The system could be adapted to
599 more specific scenarios (e.g. K-12) through further
600 fine-tuning of language models on targeted scripts.
601 More targeted adaptation is necessary for its better
602 functioning to cater to learning science purpose.

603 In terms of evaluation, although we strive to
604 provide a wide range of analysis ranging from sta-
605 tistical analysis, user study to case study, we did
606 not deploy the system in real-world usage scenario.
607 The generalizability our findings may be compro-
608 mised. In fact, how to evaluate the success of edu-
609 cation process itself is an open question and trans-
610 ferring traditional evaluation strategy to the case of
611 an LLM-based ITS may be even more challenging.
612 In addition to the ongoing debate and discussion
613 about what makes good education (Graesser et al.,
614 1995; Cohen et al., 1982), AI techniques face their
615 own special concerns such as the validity of gener-
616 ated education content and the potential bias from
617 training data (Feng et al., 2023; Nozza et al., 2022).
618 It should be also acknowledged that properly evalu-
619 ating multi-turn conversation is in itself a topic
620 worth investigating. More standard metrics should
621 be developed other than human evaluation which
622 is rather subjective.

623 Ethical Statement

624 The implemented system realizes a full-fledged in-
625 telligent education process, which can help achieve
626 efficient online personalized tutoring and in turn,

promote education equality. On the other hand, 627
the experimental system faces the risk of generat- 628
ing inaccurate information given the hallucination 629
problem in LLMs. The inherent bias may also 630
mislead the users. It is important to address these 631
issues before a practical application is launched. 632
The participants in the evaluation process are all 633
fairly paid with agreed salaries. 634

References 635

- 636 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama
637 Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
638 Diogo Almeida, Janko Altschmidt, Sam Altman,
639 Shyamal Anadkat, et al. 2023. Gpt-4 technical report.
640 *arXiv preprint arXiv:2303.08774*.
- 641 Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wen-
642 liang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei
643 Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan
644 Xu, and Pascale Fung. 2023. *A multitask, multilin-
645 gual, multimodal evaluation of chatgpt on reasoning,
646 hallucination, and interactivity*.
- 647 Aditi Bhutoria. 2022. *Personalized education and ar-
648 tificial intelligence in the united states, china, and
649 india: A systematic review using a human-in-the-
650 loop model*. *Computers and Education: Artificial
651 Intelligence*, 3:100068.
- 652 Benjamin S Bloom. 1984. The 2 sigma problem: The
653 search for methods of group instruction as effec-
654 tive as one-to-one tutoring. *Educational researcher*,
655 13(6):4–16.
- 656 Benjamin S Bloom and David R Krathwohl. 2020. *Tax-
657 onomy of educational objectives: The classification
658 of educational goals. Book 1, Cognitive domain*.
659 longman.
- 660 Tom Brown, Benjamin Mann, Nick Ryder, Melanie
661 Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind
662 Neelakantan, Pranav Shyam, Girish Sastry, Amanda
663 Askell, et al. 2020. Language models are few-shot
664 learners. In *Proceedings of NeurIPS*.
- 665 Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng,
666 Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan
667 Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al.
668 2023. Vicuna: An open-source chatbot impressing
669 gpt-4 with 90%* chatgpt quality. See [https://vicuna.
670 lmsys.org](https://vicuna.lmsys.org) (accessed 14 April 2023).
- 671 Peter A Cohen, James A. Kulik, and Chen-Lin C. Kulik.
672 1982. *Educational outcomes of tutoring: A meta-
673 analysis of findings*. *American Educational Research
674 Journal*, 19(2):237–248.
- 675 R. Costello and D.P. Mundy. 2009. *The adaptive intel-
676 ligent personalised learning environment*. In *2009
677 Ninth IEEE International Conference on Advanced
678 Learning Technologies*, pages 606–610.

679	Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao,	Arthur C. Graesser, Natalie K. Person, and Joseph P.	737
680	Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and	Magliano. 1995. Collaborative dialogue patterns in	738
681	Maosong Sun. 2023. Ultrafeedback: Boosting lan-	naturalistic one-to-one tutoring . <i>Applied Cognitive</i>	739
682	guage models with high-quality feedback. <i>arXiv</i>	<i>Psychology</i> , 9:495–522.	740
683	<i>preprint arXiv:2310.01377</i> .		
684	Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi	Arthur C. Graesser, Kurt VanLehn, Carolyn P. Rose,	741
685	Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun,	Pamela W. Jordan, and Derek Harter. 2001. Intelli-	742
686	and Bowen Zhou. 2023. Enhancing chat language	gent tutoring systems with conversational dialogue .	743
687	models by scaling high-quality instructional conver-	22(4):39.	744
688	sations. <i>arXiv preprint arXiv:2305.14233</i> .		
689	Sidney D’mello and Art Graesser. 2012. Autotutor and	Beate Grawemeyer, Manolis Mavrikis, Wayne Holmes,	745
690	affective autotutor: Learning by talking with cogni-	Sergio Gutierrez-Santos, Michael Wiedmann, and	746
691	tively and emotionally intelligent computers that talk	Nikol Rummel. 2016. Affecting off-task behaviour:	747
692	back . 2(4):1–39.	How affect-aware feedback can improve student	748
693		learning . In <i>Proceedings of the Sixth International</i>	749
694	Myroslava O. Dzikovska, Natalie B. Steinhauer, Elaine	<i>Conference on Learning Analytics & Knowledge</i> ,	750
695	Farrow, Johanna D. Moore, and Gwendolyn E. Camp-	LAK ’16, page 104–113, New York, NY, USA. As-	751
696	bell. 2014. Beetle ii: Deep natural language under-	sociation for Computing Machinery.	752
697	standing and automatic feedback generation for in-		
698	telligent tutoring in basic electricity and electronics .	Foteini Grivokostopoulou, Isidoros Perikos, and Ioan-	753
699	<i>International Journal of Artificial Intelligence in Ed-</i>	nis Hatzilygeroudis. 2013. An intelligent tutoring	754
700	<i>ucation</i> , 24:284–332.	system for teaching fol equivalence. In <i>The First</i>	755
701		<i>Workshop on AI-supported Education for Computer</i>	756
702	Vanessa Echeverria, Bruno Guamán, and Katherine	<i>Science (AIEDCS 2013)</i> , volume 20.	757
703	Chiluiza. 2015. Mirroring teachers’ assessment of		
704	novice students’ presentations through an intelligent	Foteini Grivokostopoulou, Isidoros Perikos, and Ioannis	758
705	tutor system. In <i>2015 Asia-Pacific Conference on</i>	Hatzilygeroudis. 2017. An educational system for	759
706	<i>Computer Aided System Engineering</i> , pages 264–269.	learning search algorithms and automatically assess-	760
707	IEEE.	ing student performance . <i>International Journal of</i>	761
708		<i>Artificial Intelligence in Education</i> , 27:207–240.	762
709	Gilan M El Saadawi, Eugene Tseytlin, Elizabeth		
710	Legowski, Drazen Jukic, Melissa Castine, Jeffrey	Gautier Izacard, Mathilde Caron, Lucas Hosseini, Se-	763
711	Fine, Robert Gormley, and Rebecca S Crowley. 2008.	bastian Riedel, Piotr Bojanowski, Armand Joulin,	764
712	A natural language intelligent tutoring system for	and Edouard Grave. 2021. Unsupervised dense infor-	765
713	training pathologists: Implementation and evaluation.	mation retrieval with contrastive learning .	766
714	<i>Advances in health sciences education</i> , 13:709–722.		
715		Hsieh S. J. 2014. Algorithm and intelligent tutoring	767
716	Mustafa Al Emran and Khaled Shaalan. 2014. A survey	system design for programmable controller program-	768
717	of intelligent language tutoring systems . In <i>2014</i>	ming. <i>International Journal of Advanced Manufac-</i>	769
718	<i>International Conference on Advances in Computing,</i>	<i>turing Technology</i> , 71:1099.	770
719	<i>Communications and Informatics (ICACCI)</i> , pages		
720	393–399.	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler	771
721		Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon,	772
722	Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia	Nouha Dziri, Shrimai Prabhumoye, Yiming Yang,	773
723	Tsvetkov. 2023. From pretraining data to language	Shashank Gupta, Bodhisattwa Prasad Majumder,	774
724	models to downstream tasks: Tracking the trails of	Katherine Hermann, Sean Welleck, Amir Yazdan-	775
725	political biases leading to unfair NLP models . In	bakhsh, and Peter Clark. 2023. Self-refine: Iterative	776
726	<i>Proceedings of the 61st Annual Meeting of the As-</i>	refinement with self-feedback .	777
727	<i>sociation for Computational Linguistics (Volume 1:</i>		
728	<i>Long Papers)</i> , pages 11737–11762, Toronto, Canada.	Ankit Malpani, Balaraman Ravindran, and Hema A	778
729	Association for Computational Linguistics.	Murthy. 2011. Personalized intelligent tutoring sys-	779
730		tem using reinforcement learning. In <i>FLAIRS Con-</i>	780
731	Goren Gordon, Samuel Spaulding, Jacqueline	<i>ference</i> , pages 561–562.	781
732	Kory Westlund, Jin Joo Lee, Luke Plummer,		
733	Marayna Martinez, Madhurima Das, and Cynthia	Philip M McCarthy, Vasile Rus, Scott A Crossley,	782
734	Breazeal. 2016. Affective personalization of a social	Arthur C Graesser, and Danielle S McNamara. 2008.	783
735	robot tutor for children’s second language skills .	Assessing forward-, reverse-, and average-entailer	784
736	30(1).	indices on natural language input from the intelligent	785
		tutoring system, istart. In <i>FLAIRS Conference</i> , pages	786
		165–170. Citeseer.	787
	Arthur C. Graesser, G. Tanner Jackson, Eric Mathews,	Jenny McDonald, Alistair Knott, Sarah J. Stein, and	788
	Heather H. Mitchell, Andrew M. Olney, Matthew	Richard Zeng. 2013. An empirically-based, tutorial	789
	Ventura, Patrick Chipman, Donald R. Franceschetti,	dialogue system: design, implementation and evalua-	790
	Xiangen Hu, Max M. Louwerse, and Natalie K. Per-	tion in a first year health sciences course .	791
	son. 2003. Why/autotutor: A test of learning gains		
	from a physics tutor with natural language dialog .		

792	Danielle S McNamara, Tenaha P O'Reilly, Rachel M Best, and Yasuhiro Ozuru. 2006. Improving adolescent students' reading comprehension with istart. <i>Journal of Educational Computing Research</i> , 34(2):147–171.	849
793		850
794		851
795		852
796		853
797	Phaedra Mohammed and Permanand Mohan. 2015. Dynamic cultural contextualisation of educational content in intelligent learning environments using icon. <i>International Journal of Artificial Intelligence in Education</i> , 25:249–270.	854
798		855
799		856
800		857
801		858
802	Elghouch Nihad, Yassine Zaoui Seghroucheni, et al. 2017. Analysing the outcome of a learning process conducted within the system als_corr [lp]. <i>International Journal of Emerging Technologies in Learning (Online)</i> , 12(3):43.	859
803		860
804		861
805		862
806		863
807	Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In <i>Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models</i> , pages 68–74, virtual+Dublin. Association for Computational Linguistics.	864
808		865
809		866
810		867
811		868
812		869
813		870
814	TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. <i>OpenAI</i> .	871
815		872
816	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. <i>Advances in Neural Information Processing Systems</i> , 35:27730–27744.	873
817		874
818		875
819		876
820		877
821		878
822	Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. 2017. Prerequisite relation learning for concepts in MOOCs. In <i>Proceedings of ACL</i> , pages 1447–1456.	879
823		880
824		881
825	Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. <i>arXiv preprint arXiv:2304.03442</i> .	882
826		883
827		884
828		885
829		886
830	Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shi Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bo Li, Ziwei Tang, Jing Yi, Yu Zhu, Zhenning Dai, Lan Yan, Xin Cong, Ya-Ting Lu, Weilin Zhao, Yuxiang Huang, Jun-Han Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool learning with foundation models. <i>ArXiv</i> , abs/2304.08354.	887
831		888
832		889
833		890
834		891
835		892
836		893
837		894
838		895
839		896
840		897
841		
842	Rod D Roscoe and Danielle S McNamara. 2013. Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. <i>Journal of Educational Psychology</i> , 105(4):1010.	
843		
844		
845		
846	Vasile Rus, Sidney D'Mello, Xiangen Hu, and Arthur Graesser. 2013. Recent advances in conversational intelligent tutoring systems. 34(3):42–54.	
847		
848		
	Vasile Rus and Arthur C. Graesser. 2006. Deeper natural language processing for evaluating student answers in intelligent tutoring systems. In <i>Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2, AAAI'06</i> , page 1495–1500. AAAI Press.	
	Vasile Rus, Dan Stefanescu, Nobal Niraula, and Arthur C. Graesser. 2014. Deeptutor: Towards macro- and micro-adaptive conversational intelligent tutoring at scale. In <i>Proceedings of the First ACM Conference on Learning @ Scale Conference</i> , pages 209–210. ACM.	
	Gómez S. 2014. Context-aware adaptive and personalized mobile learning delivery supported by uolmp. <i>Journal of King Saud University – Computer and Information Sciences</i> , 26:47.	
	Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.	
	Ying Tang, Joleen Liang, Ryan Hare, and Fei-Yue Wang. 2020. A personalized learning system for parallel intelligent education. <i>IEEE Transactions on Computational Social Systems</i> , 7:352–361.	
	Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. <i>arXiv preprint arXiv:2312.11805</i> .	
	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. <i>arXiv preprint arXiv:2307.09288</i> .	
	Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. <i>Educational psychologist</i> , 46(4):197–221.	
	Kurt VanLehn, Pamela Jordan, and Diane Litman. 2007. Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In <i>Workshop on Speech and Language Technology in Education</i> . Citeseer.	
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.	

A Related Work

Ever since the development of artificial intelligence techniques, methods and tools have been proposed to assist in teaching and learning process. AutoTutor (Graesser et al., 2001) is the first conversation-based intelligent tutoring system, which inspires a number of works followed (McNamara et al., 2006; D’mello and Graesser, 2012; Graesser et al., 2003; Rus et al., 2013; VanLehn, 2011; Dzikovska et al., 2014). In addition to AutoTutor’s application to various fields, enhancement of specific aspects of education are also investigated, including adaptive feedback (Dzikovska et al., 2014; Roscoe and McNamara, 2013), learning material recommendation (S., 2014; Mohammed and Mohan, 2015), and classifying learners (Grawemeyer et al., 2016; Nihad et al., 2017; J., 2014). Commonly adopted techniques include data mining (Echeverria et al., 2015), condition-action rule based (J., 2014; S., 2014), and bayesian based methods (Grawemeyer et al., 2016), and reinforcement learning (Malpani et al., 2011; Gordon et al., 2016). NLP-specific techniques like semantic analysis (Graesser et al., 2001) and textual entailment (Rus and Graesser, 2006; VanLehn et al., 2007; McCarthy et al., 2008) are also adopted. In terms of application field, existing systems often rely on well-structured knowledge bases and therefore only target a single domain, most popular among which are health (McDonald et al., 2013; El Saadawi et al., 2008), computer science (Grivokostopoulou et al., 2013; Mohammed and Mohan, 2015), and language learning (S., 2014; Emran and Shaalan, 2014).

As for applications with LLMs, with proper prompting and chaining, a number of works have exploited LLMs in following diverse instructions (Ouyang et al., 2022; Chiang et al., 2023; Ding et al., 2023; Cui et al., 2023), decomposing tasks (Wei et al., 2023), refining answers (Madaan et al., 2023; Shinn et al., 2023), using external tools (Qin et al., 2023), and simulating human behaviors (Park et al., 2023). While our work focuses on building an interactive tutoring system that works and teaches cooperatively with human users, featuring reflection on user’s behavior and adaptive course design.

B Experimental Details

B.1 Ablation Systems

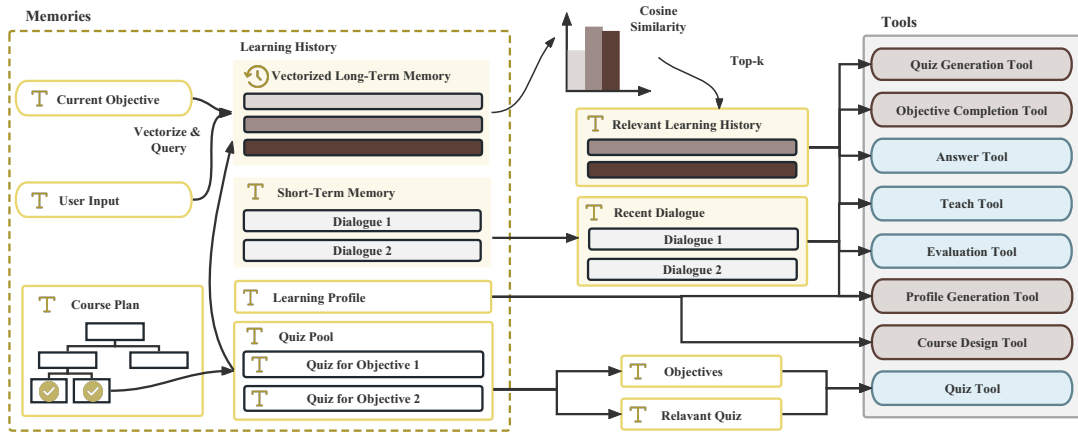
For the first ablation system (**w/o Reflection**), we remove the reflection process only. It means no

learning profile is generated throughout the whole process, and the system does not explicitly judge the objective status. Only the course plan and learning history are available for retrieval. As a result, the reaction process is triggered at a fixed time interval with limited input. For example, the course design tool has only the recent learning history as input without the learning profile, and the teach tool has no access to the current ongoing objective but can only infer from the course plan and learning history implicitly. At the same time, to mitigate the loss of learning profile, we increase short-term learning history volume to up to 10 rounds of conversation. This, however, will exceed the maximum input length for ChatGPT in some extreme cases.

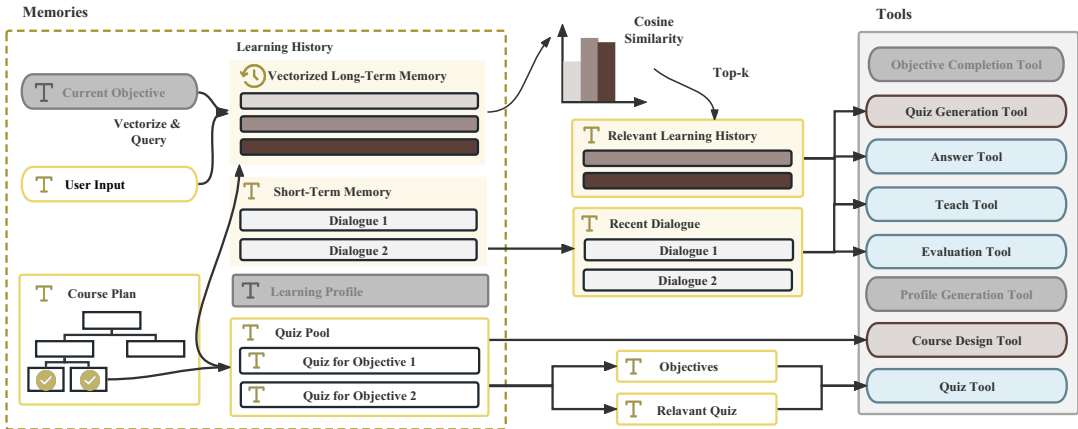
For the second ablation system (**w/o Reflection & Reaction**), both reflection and reaction processes are removed, and long-term learning history is also made unavailable. This means only short-term learning history and the initial course plan can be used in all educational activities. Figure 6 is an illustration of the differences between main system and ablation systems.

B.2 Topic collection

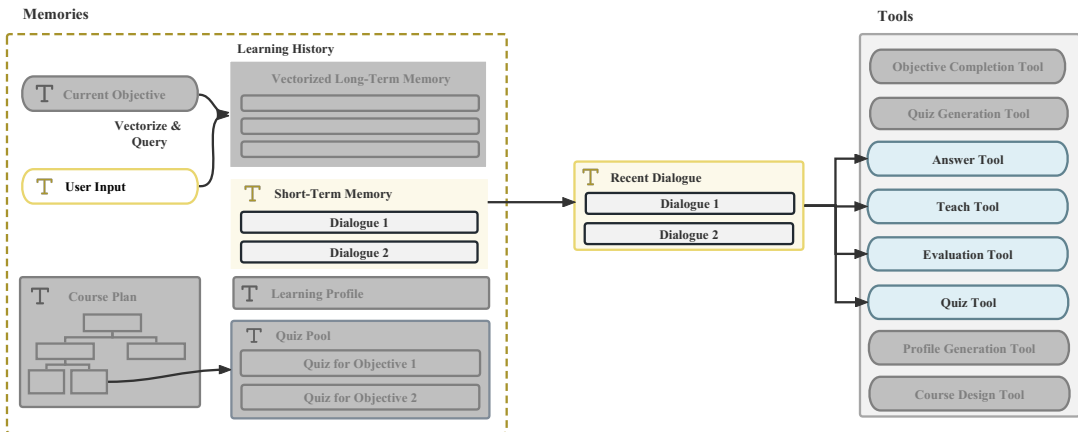
We employ GPT-4 to generate the list of main learning objectives used for system evaluation. We first ask GPT-4 to generate a set of general academic subject (e.g. Physics, Biology, Computer Science, etc.). Then we ask for generation of more fine-grained branches of discipline under each subject (e.g. “Developmental psychology” in psychology subject), and related classic concepts (e.g. “Newton’s first law” in physics subject). Besides, we also include some daily phenomena that may inspire people’s wondering. We encourage GPT-4 to generate a typical list of them in diverse language style. To demonstrate the system’s robustness in dealing with various types of learning objectives, we randomly sample from course-grained subjects, atomic concepts, and daily wondering. To make the learning process more diverse and controllable, we also design 5 difficulty levels according to Bloom’s taxonomy (Bloom and Krathwohl, 2020). The course topics are randomly assigned to difficulty levels. Specifically, we have 15 courses at difficulty level 1, 25 at level 2, 25 at level 3, 10 at level 4, and 5 at level 5. Each course is learned and evaluated with all three systems, respectively. Table 12 is a completed list of all selected topics. In reference to Bloom’s taxonomy, we manually design the difficulty-specific prompt for course de-



(a) Implementation details for the main system.



(b) Implementation details for the first ablation system (w/o Reflection).



(c) Implementation details for the second ablation system (w/o Reflection & Reaction).

Figure 6: Implementation details for ablation systems, in comparison to main system. Modules in gray means the modules are de-activated or non-existent in the system.

sign tool and teach tool for better adaptiveness and clarity. Table 5 and Table 6 present the prompts for course design tool and teach tool.

B.3 User interface

We build a frontend interface to support interaction with the backend system. Figure 7 shows the frontend design used in real system evaluation. The

main component is a chat window through which the user interacts with the system in natural language. On the right side is the current course plan, the completed objectives will be shown in blue (as shown in Figure 7b). Since both ablation systems do not conduct judgment on objective status. We change the next objective status to “completed” at each round of conversation on user interface to

1006
1007
1008
1009
1010
1011
1012
1013

Course Design Prompt

Level 1: Remembering (Knowledge) At this level, the focus is on providing basic facts. Students are required to remember facts, definitions, and concepts. Design a **very** short course.

Level 2: Understanding (Comprehension) This level involves describing and interpreting facts, definitions, and concepts. Design a course as concise as possible.

Level 3: Applying (Application) At this level, students are expected to use acquired knowledge in new and practical situations. The focus is on applying concepts to solve problems and complete tasks. Try to make the course concise and well-structured.

Level 4: Analyzing (Analysis) In this level, the course should break down information into its components to understand the relationships between parts and the overall structure. Try to make the course concise and well-structured.

Level 5: Evaluating (Evaluation) The course should assess the quality, validity, and relevance of information and arguments. Try to make the course concise and well-structured.

Table 5: Prompts for course design tool at each difficulty level.

Teach Prompt

Level 1: Teach in very simple and accessible language. Keep generated text short within a few sentences.

Level 2: Teach in simple and accessible language. Keep language and wording easy to understand.

Level 3: Teach in well-structured language and paragraph. Make information digestible. Try to use structured format to make it clearer, e.g. bullet point.

Level 4 & 5: Use precise language to explain things in a systematic way. Try to use structured format to make it clearer, e.g. bullet point.

Table 6: Prompts for teach tool at each difficulty level.

Difficulty Level	Update Interval	Max. Dialogue Round	Avg. Time Spent (mins)
1	1	10	17
2	1	15	30
3	2	20	40
4	3	25	60
5	4	30	60

Table 7: Designed update interval of learning profile and the maximum round of dialogue in system evaluation.

Course Plan

1. Relevance: The course plan is relevant to the main objective.
 2. Coherence: The course plan is coherent and logical.
 3. Compatibility: The course plan is compatible with difficulty level.
-

Instruction

4. No Repetition: There is no repeated instruction.
 5. Accessibility: The language used is easy to understand.
-

Question Answering

6. Timeliness: The learner’s questions always get immediate response.
 7. Consistency: The response is consistent with learning material.
-

Quiz

8. Relevance: The quiz questions match what has been covered.
 9. Judgment: The quiz evaluation is accurate in parsing and scoring.
-

Table 8: The complete survey questions. Learners are asked to rate the compatibility of each statement on a scale of 1~5.

make them visually indistinguishable to the main system to avoid potential influence in evaluation.

B.4 Implementation Details

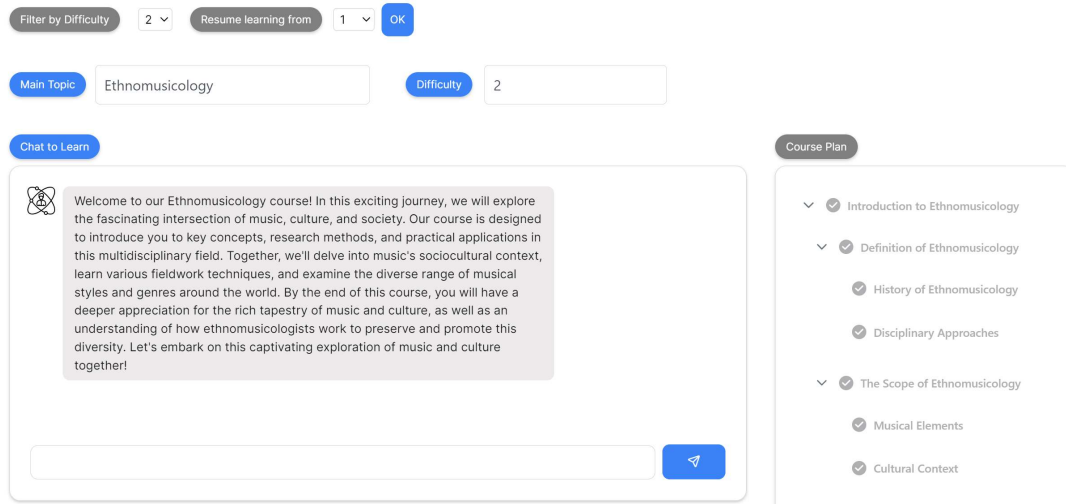
To ensure better usability and system stability, we also set different time intervals for learning profile update for each difficulty level, and control for the maximum round of conversation. Table 7 presents the detailed setting and the average time spent on courses at every difficulty level. Altogether, the evaluation for the three systems with 240 courses takes around 150 hours.

In addition to in-course quiz automatically triggered by meta agent, when all objectives have been completed or the pre-defined maximum round has been reached, the system explicitly calls for the quiz chain to generate a final quiz for the course. For final quiz generation, our system and ablation system without reflection process use similarity-based search for the most relevant 20 pieces of history as input context, while the ablation system with only interaction process uses the overall course plan. As for long-term history storage and query, we use contriever model (Izacard et al., 2021) for embedding texts and cosine similarity score for most relevant record ranking and querying. The implementation of the whole system is based on LangChain¹.

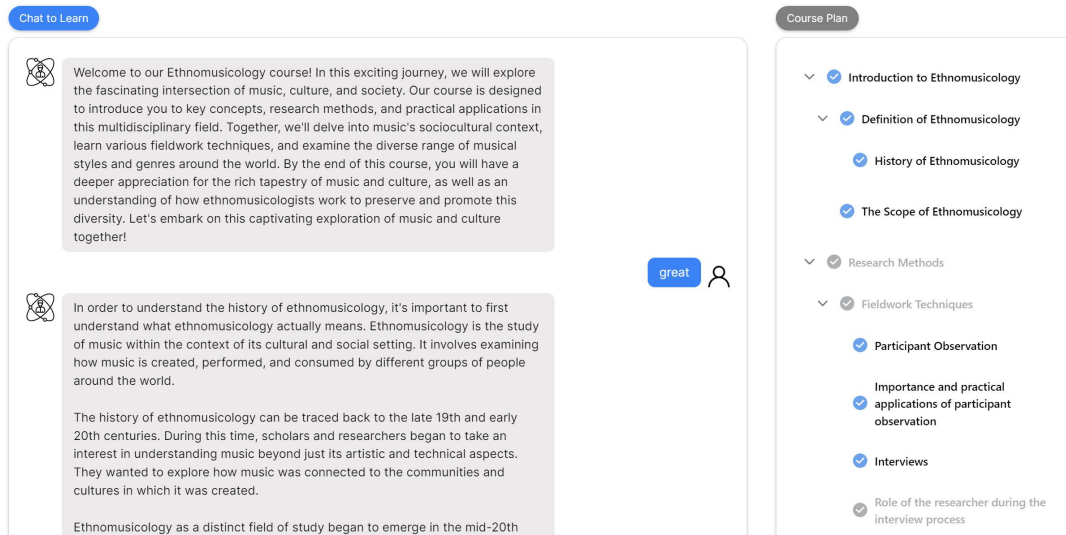
¹<https://github.com/langchain-ai/langchain>

1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040

1014
1015



(a) User interface for system evaluation at the beginning.



(b) User interface for system evaluation in progress.

Figure 7: User interface for system evaluation. The users select which course to learn and start learning by chatting to the system. The right-hand side shows the real-time course plan.

B.5 Evaluation Details

Table 8 presents the detailed statements we use in user study. As for stability annotation, each topic is learned independently with three systems. We evaluate each learning process in terms of the degree of repetition in course material, whether there is omission of sub-topics while instructing, and whether the system responds robustly to user’s random questions. Each learning is scored with 3 aspects respectively and the score can be 0, 0.5 or 1. Specifically, for response robustness, direct ignorance of the question or repeated template answer like “Let’s stay focused on the course material.” will be considered a sign of lack of robustness.

C Quiz questions

Table 9 presents sample quiz questions from the main system and the ablation system. It can be seen that the main system generates more detailed questions, while the ablation system largely relies on information available in the course plan and the generated questions are less effective in evaluating student’s ability in mastering what has been taught.

D Case Studies

We provide the details of case studies below. We take “Machine Learning” course from Stanford as a real-world example. The script is taken from (Pan et al., 2017). Table 10 shows the comparison between course plan design by CHATTUTOR and on

1069 Coursera. Figure 8 shows how the user can moti-
1070 vate more details about machine learning during the
1071 education process. Figure 9 shows CHATTUTOR
1072 hallucinates when the user asks for more details
1073 about the volcano.



Now, let's move on to the next topic: Linear Regression. Linear regression is a type of supervised learning algorithm used to predict a continuous target variable based on one or more input features. It assumes a linear relationship between the input variables and the target variable.

One important concept in linear regression is the coefficient, which represents the slope of the line that best fits the data. The coefficient tells us how much the target variable is expected to change for a one-unit change in the input variable. For example, ...

Linear regression is a powerful tool in machine learning and is widely used in various fields, such as finance, economics, and healthcare. It is a fundamental concept in understanding regression analysis and building predictive models.

Now, do you have any questions about linear regression or any specific aspect you would like to explore further?

how do we fit a linear regression model



To fit a linear regression model, we use a process called "ordinary least squares" (OLS). The goal of OLS is to find the line that best fits the data by minimizing the sum of the squared differences between the predicted values and the actual values.

Here are the steps to fit a linear regression model:

1. Preprocess the data: Before fitting the model, it's important to preprocess the data by handling missing values, scaling the features if necessary, and splitting the data into training and testing sets.
2. Define the input and target variables: Identify the input variables (features) and the target variable (the variable you want to predict). In our example of predicting house prices, the size of the house would be the input variable, and the price would be the target variable.
3. Create a linear regression model: Instantiate a linear regression model using a machine learning library like scikit-learn in Python. This model will be used to find the line that best fits the data.
4. Fit the model to the training data: Use the `fit()` function to fit the linear regression model to the training data. This process involves finding the optimal coefficients for the line that minimize the sum of squared differences.
5. Evaluate the model: Once the model is fitted, evaluate its performance using evaluation metrics such as mean squared error or R-squared. These metrics measure how well the model predicts the target variable.
6. Make predictions: After evaluating the model, you can use it to make predictions on new, unseen data. Provide the input features of the new data to the model, and it will predict the corresponding target variable.

It's important to note that the process of fitting a linear regression model may involve additional steps depending on the complexity of the data and specific requirements of the problem. However, the steps outlined above provide a basic overview of how to fit a linear regression model. I hope this helps! Let me know if you have any further questions.

Figure 8: An example of how user cam motivates more detailed information.



Now, let's move on to the next topic: shield volcanoes. Shield volcanoes are a type of volcano that is formed from the eruption of fluid lava. They have gentle slopes and a broad, flat shape, resembling a warrior's shield, which is how they got their name. Shield volcanoes are typically found at divergent plate boundaries and hotspots.

The lava erupted by shield volcanoes is low in viscosity, meaning it is very fluid and can flow easily. This is because the magma that forms shield volcanoes has a low silica content, which gives it a low viscosity. As a result, the lava can travel long distances before solidifying.

One famous example of a shield volcano is Mauna Loa in Hawaii. Mauna Loa is one of the largest volcanoes on Earth and is located on the Big Island of Hawaii. It has a gentle slope and has been erupting for thousands of years. The lava flows from Mauna Loa can travel for many miles, creating vast fields of solidified lava known as lava fields.

...

tell me some details about Mauna Loa in Hawaii



Mauna Loa is one of the largest volcanoes on Earth and is located on the Big Island of Hawaii. It has a gentle slope and has been erupting for thousands of years. The lava flows from Mauna Loa can travel for many miles, creating vast fields of solidified lava known as lava fields.

tell me more please



Sure! Mauna Loa is considered one of the most active volcanoes in the world. It has erupted **33 times** since its first well-documented eruption in 1843. The eruptions of Mauna Loa are typically characterized by slow-moving lava flows that can reach outlying areas, but they are generally not considered highly explosive or dangerous to human settlements.

One notable eruption of Mauna Loa occurred in 1950, when lava flows threatened the city of Hilo. Fortunately, the lava diverted its path just in time, sparing the city from destruction. Another significant eruption took place in 1984, when lava flows came within 4 miles of the town of **Kalapana**, destroying homes and covering roads. Despite these occasional threats, Mauna Loa eruptions are closely monitored, and evacuation plans are in place to ensure the safety of nearby communities.

It's important to note that while Mauna Loa is an active volcano, it is not currently erupting. However, scientists continue to monitor the volcano closely for any signs of activity to ensure the safety of the surrounding areas.



Now, let's move on to the new addition to our course plan, Mauna Loa in Hawaii.

...

One interesting fact about Mauna Loa is that it holds the record for the largest continuous eruption in recorded history. **From 1983 to 2018, Mauna Loa experienced an eruption that lasted for 35 years.** During this time, the volcano produced lava flows that covered an area of more than 100 square kilometers. Fortunately, the eruption occurred in a relatively uninhabited area, so there was minimal impact on human settlements.

...

Figure 9: An example of hallucination. Hallucinated generation as been colored red.

Topic	Gravity Waves
Course plan	<ul style="list-style-type: none"> 1. Introduction to Gravity Waves <ul style="list-style-type: none"> a. Definition and Key Concepts <ul style="list-style-type: none"> i. Differentiation between Gravity Waves and Gravitational Waves ii. Causes and Influences of Gravity Waves b. Factors Affecting Gravity Wave Formation <ul style="list-style-type: none"> i. Atmospheric Stability ii. Wind Shear iii. Topography and Surface Obstacles 2. Characteristics and Properties of Gravity Waves <ul style="list-style-type: none"> a. Wave Amplitude and Wavelength <ul style="list-style-type: none"> i. criteria for wave breaking ii. energy transfer and propagation b. Wave Speed and Frequency <ul style="list-style-type: none"> i. dispersion relation ii. role of buoyancy and ambient fluid properties 3. Observation and Detection of Gravity Waves <ul style="list-style-type: none"> a. Remote Sensing Techniques <ul style="list-style-type: none"> i. satellite imagery ii. lidar and radar <p>...</p>
Main system	<ul style="list-style-type: none"> 1. What is amplitude in the context of gravity waves? <ul style="list-style-type: none"> a) The distance between successive crests or troughs of a wave b) The maximum displacement of particles within a wave c) The rate at which wave energy is transferred vertically d) The disrupted, turbulent state of a wave 2. How does wavelength influence the behavior of gravity waves? <ul style="list-style-type: none"> a) It determines the spatial scale of the wave b) It represents the distance over which the wave repeats itself c) It influences the rate at which wave energy is transferred vertically d) It determines the amplitude of the wave 3. When does wave breaking occur in gravity waves? <ul style="list-style-type: none"> a) When the wave reaches its highest crest b) When the wave transitions from a regular pattern to a turbulent state c) When the force of gravity restores the disturbed equilibrium d) When the amplitude or wavelength exceeds certain thresholds
Ablation system	<ul style="list-style-type: none"> 1. What is the main focus of the course "Introduction to Gravity Waves"? <ul style="list-style-type: none"> a. Gravitational Waves b. Atmospheric Stability c. Causes and Influences of Gravity Waves d. Topography and Surface Obstacles 2. What are the factors affecting gravity wave formation? <ul style="list-style-type: none"> a. Wave Speed and Frequency b. Wind Shear c. Wave Amplitude and Wavelength d. Dispersion Relation 3. Which remote sensing techniques are used for detecting gravity waves? <ul style="list-style-type: none"> a. Satellite imagery b. In-situ measurements c. Balloon-borne instrument packages d. Lidar and radar

Table 9: Example quiz questions generated by different systems on "Gravity Waves".

Coursera	CHAT TUTOR
<ul style="list-style-type: none"> 1. introduction <ul style="list-style-type: none"> a. welcome to machine learning b. welcome c. supervised learning d. unsupervised learning 2. linear regression with one variable <ul style="list-style-type: none"> a. model representation b. cost function c. cost function intuition i d. cost function intuition ii e. gradient descent f. gradient descent intuition g. gradient descent for linear regression 3. linear algebra review <ul style="list-style-type: none"> a. matrices and vectors b. addition and scalar multiplication c. matrix vector multiplication d. matrix matrix multiplication e. matrix multiplication properties f. inverse and transpose 4. linear regression with multiple variables <ul style="list-style-type: none"> a. multiple features b. gradient descent for multiple variables c. gradient descent in practice i feature scaling d. gradient descent in practice ii learning rate e. features and polynomial regression f. normal equation g. normal equation noninvertibility h. working on and submitting programming assignments 5. octave matlab tutorial <ul style="list-style-type: none"> a. basic operations b. moving data around c. computing on data d. plotting data e. control statements for while if statement f. vectorization 6. logistic regression <ul style="list-style-type: none"> a. classification b. hypothesis representation c. decision boundary d. cost function e. simplified cost function and gradient descent f. advanced optimization g. multiclass classification one vs all 7. regularization <ul style="list-style-type: none"> a. the problem of overfitting b. cost function c. regularized linear regression d. regularized logistic regression ... 	<ul style="list-style-type: none"> 1. Introduction to Machine Learning <ul style="list-style-type: none"> a. Definition of Machine Learning b. Importance and Applications of Machine Learning c. Types of Machine Learning 2. Supervised Learning <ul style="list-style-type: none"> a. Definition and Explanation b. Classification <ul style="list-style-type: none"> i. Binary Classification ii. Multiclass Classification c. Regression <ul style="list-style-type: none"> i. Linear Regression ii. Polynomial Regression 3. Unsupervised Learning <ul style="list-style-type: none"> a. Definition and Explanation b. Clustering <ul style="list-style-type: none"> i. K-Means Clustering ii. Hierarchical Clustering c. Dimensionality Reduction <ul style="list-style-type: none"> i. Principal Component Analysis (PCA) ii. t-Distributed Stochastic Neighbor Embedding (t-SNE) 4. Evaluation and Validation <ul style="list-style-type: none"> a. Training, Testing, and Validation Data b. Accuracy, Precision, Recall, and F1-Score c. Cross-Validation 5. Model Selection and Regularization <ul style="list-style-type: none"> a. Bias-Variance Tradeoff b. Overfitting and Underfitting c. Regularization Techniques 6. Introduction to Neural Networks <ul style="list-style-type: none"> a. Definition and Explanation b. Perceptron and Activation Functions c. Backpropagation Algorithm 7. Deep Learning <ul style="list-style-type: none"> a. Introduction to Deep Learning b. Convolutional Neural Networks (CNN) c. Recurrent Neural Networks (RNN) 8. Model Deployment and Future Trends <ul style="list-style-type: none"> a. Deploying Machine Learning Models b. Ethics and Bias in Machine Learning c. Current and Future Trends in Machine Learning






Table 10: Course plan comparison for “Machine Learning”.

Table 11: A complete list of course topics and difficulty level by category used for system evaluation.

Category	Main Learning Objective	Difficulty Level
Art	Baroque architecture	2
	Installation art	2
	Impressionism	3
Psychology	Cognitive biases	2
	Obedience and conformity	3
	Operant conditioning	1
	Developmental psychology	5
Geology	Paleontology	2
	Earth's mantle	2
	Igneous rocks	2
	Mineralogy	2
	Volcanoes	2
	Geologic formations	3
	Geothermal energy	3
Erosion	1	
Computer Science	Sorting algorithms (e.g., bubble sort, merge sort)	2
	Linked lists	3
	Web development	4
	Computer architecture	4
Chemistry	pH scale	2
	Lewis dot structure	3
Daily Wondering	How do bees communicate and find their way back to the hive?	2
	Biological Circadian Rhythms (sleep-wake cycles)	2
	Gravity Waves (not to be confused with gravitational waves)	2
	Volcanic Eruptions	2
	Predator-prey relationships	2
	Formation of Coral Reefs	2
	How do magnets work, and what makes some materials magnetic?	2
	Change of Seasons	3
	Boom-and-bust cycles	3
	Reflection and Refraction of Light	3
	Formation of Crystals	3
	How does our brain store and retrieve memories?	3
	Hibernation	3
	Bioluminescence in Deep-Sea Creatures	3
	How do plants "know" which way to grow toward light?	3
	Why do some people have allergies, and how do they develop?	3
	Formation of Rainbows	3
	Cyberbullying	3
	How do rainbows form and why do they have different colors?	3
	Bioluminescence (e.g., glowing organisms in the ocean)	1
	Entanglement	1
	Aurora Borealis (Northern Lights)	1
	How do airplanes stay up in the air and maintain flight?	1
	Formation of Rain and Snow	1
	Why do some people have left-handed dominance while others are right-handed?	1
	Creation and Dissipation of Fog	1
	The uncertainty principle	4
Hurricane formation	4	
Personalized ads	4	
Umami taste	4	
Jet lag	5	
Fight-or-flight response	5	
Music	Ethnomusicology	2
	Rhythm and meter	2
	Baroque composers (e.g., Johann Sebastian Bach, Antonio Vivaldi)	2
	Polyphony	1
Physics	Quantum physics	2
	Newton's second law of motion	1
	Newton's first law of motion	1
	Snell's law of refraction	5
History	Medieval history	2

Category	Main Learning Objective	Difficulty Level
	Mongol Empire	3
	Ancient Egyptian civilization	3
	American history	4
Philosophy	Aesthetics	2
	Pragmatism	1
	Syllogism	5
Mathematics	Permutations and combinations	2
	Mathematical modeling	4
	Differential equations	4
Biology	Food chains and food webs	2
	Nitrogen fixation	3
	Plant biology	3
Sociology	Deviance and social control	3
	Social networks	3
Literature	Postcolonial literature and writers	3
	Literary theory	1
	Stream of consciousness	1
	Symbolism in poetry	4

Table 12: An example of dialogue and corresponding real-time course plan.

Dialogue	Course Plan
<p> Welcome to our Ethnomusicology course! In this exciting journey, we will explore the fascinating intersection of music, culture, and society. Our course is designed to introduce you to key concepts, research methods, and practical applications in this multidisciplinary field. Together, we'll delve into music's sociocultural context, learn various fieldwork techniques, and examine the diverse range of musical styles and genres around the world. By the end of this course, you will have a deeper appreciation for the rich tapestry of music and culture, as well as an understanding of how ethnomusicologists work to preserve and promote this diversity. Let's embark on this captivating exploration of music and culture together!</p>	<ol style="list-style-type: none"> 1. Introduction to Ethnomusicology <ol style="list-style-type: none"> a. Definition of Ethnomusicology <ol style="list-style-type: none"> i. History of Ethnomusicology ii. Disciplinary Approaches b. The Scope of Ethnomusicology <ol style="list-style-type: none"> i. Musical Elements ii. Cultural Context
<p> In order to understand the history of ethnomusicology, it's important to first understand what ethnomusicology actually means. Ethnomusicology is the study of music within the context of its cultural and social setting. It involves examining how music is created, performed, and consumed by different groups of people around the world.</p> <p>The history of ethnomusicology can be traced back to the late 19th and early 20th centuries. During this time, scholars and researchers began to take an interest in understanding music beyond just its artistic and technical aspects. They wanted to explore how music was connected to the communities and cultures in which it was created.</p> <p>Ethnomusicology as a distinct field of study began to emerge in the mid-20th century. It was influenced by various disciplines such as anthropology, folklore studies, and musicology. Ethnomusicologists started conducting field research, traveling to different parts of the world to study and document traditional music practices.</p> <p>This approach differed from the traditional Western musicology, which largely focused on the study of classical music. Ethnomusicologists sought to understand and appreciate the diversity of musical traditions found globally. They recognized that music was not just a universal language, but also a reflection of the values, beliefs, and identities of the communities that created it.</p> <p>Ethnomusicologists use a wide range of research methods to study music in its cultural context. They may conduct interviews with musicians and community members, observe musical performances, and analyze the historical and social factors that shape musical practices. Through their work, ethnomusicologists aim to preserve and promote the rich tapestry of musical traditions from around the world.</p> <p>By studying the history of ethnomusicology, we gain a deeper understanding of how our appreciation and understanding of music has evolved over time. It helps us recognize the importance of cultural diversity and the role music plays in shaping our collective human experience.</p>	<p style="text-align: center;">great </p> <ol style="list-style-type: none"> 2. Research Methods <ol style="list-style-type: none"> a. Fieldwork Techniques <ol style="list-style-type: none"> i. Participant Observation ii. Interviewing b. Analysis and Interpretation <ol style="list-style-type: none"> i. Transcription ii. Contextual Analysis 3. Key Concepts <ol style="list-style-type: none"> a. Musical Diversity <ol style="list-style-type: none"> i. Cultural Differences ii. Music Styles and Genres b. Music, Identity, and Representation <ol style="list-style-type: none"> i. Social Identity ii. Music as Cultural Expression 4. Ethnomusicology in Practice <ol style="list-style-type: none"> a. Applied Ethnomusicology <ol style="list-style-type: none"> i. Community-based Projects ii. Advocacy and Activism b. Teaching and Dissemination <ol style="list-style-type: none"> i. Academic Institutions ii. Public Outreach <p style="text-align: center;">ok </p>
<p> Now that we have a better understanding of the history of ethnomusicology, let's explore the scope of this fascinating field. Ethnomusicology encompasses a wide range of topics and approaches, allowing us to delve into various aspects of music and culture.</p> <p>One important aspect of ethnomusicology is the study of musical instruments. Different cultures have unique instruments, each with its own distinct sounds, techniques, and meanings. Ethnomusicologists examine how these instruments are made, how they are played, and how they contribute to the overall musical experience.</p> <p>Another area of interest in ethnomusicology is the study of musical genres. Just as different cultures have their own instruments, they also have their own musical styles and genres. From traditional folk songs to contemporary pop music, ethnomusicologists analyze the characteristics, cultural significance, and historical development of these genres.</p> <p>Ethnomusicology also involves studying the role of music in different social contexts. Music has the power to bring people together, express emotions, and communicate cultural identities. Ethnomusicologists explore how music is used in rituals, celebrations, ceremonies, and everyday life, shedding light on the social functions and meanings behind musical practices.</p> <p>In addition to studying music within specific cultures, ethnomusicology also examines the interactions between different cultures and musical traditions. With globalization and migration, musical styles and practices often intersect and blend. Ethnomusicologists investigate these cross-cultural exchanges, exploring how music travels and adapts to new environments while retaining its original cultural and artistic elements.</p> <p>Lastly, ethnomusicology explores the significance of music in identity formation and cultural preservation. Music plays a crucial role in shaping individual and collective identities, reflecting the values, beliefs, and aspirations of communities. Ethnomusicologists study how music is used as a form of cultural expression and as a means of preserving heritage and traditions.</p> <p>By exploring the diverse scope of ethnomusicology, we can gain a deeper appreciation for the richness and complexity of music and its connection to culture. This multidimensional approach allows us to understand how music serves as a bridge between different societies and as a powerful tool for social change and understanding.</p>	



Now, let's discuss another important aspect of ethnomusicology: fieldwork techniques or research methods. Ethnomusicologists often engage in fieldwork, which involves immersing themselves in a particular culture or community to observe and document musical practices firsthand. Through fieldwork, ethnomusicologists are able to gather valuable information and insights about the music and its cultural context.

One fieldwork technique commonly used in ethnomusicology is participant observation. This method involves actively participating in the musical activities and events of a particular community while also observing and taking notes. By being a part of the musical experience, ethnomusicologists can gain a deeper understanding of the music's cultural significance and how it is performed within its social context.

In addition to participant observation, ethnomusicologists also conduct interviews as part of their fieldwork. They engage in dialogue with musicians, community members, and experts to gather information about the music, its origins, and its role in society. These interviews provide valuable insights into the cultural meanings and practices associated with the music, as well as the personal experiences and perspectives of those involved.

Ethnomusicologists may also employ audio and video recording techniques during their fieldwork. Recording performances, rituals, or everyday musical encounters allows for a more detailed analysis and documentation of the music. These recordings capture the nuances of the music, including its melodies, rhythms, and vocal expressions, which can be analyzed later in the research process.

In addition to these techniques, ethnomusicologists often utilize archival research. This involves studying historical records, manuscripts, recordings, and other existing materials related to the music and culture of interest. Archival research helps ethnomusicologists establish historical context, trace the evolution of musical practices, and understand how music has been shaped by social, political, and cultural forces over time.

Another important tool in ethnomusicological fieldwork is ethnographic writing. Ethnography involves writing detailed descriptions and narratives based on the researcher's observations and interactions in the field. Ethnomusicologists use ethnographic writing to convey the rich sensory experiences of music and to provide cultural context for their findings. This form of writing allows ethnomusicologists to share their research with others and contribute to the broader understanding of music and culture.

By utilizing these fieldwork techniques and research methods, ethnomusicologists are able to explore the complexities and intricacies of music in its cultural context. Through participant observation, interviews, recordings, archival research, and ethnographic writing, they gain a comprehensive understanding of the music's meaning, function, and impact on society. Fieldwork techniques are integral to ethnomusicology as they enable researchers to engage with music and culture in a holistic and immersive manner.

Got it



1. Introduction to Ethnomusicology
 - a. Definition of Ethnomusicology
 - i. History of Ethnomusicology
 - b. The Scope of Ethnomusicology
2. Research Methods
 - a. Fieldwork Techniques
 - b. Analysis and Interpretation
3. Key Concepts
 - a. Musical Diversity
 - b. Music, Identity, and Representation
4. Ethnomusicology in Practice
 - a. Applied Ethnomusicology
 - b. Teaching and Dissemination

ok so why is this subject important? Why can't we just appreciate music?



That's a great question! Ethnomusicology is an important subject because it allows us to go beyond simply appreciating music and understand it in a deeper and more meaningful way. Here are a few reasons why studying ethnomusicology is important:

1. **Cultural Understanding:** Ethnomusicology helps us understand how music is connected to a culture's history, values, beliefs, and social practices. By studying different musical traditions, we gain insight into the diverse ways in which people express themselves artistically and culturally. This understanding promotes cultural appreciation and fosters empathy and respect for different societies and their musical traditions.
2. **Preservation of Cultural Heritage:** Ethnomusicology plays a crucial role in preserving and documenting musical traditions that may be endangered or at risk of being lost. By studying and documenting these traditions, ethnomusicologists contribute to the preservation of cultural heritage and history. This ensures that future generations can learn from and appreciate these rich musical traditions.
3. **Cross-Cultural Dialogue:** Music has the power to bridge gaps and foster understanding between different cultures. Studying ethnomusicology allows us to explore the similarities and differences in musical practices across cultures. By engaging in cross-cultural dialogue and understanding, we promote global awareness and create opportunities for collaboration and exchange.
4. **Shedding Light on Social Issues:** Music often reflects the social, political, and cultural realities of a society. Ethnomusicology allows us to analyze and understand how music can be used to express and address social issues such as identity, inequality, and resistance. By examining the ways in which music is used as a form of cultural expression and activism, we can gain insight into the experiences and struggles of different communities.
5. **Personal Growth and Enrichment:** Studying ethnomusicology can also be personally enriching. It deepens our appreciation for the diversity of human creativity and expression, and allows us to develop a broader perspective on music and its cultural significance. It enhances our ability to critically analyze and interpret different musical traditions, and can also inspire creativity and innovation within our own musical pursuits.

Overall, studying ethnomusicology is important because it provides us with a deeper understanding and appreciation of music's role in shaping our world. It helps us recognize the power of music as a cultural and social force, and allows us to engage with diverse musical traditions in a meaningful and respectful manner.

alright, talk about some practical application then



Sure! Let's talk about a research method called participant observation, which is commonly used in ethnomusicology fieldwork.

Participant observation is a qualitative research technique that involves immersing oneself in a specific musical community or cultural context. Ethnomusicologists who use this method actively participate in the musical activities and events of the community they are studying, while also observing and taking notes on their experiences.

By actively participating in the musical practices of a community, researchers gain a firsthand understanding of the cultural and social significance of music within that particular context. This method allows ethnomusicologists to not only observe but also experience the music in its intended cultural and social setting.

Through participant observation, researchers can develop close relationships with community members, which can provide valuable insights into their perspectives, beliefs, and experiences. By engaging with the community as a participant, researchers are able to build trust and rapport, which leads to more authentic and meaningful data collection.

In addition to observing and participating, ethnomusicologists also document their experiences through fieldnotes. These fieldnotes capture their observations, thoughts, and reflections on the musical practices, performances, and interactions within the community. Fieldnotes provide a detailed record of the researcher's experiences, which can be analyzed and interpreted to gain a deeper understanding of the music and its cultural significance.

Participant observation allows researchers to uncover the nuances and intricacies of a musical tradition. It provides rich and detailed data that goes beyond mere descriptions, allowing for a more holistic analysis of the music and its cultural context.

This method also enables researchers to gain a more nuanced understanding of how music is learned, transmitted, and practiced within a specific community. By actively participating, researchers can witness firsthand the teaching methods, learning processes, and social dynamics that shape the musical practices of the community.

Overall, participant observation is a valuable research method in ethnomusicology because it allows researchers to immerse themselves in a musical community, experience the music firsthand, build connections with community members, and gain a deep understanding of the music's cultural and social significance.



Alright! Let's delve further into the importance and practical applications of participant observation in the field of ethnomusicology.

One of the key benefits of participant observation is its ability to capture the dynamic nature of music within a cultural context. By actively immersing themselves in the musical community, ethnomusicologists can witness the evolving nature of music and how it adapts to changes in society. This method allows researchers to identify and understand the cultural and historical factors that influence the development and transformation of musical traditions.

Another practical application of participant observation is its role in debunking stereotypes and misconceptions surrounding different musical communities. By directly engaging with the community as a participant, researchers can challenge preconceived notions and stereotypes, gaining a more accurate and nuanced understanding of the musical practices and their social significance. This can help in breaking down cultural barriers and fostering cross-cultural understanding and appreciation.

Participant observation also enables researchers to explore the social dynamics and power structures within a musical community. By immersing themselves in the community, ethnomusicologists can observe how music is used as a means of social interaction, communication, and identity formation. This method allows researchers to uncover the hierarchical relationships, gender dynamics, and other social dimensions that shape the musical practices and their broader cultural context.

Furthermore, participating in musical events and activities provides researchers with valuable opportunities for data collection and analysis. By actively engaging in the music-making process, researchers can gather detailed information about musical techniques, performance styles, and the underlying meanings behind the music. This firsthand experience enhances the authenticity and depth of research findings.

Participant observation can also contribute to the preservation and documentation of musical traditions. By actively participating in the community's musical activities, researchers can help to safeguard cultural heritage by documenting and recording performances, rituals, and traditions that might otherwise be lost or forgotten. This documentation serves as a valuable resource for future generations, ensuring the continuity and appreciation of diverse musical practices.

Lastly, participant observation can be a catalyst for meaningful collaborations and partnerships between researchers and community members. By immersing themselves in the community, ethnomusicologists can foster trust and mutual respect, leading to collaborative research projects that address the concerns and interests of the community. This participatory approach allows for a more ethical and inclusive research process, where the voices and perspectives of the community are represented and respected.

In conclusion, participant observation is a vital research method in ethnomusicology that provides researchers with a unique and in-depth understanding of musical traditions and their cultural significance. Through active participation, observation, and documentation, this method allows for a holistic analysis of music within its cultural context, while also promoting cross-cultural understanding, breaking down stereotypes, and ensuring the preservation of diverse musical practices.

good



1. Introduction to Ethnomusicology
 - a. Definition of Ethnomusicology
 - i. History of Ethnomusicology
 - b. The Scope of Ethnomusicology
2. Research Methods
 - a. Fieldwork Techniques
 - i. Participant Observation
 - ii. Importance and practical applications of participant observation
 - iii. Interviews
 - iv. Importance and practical applications of interviews
 - v. Audio and Video Recording Techniques
 - vi. Importance and practical applications of audio and video recording techniques
 - vii. Archival Research
 - viii. Importance and practical applications of archival research
 - ix. Ethnographic Writing
 - x. Importance and practical applications of ethnographic writing
 - b. Analysis and Interpretation
3. Key Concepts
 - a. Musical Diversity
 - b. Music, Identity, and Representation
4. Ethnomusicology in Practice
 - a. Applied Ethnomusicology
 - i. Practical applications of ethnomusicology in cultural understanding, preservation of cultural heritage, cross-cultural dialogue, shedding light on social issues, and personal growth and enrichment
 - b. Teaching and Dissemination

thx for your information, now maybe just give a quick overview of each method



Certainly! Here is a quick overview of the different research methods used in ethnomusicology fieldwork:

1. Participant observation: Involves actively participating in the musical activities and events of a community while also observing and taking notes. Provides a firsthand understanding of the cultural and social significance of music within a specific context.
2. Interviews: Engages in dialogue with musicians, community members, and experts to gather information about the music, its origins, and its role in society. Provides valuable insights into cultural meanings and practices.
3. Audio and video recording: Utilizes recording techniques to capture performances, rituals, and everyday musical encounters. Allows for detailed analysis and documentation of the music's nuances and characteristics.
4. Archival research: Involves studying historical records, manuscripts, recordings, and other existing materials related to the music and culture of interest. Establishes historical context and traces the evolution of musical practices.
5. Ethnographic writing: Involves writing detailed descriptions and narratives based on the researcher's observations and interactions in the field. Conveys the richness and cultural context of the music.

These methods, when used in combination, provide a comprehensive understanding of the music's meaning, function, and impact on society. They allow for an immersive and holistic exploration of music and its cultural context within ethnomusicology research.