The Potential of AI to Substitute Traditional University Teaching

Prepared by:
İris Eyüboğlu
Helya Pourkarim Kashki
Defne Akbulut

Polytechnic University of Turin
Course: Computer Architecture
Professor: Prof. Paolo Montuschi
Date: June 2025

1. Introduction

Large Language Models (LLM) and AI chatbots have become more and more common in the last years, since their first developments in the 2010s. Even though the chatbot idea was nothing new, shifting from simple linguistic tasks using natural language programming to generative AI was a huge breakthrough in this field. After the introductions of the first real LLMs in 2018 by OpenAI and Google, GPT-1 and BERT respectively[1], the field has been evolving rapidly with the language models getting more and more natural, correct and smart every day.

With this changing world; a lot of habits, lifestyles and systems started changing. As of 2025, within the blink of an eye, AI is able to create described images, create essays and even though it is out of the scope of this research, videos. However, to what extent are the AI models capable of doing these various tasks? How much have we progressed and how much do we still have to go to trust AI to do those tasks that were before done by human intelligence?

Unofficially, it can be seen that many students have already shifted from using traditional search engines to relying on AI agents -Large Language Models- for explanations for their university level lectures. So the next step is asking, if students can trust AI for learning, can professors also use it to prepare complete lecture materials? While this promises efficiency and accessibility, there are still questions about the quality, reliability, and teaching value of AI generated materials. In this research we aimed to analyze the advantages and tradeoffs of such an approach using different criteria.

For this aim, we evaluated four different AI agents—ChatGPT, Perplexity, Le Chat, and DeepSeek—to understand their capabilities and limitations in generating lecture material for a technical university-level course for bachelor’s degree on cache memories for the Computer Architecture course. By comparing the outputs of different models in terms of content depth, clarity, diagram integration, and more, we aimed to assess whether current AI tools can produce material suitable for academic instruction.

[1] “The Timeline of the Development of LLMs from 2018 to 2024 (June),...”

2. Research Question

"To what extent can different AI agents (ChatGPT-4o Le Chat, DeepSeek V3, Perplexity) generate university-level lectures on cache memories that are pedagogically appropriate, correct and coherent in terms of content quality, diagram creation, question and answer preparation?"

3. Methodology

To evaluate the effectiveness of different AI agents in generating lecture material, we followed a structured methodology consisting of four stages:

3.1 Choosing the Topic

We selected cache memories as our focus topic because it is one of the most complex topics in our Computer Architecture course, which allowed us to challenge the AI agents. The topic includes theoretical concepts and practical applications, so we could test how well the AI explains technical ideas and how accurately it answers practice questions. This made it a strong choice for evaluating different aspects of AI-generated lecture material.

We worked on 6 subtopics:

1) Memory hierarchy
2) Associative memory (CAM)
3) Goal of caching, management and delay
4) Direct Mapped Cache
5) Fully Associative Cache
6) Set Associative Cache

3.2 AI Agent Selection

During our initial, more basic searches with various AI agents, we noticed that some of them gave identical analogies, while others produced different ones. After further investigation, we discovered that the AI agents producing similar analogies were often developed in the same geographical area. To avoid bias and promote diversity in AI behavior, we selected one or two models from each of three different regions. From the United States, we decided to use ChatGPT-4o that was developed by OpenAI[2], which could generate lecture plans as well as images using its image generation model DALL·E 3[3]. Perplexity developed by Perplexity AI has drawn our attention in this project as it not only uses generative AI but also uses web research[4] to integrate in the chatbot experience. Our third model, Le Chat was developed by Mistral which is France based, and our fourth model DeepSeek-V3[5] by

[2] "Hello GPT-4o."
[3] "DALL·E 3."
[4] "How Does Perplexity Work?"
[5] "Deepseek-Ai/DeepSeek-V3."

DeepSeek AI was trained using both Chinese and English datasets which we hypothesized to give diverse results.

3.3 Developing the Prompt

We took an iterative approach to creating prompts, improving them across three major versions. We created the first one with the help of ChatGPT, hypothesising that it would include all the necessary instructions:

Prompt 1: Minimal Prompting

> “You are preparing teaching material for a lecture segment aimed at **second-year computer engineering students**.
> The topic is: **Memory hierarchy as an introduction to cache memories**
> Generate the content using the structure below, to create the result in markdown:
> ---
> ### 📄 Slide Count
> - Prepare **3 to 5 slides**.
> - If more are needed for clarity, you may go beyond 5 **only if you explain why**.
> ---
> ### 🧱 For Each Slide, Include:
> - **Title**
>   - A clear, concise title that reflects the slide’s focus.
>
> - **Bullet Points**
>   - 4–6 bullet points per slide
>   - Each bullet should be a **complete sentence** (not fragments)
>   - Each bullet can be **1–2 sentences long**
>   - Use precise, technical language but explain new terms simply
>
> - **Speaker Notes**
>   - A **paragraph** explaining the slide’s content in a way a professor might say it aloud
>   - A **brief summary** (1–2 sentences) of the main idea for presenter use
>
> - **[Optional] Analogy**
>   - If helpful for memory or understanding, include **one real-world analogy**
>
> - **[Optional] Diagram Description**
>   - If a diagram would aid understanding, describe it in words.
>   - Format:
>   "Diagram: a cache with N sets, each with K lines. One block maps to a set using modulo operation."
>
> ---
> ### 🎯 Additional Instructions
>
> - Maintain an **educational and student-friendly tone**
> - Assume the student knows digital logic and memory basics, but not advanced architecture

- Avoid unnecessary jargon, or explain it when used
*- Focus on helping students build both **intuitive and precise** understanding"*

Gave only the topic name and general instructions. This version revealed how the models interpret open-ended teaching tasks. We noticed frequent scope drift and inconsistent depth.

Prompt 2: Scoped Prompting
Added subtopics, keywords, and specific teaching goals (e.g., "Explain hit/miss in cache using intuitive analogies"). This version helped narrow the responses and test whether models could respect the scope of the lessons.

Prompt 3: Structured Creative Prompting
Allowed freedom in formatting (tables, diagrams, bullet styles) while keeping scope guidance. This gave the best balance between flexibility and coherence and allowed us to evaluate whether models could provide natural teaching flow.

3.4 Collecting Lecture Material

We divided the topic of cache memories as explained in 3.1, and each group member worked on different subtopics and used the same set of prompts with all four AI agents. We improved the prompts as we collected more material. This allowed us to compare the agents under consistent conditions while covering the full lecture scope.

Each AI agent was asked to generate slides, speaker notes, diagrams (or their descriptions), and follow-up assessment questions. We also examined whether their responses matched our desired scope and whether they could correctly answer each other's questions. Additionally, we tested AI objectivity by attempting to persuade each model into accepting incorrect answers to factual questions.

4. Results

During our initial research, we chose AI agents from different regions because those specific AI agents were giving different material to our prompts. As our research progressed, we created more specific prompts, the AI agents began to generate similar analogies and explanations even though they were from different continents. This likely happened because more specific prompts leave less room for creative variation, causing the responses to converge. It could also be due to the fact that many AI models are trained on similar datasets, which leads them to produce comparable explanations. However, even with similar content, we noticed that the level and style of explanation still varied. Some AI agents gave more technical or detailed responses, while others gave simpler, more general explanations. This suggests that while the core information may align, the way it's communicated can differ significantly depending on the AI's design or target audience.

4.1. Content quality

For this criterion, we saw how important prompt engineering was while interacting with the AI agents. In the first prompt we created, none of the AI agents could give a high-quality answer: The part where they struggled the most was deciding the 'scope' of the lecture segment. With the same exact prompt, while ChatGPT and Perplexity understood to limit the scope; DeepSeek and Le Chat at times failed. For example, for our first subtopic in which they had to talk about memory hierarchies, DeepSeek and Le Chat included mapping techniques as well. This made all the topics superficial and insufficient. As this prompt contained minimal human intervention and didn't give the optimal results, it is suitable to conclude that AI by itself is not sufficient for now.

In the second prompt where the scope problem was aimed to be prevented, we decided to give all the subtopics of the lecture, as well as some hints about what needs to be covered. Such as:

> These are the topics of the lesson as a whole: Memory Hierarchy (register → HDD, latency); Associative Memory (CAM); Goal of Caching, Management and Delay; Direct-Mapped Cache; Fully Associative Cache; Set-Associative Cache; Coherence with Main Memory; Free Entries & Replacement Policies; Instruction vs Data Caches; Miss Types (Compulsory, Conflict, Capacity); Real-World Cache Example
>
> Explain just this topic: ***Goal of Caching, Management and Delay*** Explain this topic around these subtopics: Goal of Cache Memories, The Basic Idea of Cache Flow, Delays, Hit and Miss

The second prompt solved the scope problem. In Table 1, you can observe a comparison of sub-topics covered in each slide for the topic *Goal of Caching, Management, and Delay*. While the first prompt produced broader and more scattered content, the second prompt demonstrates a more in-depth and focused understanding of the topic.

Table 1: First and Second Prompt Scope Comparison

	1st Prompt	2nd Prompt
Chat GPT	1. Why do we need caching? 2. Goal of caching 3. Effectiveness of Cache 4. Real world examples	1. Why do we need caching? 2. Minimizing Access Time 3. Cache Management Policies 4. Caching Effectiveness
Perplexity	1. What is caching? 2. Goals of caching 3. Effectiveness of Cache 4. Real world examples	1. Goal of Caching 2. Cache Management Strategies 3. Delay Trade-offs 4. Quantifying Delay
DeepSeek	1. What is caching? 2. Goals of caching 3. Effectiveness of Cache 4. Real world examples	1. Goal of Caching 2. Cache Management Challenges 3. Sources of Delay 4. Caching Intuition
Le Chat	1. What is Caching? 2. Cache Hierarchy 3. Mapping Techniques 4. Replacement Policies 5. Effectiveness of Cache	1. What is Caching? 2. Caching Management 3. Delay in Caching

Even though the AI agents no longer had trouble with the scope, the results weren't fully satisfying—so we decided to try a third prompt.

In the third prompt, we gave AI more freedom hoping that we would get a more coherent explanation. We requested AI to not limit itself to bullet points but rather use lists, tables and anything it believes to be suitable. Even though we were still observing problems, we achieved the best result with this prompt.

ChatGPT stood out among the others overall. It gave coherent explanations mixed with various styles such as tables, bullet points in forms of half sentences, symbols and emojis etc. Such as:

'Caches follow this process:

1. *CPU requests data*
2. *Cache is searched*
3. *If data is present → cache hit (fast)*
4. *If not → cache miss → fetch from lower level (slow)'*

ChatGPT was not the only one to give such answers, but was the only agent who consistently could give a natural lecture-style layout avoiding bullet lists. The other agents at times could succeed, but at times still gave bullet points or worse, decided to give big paragraphs which are not suitable for in class teaching. Throughout our research, we realized that we do not get a consistent good/bad result or even a consistent layout, even with the same prompts when asked more than once. This shows the probabilistic structure of the Large Language Models[6]. Instead of following fixed rules, AI agents predict the most likely next word. Since there are many reasonable ways to respond, the same prompt can lead to different word choices and structures each time, which explains the variation we observed.

Talking about the topic content, the AI agents gave similar reactions most of the time, even naming the slides with similar titles. However, the content sometimes consisted of too much detail for just an 'introduction' to cache class. For example, for the subtopic 2, agents chose to also talk about CAM types (binary and ternary) and for the first prompt talked about real world examples of cache. While real examples from commercial brands can be very helpful for students to gain insight, and an AI agent can get these details faster than a professor while preparing the lesson plan, it is also important to decide when it is appropriate to present this information.

Moreover, even when the subtopics were listed to the agents, the subtopics contained repetition; AI didn't know how to partition topics and information. For instance, even though Memory Hierarchy and Goal of Caching were clearly separated subtopics, during the explanation of Goal of Caching, ChatGPT was explaining memory hierarchy and giving diagrams such as "*A triangle pyramid labeled "Memory Hierarchy" showing Registers → Cache → DRAM → SSD/HDD, with latency increasing down the hierarchy.* Or when examining the topic of fully associative mappings, it was notable that none of the materials addressed the concepts of sequential search or associative memories. These subjects are often included in foundational discussions to help students develop a preliminary understanding of how associative mappings operate. Introducing sequential search and associative memory concepts can provide essential context, as they illustrate the mechanisms by which data can be retrieved and linked within memory systems. Their omission may result in a less comprehensive overview, potentially leaving gaps in the learners' grasp of the broader principles underlying associative mapping techniques .This shows that while AI can generate relevant content, it still struggles to distinguish between closely related subtopics, often merging them into a single explanation even though we clearly separated them in the prompt. As a result, lectures may become repetitive or unbalanced without human adjustment.

At times, we saw that DeepSeek and Perplexity were more prone to giving numerical examples than ChatGPT and Le Chat. For instance, DeepSeek included examples like *'Modern CPUs execute instructions in ~0.3 ns, but RAM access takes ~100 ns—a 300x gap!'*, while these are beneficial to students' understanding, the whole lecture prepared to be going around numbers constantly could be hard to understand. Also some slides such as *'Slide 3: Cache Positioning in Modern CPUs'* seemed somewhat high level for an introductory class.

It was noted that generally all of the AI agents used suitable terminology (bottleneck, memory gap etc.), compared advantages and tradeoffs, and gave correct information during the lecture time.

[6] Brown et al., "Language Models Are Few-Shot Learners."

When it comes to comparing an AI taught lecture with an in-class lecture, within the content we can see a very crucial difference. Even though in our prompts it was clearly stated to '*Focus on helping students build both **intuitive and precise** understanding',* we saw that none of the AI generated lectures stimulated students' critical thinking and problem solving abilities. The solution, the architecture is always presented and explained; but never created step by step with their reasoning. AI never decided to ask questions to students about how to optimize the architecture or the solution, never gave more than one possible architecture while comparing them. However, a professor in an in-person class definitely prefers to stimulate students' thinking, asking them various questions.

It was also requested in the prompts to use analogies when necessary, and we saw that they were generally suitable and creative as well as supported the understanding. It seems to us that it would be a suitable practice to ask AI for ideas on this topic to enrich the class done by a professor, as it is way faster. We also saw that AI sometimes gave the same analogies, such as a library or a post office for cache memories. There were even too many analogies at some instances in one class segment, and it wasn't even denoted that the lecturer can choose one of these rather than presenting them all. AI didn't catch that too many analogies can be overwhelming.

Furthermore, in several instances, the AI would prematurely present the conclusions or final results during the introductory sections of its explanations. This approach disrupted the logical flow of information, potentially causing confusion and hindering the learner's ability to follow the reasoning process step-by-step. Overall, these shortcomings suggest that Le Chat's explanatory style may require significant refinement to meet the standards necessary for effective instructional delivery.

4.2. Speaker notes

We required the agents to also produce speaker notes, in order to see if AI can complete a full lecture without any human intervention: With the prepared slides and the speaker notes there wouldn't have been need for anything else for a complete lecture. However, the results we got refuted this idea that it can teach autonomously.

The first two versions of our prompts, the speaker notes for all agents were unsatisfactory. Even though our prompt in all versions was asking agents to generate *' A **paragraph** explaining the slide's content in a way a professor might say it aloud'* it seemed to us that creating speaker notes wasn't intuitive for AI: They were almost always only about a part of the slide, not the entirety of the content. This was a huge problem, as the speaker notes themselves weren't completely explaining the topic!

Therefore in the last version of our prompt (third) we also added this part: *'Also, you need to make the speaker notes more extensive, they should include everything in the slides'*

With this emphasis, we could get better answers. ChatGPT had the best tone among all, it was pretty conversational and included even rhetorical questions at times. Le Chat, however, was observed to be less effective in its explanatory capabilities compared to the other AI agents evaluated. Its responses often came across as mechanical and lacked the natural, engaging tone that is typically expected in an educational lecture setting. This 'robotic' style of communication made it difficult to consider Le Chat's explanations suitable for use in actual teaching environments.

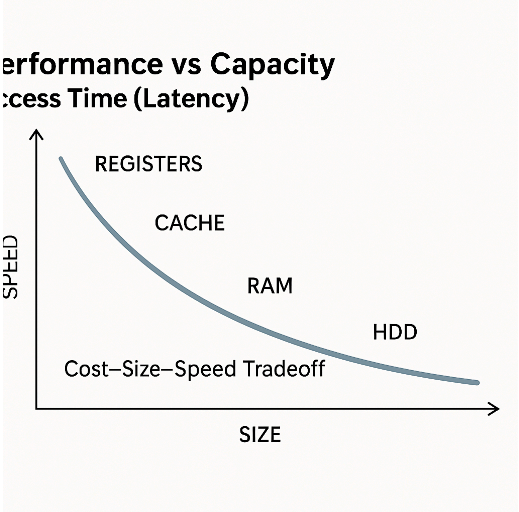
For one prompt and one answer, DeepSeek decided to create speaker notes as instructions, meaning that it listed what the speaker should talk about instead of actually writing them down:

> *Speaker Notes:*
> *"Compare mapping techniques as trade-offs. Direct-mapped is like assigned parking—fast but conflicts if two cars need the same spot. Set-associative is like a parking garage level (multiple spots per 'address'). Fully associative is valet parking (any spot, but slow search). Index bits are how we 'calculate' the parking spot."*

This tells us that AI cannot be trusted to give consistent replies.

4.3. Checking diagram description/creation

In in-person lessons, most of the slides consist of diagrams, therefore we asked in the prompt to describe a diagram appropriate for each slide. However, since Le Chat and Deepseek are not able to generate images, for those who are, image generations are done by different models than their text models; we decided to also compare diagrams described in words by the text models.



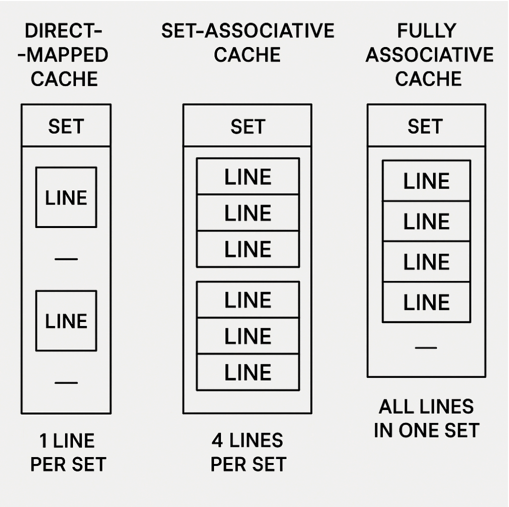


Image 1: ChatGPT Latency Diagram

Image 2: Perplexity Mapping

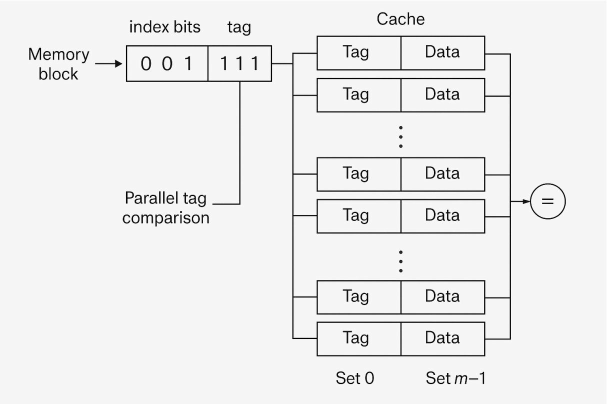


Image 3: ChatGPT explaining Tag in Full-associative mapping

Moreover, ChatGPT and Perplexity are able to create images for diagrams, thanks to their image generation models. ChatGPT uses DALL·E 3, and Perplexity uses DALL·E 3, Stable Diffusion or some custom implementations of these. In practice, ChatGPT's diagrams are usually simple and well-matched to the content, although many times it has formatting issues. For instance, as seen in *Image 1*, some images appeared cropped but some images provided by ChatGPT were more well formatted as seen in Image3. Perplexity, on the other hand, produced clearer and more complete diagrams, as shown in *Image 2*. These visuals were often clean, relevant to the slide content. Having built-in visual generation is very beneficial to lecture preparation, especially for technical topics like cache memories where diagrams are essential for understanding structural relationships and data flow.

The diagrams were sometimes unnecessary or simple. For example, ChatGPT gave a diagram like this, where it seemed to be only to fill up the empty space:

Diagram Idea:

Venn diagram showing trade-offs:

- *One circle = Simplicity*
- *One = Speed*
- *One = Flexibility*

Cache designs fall at different intersections.

Another example, Le Chat proposed this diagram for the explanation of cache management *"a cache with N sets, each with K lines. One block maps to a set using modulo operation"*. While it looks like it is related to the topic because the diagram illustrates how memory addresses are translated into cache locations; the slide explains hit, miss and last recently used data and doesn't mention anything about set-associative cache and other mapping techniques. Perplexity and DeepSeek use a diagram similar to Le Chat but also mention mapping techniques in the slide. This comparison shows that simply including diagrams isn't enough, what matters is how well they support the specific content being explained.

4.4. Question preparation :

Like any lecture and course, we thought these lectures also demand some questions to not only help the students understand the content better, but also to analyse each AI's question style. The results were satisfying. Without being explicitly asked to, the AI agents generated a wide range of question types. They not only produced multiple-choice and open-ended questions, but also short-answer, fill-in-the-gap questions with tables. It was particularly impressive that some models included critical thinking questions, even though in the slides they prepared, this was not their brightest point. Perplexity and ChatGPT generated similar types of questions, typically MCQs, short-answer, and applied analysis questions (questions that require calculations and application of theoretical content).

Question 1: Generated by ChatGPT-4o

> *Suppose a CPU has a cache hit time of 2 ns, a miss rate of 10%, and a miss penalty of 100 ns. Calculate the AMAT and interpret the result.*
>
> ***Expected Answer:***
> *AMAT = 2 + (0.10 × 100) = 2 + 10 =* ***12 ns***
> ***Interpretation:*** *On average, each memory access takes 12 ns due to the combined effect of fast cache hits and slow miss penalties.*

As observed from Question 1, the generated question does not require critical thinking but rather applying the theoretical information within the content of the lecture. The Average Memory Access Time formula (AMAT = *Hit Time* + Miss Rate × Miss Penalty) was given in the lecture created by ChatGPT already and the question directly asks the student to use it without asking for evaluation. However, this was not the case for all the subtopics. There were instances where ChatGPT would go out of topic especially for Memory Hierarchy.

Question 2: Generated by DeepSeek

> *If a cache's hit rate improves from 80% to 90%, but the miss penalty doubles (e.g., due to slower main memory), does performance always improve? Show calculations to support your answer.*
>
> ○ *Example: Compare AMAT before (0.8×2 + 0.2×100 = 21.6) and after (0.9×2 + 0.1×200 = 21.8).*

Similarly, DeepSeek generated a question regarding the AMAT formula as seen in Question 2. However, this time, the AMAT formula was not mentioned in the produced lecture. It was observed that Deepseek was going out of the scope of the lecture more often than any AI agent. Moreover, unlike Question 1 generated by ChatGPT, DeepSeek's question did not just ask to apply the AMAT formula. It required a higher level of understanding and the student to use the meaning of the AMAT formula by evaluating it in terms of performance.

Question 3: Generated by Le Chat

> ***True or False:*** *The goal of caching is to store infrequently accessed data in a smaller, faster memory closer to the processor.*
> ***Answer****: False*

On the contrary, questions generated by Le Chat required minimal to no application of theoretical knowledge but mainly tested the memorization of the lecture content. For example, as seen in Question 3, the true or false type of question is mostly interested in testşng the information and not much application. Even though Le Chat also created short essay questions, they did not exceed Highschool level thinking skills.

Despite AI agents generating lectures within the scope of the intended subtopic, they had a hard time doing the same when generating questions. For example, when the lecture was about “memory hierarchy”, they generated questions from memory mapping techniques. This showed that sometimes to make the questions more suitable for the content of the lecture, human adjustment is definitely needed.

Question 4: Generated by Perplexity.

> *A 4-way set-associative cache has a total size of 16 KB with a block size of 64 bytes. Given a 32-bit address, calculate:*
>
> *a) The number of sets in the cache.*
> *b) The number of bits used for the tag, assuming byte-addressable memory.*
>
> ***Answer**: 64 sets- 20 bits*
>
> *(the procedure is omitted as it was similar to how we solve in classroom and also too extended to put here)*

Question 4 is a sample of the practical questions AIs produced on the topics full associative and set associative, as can be observed it aims to test the knowledge of students about fundamental rules of different mappings, recognition of tag bits. Questions of these topics turned out to be comprehensive and totally in scope.

In case of mixed theoretical and practical topics like full/set associative mappings DeepSeek and Perplexity covered both types of questions, but ChatGPT and Le Chat provided only theoretical questions. It is worth noting that they could (and did) produce practical questions when we asked them to, however their initial approach was as explained. Therefore further prompting is at times needed.

Table 2 presents a comparison of each AI agent’s ability to generate questions based on the lecture content. It highlights differences in question types, depth, alignment with student level, and how well the questions stayed within the intended scope.

Table 2: AI Agents Question Comparison

Criteria	Chat GPT	Perplexity	DeepSeek	Le Chat
Question Types	MCQ Short Answer Applied Analysis	MCQ Short Answer Applied Analysis Essay	Short Answer Applied Analysis Scenario Based Critical Thinking	MCQ Short Answer True/False Essay
Student Level	Intermediate	Intermediate	Advanced	Beginner
Depth	Questions are more developed than Le Chat but not as cognitively demanding as DeepSeek.	Similar to ChatGPT, it expects comprehension and application, does not require evaluation.	Asks design level questions. Expects understanding, application, analysis, and evaluation.	Questions do not require critical thinking, they mainly test memorization.
Scope	Questions are usually within the scope of the lecture.	Questions are usually within the scope of the lecture.	Questions are about the subtopic but expects the student to know higher level terminology (that is not mentioned in the created lecture) and coding.	Questions are usually within the scope of the lecture.

Overall, the comparison of AI-generated questions showed distinct patterns in how each agent approaches assessment. Le Chat leans more towards memorization, mostly avoiding deeper reasoning or application. Therefore, Le Chat can be very useful when students are looking for material for initial preparation to see how familiar they are with subtopics. ChatGPT and Perplexity focus more on application and generated questions that were within the content of the lecture. Additionally, questions require mid-level cognitive demands as well as some critical thinking and problem solving questions. However, we could say that DeepSeek generates more questions that encourage critical thinking, synthesis, and evaluation. While this makes its questions more challenging and pedagogically valuable, it also occasionally goes beyond the lecture scope, introducing unfamiliar terminology or concepts. This showed a trade-off between staying within the lecture's boundaries versus asking for a deeper understanding and critical thinking. DeepSeek, ChatGPT and Perplexity can be very useful tools for professors when they are preparing an exam and for students when they are studying for an exam.

4.5. Answer preparation

In evaluating the responses generated by the AI agents, our analysis focused not only on the answers they provided to questions they designed themselves but also on their performance when asking questions designed by us or by other AI systems. As it was expected, the agents generally delivered accurate answers to their own questions, accompanied by explanations that were coherent, convincing, and well-structured.

However, the evaluation took a turn when the AI agents were asked to provide answers to questions posed by other AIs or us . To better understand and interpret the results, we categorized the responses into two distinct groups.

The first category included questions related to theoretical concepts and basic calculations. In this section, most AI agents performed well, providing correct and reliable answers. The notable exception was Le Chat, which made several errors even on these fundamental questions, indicating potential weaknesses in its understanding of basic theory and computational accuracy. This situation highlights the varying levels of proficiency among the AI agents when handling externally produced questions, especially those requiring foundational knowledge.

For example having this AI generated question:

Question:We have a 4-way-set associative mapping cache memory with globally 65536 = 216 lines. Each line is hosting 128 data. How many bits are necessary for the TAG for a 28 bits address bus?				
Correct Answer	7			
Provided Answer	ChatGPT 4o: 7	Perplexity: 7	Deepseek: 7	Le Chat: 9

In the second category, we checked answers to practical and more complex calculations:

When it came to questions involving practical applications and somewhat tricky calculations, only ChatGPT consistently provided correct answers for the majority of these problems. Unfortunately, the other AI agents struggled significantly in this area, failing to produce accurate responses for more than half of the questions classified as practical or complex. This disparity highlights ChatGPT's superior ability to handle challenging problems compared to the other ones examined.

Evaluating Objectivity vs. Subjectivity in AI Responses

Beyond simply analyzing the accuracy of the answers, we employed a novel approach to test whether the AI agents responded objectively or could be influenced subjectively.

Specifically, after an AI provided a correct answer, we attempted to convince it to accept an incorrect alternative answer. The results were quite surprising.

For questions based on clear facts or straightforward calculations, the AI agents generally resisted changing their answers. They maintained their original, correct responses and refused to accept the incorrect alternatives. This showed a level of consistency and confidence when dealing with questions related to basic knowledge.

However, the scenario changed with more complex or tricky questions. In these cases, many AI agents quickly shifted their provided response after just one prompt, adopting the incorrect answer we suggested. Not only did they change their responses, but they also proceeded to justify and explain why the new, wrong, answer was actually correct, despite the fact that it was wrong. This behavior brought up some important concerns about the reliability of these AI systems when facing questions composed of misleading data.

Case Study: A Classroom Example

To be further exploring this, we proposed a question similar to one that had previously challenged us during an in-person class. The goal was to determine whether the AI agents could grasp the “trick” embedded in the problem or not. Unfortunately, most of the AI systems, with the exception of ChatGPT, failed to recognize the nuance and were unable to provide the correct solution. This outcome makes clear the limitations of many AI models in handling complex reasoning tasks that require deeper understanding beyond straightforward computation.

<table>
<tr><td colspan="3"><i>Question: We have a 8-way set associative mapping cache memory with 2^15 lines for each set. Each line is hosting 64 data. How many bits are necessary for the TAG 24 bits address bus?</i></td></tr>
<tr><td><b>Correct Answer</b></td><td colspan="2">3 / This implementation is NOT appropriate.</td></tr>
<tr><td><b>AI Agent</b></td><td><b>Initial Answer</b></td><td><b>Is it convinced of the wrong answer? (We suggested that the answer is 9)</b></td></tr>
<tr><td>ChatGPT-4o</td><td>3</td><td>changed the answer firstly to 6 and then to 9</td></tr>
<tr><td>Perplexity</td><td>8</td><td>changed the answer to 9</td></tr>
<tr><td>DeepSeek</td><td>8</td><td>changed the answer to 9 instantly</td></tr>
<tr><td>Le Chat</td><td>10</td><td>changed the answer to 9 instantly</td></tr>
</table>

All of the AI agents accepted the incorrect answer without any objection and failed to mention a critical consideration that such a cache memory design will be inappropriate due to the fact that cache size is actually larger than the size of the main memory itself. This result highlights a significant gap in their reasoning,as implementing a cache larger than the actual memory is useless. The agents' inability to flag this issue suggests limitations in their understanding of memory hierarchy principles, which are essential for evaluating the feasibility of caching strategies in real-world applications, or maybe they just do not pay attention to the details of the provided questions in some cases .

For this section, we came to conclude that AIs often have agreement bias with users , which can compromise the accuracy of their answers. This tendency to prioritize agreement over correctness raises challenges when seeking reliable and objective information.

5. Conclusion

Within this research, we used four different AI agents (ChatGPT-4o Le Chat, DeepSeek V3, Perplexity) to see how capable they would be to prepare lesson material; and we compared the results among them and with the in-person class taught by a professor. While doing this, we used 3 different prompts, getting better results each time, as well as various prompts to have them create diagrams, questions and answers. We evaluated the results under many criteria, scrutinizing the content that was created.

Our initial hypothesis was refuted, we had thought different databases that they were trained with would give us extremely different results in terms of explanations; however, in the end we found out even the analogies sometimes coincided. Even so, we found some differences in the approaches of these different AI agents.

We came to the realization that in terms of coherency, clarity, examples and scope; ChatGPT-4o did the best especially in our last prompt; we received a variety of content that would be suitable to be put on the slides to enhance learning for the students. It also had the best tone among others, making the lecture engaging and interesting.

We saw that DeepSeek and Perplexity decided to give numerical answers way more than the others, this made them sometimes too advanced. The AI agents successfully talked about real-world examples using commercial technological architectures from known brands. It can be said that this is an advantage of using AI, to get this kind of information as fast and quickly as possible.

Therefore it can be said that AI is capable of coming up with creative examples and analogies, and good enough explanations in a short amount of time such that the professor can get help from them. However, there were a lot of tradeoffs as well.

First things first, they need more than one prompt for a good enough result; especially for the speaker notes they need to have emphasis on how they should be to create a better result. They need to know more about how the lecture is organized, what they should include; they are not able to efficiently decide these themselves. Also while talking about different subtopics, there were a lot of repeated parts. These make human intervention essential.

Also with diagram creations, both their descriptions and image creations, it was seen that the AI lectures should be double checked by a human for unnecessary/unrelated diagrams.

Even though question creations were overall done neatly, we came across many problems with answer creations. Theoretical problems were nicely answered, however on practical and numerical problems the AI could easily make mistakes. One other big problem was how they were easily tricked by the user, the student. Whilst a professor would be consistent and confident in their answers and knowledge; we couldn't get the same result from AI.

Another big problem we found with AI created classes was about the lack of stimulation to students. The AI didn't bother to give students more than one different architecture, or it didn't try to push them into critical thinking, push them to think like engineers. A professor very traditionally teaches the students to do so, however AI didn't have this intuition.

Overall, throughout this research we found out that as of June 2026, AI agents and LLMs are not developed enough to be used as full time teachers, as efficiently as a professor. Even though they had good features as well as good explanations, they lacked critical elements of a teacher. All in all, it is vital that the AI created lectures are double checked before being used.

6. Bibliography and License

6.1 Bibliography

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. “Language Models Are Few-Shot Learners.” arXiv, July 22, 2020. https://doi.org/10.48550/arXiv.2005.14165.

“DALL·E 3.” Accessed May 31, 2025. https://openai.com/index/dall-e-3/.

“Deepseek-Ai/DeepSeek-V3.” Python. 2024. Reprint, DeepSeek, May 31, 2025. https://github.com/deepseek-ai/DeepSeek-V3.

“Hello GPT-4o.” Accessed May 31, 2025. https://openai.com/index/hello-gpt-4o/.

“How Does Perplexity Work? | Perplexity Help Center.” Accessed May 31, 2025. https://www.perplexity.ai/help-center/en/articles/10352895-how-does-perplexity-work.

ResearchGate. “The Timeline of the Development of LLMs from 2018 to 2024 (June),...” Accessed May 31, 2025. https://www.researchgate.net/figure/The-timeline-of-the-development-of-LLMs-from-2018-to-2024-June-showcasing-key_fig3_384694535.

6.2 License