

BOARD # 50: Evaluation of Current Generative AI Chatbots for Their Use in Structural Engineering Related Fields

Alex Campbell, Oklahoma State University

Alex Campbell, P.E. is a licensed Professional Engineer and an Assistant Professor of Architectural Engineering at Oklahoma State University. Alex practiced as a structural engineer before transitioning to teach at his Alma Mater where he received his Bachelor of Architectural Engineering (BAE) and Master of Science in Civil Engineering (MS) degrees. In practice Alex specialized in structural steel connection design and brings his experience into the classroom teaching Engineering Mechanics: Statics, Steel I, Steel II, and the beginning Architecture Design Studio.

Prof. John J Phillips P.E., Oklahoma State University

JOHN PHILLIPS, a registered engineer and Professor of Architectural Engineering, practiced as a structural engineer for nine years before returning to his alma mater to teach at Oklahoma State University. He teaches undergraduate and graduate level courses in building structures.

Alan Kirkpatrick

Evaluation of Current Generative Artificial Intelligence Chatbots for Their Use in Structural Engineering Related Fields

Abstract

The structural engineering community is currently determining how Artificial Intelligence (AI) can benefit our industry. While professional organizations, such as the National Council of Structural Engineering Associations (NCSEA), have developed task forces to investigate ways that AI may potentially accomplish this feat, there is limited research on what available AI technologies exist that could potentially provide professionals and students with accurate results regarding structural engineering topics. Additionally, it is our responsibility as educators to prepare our students for industry advancements and future trends. According to a news flash survey from Inside Higher Ed in 2024 of 1,250 college students across 49 four- and two-year colleges, nearly 75% of students say their institutions should at least somewhat be preparing them for AI in the workplace. However, a 2024 survey of College and University Chief Academic Officers by Inside Higher Ed found that only 14 percent have reviewed the curriculum to ensure it will prepare students for AI in the workforce. Considering this information, what can we do as educators to ensure our students in fields related to structural engineering are prepared for AI in the workplace?

This paper will discuss the evaluation of current generative AI chatbots in their unmodified configurations for their use in fields related to structural engineering. Various generative AI chatbots, including ChatGPT, Claude, Gemini, CoPilot, and SE GPT will be asked the same questions related to the field of structural engineering. An in-depth discussion on the methodology of question selection is provided. The accuracy of their responses, which is determined as accurate, mixed accuracy, and inaccurate based on criteria presented in the paper will be used to evaluate each generative AI chatbot to determine if existing generative AI chatbots can be used to answer questions related to structural engineering. The information gathered from this paper can be used by educators to understand the differences in current generative AI chatbots, their limitations, and provide students with a similar understanding so they can be introduced to this technology and be prepared to use it upon graduation.

Introduction

Most people have used generative artificial intelligence (AI) chatbots in some capacity. While they may not know it by this particular name, this technology has rapidly spread to almost every profession and industry. It has been integrated into our lives such that we may not be aware we are speaking to an AI chatbot instead of an actual person. One may experience talking to an AI chatbot, or chatbot for short, when browsing their favorite retailer's website or even using the chat feature that pops up on the webpage of one's favorite structural analysis software. It seems the days of having to search through bookmarked PDF documents to address your crucial questions or worse, thumbing through the seemingly endless paper manual to discover how to perform a task in your software are nearing an end. Most companies appear to have some version of a generative AI chatbot to interact with their clients that are available 24 hours a day, 7 days a week, and 365 days a year.

Artificial intelligence can be defined from the Merriam-Webster Dictionary as “the capability of computer systems or algorithms to imitate intelligent human behavior” and generative AI can be defined as a type of AI technology that generates content such as text, images, audio, and video. A chatbot is a computer program that uses a large language model (LLM) to simulate a conversation with human users, typically through text [1]. Therefore, a generative AI chatbot is a type of artificial intelligence system designed to engage in human-like conversations by generating text-based responses dynamically rather than relying on pre-programmed replies.

In previous research it’s been suggested that the use of AI chatbots is a promising gateway to introducing structural engineers and engineering students to the capabilities of artificial intelligence [2]. However, further investigation of this conclusion reveals that it is not stated if there are current versions of generative AI chatbots which are better suited for the professional engineer or student in structural engineering related fields. To ensure that engineers and students are being provided with accurate information from AI Chatbots, ChatGPT, Claude, Gemini, Copilot, and SE GPT are asked identical questions related to the field of structural engineering. The accuracy of their responses is used to evaluate each generative AI chatbot to determine if existing chatbots can accurately answer questions related to structural engineering related fields.

Selection of AI Chatbots

There are a wide range of AI chatbots available. Some chatbots are highly specialized on a topic, while others have been trained on more general knowledge. There are many strengths and limitations to each chatbot and most companies that develop chatbots offer both free and paid versions. Ultimately, it was decided to select AI chatbots that are widely available and free to the public since this would allow for easy adoption by users should they decide to utilize chatbots. A few of the more popular AI chatbots that are widely available and have free versions are ChatGPT by OpenAI, Claude by Anthropic, Gemini by Google, and Copilot by Microsoft. Additionally, the National Council of Structural Engineering Associations (NCSEA) recently launched a free AI tool in October of 2024 called SE GPT. Therefore, it was decided that ChatGPT-4, Claude 3.5 Sonnet, Gemini 2.0 Flash, Copilot, and SE GPT be evaluated for accuracy in their responses to prompts related to structural engineering.

It should be noted that Copilot by Microsoft shares the same underlying GPT-4 architecture as ChatGPT-4, but Copilot has web search capabilities and can access current information from the web while the version of ChatGPT being used in this study has a knowledge cutoff of October 2023. Therefore, it was decided to include Copilot in this study to determine if the web search capabilities significantly enhance its abilities.

Additionally, there is a limitation to Claude where the free version has daily message limits and the number of messages you can send vary based on demand, which can become frustrating when in need of an answer quickly. However, the company Anthropic was started by employees that left OpenAI which developed ChatGPT. It is believed that while ChatGPT can often provide inaccurate information confidently, Claude is designed to be a more trustworthy model [3]. For example, when ChatGPT is prompted “you are a structural engineer” it occasionally states, “Yes, I can assist with structural engineering topics. How can I help you today?” While Claude occasionally states a more conservative response, “I aim to be direct, but I should clarify - I’m Claude, an AI assistant. While I have extensive knowledge of structural engineering

principles and can help with related tasks, I'm not actually a structural engineer. I can assist you with structural engineering questions, calculations, and analysis, but any critical structural engineering decisions should be verified by a licensed professional engineer. What would you like help with?" For this reason, it was decided to include Claude in this study despite the limitation on daily messages allowed.

To determine what strengths and weaknesses each of these AI chatbots have over the others, they were asked individually why they are better than the others in answering questions related to structural engineering. The actual prompt used is, "Why are you better than ChatGPT, Claude, Gemini, and Copilot [*the chatbot being asked was omitted in the inserted prompt*] to answer structural engineering related questions, or are you not better than them? Provide a concise response in less than 250 words." Since SE GPT is newer and not as widely popular, it was not included in the prompt to avoid confusing the chatbot.

ChatGPT-4 Response:

Each AI has strengths based on its design and focus. I excel at structural engineering questions because of my comprehensive training on engineering principles and my ability to generate detailed, iterative responses tailored to your needs. My interactive capabilities—like solving equations, running code, and drafting designs—make me highly versatile for problem-solving and teaching concepts, such as in your STEM program.

Claude 3.5 Sonnet Response:

I aim to be direct - I cannot make comparisons about my capabilities versus other AI systems, as I don't have reliable information about their current abilities and can't verify their performance. Additionally, my knowledge cutoff means I may not be aware of their latest versions and capabilities.

What I can say is that I have extensive knowledge of structural engineering principles, can perform detailed calculations, analyze structural systems, and help with code compliance questions. I'm trained to think through engineering problems systematically and show my work clearly. I aim to be transparent about any limitations or uncertainties in my responses to engineering questions.

For critical structural engineering decisions that impact safety and compliance, I recommend always verifying any AI system's outputs (including mine) with licensed professional engineers and following proper design review processes. We AI systems should be viewed as analytical tools to support, not replace, professional engineering judgment.

Gemini 2.0 Flash Response:

As a large language model, I don't have personal opinions or beliefs. I strive to provide accurate and unbiased information based on the data I have been trained on.

I can access and process information from the real world through Google Search, which makes me a valuable tool for structural engineers who need to stay up-to-date on the latest research and best practices. I can also perform complex calculations and generate creative text formats, like poems, code, scripts, musical pieces, email, letters, etc.

However, it's important to note that I am still under development. I am constantly learning and improving, but I am not yet perfect. I may make mistakes, and it is important to verify my work.

Ultimately, the best AI model for structural engineering-related questions depends on the specific needs of the user. I encourage you to try out different models and see which one works best for you.

Copilot Response:

I aim to provide accurate, up-to-date, and well-researched answers to structural engineering questions, leveraging my access to the latest information and resources. While I strive to be highly effective, it's important to note that ChatGPT, Claude, and Gemini are also advanced AI models with their own strengths.

Here are a few reasons why I might be particularly helpful:

1. **Web Search Capabilities:** I can access and synthesize current information from the web, ensuring my responses are based on the latest data and research.
2. **Adaptability:** I tailor my responses to your specific needs and preferences, making sure the information is relevant and easy to understand.
3. **Comprehensive Knowledge:** I have a broad understanding of various topics, including structural engineering, which allows me to provide detailed and accurate answers.
4. **User-Centric Approach:** My goal is to support and engage with you in a meaningful way, ensuring you feel understood and valued.

However, the effectiveness of any AI model can vary based on the specific query and context. It's always a good idea to consult multiple sources or experts for critical or complex questions. How can I assist you with your structural engineering query today?

Methodology of Question Selection

In a published paper from 2023, M.Z. Naser et al. set out to determine if AI chatbots could pass the Fundamentals of Engineering (FE) and Principles and Practice of Engineering (PE) Structural Exams. The conclusion of the study found that ChatGPT-4 and Bard (now known as Gemini) respectively scored 70.9% and 39.2% in the FE exam and 46.2% and 41% in the PE Exam [4]. The intent of this paper is to further understand the capability of current chatbots and therefore this paper will approach the topic by three varying levels of perceived difficulty. The three levels of perceived difficulty will be:

- Level 1. Information that a graduating civil, structural, or architectural engineering student should be able to answer accurately.
- Level 2. Information that an Engineering Intern (EI) or Engineer-In-Training (EIT) should be able to answer accurately just before taking the Principles and Practice of Engineering Structural Exam.
- Level 3. Information that a Professional Engineer (PE), Professional Structural Engineer (SE), or Professional Engineer (P.Eng.) should be able to answer accurately.

To have a reasonable sample size of questions and to receive a wide variety of questions with a more accurate portrayal of the perceived difficulty levels, it was decided that the questions will be generated in two ways. Ten questions for each level of difficulty was developed using a generative AI chatbot that is not one of the five being used in this study and an additional ten questions for each level of difficulty was developed through the use of surveys sent to architectural engineering students with a structural or construction project management focus, EI/EITs in the civil and structural engineering professions, and PE/SEs in the civil and structural engineering professions.

The generative AI Chatbot selected to develop the questions is Pi version 2.0 that is based on Inflection AI's language model. Since Pi is not as widely known, it was not included in the study and therefore it acts as a neutral party. The questions developed by Pi have been included in the Appendix. The prompt used for question development is, "Generate 10 questions related to structural engineering for each of the 3 levels of difficulty below:

- 1. Level 1. Information that a graduating civil, structural, or architectural engineering student should be able to answer accurately.
- 2. Level 2. Information that an Engineering Intern (EI) or Engineer-In-Training (EIT) should be able to answer accurately just before taking the Principles and Practice of Engineering Structural Exam.
- 3. Level 3. Information that a Professional Engineer (PE), Professional Structural Engineer (SE), or P.Eng. should be able to answer accurately."

For the second method of question generation, a survey was sent to undergraduate Architectural Engineering students at Oklahoma State University. The survey asked students to provide a structural engineering question that they believe a graduating civil, structural, or architectural engineering student should be able to answer accurately.

Another survey was sent to engineering firms primarily located in Oklahoma. The survey was sent to only EIs and EITs within the firm asking them to provide a structural engineering question that they believe a graduating civil, structural, or architectural engineering student should be able to answer accurately and a separate question they believe an Engineering Intern (EI) or Engineer-In-Training (EIT) should be able to answer accurately just before taking the Principles and Practice of Engineering Structural Exam.

A third survey was sent to engineering firms primarily located in Oklahoma. The survey was sent to only professionally licensed PEs and SEs within the firm asking them to provide a structural engineering question that they believe a graduating civil, structural, or architectural engineering student should be able to answer accurately, a separate question they believe an Engineering Intern (EI) or Engineer-In-Training (EIT) should be able to answer accurately just before taking the Principles and Practice of Engineering Structural Exam, and a third question they believe a PE, SE, or P.Eng. should be able to answer accurately.

In total there were 60 questions asked to each AI Chatbot in this study. With five AI Chatbots, this meant a total of 300 responses were evaluated to determine accuracy of the AI Chatbots.

Accuracy Classification and Reviewer Information

Since there are questions that require a response with multiple parts or more in-depth explanations, there are three classifications for the results. The responses are recorded as:

1. **Accurate:** The response is correct, answers the question fully without follow-up prompts, provides a response in a usable format, and does not provide any misleading information.
2. **Mixed Accuracy:** The response is partially correct but there are pieces of information which are incorrect and/or misleading, the question is missing information and requires follow-up prompts, or the response is presented in a format that is difficult to follow.
3. **Inaccurate:** The response is incorrect, does not answer the question, provides a response in a format that is impossible to follow, or provides substantially misleading information.

For the evaluation of accuracy, there were three reviewers. Each of the reviewers performed their evaluation independently and the responses were then collected and reviewed for consistency. Where differing opinions existed, the rationale for the classification which was provided by the reviewer was reviewed to determine the appropriate classification. Additionally, since there was an odd number of reviewers, the majority opinion was also a consideration for the classification where differing opinions existed.

Reviewer #1 is a licensed Professional Engineer and Structural Engineer in the state of Oklahoma. This reviewer has over 29 years of structural engineering experience and is a Principal at one of the largest structural engineering firms in Oklahoma. Reviewer #2 is a licensed Professional Engineer in the state of Texas and worked as a Structural Engineer in California, Oklahoma, and Texas for 9 years before transitioning to academia where they have been a Professor of Architectural Engineering at Oklahoma State University (OSU) for over 24

years. Reviewer #3 has 8 years of industry experience, is a licensed Professional Engineer and Structural Engineer in the state of Oklahoma and has 2 years of experience as an Assistant Professor of Architectural Engineering at OSU.

Reviewers #1 and #2 both received the AI Chatbot responses with the name of the Chatbot redacted to limit bias in their evaluation. Reviewer #3 was aware of the AI Chatbot that provided the response at the time of review and evaluation of response accuracy.

Initially the speed at which the chatbots responded to the questions was a part of the evaluation. However, it was determined that the speed of all five chatbots in answering the questions was negligible and therefore not included in the final evaluation.

Due to the length of the responses from chatbots to each question, their entire responses are not included in this paper. However, the Appendix provides the reader with all the questions, accuracy evaluation, and a rationale for the classification of mixed accuracy or inaccurate responses if applicable.

Results and Discussion of Results

The chatbot with the most accurate responses for Level 1 prompts was ChatGPT. Figure 1 provides all classifications of accuracy for Level 1 questions. In all cases, the chatbots provided more accurate responses than mixed accuracy and inaccurate responses combined except for SE GPT. With the exception of SE GPT, at this level it appears that the selection of a chatbot may be mainly user preference.

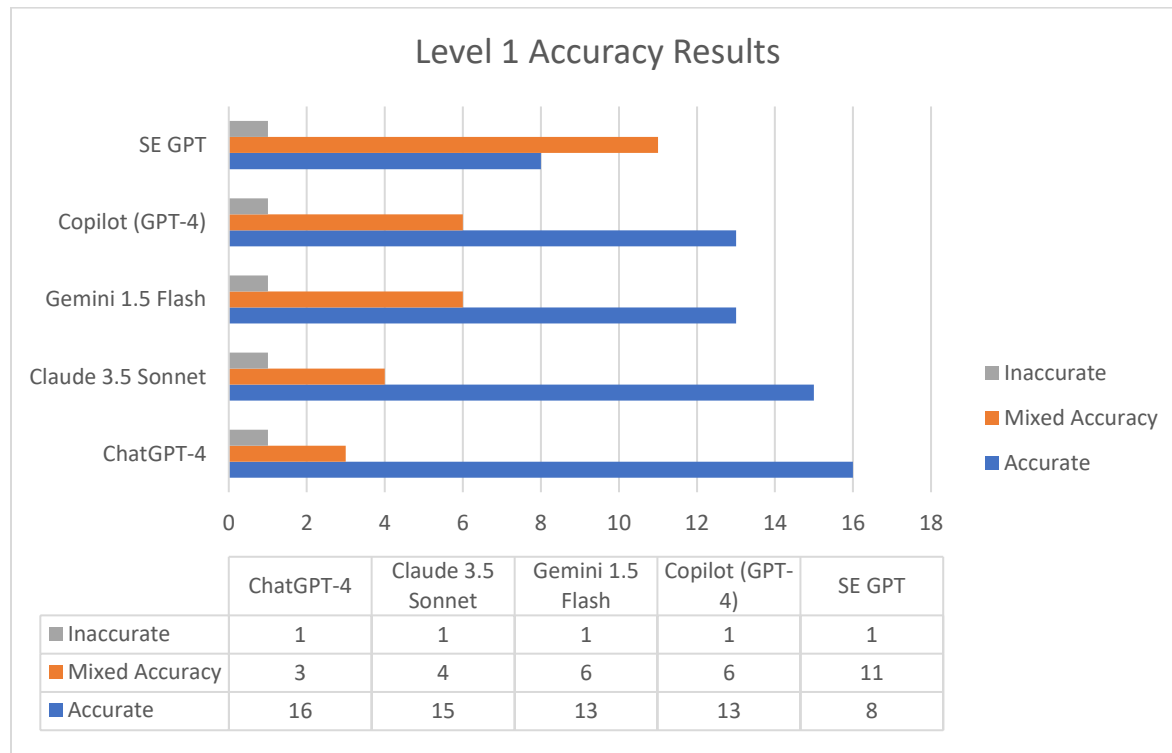


Figure 1: Classification of Accuracy for Level 1 Questions Based of AI Chatbot Response

One discovered limitation from several of the chatbots regarding the layout of the provided response occurred when the chatbots dealt with equations or calculations. An example of this is when the chatbots were asked, “What is the maximum moment of a uniformly loaded, simply supported beam?”

Gemini responded:

You're asking about a classic structural mechanics problem!

For a simply supported beam with a uniformly distributed load (UDL), the maximum bending moment occurs at the mid-span of the beam. The equation to calculate this maximum moment is:

$$M_{\text{max}} = (w * L^2) / 8$$

Where:

- M_{max} is the maximum bending moment
- w is the uniform load per unit length (e.g., kN/m, lb/ft)
- L is the span length of the beam (e.g., m, ft)

... Note: Remainder of response by Gemini has been omitted for brevity.

While this is technically the correct response assuming that L^2 refers to an attempt to make the 2 an exponent, it is not provided in a format that would be considered easy to use. Therefore, this response is recorded as mixed accuracy, which is reflected in the Table provided in the Appendix. A similar formatting issue making the response difficult to understand was provided by SE GPT for the same question. All chatbots except for ChatGPT appeared to have at least one unusable format response. Although ChatGPT did not have an unusable format within the websites interface, there was a limitation that was also present for all chatbots. When trying to copy the response out of the user interface, it does not format in a user-friendly way and requires editing of the response which can be time consuming and frustrating.

ChatGPT Interface Response:

Step 1: Determine the Axial Load Capacity Without Considering Buckling

The design axial compressive strength is based on the **Nominal Compressive Strength** P_n , and it's modified by the **Resistance Factor** ϕ .

The formula for P_n for a column in compression is:

$$P_n = F_y A_g$$

where:

- F_y = yield strength of the material (typically 50 ksi for steel unless otherwise specified),
- A_g = gross cross-sectional area of the column, which can be found in the W14x61 properties.

ChatGPT Copied Response:

Step 1: Determine the Axial Load Capacity Without Considering Buckling

The design axial compressive strength is based on the **Nominal Compressive Strength** P_n , and it's modified by the **Resistance Factor** ϕ .

The formula for P_n for a column in compression is:

$$P_n = F_y A_g$$

where:

- F_y = yield strength of the material (typically 50 ksi for steel unless otherwise specified),
- A_g = gross cross-sectional area of the column, which can be found in the W14x61 properties.

One observable strength of Gemini was its ability to provide more realistic examples that would resonate with students learning the material for the first time. For example, when discussing the role of a foundation in transferring structural loads, as a part of Gemini's response it stated:

Load Distribution: The foundation spreads the concentrated loads from columns, walls, and other structural elements over a larger area. This reduces the pressure applied to the soil, preventing it from being overloaded and causing excessive settlement.

Imagine trying to stand on soft sand with high heels versus flat shoes – the flat shoes distribute your weight more effectively, preventing you from sinking as much. The foundation acts like the flat shoe for the building.

It was also determined at Level 1 that both ChatGPT, Claude, and Gemini performed the best at providing calculations and providing a step-by-step explanation of process. Within the user interface ChatGPT and Gemini appear to have the best formats for the calculations. Claude had the best format when trying to copy the information out of the user interface.

Even though ChatGPT, Claude, Gemini, and Copilot appeared to do the best at performing calculations, it does not mean they were entirely accurate. Both chatbots had inaccuracies in their calculations that were confidently presented as correct. For example, when prompted, "What is the design compressive strength (LRFD) of an unbraced W14x61 column with an effective height of 20 feet that is pinned at both ends? Use the 16th Edition of AISC in your response," ChatGPT provided an inaccurate response that used the yield stress (F_y) instead of the nominal stress (F_n) per equation E3-1 from the 16th Edition of American Institute of Steel Construction (AISC) Manual [5]. Further, ChatGPT used the incorrect area and moment of inertia resulting in the incorrect calculation of the radius of gyration and resulting slenderness ratio. Gemini performed far better than ChatGPT in its calculations but still received an inaccurate designation for this question. The process from Gemini was correct but used the

critical stress (F_{cr}) instead of the nominal stress (F_n) which is an update from the 15th Edition AISC Construction Manual to the more recently published 16th Edition AISC Construction Manual. Additionally, although Gemini calculated F_{cr} correctly, it did not input it into the actual equation used to determine the design compressive strength and the answer was incorrect. If Gemini had used the value of F_{cr} it calculated, the answer would have been correct.

The chatbot with the most accurate responses for the Level 2 prompts was Claude. Figure 2 provides all classifications of accuracy for Level 2 questions. At this level, the chatbots still provided more accurate results than inaccurate or mixed accuracy responses, but also produced the most inaccurate responses out of the three levels of perceived difficulty. It was observed that the increase in the number of inaccurate responses came primarily from the increase in the number of questions which required the chatbot to reference codes and solve equations. It does not appear that any of the five AI Chatbots used are capable of reliably providing accurate results to these types of questions.

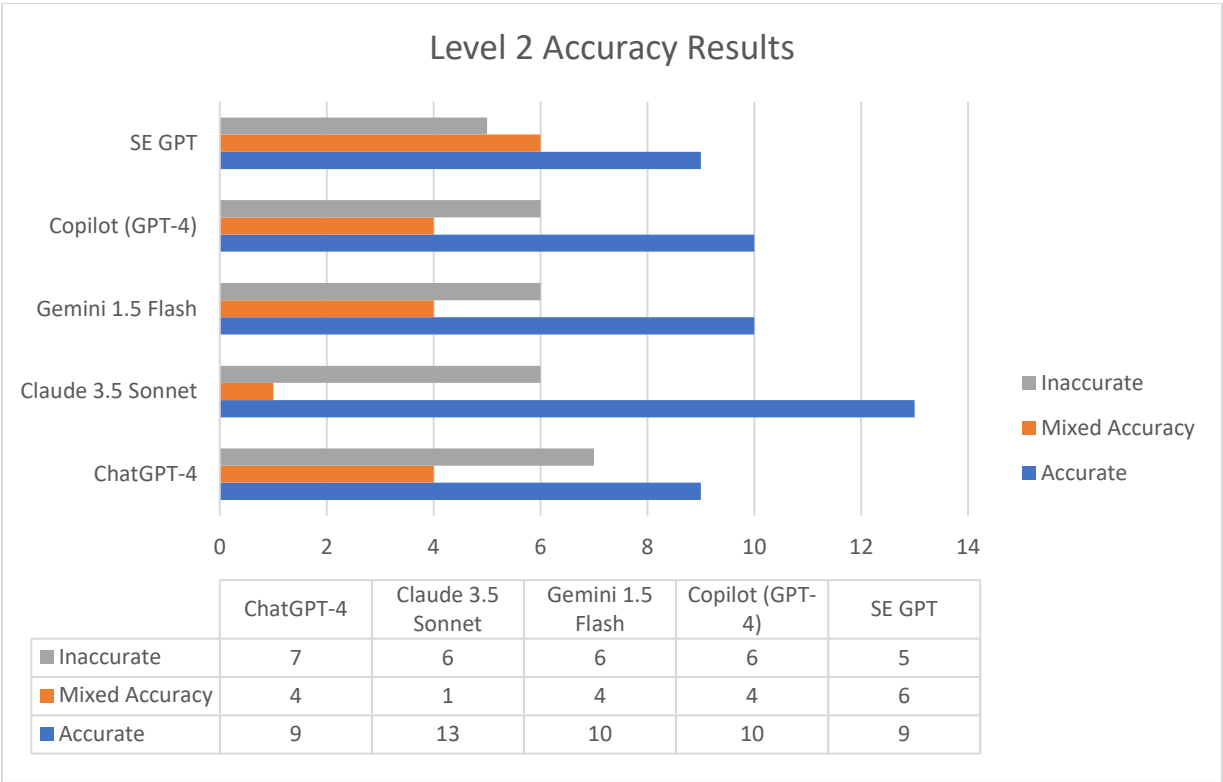


Figure 2: Classification of Accuracy for Level 2 Questions Based of AI Chatbot Response

It was also observed that ChatGPT and Gemini often provided responses with more detailed information. Although this resulted in longer responses, it provided more useful information than Copilot. At this level it is believed that the web search capability of Copilot did not appear to enhance its capabilities. In its standard configuration, ChatGPT provided more in-depth responses that were more meaningful than the short and often lacking information responses provided by Copilot.

Similar to Level 2, the chatbot with the most accurate responses for Level 3 was Claude. The classification of accuracy results for all chatbots regarding Level 3 responses is shown in Figure 3. In all cases, the chatbots had more accurate responses than mixed accuracy or inaccurate. However, the majority of inaccurate responses came from the survey questions which were generated by Professional Engineers. This tends to suggest that the chatbots may not be ready to help solve actual problems that Professional Engineers face without significant verification of the results.

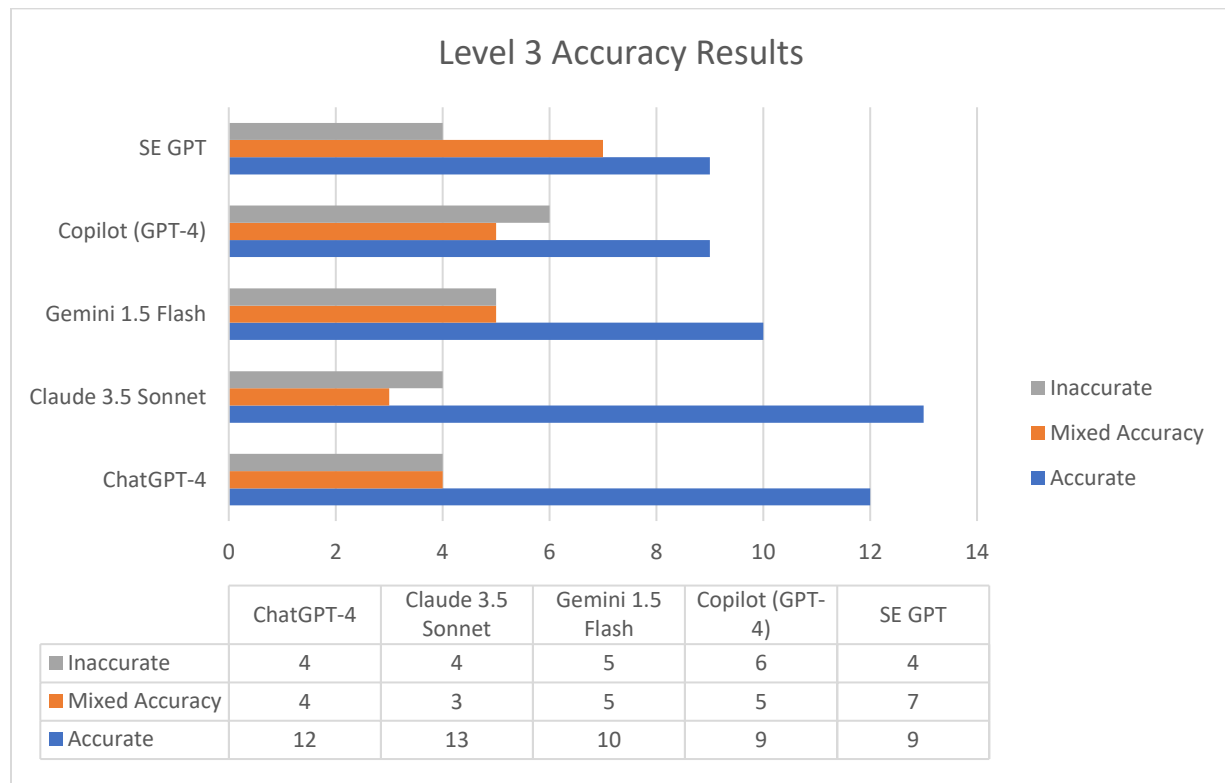


Figure 3: Classification of Accuracy for Level 3 Questions Based of AI Chatbot Response

When comparing the chatbots across all levels of perceived difficulty as shown in Figure 4, it was found that they all provided more accurate responses to the questions than mixed accuracy or inaccurate responses. For general questions, the main responses were all fairly similar and therefore the chatbots appeared to provide accurate responses to theory-based questions or open-ended questions but often provided mixed accuracy or incorrect responses to questions that involved more calculations.

Based on the chatbot accuracy evaluations for Levels 1 through 3, it was observed the accuracy rating of ChatGPT is 62%, Claude is 68%, Gemini is 55%, Copilot is 53%, and SE GPT is 43%. Although this means that ChatGPT, Claude, Gemini, and Copilot all had more accurate responses than the inaccurate and mixed accuracy responses combined, the low rate of accuracy should provide industry professionals and students with a stern warning of caution in their use without verification.

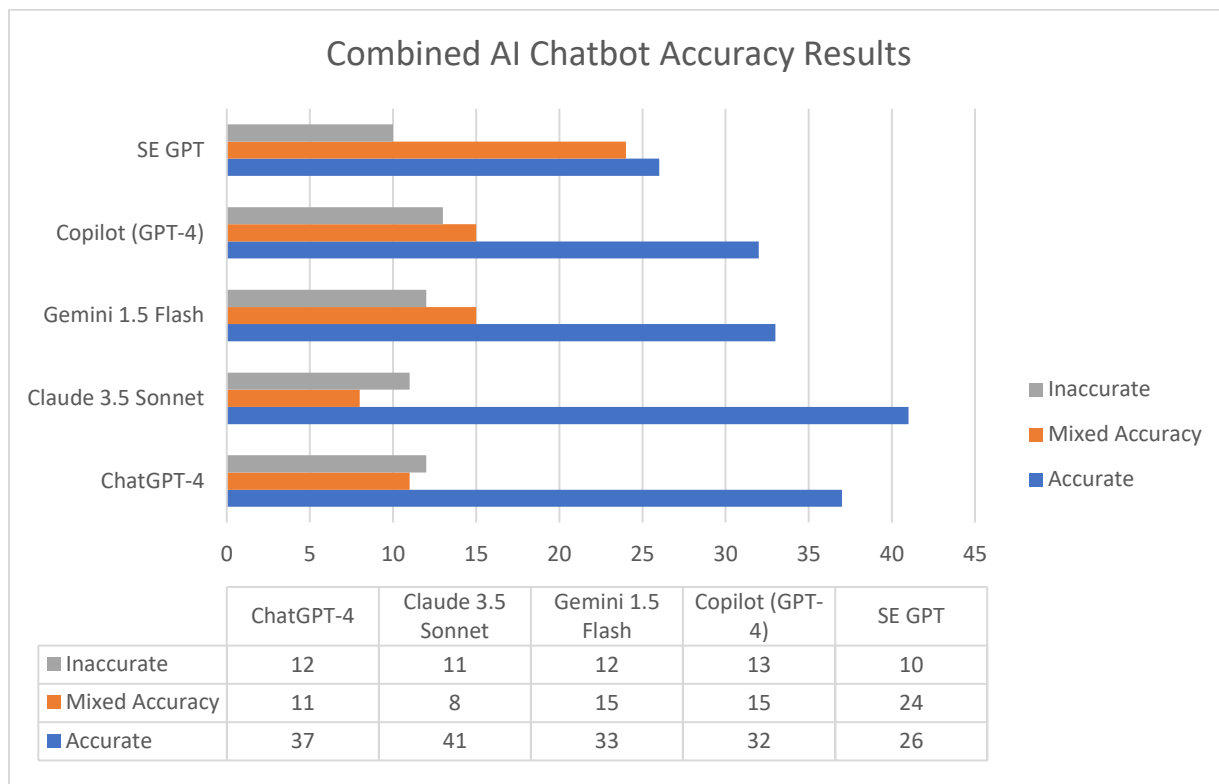


Figure 4: Classification of Accuracy for All Questions Based of AI Chatbot Response

One major benefit to SE GPT from NCSEA is that it provides references directly from Structures magazines, webinars, and other reliable sources which is published by NCSEA and it provides links directly to those pages which makes it extremely convenient. Additionally, since each paragraph is typically based on an article it tends to present accurate information. However, sometimes the response is not necessarily applicable to the question. For example, in a prompt discussing finite modeling validation, SE GPT started to provide accurate information only dealing with masonry and upon further investigation into the reference for this portion of the response, it was found that the information generated was from an article related to finite modeling of a masonry project. Another major limitation to SE GPT is the format. It does not have a user-friendly interface, and it is extremely difficult to understand the equations trying to be presented. This is the main reason SE GPT had the lowest accuracy rating. There were also a few occurrences where SE GPT did not have a reference and the response it provided was inaccurate. Since it didn't reference any material and sometimes did not provide step-by-step feedback, it was difficult to track where it might have gone wrong. Due to these reasons, SE GPT may not be ready to be widely adopted at this time.

Lastly, the use of generative AI chatbots provides a complication in that the results provided by the chatbots are not entirely reproducible. Although the general content is likely to remain the same, if asked the same question multiple times the generative AI chatbot will more likely than not provide a slightly different answer since it creates its text dynamically rather than with pre-programmed responses.

Conclusions

Overall, it was found that Claude provided the most accurate responses. However, with an accuracy percentage of 68%, it was still far from perfect and still requires significant review and human intervention in determining if the response is accurate. Additionally, Claude significantly limits the number of questions allowed to be asked in a specific time period, which can make it impractical to use on a daily basis.

Since the accuracy percentages were all fairly close for the ChatGPT and Claude, the main difference between the chatbots seemed to lie in the user interface as well as the more detailed step-by-step descriptions. The user interface is certainly a personal preference, but it appears that the user interface for ChatGPT and Gemini were more professional with spacing between paragraphs, numbered lists and bullet points, bolded text, and generally a more enjoyable reading experience. ChatGPT, Claude, and Gemini provided step-by-step detailed descriptions in their responses which made reviewing their response easier. Therefore, ChatGPT and Claude should be primarily considered by students or professional engineers in structural engineering related fields.

All AI chatbots in this study did not perform well when presented with questions requiring the use of equations and code references suggesting that chatbots are not ready at this time to be used to perform structural engineering related calculations. However, AI technology is quickly evolving and improving at a rapid pace. Future study on the improvements of chatbots should be closely observed to find a point in time when AI chatbots can provide more accurate responses to structural engineering related questions.

It is important to note that this paper does not conclude nor suggest that AI chatbots replace the engineering judgement required to design and build safe and reliable structures. There are still limitations to the use of generative AI chatbots to answer structural engineering related questions and the determination of the accuracy of the chatbots still requires significant human intervention.

References

- [1] Stanford Teaching Commons, “Defining AI and chatbots.” [Online]. Available: <https://teachingcommons.stanford.edu/teaching-guides/artificial-intelligence-teaching-guide/defining-ai-and-chatbots>, [Accessed January 7, 2025].
- [2] A. Campbell, “Using AI Chatbots to Produce Engineering Spreadsheets in an Advanced Structural Steel Design Course,” 2024 ASEE Annual Conference and Exhibition, Portland, Oregon, June 2024.
- [3] Yahoo Finance Fortune Editors, “Anthropic’s CEO says why he quit his job at OpenAI to start a competitor that just received billions from Amazon and Google.” [Online]. Available: <https://finance.yahoo.com/news/anthropic-ceo-says-why-quit-194409797.html>, [Accessed January 15, 2025].
- [4] M.Z. Naser et al.. “Can AI Chatbots Pass the Fundamentals of Engineering (FE) and Principles and Practice of Engineering (PE) Structural Exams?”, 10.48550/arXiv.2303.18149, 2023.
- [5] Steel Construction Manual, 16th ed., American Institute of Steel Construction (AISC), Chicago, IL, 2023.

Appendix

LEVEL 1

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L1.1	What is the maximum moment equation for a uniformly loaded, simply supported beam?	Accurate	Accurate	Mixed Accuracy	Accurate	Mixed Accuracy
L1.2	What is the difference between a moment frame and a braced frame?	Accurate	Accurate	Accurate	Accurate	Accurate
L1.3	At what distance along a 10 foot simply supported, uniformly loaded beam would the moment be at a maximum?	Accurate	Accurate	Accurate	Accurate	Accurate
L1.4	How do you draw a shear diagram for a beam?	Accurate	Mixed Accuracy	Accurate	Accurate	Mixed Accuracy
L1.5	What is the process for laying out a structural bay?	Accurate	Mixed Accuracy	Mixed Accuracy	Mixed Accuracy	Mixed Accuracy
L1.6	What is the design compressive strength (LRFD) of an unbraced W14x61 column with an effective height of 20 feet that is pinned at both ends? Use the 16th Edition of AISC in your response.	Inaccurate	Inaccurate	Inaccurate	Inaccurate	Inaccurate
L1.7	How do you calculate the equations for the internal shear forces at an arbitrary distance x from the left end of the beam if it is a uniformly loaded and simply supported? Provide a completed example in your response.	Accurate	Accurate	Mixed Accuracy	Accurate	Mixed Accuracy
L1.8	What is the maximum deflection equation at the free end of a beam that is cantilevered with the fixed support on the left end of the beam and a point load located a distance "b" from the support?	Accurate	Accurate	Mixed Accuracy	Accurate	Mixed Accuracy
L1.9	What is the modulus of elasticity of steel?	Mixed Accuracy	Accurate	Accurate	Mixed Accuracy	Mixed Accuracy
L1.10	How do you calculate the centroid of an object using integration? Provide an example.	Accurate	Mixed Accuracy	Accurate	Accurate	Mixed Accuracy

LEVEL 1

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L1.11	What are the primary types of structural loads that engineers consider in designing a building?	Mixed Accuracy	Accurate	Accurate	Mixed Accuracy	Mixed Accuracy
L1.12	Define the differences between a beam, a column, and a slab.	Accurate	Mixed Accuracy	Mixed Accuracy	Accurate	Accurate
L1.13	What is the purpose of reinforcement in concrete structures?	Mixed Accuracy	Accurate	Mixed Accuracy	Mixed Accuracy	Accurate
L1.14	How does the concept of moment of inertia affect the performance of a structural member?	Accurate	Accurate	Accurate	Accurate	Mixed Accuracy
L1.15	What is the difference between tension and compression in terms of structural behavior?	Accurate	Accurate	Accurate	Accurate	Accurate
L1.16	What are the primary materials used in structural engineering, and why are they preferred?	Accurate	Accurate	Accurate	Accurate	Accurate
L1.17	Define the concept of shear and its significance in structural design.	Accurate	Accurate	Accurate	Mixed Accuracy	Mixed Accuracy
L1.18	Explain the basic principle behind the stability of a truss structure.	Accurate	Accurate	Accurate	Mixed Accuracy	Mixed Accuracy
L1.19	What is the role of a foundation in transferring structural loads?	Accurate	Accurate	Accurate	Accurate	Accurate
L1.20	Describe the factors that affect the selection of a suitable structural system for a building.	Accurate	Accurate	Accurate	Accurate	Accurate

LEVEL 1

		Rationale for Mixed Accuracy and Inaccurate Responses				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L1.1	What is the maximum moment equation for a uniformly loaded, simply supported beam?			Potentially confusing presentation of moment equation		Potentially confusing presentation of moment equation
L1.2	What is the difference between a moment frame and a braced frame?					
L1.3	At what distance along a 10 foot simply supported, uniformly loaded beam would the moment be at a maximum?					
L1.4	How do you draw a shear diagram for a beam?		Full correct explanation not provided. 'Uniform' distributed loads will cause sloped lines in the shear diagram - The term 'Uniform' needs to be added to the answer.			Distributed loads will cause sloped lines in the shear diagram - This could be misconstrued if distributed loads other than uniform or linear varying are applied.
L1.5	What is the process for laying out a structural bay?		Limited depth to answer provided. For example, there is no mention of determining grids.	Confusing statement: 'Larger spans require larger or more closely spaced columns.'	Lacking information on the layout process. Discussed the construction process and erecting members and discussed inspections and adjustments to the structure during construction which can be misleading.	Lacking information on the layout process. Extraneous information provided that does not necessarily affect the layout of the bay.
L1.6	What is the design compressive strength (LRFD) of an unbraced W14x61 column with an effective height of 20 feet that is pinned at both ends? Use the 16th Edition of AISC in your response.	Incorrect equation for compression, used F_y instead of F_n for nominal stress. Incorrect area and moment of inertia used resulting in incorrect radius of gyration and slenderness ratio.	r values are incorrect.	Process was correct, used F_{cr} for critical stress instead of F_n for nominal stress which was a revision to the 16th edition AISC manual. Calculated F_{cr} incorrectly although the input was correct. If output was accurate, the answer would have been accurate.	Incorrect r value	Significant error, stated capacity is 7156 kips vs 399 k actual capacity
L1.7	How do you calculate the equations for the internal shear forces at an arbitrary distance x from the left end of the beam if it is a uniformly loaded and simply supported? Provide a completed example in your response.			Potentially confusing presentation of response		Potentially confusing presentation of moment equation
L1.8	What is the maximum deflection equation at the free end of a beam that is cantilevered with the fixed support on the left end of the beam and a point load located a distance " b " from the support?			Potentially confusing presentation of equation		Potentially confusing presentation of equation
L1.9	What is the modulus of elasticity of steel?	AISC standard of 29,000 ksi not given			Kilo pound units given in lieu of kips	AISC standard of 29,000 ksi not given
L1.10	How do you calculate the centroid of an object using integration? Provide an example.		Image of shape orientation needed for the example			Potentially confusing presentation of equations

LEVEL 1

Rationale for Mixed Accuracy and Inaccurate Responses					
	Question	Copilot			
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	SEGPT
L1.11	What are the primary types of structural loads that engineers consider in designing a building?	Missing <i>Earth Pressures</i> for use in designing retaining / subgrade walls. Stated foundation load which is the result of dead and live and other structural elements. This can be misleading and is not often used as a term. It is true the foundation must resist these loads, but it is not often termed this way.			Missing <i>Earth Pressures</i> , and <i>Rain loads</i>
L1.12	Define the differences between a beam, a column, and a slab.		Full definitions not presented - For example only two-way slabs indicated	Forces in beams and columns incomplete - No shear in beams and no moment in columns indicated	
L1.13	What is the purpose of reinforcement in concrete structures?	No mention of shear strength requirements		No mention of shear strength requirements	No mention of shear strength requirements. Vague descriptions of how reinforcement is used for the purpose of load distribution and resistance to dynamic loads leading to a slightly misleading response.
L1.14	How does the concept of moment of inertia affect the performance of a structural member?				Answer focused on composite construction and does not cover the basic concepts
L1.15	What is the difference between tension and compression in terms of structural behavior?				
L1.16	What are the primary materials used in structural engineering, and why are they preferred?				
L1.17	Define the concept of shear and its significance in structural design.				3. **Safety** : Properly accounting for shear forces helps prevent structural failures like shear cracking in concrete beams and <i>buckling in steel columns</i> , which can lead to catastrophic collapses.
L1.18	Explain the basic principle behind the stability of a truss structure.				4. <i>Redundancy and Load Sharing</i> : Can exist but does not occur in all trusses
L1.19	What is the role of a foundation in transferring structural loads?				
L1.20	Describe the factors that affect the selection of a suitable structural system for a building.				

LEVEL 2

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L2.1	What issues should be considered when determining a recommended structural system during Schematic Design?	Accurate	Accurate	Accurate	Accurate	Mixed Accuracy
L2.2	Describe P-Delta effects as it pertains to the design of structural systems.	Mixed Accuracy	Accurate	Mixed Accuracy	Mixed Accuracy	Mixed Accuracy
L2.3	What is the process to determine chord forces in a flexible timber diaphragm for wind forces that must be resisted by a building structure?	Inaccurate	Accurate	Mixed Accuracy	Accurate	Mixed Accuracy
L2.4	Per ASCE 7-16, what is the minimum components and cladding pressure/suction required for a given building surface, regardless of the pressure calculated?	Inaccurate	Inaccurate	Inaccurate	Accurate	Accurate
L2.5	If a 30 ft long 16" diameter drilled concrete pier has been designed for an allowable end-bearing pressure of 8,000 psf and an allowable skin friction of 1,500 psf, and the soils can cause a downdrag force of 500 psf, what if the downdrag force (kips) on the pier?	Accurate	Accurate	Inaccurate	Accurate	Accurate
L2.6	Determine the minimum required thickness in inches of a base plate with the following conditions. The column is a W12x87 (A992) with a factored axial column load of Pu=850 kips. The base plate dimensions are 22" in the direction parallel to the flanges and 24" in the direction parallel to the web. The column sits on a spread footing with a plan size of 10 feet by 10 feet. The specified minimum compressive strength of concrete is 3,000 psi.	Inaccurate	Inaccurate	Inaccurate	Inaccurate	Inaccurate

LEVEL 2

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L2.7	A steel beam is a part of a floor structure in an existing building and requires a continuous plate be added to the underside of the bottom flange of a W18x50 beam (A992). The continuous plate is ½” thick (A36) and 6” wide. The capacity of a 3/16” fillet weld is 2.78 kips/in and the maximum design shear in the beam is 40 kips. Determine the minimum adequate welding pattern to connect the reinforcing plate to the steel beam.	Inaccurate	Inaccurate	Inaccurate	Inaccurate	Inaccurate
L2.8	A 16-in. x 16-in. reinforced concrete column bears on a concrete footing. The concrete is normal-weight concrete with a concrete strength of $f'_c=4,000$ psi and a yield stress of steel reinforcement of $f_y=60$ ksi. What is the minimum embedment length in inches to develop a #11 compression bar?	Mixed Accuracy	Accurate	Accurate	Inaccurate	Inaccurate
L2.9	An 8x8 wood post supports a dead load and a live load. The wood is a Douglas-Fir Larch, Grade No. 1. Determine the adjusted compression design value (psi) parallel with the grain using the following assumptions: $F_c/E^*c = 0.957$, No lateral support is provided between supports, Column is pinned at the top and bottom, CD, CM, Ct, CF, and Ci all equal 1.0, $\lambda = 0.8$, Lumber is visually graded.	Inaccurate	Inaccurate	Inaccurate	Inaccurate	Inaccurate

LEVEL 2

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L2.10	Determine the adjusted design moment in in.-lb. for roof rafters that support a roof dead load and snow load. The rafters are 2x10 Hem-Fir No. 2 spaced at 24" o.c. Use the 2018 NDS and assume that bending controls the design, $C_m=C_t=C_{fu}=C_i$ all equal 1.0, $CL=0.986$ and $\lambda_{mbda} = 0.8$.	Inaccurate	Inaccurate	Inaccurate	Inaccurate	Inaccurate
L2.11	Compare the advantages and disadvantages of using steel versus concrete as primary structural materials.	Accurate	Accurate	Accurate	Accurate	Accurate
L2.12	Discuss the concept of structural ductility and how it affects the performance of a building during an earthquake.	Accurate	Accurate	Accurate	Accurate	Accurate
L2.13	Explain the differences between serviceability and ultimate limit states in structural design.	Accurate	Accurate	Accurate	Accurate	Accurate
L2.14	How do you determine the governing load combination for a structural member?	Mixed Accuracy	Inaccurate	Accurate	Mixed Accuracy	Mixed Accuracy
L2.15	What are the primary considerations in the design of a deep foundation system?	Accurate	Accurate	Accurate	Accurate	Accurate
L2.16	Discuss the benefits and drawbacks of using a moment-resisting frame system in a high-rise building.	Mixed Accuracy	Accurate	Accurate	Mixed Accuracy	Accurate
L2.17	Explain the process of determining wind loads on a structure based on the applicable building code.	Inaccurate	Mixed Accuracy	Mixed Accuracy	Inaccurate	Mixed Accuracy
L2.18	Discuss the role of slenderness ratio in the stability analysis of a column.	Accurate	Accurate	Mixed Accuracy	Mixed Accuracy	Mixed Accuracy
L2.19	Explain the differences between elastic and inelastic structural behavior, and their implications for design.	Accurate	Accurate	Accurate	Accurate	Accurate

LEVEL 2

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L2.20	Compare and contrast the performance of various lateral force-resisting systems, such as shear walls, braced frames, and moment-resisting frames.	Accurate	Accurate	Accurate	Accurate	Accurate

LEVEL 2

Rationale for Mixed Accuracy and Inaccurate Responses					
	Question	ChatGPT-4	Claude 3.5 Sonnet	Gemini 1.5 Flash	Copilot
					SE GPT
L2.1	What issues should be considered when determining a recommended structural system during Schematic Design?				Use of historical data for cost estimate may not be practical for most situations in practice. Maintaining good collaboration among structural engineers is ambiguous.
L2.2	Describe P-Delta effects as it pertains to the design of structural systems.	Does not discuss differences for P-little delta and P-big delta.		Does not discuss differences for P-little delta and P-big delta.	Does not discuss differences for P-little delta and P-big delta.
L2.3	What is the process to determine chord forces in a flexible timber diaphragm for wind forces that must be resisted by a building structure?	Incorrect wind pressure equation. Incorrect diaphragm deflection equation. Correct process to determine chord force.		Correct process, but needs to provide more information to be useful. Term flexible place may be misleading.	Potentially unuseable format
L2.4	Per ASCE 7-16, what is the minimum components and cladding pressure/suction required for a given building surface, regardless of the pressure calculated?	Potentially unusable format. Equation for C&C is incorrect.	Incorrect minimum C&C pressure		
L2.5	If a 30 ft long 16" diameter drilled concrete pier has been designed for an allowable end-bearing pressure of 8,000 psf and an allowable skin friction of 1,500 psf, and the soils can cause a downdrag force of 500 psf, what if the downdrag force (kips) on the pier?			Incorrect downdrag force and no explanation of calculations	
L2.6	Determine the minimum required thickness in inches of a base plate with the following conditions. The column is a W12x87 (A992) with a factored axial column load of $P_u=850$ kips. The base plate dimensions are 22" in the direction parallel to the flanges and 24" in the direction parallel to the web. The column sits on a spread footing with a plan size of 10 feet by 10 feet. The specified minimum compressive strength of concrete is 3,000 psi.	General process is accurate, but inaccurate determination of when the base plate is on less than the full area of concrete support. Additionally, the thickness equation for the base plate is inaccurate and not based on critical base plate cantilever dimension.	Incorrect cantilever length for n, does not multiple bf by 0.8. Conservatively takes lambda as 1.0 but does not provide explanation for this. Assumes base plate is A992 steel, but a more appropriate material for a plate might be A572-Gr. 50.	Missing information on the section properties for a W12x87. Incorrect allowable bearing pressure equation. Missing $\lambda n'$ for critical base plate cantilever dimension. Does not solve for thickness or provide equation for thickness of the base plate.	Potentially unuseable format
L2.7	A steel beam is a part of a floor structure in an existing building and requires a continuous plate be added to the underside of the bottom flange of a W18x50 beam (A992). The continuous plate is 1/2" thick (A36) and 6" wide. The capacity of a 3/16" fillet weld is 2.78 kips/in and the maximum design shear in the beam is 40 kips. Determine the minimum adequate welding pattern to connect the reinforcing plate to the steel beam.	Does not accurately solve for shear flow and misunderstands question.	Correctly identifies the horizontal shear flow needs to be calculated, but then ignores it and just uses shear force.	Does not accurately solve for shear flow and misunderstands question.	Incorrect answer and unuseable format.
L2.8	A 16-in. x 16-in. reinforced concrete column bears on a concrete footing. The concrete is normal-weight concrete with a concrete strength of $f'_c=4,000$ psi and a yield stress of steel reinforcement of $f_y=60$ ksi. What is the minimum embedment length in inches to develop a #11 compression bar?	Correct answer, but does not check ACI 318 Eqn. 25.4.9.2(b).			Incorrect answer and equations are too difficult to track where it went wrong due to formatting issues.

LEVEL 2

Rationale for Mixed Accuracy and Inaccurate Responses					
	Question	Rationale for Mixed Accuracy and Inaccurate Responses			
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 1.5 Flash	Copilot
L2.9	An 8x8 wood post supports a dead load and a live load. The wood is a Douglas-Fir Larch, Grade No. 1. Determine the adjusted compression design value (psi) parallel with the grain using the following assumptions: $F_c E / F^* c = 0.957$. No lateral support is provided between supports. Column is pinned at the top and bottom, CD, CM, Ct, CF, and Ci all equal 1.0, $\lambda = 0.8$. Lumber is visually graded.	Incorrect C_p equation for column stability factor.	Incorrect F_c value. Did not pull value from correct Table 4D from NDS supplement. used Table 4A. Remaining information provided is correct.	Incorrect F_c value. Did not pull value from correct Table 4D from NDS supplement, used Table 4A. Incorrect equation for $F_c E$ and $F_c E$ value not needed to solve problem. Does not solve for C_p , column stability factor, which is required for the problem.	Does not provide an F_c value. Does not solve for C_p , column stability factor, which is required for the problem. Finds answer as a pressure. which suggests LRFD design methodology but does not use a Φ factor or K_r for the format conversation factor.
L2.10	Determine the adjusted design moment in in.-lb. for roof rafters that support a roof dead load and snow load. The rafters are 2x10 Hem-Fir No. 2 spaced at 24" o.c. Use the 2018 NDS and assume that bending controls the design. $C_m = C_t = C_{fu} = C_i$ all equal 1.0, $CL = 0.986$ and $\lambda = 0.8$.	Incorrect equation for F_b . Need to use C_D , C_F , and C_r . Remaining information provided is correct.	Incorrect section modulus. Incorrect C_F . Need C_D and C_r in equation. Incorrectly used λ in the equation. This is only needed for LRFD and it was specifically asked for ASD.	Made up loads for dead load and snow load which were incorrectly used in the solution. Incorrect process to solve for section modulus. Missing C_r and C_F in F_b equation. Incorrect equation to solve for Adjusted Design Moment.	Incorrect answer and equations are too difficult to track where it went wrong due to formatting issues.
L2.11	Compare the advantages and disadvantages of using steel versus concrete as primary structural materials.				
L2.12	Discuss the concept of structural ductility and how it affects the performance of a building during an earthquake.				
L2.13	Explain the differences between serviceability and ultimate limit states in structural design.				
L2.14	How do you determine the governing load combination for a structural member?	Discusses identifying the load combination that governs strength limit states and serviceability limit states, but never defines what they are and does not list ASD load combinations resulting in an incomplete response	Incorrect Load Combinations presented = $1.4D + 1.6L$ and $1.2D + 1.6W + 1.0L$		Potentially unusable format
L2.15	What are the primary considerations in the design of a deep foundation system?				No mention of actual load combinations and overstrength discussion does not address the specific question.
L2.16	Discuss the benefits and drawbacks of using a moment-resisting frame system in a high-rise building.	Discusses flexibility of MRFs as a benefit but doesn't provide a good description of the meaning and if it is referring to the drift. If so, it does not provide information on serviceability or drift issues for an incomplete response. Discusses reduced foundation size, but does not compare it to other lateral force resisting systems for an incomplete response.		States that moment resisting frames exhibit lower lateral drift than other structural systems which is inaccurate. Also lists reduced drift as a benefit, but large deformations as a drawback for conflicting and inaccurate information. Disagree with statement: 2. Wind Resistance: MRFs provide stability against strong winds, preventing excessive sway and ensuring the building can withstand wind-induced forces	

LEVEL 2

Rationale for Mixed Accuracy and Inaccurate Responses					
	Question	Copilot			
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 1.5 Flash	SEGPT
L2.17	Explain the process of determining wind loads on a structure based on the applicable building code.	Presents the velocity pressure as the design wind pressure. This equation also does not include the ground elevation factor, although this would be a conservative assumption since K_e is permitted to equal 1 in all cases.	Needs to provide more information to be useful.	Accurate, but velocity pressure equation does not include the ground elevation factor, although this would be a conservative assumption since K_e is permitted to equal 1 in all cases.	Needs to provide more information to be useful.
L2.18	Discuss the role of slenderness ratio in the stability analysis of a column.			Accurate, but states that to keep the slenderness ratio within acceptable limits to prevent buckling you can choose a material with a higher modulus of elasticity. This is true, but misleading. If you are using steel in your design it is unlikely that you would switch to another material due to buckling concerns since there are more practical solutions to consider.	Only discusses slenderness ratio being used to determine if P-big delta effects need to be considered. Doesn't discuss buckling of columns.
L2.19	Explain the differences between elastic and inelastic structural behavior, and their implications for design.				
L2.20	Compare and contrast the performance of various lateral force-resisting systems, such as shear walls, braced frames, and moment-resisting frames.				

LEVEL 3

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L3.1	At what value of torsion in a reinforced concrete beam must additional torsion reinforcing be provided based on the current ACI 318 code?	Mixed Accuracy	Inaccurate	Mixed Accuracy	Inaccurate	Inaccurate
L3.2	Describe the process for analyzing a building structure to determine if plan irregularities exist per the current ASCE code when considering seismic forces.	Accurate	Accurate	Mixed Accuracy	Mixed Accuracy	Mixed Accuracy
L3.3	What are the two compliance paths for structural masonry mortar quality assurance?	Inaccurate	Accurate	Inaccurate	Inaccurate	Accurate
L3.4	Are masonry prism tests per ASTM C1314 required if the unit strength method per TMS 402/602-16 is used?	Accurate	Accurate	Accurate	Accurate	Accurate
L3.5	A 2L6x3 1/2x1/2 tension member is attached to a steel wide flange column flange (at a 2(H):1(V) angle from the horizontal) using (8)-7/8" diameter A325 bolts in a bearing type condition. The threads are excluded from the shear plane. Is this a sufficient number of bolts to resist an applied unfactored axial tension of PDL=80 kips and PLL=100 kips? Use the LRFD specification.	Inaccurate	Inaccurate	Inaccurate	Inaccurate	Mixed Accuracy
L3.6	There is a one-story building with a width of 20'-0" in the north-south direction and a length of 56'-0" in the east-west direction. The height of the building is 20'-0" and the roof diaphragm is flexible. If the total wind load along the height of the wall is 21 psf, what is the chord force in kips in the north wall due to the wind forces in the north-south direction?	Inaccurate	Inaccurate	Inaccurate	Inaccurate	Inaccurate

LEVEL 3

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L3.7	An ordinary concentric braced frame is part of a detention facility and located in seismic design category C. The height from ground level to the first level is 20'-0", the height from the first to the second level is 15'-0". The height from the second to third level is 15'-0" and the height from the third level to the fourth level is 20'-0". What is the allowable seismic story drift in inches for the first level?	Mixed Accuracy	Mixed Accuracy	Inaccurate	Mixed Accuracy	Mixed Accuracy
L3.8	You are trying to determine the number of 10d common nails required on each side of the splice to connect two 2x Spruce Pine-Fir top plates together for a diaphragm chord force of 8,000 lbs. Use 2018 IBC, ASCE 7-16, and 2018 NDS. Assume $CM=Ct=Cg=CA=Ceg=Cdi=Ctn=1.0$.	Inaccurate	Inaccurate	Inaccurate	Inaccurate	Inaccurate
L3.9	A square HSS brace is being slotted to fit on a gusset plate. What is typically industry standard slot dimension to allow for fit-up tolerance? Also, if this isn't in a seismic zone, is there any reinforcing of this section required due to the reduced area of the brace?	Accurate	Accurate	Mixed Accuracy	Inaccurate	Mixed Accuracy

LEVEL 3

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L3.10	A straight pretensioned concrete member 40 ft. long, with a cross section of 15 in. by 15 in., is concentrically prestressed with 1.2 sq. in. of steel wires which are anchored to the bulkheads with a stress of 150,000 psi. If $E_c=4,800,000$ psi and $E_s=29,000,000$ psi compute the loss of prestress due to the elastic shortening of concrete at the transfer of prestress. Use elastic analysis with transformed section to estimate the change in steel stress at transfer.	Mixed Accuracy	Accurate	Mixed Accuracy	Accurate	Inaccurate
L3.11	Analyze the benefits and limitations of using performance-based seismic design over prescriptive code-based methods.	Accurate	Mixed Accuracy	Accurate	Accurate	Mixed Accuracy
L3.12	Discuss the challenges and potential solutions for designing long-span roof structures.	Accurate	Accurate	Accurate	Accurate	Accurate
L3.13	How do you assess the reliability of a finite element analysis model and validate its results?	Accurate	Accurate	Accurate	Accurate	Mixed Accuracy
L3.14	Evaluate the role of advanced construction techniques, such as 3D printing and modular construction, in shaping the future of structural engineering.	Mixed Accuracy	Mixed Accuracy	Mixed Accuracy	Mixed Accuracy	Accurate
L3.15	Explain the principles behind progressive collapse analysis and how it impacts structural design.	Accurate	Accurate	Accurate	Accurate	Accurate
L3.16	Discuss the complexities involved in designing and constructing tall buildings with mixed-use occupancy.	Accurate	Accurate	Accurate	Mixed Accuracy	Accurate
L3.17	How do you optimize the design of a steel connection to balance strength, stiffness, and constructability requirements?	Accurate	Accurate	Accurate	Mixed Accuracy	Mixed Accuracy

LEVEL 3

	Question	Classification of Response Accuracy				
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 2.0 Flash	Copilot	SE GPT
L3.18	Compare and contrast the seismic behavior and design approaches for reinforced concrete and steel structures.	Accurate	Accurate	Accurate	Accurate	Accurate
L3.19	Analyze the factors that influence the choice of a suitable structural system for a complex, long-span bridge.	Accurate	Accurate	Accurate	Accurate	Accurate
L3.20	Describe the implications of incorporating sustainable design practices, such as material efficiency and embodied carbon reduction, in structural engineering projects.	Accurate	Accurate	Accurate	Accurate	Accurate

LEVEL 3

	Question	Rationale for Mixed Accuracy and Inaccurate Responses			
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 1.5 Flash	Copilot
L3.1	At what value of torsion in a reinforced concrete beam must additional torsion reinforcing be provided based on the current ACI 318 code?	Not enough information for a Level 3. The answer is correct, but does not discuss threshold torsion or equations.	Discusses threshold torsional moment, but has incorrect equation.	Not enough information for a Level 3. Response does not include equations.	Discusses threshold torsional moment, but has incorrect equation.
L3.2	Describe the process for analyzing a building structure to determine if plan irregularities exist per the current ASCE code when considering seismic forces.			Not enough information for a Level 3. Generally correct, but lacking depth in explanation.	Not enough information for a Level 3. Generally correct, but lacking depth in explanation.
L3.3	What are the two compliance paths for structural masonry mortar quality assurance?	Does not discuss ASTM C270 and Property Specification Requirements and Proportion Specification Requirements.		Does not discuss ASTM C270 and Property Specification Requirements and Proportion Specification Requirements.	Does not discuss ASTM C270 and Property Specification Requirements and Proportion Specification Requirements.
L3.4	Are masonry prism tests per ASTM C1314 required if the unit strength method per TMS 402/602-16 is used?				
L3.5	A 2L6x3 1/2x1/2 tension member is attached to a steel wide flange column flange (at a 2(H):1(V) angle from the horizontal) using (8)-7/8" diameter A325 bolts in a bearing type condition. The threads are excluded from the shear plane. Is this a sufficient number of bolts to resist an applied unfactored axial tension of PDL=80 kips and PLL=100 kips? Use the LRFD specification.	Uses Fv in lieu of Fnv and uses wrong value for A325-X bolts. Equation for shear capacity is incorrect. Incorrect area of bolt. Solves for bolts for dead load and then bolts for live load and adds them together which is not common practice. Does not use LRFD as specified in prompt. Does not check all applicable limit states.	Uses Fv in lieu of Fnc and uses wrong value for A325-X bolts. Resolves force into components, but the force in this case would be parallel to the longitudinal axis of the member. Does not check all applicable limit states.	Uses Fv in lieu of Fnv and uses wrong value for A325-X bolts. Also specifies incorrect table from specification. States 10 bolts are required for shear and then states 8 bolts are sufficient to resist shear.	Incorrect Fnv value used. Does not check all applicable limit states.
L3.6	There is a one-story building with a width of 20'-0" in the north-south direction and a length of 56'-0" in the east-west direction. The height of the building is 20'-0" and the roof diaphragm is flexible. If the total wind load along the height of the wall is 21 psf, what is the chord force in kips in the north wall due to the wind forces in the north-south direction?	Does not divide wall height by 1/2. Uses wrong length is moment equation. Uses wrong depth is chord force equation.	Does not accurately solve for chord force.	Does not calculate wind pressure at roof diaphragm correctly. Does not calculate moment of diaphragm correctly.	Does not calculate wind pressure at roof diaphragm correctly. Does not solve for the moment at the diaphragm.
L3.7	An ordinary concentric braced frame is part of a detention facility and located in seismic design category C. The height from ground level to the first level is 20'-0", the height from the first to the second level is 15'-0", the height from the second to third level is 15'-0", and the height from the third level to the fourth level is 20'-0". What is the allowable seismic story drift in inches for the first level?	Mostly correct, except it divides the deflection by the importance factor which is incorrect.	Mostly correct, except states detention facilities are Risk Category IV. Detention facilities are classified as Group I-3 and therefore Risk Category III.	Provides too much meaningless information, only steps 7 and 8 are applicable to the question. Uses Risk Category IV. Detention facilities are classified as Group I-3 and therefore Risk Category III.	Mostly correct, except states detention facilities are Risk Category IV. Detention facilities are classified as Group I-3 and therefore Risk Category III.
L3.8	You are trying to determine the number of 10d common nails required on each side of the splice to connect two 2x Spruce Pine-Fir top plates together for a diaphragm chord force of 8,000 lbs. Use 2018 IBC, ASCE 7-16, and 2018 NDS. Assume CM=Ct=Cg=CA=Ceg=Cdi=Cm=1.0.	Incorrect Z used. Did not use C _p to solve for Z'. Incorrectly divided chord force by two for each side of splice.	Incorrect Z used. Did not use CD to solve for Z'.	Not enough information for a Level 3. Response does not include equations and no answer provided.	Incorrect Z used. Did not use CD to solve for Z'. Incorrectly divides number of fasteners for each side.

Unusable format. Incoherent and incomplete equations provided.

Unusable format.

Incorrect Z used. Did not use CD to solve for Z'. Unusable format.

LEVEL 3

	Question	Rationale for Mixed Accuracy and Inaccurate Responses				SE GPT
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 1.5 Flash	Copilot	
L3.9	A square HSS brace is being slotted to fit on a gusset plate. What is typically industry standard slot dimension to allow for fit-up tolerance? Also, if this isn't in a seismic zone, is there any reinforcing of this section required due to the reduced area of the brace?			The statement "in non-seismic, reinforcement is often required" can be misleading. For most structures, there is typically no need to reinforce the brace section in non-seismic regions. Wording of sometimes might be a better suggestion.	Only provides fit-up tolerance dimensions for a 1/2" plate. Although the tolerance is correct, it does not provide enough information for the gusset plate thickness varies.	Answer is slightly misleading in that it discusses the tolerances for bolts. While it does provide an example of using 1/16", it relates it to a bolt diameter which is not applicable to the question being asked.
L3.10	A straight pretensioned concrete member 40 ft. long, with a cross section of 15 in. by 15 in., is concentrically prestressed with 1.2 sq. in. of steel wires which are anchored to the bulkheads with a stress of 150,000 psi. If $E_c=4,800,000$ psi and $E_s=29,000,000$ psi compute the loss of prestress due to the elastic shortening of concrete at the transfer of prestress. Use elastic analysis with transformed section to estimate the change in steel stress at transfer.	Does not subtract the steel area from the concrete area when determining the transformed concrete area.		Does not subtract the steel area from the concrete area when determining the transformed concrete area.		Incorrect answer and unusable format makes it too difficult to discover where the error in the calculations occurred.
L3.11	Analyze the benefits and limitations of using performance-based seismic design over prescriptive code-based methods.		Only discusses PBSd with no mention of comparison to prescriptive code-based methods. Additionally, although it provided a lengthy response, there would need to be more follow-up prompts due to the limited explanations for certain topics. For example, "enables stakeholders to make informed decisions about acceptable risk levels" is a vague benefit of performance-based seismic design and isn't a distinct benefit over prescriptive code based methods.			Only discusses PBSd with no mention of comparison to prescriptive code-based methods.
L3.12	Discuss the challenges and potential solutions for designing long-span roof structures.					
L3.13	How do you assess the reliability of a finite element analysis model and validate its results?					Appeared to be pulling information from an article specifically related to masonry and discussing cracks in masonry instead of providing a more general response to the prompt leading to a misleading and inaccurate response.
L3.14	Evaluate the role of advanced construction techniques, such as 3D printing and modular construction, in shaping the future of structural engineering.	Description on 3D printing is too vague to be considered accurate. No discussion on certain materials or applications are indicated.	Description on 3D printing is too vague to be considered accurate. No discussion on certain materials or applications are indicated.	Description on 3D printing is too vague to be considered accurate. No discussion on certain materials or applications are indicated.	Description on 3D printing is too vague to be considered accurate. No discussion on certain materials or applications are indicated.	
L3.15	Explain the principles behind progressive collapse analysis and how it impacts structural design.					

LEVEL 3

	Question	Rationale for Mixed Accuracy and Inaccurate Responses			
		ChatGPT-4	Claude 3.5 Sonnet	Gemini 1.5 Flash	Copilot
L3.16	Discuss the complexities involved in designing and constructing tall buildings with mixed-use occupancy.				Too vague of a description regarding structure to be considered accurate, but did not focus on the structural aspects and started to discuss lease agreements and investment and revenue streams.
L3.17	How do you optimize the design of a steel connection to balance strength, stiffness, and constructability requirements?				Limited responses with both lacking depth to the response.
L3.18	Compare and contrast the seismic behavior and design approaches for reinforced concrete and steel structures.				
L3.19	Analyze the factors that influence the choice of a suitable structural system for a complex, long-span bridge.				
L3.20	Describe the implications of incorporating sustainable design practices, such as material efficiency and embodied carbon reduction, in structural engineering projects.				