Group: Reflections on Vampire Diaries(Shanon Marietta Badger
sbad@itu.dk, Benjamin Storm Larsen bsla@itu.dk, Yuliia Storm Larsen yuls@itu.dk)

Introduction

Currently generative artificial intelligence is at the peak of its popularity, especially in academia. However, the ability to produce factually accurate and well-supported content remains uncertain. As the use of generative AI becomes part of everyday lives, the question arises: “Can we actually deem the generated results credible?”. This study evaluates the reliability of Perplexity Pro, ChatGPT, and Claude AI in generating a multiple-choice exam on linear regression.

Linear regression is a foundational topic in predictive modeling and statistical learning, requiring mathematical precision and credible sourcing. This report evaluates the accuracy of AI-generated exam questions, the credibility and diversity of sources cited, and potential issues such as misinformation, citation biases, and unsupported claims, providing insights into AI's practical reliability in academic settings. The exact prompt provided to all AI’s was: *“Please create a multiple choice mock exam for first year bachelor students in Machine Learning. The subject should be about linear regression and have forty questions. Include the correct answers and the sources for each answer”.* However, ChatGPT did not provide all forty questions, so an additional prompt was asked to produce the remainder: *“Please provide the remaining 20 questions”.* The AI artifacts can be found at our github: https://github.com/jsolomkk0/Reflections-on-Data-Science-Spring-2025---Reflections-on-Vampire-Diaries.

Methodology

To evaluate the reliability of AI-generated educational content, we conducted three key analyses: source credibility analysis, source use analysis, and claim analysis. Each analysis assessed how AI models sourced and supported exam content.

A credibility rating system was developed to assess the reliability of cited sources. Each reference was classified based on its origin:

- Academic publications (1.0) – Peer-reviewed journal articles, textbooks, and research papers from universities or recognized institutions.
- Industry sources (0.8) – Technical documentation, company reports, and authoritative publications from professional organizations or research institutions.
- Public educational content (0.6) – Open-access, peer-editable platforms such as Wikipedia and community-driven educational sites.
- Social media and informal discussions (0.5) – Reddit threads, YouTube videos, and other non-expert sources.

It is important to note that the assessment conducted by each group member will slightly differ from the defined rating system, depending on the variety and classification of each source.
Assigning numerical values to sources allowed us to calculate a weighted credibility score, showing whether the AI relied primarily on highly credible academic materials or leaned toward less formal, potentially unreliable sources.

Beyond credibility, we analyzed citation distribution, recording the number of unique sources, citation frequency, and patterns in selection to determine whether AI models referenced a wide range of expert perspectives or recycled a limited pool of sources.

For claim verification, we assessed explicit statements against cited sources, determining accuracy and allowing multiple references to validate a claim when applicable. This revealed how often AI models produced unsupported claims, misattributed facts, or inconsistencies, highlighting their strengths and weaknesses in sourcing.

The Perplexity model used version Pro Research, ChatGPT used version GPT-4o, and Claude was version 3.7 Sonnet.

Perplexity Pro
Source Credibility Analysis

Perplexity Pro's sources had an average credibility rating of 0.774 across 72 references, indicating moderate reliability but a reliance on lower-ranked materials. Industry-standard sources (0.8) were overweighted, favoring professional consensus over academic rigor. While this improves practical relevance, it may weaken theoretical depth, particularly where industry practice diverges from foundational research. As Figure A illustrates, citation distribution reinforces this preference for industry sources over purely academic references.

Source Use Analysis

Perplexity Pro's citation distribution showed significant inconsistencies. Although 72 sources were listed, only 62 citations were used, with just five unique sources accounting for all references (see Figure B). This resulted in an inflated source count, highly selective citation practices, and limited source diversity, as industry standards and public educational platforms were prioritized over academic sources. The inclusion of Berkley University may suggest scholarly depth, yet its limited citation frequency indicates minimal reliance on peer-reviewed material. While the credibility score increased to 0.84, this improvement is overshadowed by citation repetition and overdependence on a small subset of sources, reducing the depth and reliability of the exam content. A higher credibility score doesn't offset limited source diversity, which may cause gaps in accuracy.

Claim Analysis

A review of Perplexity Pro's citations found that 11 out of 62 could not be manually confirmed, resulting in a high inaccuracy rate of 17.74%. Many references either lacked the claimed information, were misattributed, or exhibited factual inconsistencies, with some appearing to be entirely fabricated. Despite these citation issues, all quiz answers were correct, suggesting that the model may have drawn from additional sources without citing them. This raises concerns about its transparency in sourcing, as well as the reliability of its citation practices versus its answer accuracy. With nearly one in five citations failing validation, these inconsistencies significantly undermine confidence in the model's ability to generate fact-based educational content.

Additional Issues/Findings

Perplexity Pro creates an illusion of exhaustive research, but closer inspection reveals inflated sources, possibly uncited sources and unverifiable citations. While the Pro version might suggest improved source reliability, credibility appears unchanged, meaning users may not receive better-vetted references. The extensive reference list lends initial credibility, yet citation repetition and unverifiable claims undermine its depth. Users may falsely assume well-researched content due to a premium service and broad source list, despite gaps in accuracy and diversity.

ChatGPT

Source Credibility Analysis

Disappointingly, the ChatGPT only referenced 7 sources for its 40 questions, opting to reuse them across the questions. The most notable difference from the other AI's is the fact that it had to be prompted twice to generate the full sheet of questions. The sources has an average credibility score of 0.76, with the lowest rated being a user generated quizlet and the highest stemming from Dalhousie University in Canada. The AI was expected to provide sources for the knowledge that it displayed, instead it decided to provide a mix of sources where all besides one references pre-made quizzes on the subject and the other being an actual article. Figure C provides a chart of the percentage distribution of the ratings.

Source Use Analysis

The most used source was Investopedia with 34% occurrences, which is an article on multiple linear regression and is the only source which is not a pre-made quiz. This source has the most text and explanations, but not direct questions for the AI to copy to its own text. The most used quiz is the Geeks For Geeks quiz which is used in 8 questions. The rest of the quizzes had between 1 and 5 occurrences. However, they contain more specific text for the AI to use, so are more likely to be directly correlated to the questions produced.

Claim Analysis

The generated text had 65% of its sources which did contain information about the written question. It appears that the AI hallucinated by just adding a random previously used source. For most of the sources the contents do not remotely contain similar information to that in the question. The rest of the sources matched the questionnaire with all claims besides one also being valid and correct in the source. The most reliable source was the Geeks For Geeks source with all 8 occurrences matching the question and only 1 was incorrect. As expected the AI performed best in referencing the sources that were already made as a quiz, as it could probably "copy" most of it to its own questions.

By validating the questions provided and the answer associated, the AI produced an impressive 38 correctly answered questions out of 40. This clearly illustrates that the AI is not trained or built to provide the sources for what it writes, but rather provide text based on its enormous library of text. However, given how the generative AI works this is to be expected.

Additional Issues/Findings

Interestingly the ChatGPT decided to include 5 math questions which were simple calculations related to the topic, it did not attribute any claims to these questions. Instead they were all validated to be correctly calculated.

As a result of having to prompt ChatGPT a second time, it somewhat forgot the questions it had created previously. This resulted in 4 duplicate questions, which is very interesting, especially as memory was turned on and the question was asked within the same chat/scope.

Claude AI

Claude AI is a bit "younger" AI model in the expanding landscape of large language models, developing its capabilities through ongoing training and refinement. As a newer entrant compared to some established systems, Claude aims to balance its developing experience with strong performance across various tasks. Despite its relative youth, Claude brings a fresh perspective to AI assistance.

Source Credibility Analysis

Based on the provided data, a critical assessment of how Claude 3.7 Sonnet created the mock exam on linear regression as an answer for aforementioned prompts can be offered:
The mock exam demonstrates excellent source credibility management, which, surprisingly, used information exclusively from high-quality academic materials. All 40 questions were sourced from materials rated as either "Very High" (55.6%) or "High" (33.3%) or "Medium" (11.1%) credibility, with no content from low-credibility sources. This shows Claude carefully prioritized authoritative textbooks, academic papers, and university courses like "An Introduction to Statistical Learning" and MIT OpenCourseWare materials when generating questions. Appendix D

Figure D : Credibility Rating for Claude Sources

Source Use Analysis

The source use analysis reveals balanced utilization across the resource library. The most-cited source (Introduction to Statistical Learning) only accounted for 15% of the exam content, indicating Claude avoided over-reliance on any single resource. Additionally, Claude demonstrated strategic source selection, using 11 different high-credibility sources to create a comprehensive examination covering the full spectrum of linear regression topics.

Figure E : Source Usability Analysis for Claude

Claim Analysis

The claim analysis reveals some discrepancies in how Claude handled source attribution. While the original data shows all claims were found in their cited sources (with "Y" in the "Claim Found in Source?" column), the additional data that Claude Provided as “Additional Academic Sources” - suggests there were citation issues. According to my analysis, only 67% of sources were cited correctly, with 33% having some form of miscitation. This suggests Claude may have struggled with proper attribution for some content, potentially drawing from sources without explicitly acknowledging them or incorrectly attributing content across sources. This highlights an area where Claude's approach to information integrity could be improved to ensure each question and answer is properly supported by its cited reference material.

Figure F : Claude Citation Credibility

Additional Issues/Findings

One limitation of the exam is its difficulty level organization. While Claude created questions that are technically accurate, the exam lacks a thoughtful progression from simpler to more complex concepts that would better serve first-year students. Freshmen-year students typically benefit from starting with basic definitions and gradually working toward more challenging applications. A better approach would have been to structure the exam with earlier questions covering fundamental concepts like the basic linear regression equation and assumptions, before moving to more advanced topics like regularization techniques. This kind of structured approach helps students build confidence and reinforces learning in a more effective way than questions arranged without consideration for the learning journey.

Additionally, I would like to add the note regarding the bias in my assessment process, therefore the validation of my finding would be appropriate by my teammates. Based on personal background I may have evaluated the sources according to my preferences and cultural baselines that could be different from others.

Comparison of the AI

All of the generative AI models appear to favour medium-credibility sources when providing references. Among them, Perplexity stands out by offering a higher number of references, which aligns with its core system as a retrieval-augmented system. In contrast, Claude and ChatGPT tend to provide fewer citations and often reuses them across the questions. However, all systems seem to suffer from hallucination by sometimes referencing a claim which either does not relate to the question or is factually incorrect.

The most notable difference between the three answers is their approach to gathering sources. Perplexity and Claude’s answers are inline with the expectation that the AI will attempt to attach one or more sources to each question. However, in complete contrast,

ChatGPT chose to find pre-made quizzes online to use for its referencing, which is an interesting and unexpected approach.

The differences may take the root in the sources that the modes are trained on, as ChatGPT is “famous” for its failures and citing non-existing sources. The other reason for the major differences could be the usage of the so-called benefits of the different versions of the AI. For example, we used the Perplexity AI Pro version but both - ChatGPT and Clause were free versions. Following conditions could affect the differences in the outcomes which will also signify the consumer-related aspect of the results: pay more - get more accurate results.

Conclusion

In conclusion, using generative AI models for generating exams is not a reliable method. Throughout our evaluation, we observed that these systems frequently hallucinate facts, misattribute sources. This makes them unreliable and untrustworthy for direct use in an educational setting.

Using Artificial Intelligence for creating the learning materials for an “uncertain” group of people, where the model is unsure about the materials that target group covered or the intended learning outcomes is quite misleading in the baseline of the prompt. Therefore this gives room for the future research and following analysis about how these three models will behave themselves based on the true historical data that we can provide for the creation of the quizzes. Feeding the course materials, supplementary literature and the previous exams could create more stable results. This will allow “training” for the model on the real-case data that will be used to create a more stable and academically-friendly exam sheet.

Even though presented AI models interpret the same prompt differently, each and every of the language models affects the reliability of their generated responses. This reminds a bit of the human perception and the analysis that every human provides will have differences based on personality traits and cultural development.

While the models could inspire quiz questions, an academia representative should fact-check the content before use. Future evaluations could explore how the models handle more complex machine learning topics where pre-made quizzes are less available for the public.

Group: Reflections on Vampire Diaries(Shanon Marietta Badger sbad@itu.dk, Benjamin Storm Larsen bsla@itu.dk, Yuliia Storm Larsen yuls@itu.dk)

Appendix

For supplementary materials that were used in this research(Mock Exam samples from each AI large language model and the following analysis of the sources that was conducted by each group member individually and after validated collectively) you may use this link to the github repository or in the submission zip-file:
https://github.com/jsolomkk0/Reflections-on-Data-Science-Spring-2025---Reflections-on-Vampire-Diaries/blob/main

Figure A : Credibility Rating for Perplexity Sources

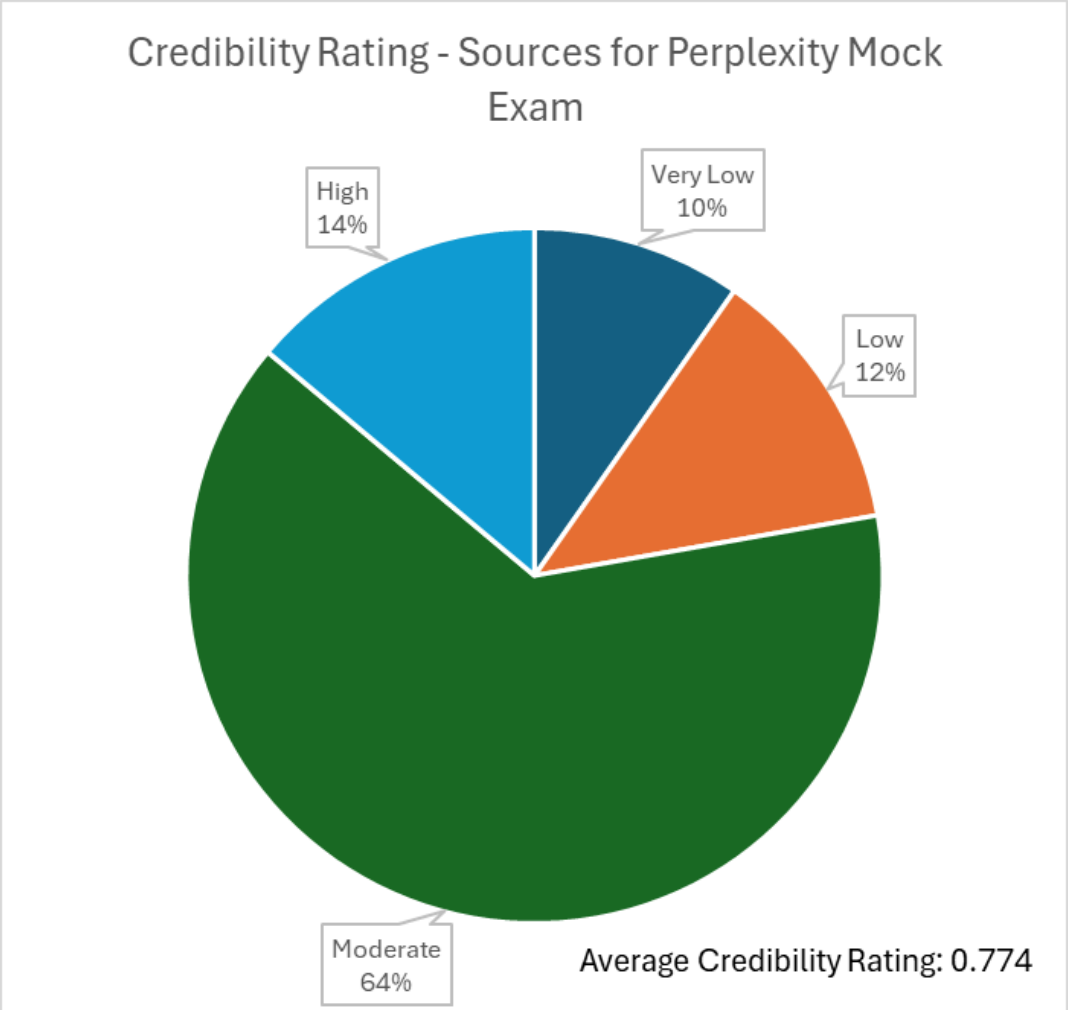

Figure B : Perplexity Actual Citations

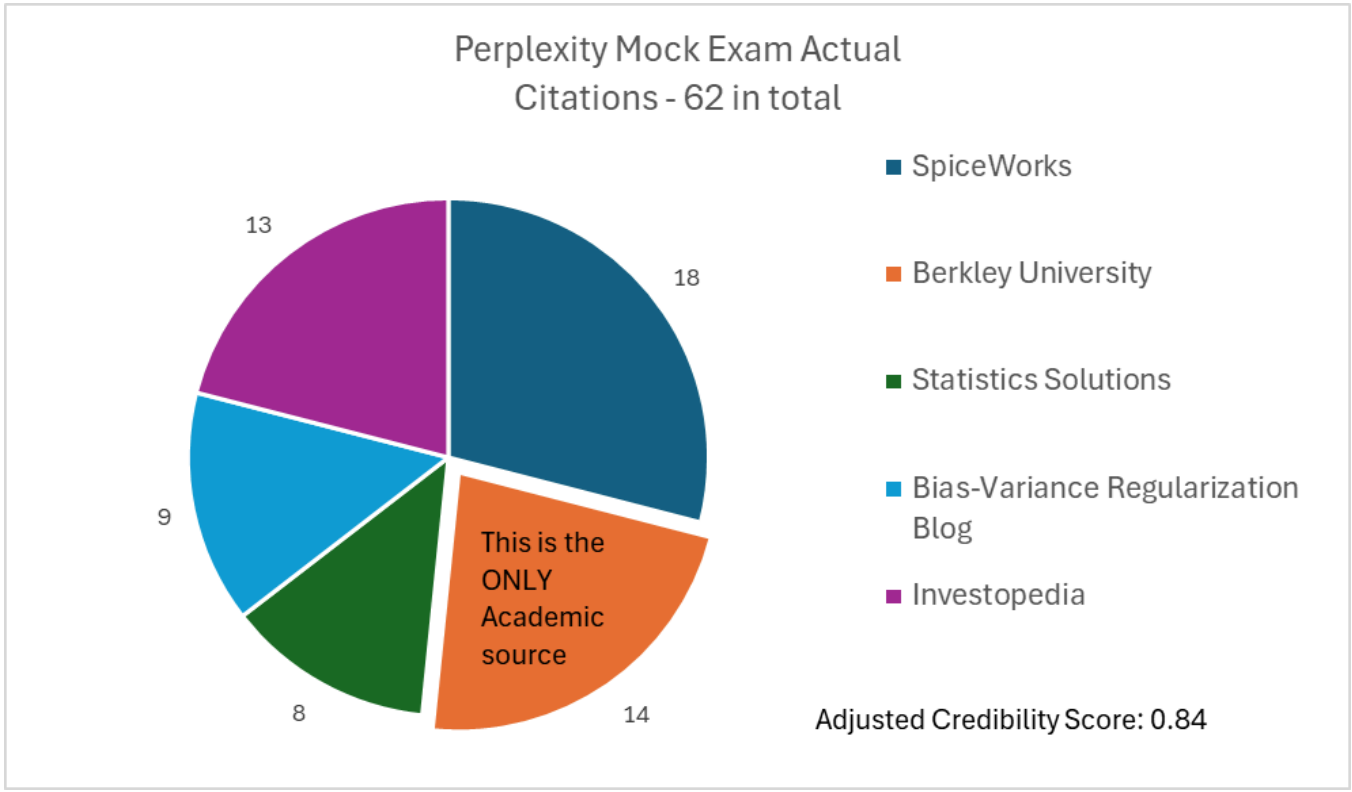

Figure C : Credibility Rating for ChatGPT Sources

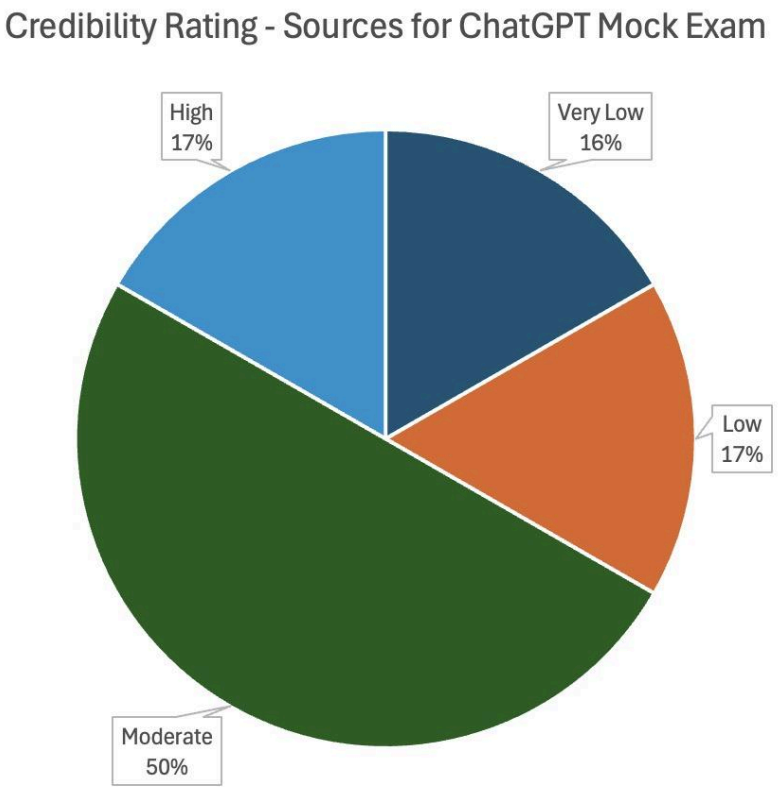

Figure D : Credibility Rating for Claude Sources

Group: Reflections on Vampire Diaries(Shanon Marietta Badger sbad@itu.dk, Benjamin Storm Larsen bsla@itu.dk, Yuliia Storm Larsen yuls@itu.dk)

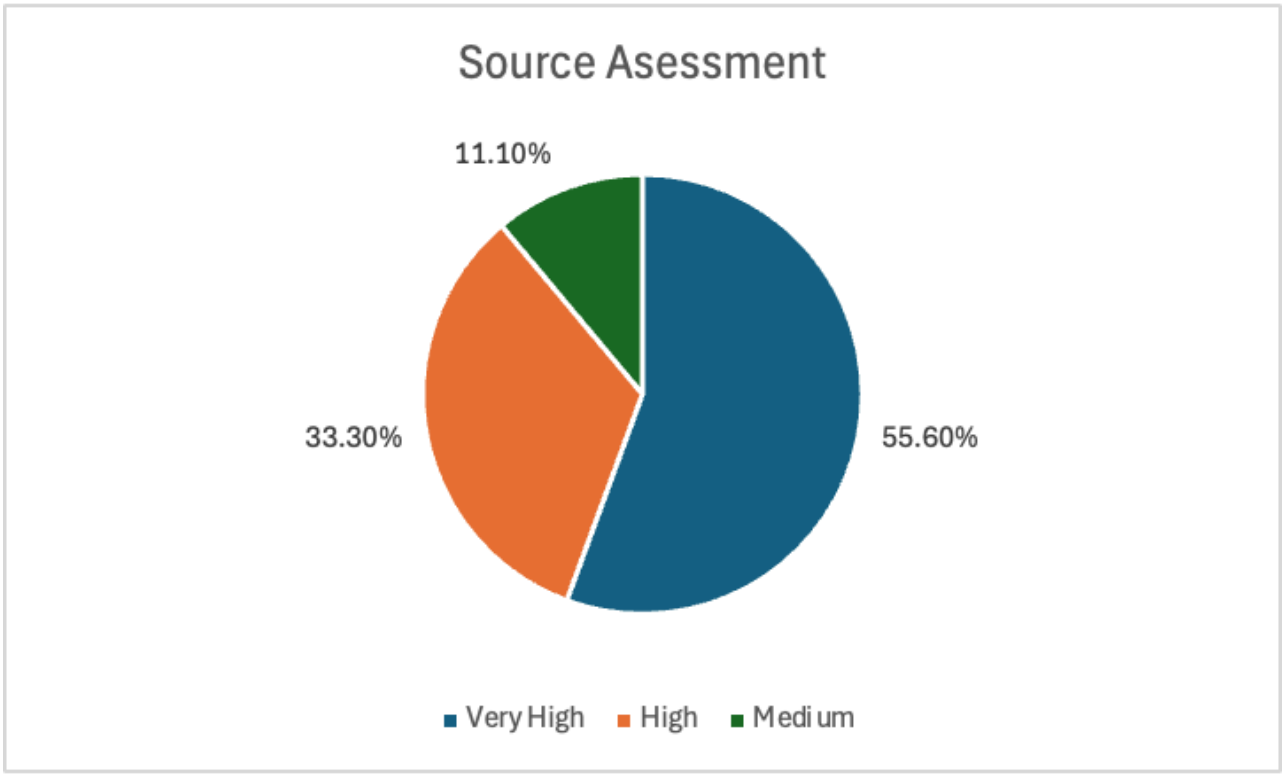

Figure E : Source Usability Analysis for Claude

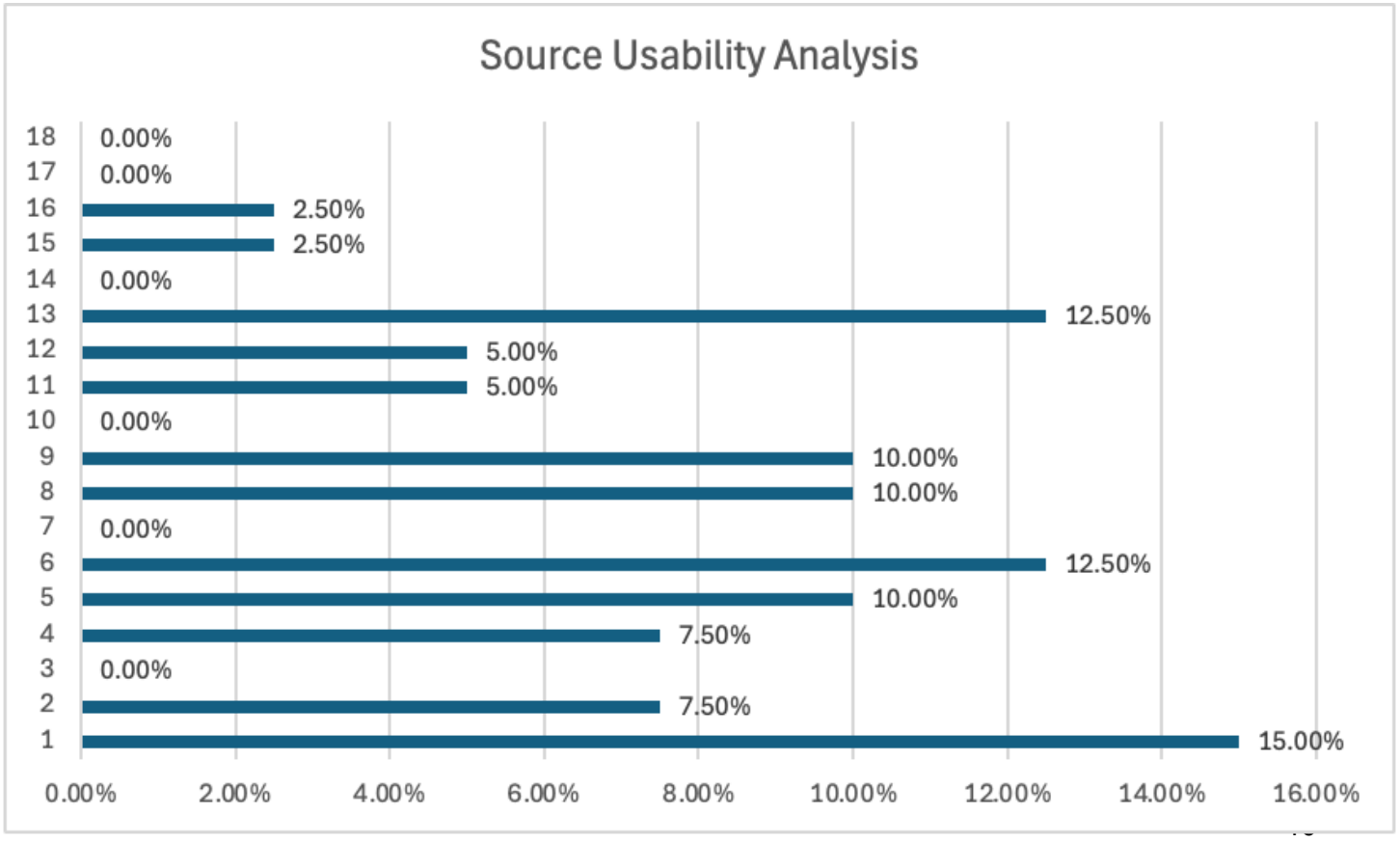

Figure F : Claude Citation Credibility