DLINE JOURNALS

Journal of Digital Information Management

ISSN: 0972-7272

JDIM 2025: 23 (3)

# Measuring Similarity, Credibility and Value of Information Content of Google and Generative AI Platforms

Preethi Pichappan
London Metropolitan University
Holloway Rd, London N7 8DB. UK

Pit Pichappan
Digital Information Research Labs
Chennai 600017. India
pichappan@dirf.org

## ABSTRACT

*This study examines the similarity, credibility, and value of information content generated by Google and various Generative AI (GAI) platforms, such as ChatGPT, Qwen, Perplexity, and Poe. Using the Jaccard Similarity Index, the study measures content overlap between Google search results and AI-generated responses, resulting in two clusters. The first one is the low correlation between the Google and AI outputs. Second is the consistency among AI-based models in terms of similarity. Thus, AI platforms reflect higher internal similarity, suggesting a high degree of homogeneity among them. The research also evaluates credibility, finding Google to have higher average credibility scores (0.82) compared to Perplexity (0.59) in reference analysis, with GAI platforms often lacking reliable citations and traceability. Content credibility scores, assessed using criteria such as source credibility, content accuracy, completeness, objectivity, and verifiability, indicate that Google excels the AI platforms. Google scored 6.54, whereas Perplexity, Qwen, ChatGPT, and Poe scored 4.72, 3.75, 2.81,* and *2.26, respectively, in user evaluation. In terms of credibility, Google surpasses generative AI platforms; while AI tools provide useful summaries, they lack the depth, evidence base, and scholarly context found in Google search results. The study concludes that while AI enhances communication efficiency, it raises concerns about reliability, knowledge foraging, and scientific integrity. Future research aims to analyze content volume and length as measures of value, emphasizing the need for transparency and human-centered AI development.*

**Subject Categories and Descriptors:** **[H.3.3 Information Search and Retrieval];** Information Filtering and Retrieval models **[H.3.1 Content Analysis and Indexing]; [H.1.1 Systems and Information Theory];** Value of Information

**General Terms:** LLM, Google Search, AI Platforms, Information Value, Content Similarity

**Keywords:** OpenAI, Generative AI models, Google Search, Content Credibility, Content Value, Content Similarity, Jaccard's Index

**Received:** 3 June 2025, Revised 29 July 2025, Accepted 3 August 225

**Review Metrics:** Review Scale: 0/6, Review Score: 4.95, Inter-reviewer consistency: 75.4%

**DOI:** https://doi.org/10.6025/jdim/2025/23/3/149-170

## 1. Introduction

AI-generated models are on the rise, offering numerous benefits to society, ranging from information generation to coding in a language. Generative models perform several tasks, such as answering user requests to reviewing content, and their responses are convincing to many users, particularly the younger generation. In many situations, evaluators struggle to determine whether the content is human-generated or AI-generated. However, their value remains a challenge in information extraction, as well as their reliability. Most AI-generated models are homogeneous in terms of content generation, derivation, and delivery. They have an inbuilt synthetic character and are prone to questions about reliability and accuracy.

Before we accept AI-generated content, more empirical research should be conducted to validate it. AI platforms summarise the content effectively, whereas interpreting, critically evaluating, and extracting the core concepts from literature is often ambiguous. LLMs have good grammar and structure but lack creativity, personality and insight. Creativity is not limited to internal thinking or cognition, but rather the blend of experience, observation, and external knowledge sensing. However, the most influential issue is cognition and perception. The mind synthesises issues, resulting in the birth of new ideas and contributions.

Large language models (LLMs) are the outcome of algorithmically driven content. [1] (Lindsey) They have the potential to transform work and learning, but their success depends on building trust through credibility, transparency, and explainability, especially in knowledge building. Achieving this requires combining technical improvements, ethical regulations, and open societal discussions to ensure AI remains a responsible and human-centred tool. [2] (Bozkurt)

## 2. Background

The AI tools perform several tasks, from writing and editing text, from emails to documents, helping with code and programming tasks, solving problems through step-by-step reasoning, and providing detailed answers and explanations. One potential application is the extraction of answers for queries, which is more widely used than for other purposes.

During the content extraction process, users tend to heavily rely on and use the output without considering the value or credibility of the generated content. Typically, these tools may provide inaccurate or deceptive information, and users can experience LLM hallucination, which convinces them to accept faulty or erroneous information that is misleading or not grounded in facts. Thus, these models fail to meet the requirements of academic research, where acceptance relies on accuracy and honesty. [3] (Javier Luna)

GAI lacks accountability (except for a few, such as Perplexity.ai), as they fail to acknowledge the original source of content, raising concerns about reliability and credibility. Wikipedia presents human-generated and

curated content, while generative AI creates content without human intervention and is known to generate erroneous output. The limitations, shortcomings, and erroneous output of OpenAI are documented in numerous studies, accompanied by a handful of illustrations. [4]. To what extent Open AI content is related to human-written content is a crucial question, many users are interested in knowing. [5].

## 3. Related Work/Credibility Issues

There remains an ongoing threat to the credibility of content, given that generative AI platforms such as ChatGPT have a tendency to create false or misleading information. [6, 7, 8, 9] Users employ various motivations and strategies to assess the authenticity of AI-generated content. [10] A comparison of credibility perceptions of human-generated and computer-generated content in different user interface (UI) settings was conducted through a survey by Martin. Unfortunately, survey participants attributed trustworthiness to AI-generated content more than to human-generated content, partly due to the impact of hallucination, where AI content is perceived as more transparent, engaging, and unambiguous and produced a positive attitude towards AI [11]. AI-generated content often makes human-generated content appear convincing and polished. [12]. When generative AI is used for user-generated content, the quality of the content tends to decrease, particularly for average users. [13] (Samsun Knight,) The perceived intelligence, transparency, empathy, and hallucination of knowledge affect trust in AI-generated content and, ultimately, the choice to adopt. [14].

Although generative AI can assist with writing summaries, the credibility and accuracy of AIGC are not guaranteed to be 100% [15, 16]. Credibility, reliability and value of retrieved content are the most critical issues when using AI-generated text. Scholars and experts do not validate AI-generated content, and there is a major issue about its dependence. AI content is proliferating without established standards or evaluation, and it is tempting users, particularly the younger generation, to use it indiscriminately.

AI content is prone to errors, misperceptions, and even hallucinatory behaviour persists within these applications [17,18] (Jang, Shen). AI-generated content often exhibits signal manipulation, semantic inconsistencies, logical incoherence, and psychological manipulation. [19]

Different topics arise from concerns about content quality, validation, and copyright when exploring generative AI platforms such as ChatGPT and Midjourney for marketing purposes [20]. GenAI models, for instance, can identify low-credibility information with accuracy; however, their explanation is linguistically motivated, without any understanding of the veracity.[21]

It highlights the evolution of credibility constructs, measurement inconsistencies, and challenges in adapting traditional models to the online environment. The authors propose moving toward a dual-credibility model that distinguishes only between sources and messages, rather than maintaining the traditional trinity of source, media, and message credibility. [22]. The characteristic features of Open AI and human-created content are compared by Biyang Guo et al. [23].

## 4. How to Operationalise the Similarity, Credibility, and Value of Information Content?

### 4.1 Similarity

The primary intention of this research is to investigate the relationship between Google-retrieved content and

GAI in terms of content, which can provide a signal to end users to judge the ability and dependability of these platforms. The ideal way to measure content similarity is through text matching and calculating similarity scores. We have utilised Jaccard's similarity measure to determine the relatedness between the Google and GAI platforms. The Jaccard similarity index is often used in science and technology as a means to express the similarity between two sets in numerical terms. [24] (Gonazalo Travieso). We subject the responses in both platforms to text analysis using Jaccard's similarity, which provides the percentage of commonality.

For two sets A and B, the Jaccard similarity *J (A, B)* is *e*xpressed as

$$J(A, B)=|A\cap B||A\cup B|$$

Where:

· $|A\cap B|$ is the number of elements common to both sets (intersection)

· $|A\cup B|$ is the total number of distinct elements in either set (union)

While measuring similarity, we identify outlier values in the selected prompts to determine if any values deviate from the other patterns. We further measured the mean similarity values of the tested prompts and pairs to inform our core decision.

The queries/prompts are randomly generated and fed into Google and four GAI platforms: Perplexity, ChatGPT, Qwen, and Poe. The ten queries that we used when fed into Google yielded top hits, which were subjected to analysis, and this top range spans from seven to ten.

## 4.2 Credibility and Value

Credibility is measured by Talaver [25] through the verification of facts in texts by identifying key statements, retrieving trusted evidence, and revising content to ensure both accuracy and educational value. The system uses advanced language tools to extract facts, analyse evidence from reliable sources, and label it as true, false, or unclear. Measuring credibility and value is a complex issue, and several issues arise as we begin this process. Can we integrate user perceptions of fairness with its technical definitions? How might it trade off with other important concerns? Can existing transparency methods be used for fair algorithms? How is core information derived when searching content or making prompts?

The extent to which users can rely on content generated by platforms can be objectively measured through the following parameters. Operationalising the term 'credibility' is somewhat challenging. Experts' opinions are the traditional way of soliciting participants' responses and arriving at decisions based on their input.

We employed three primary mechanisms to assess credibility, including *user perceptions, the value of the references used by Google versus Perplexity, and content credibility* as measured by the GAI platform and ChatGPT.

Google search and the Perplexity (the only GAI that we studied) list the references used explicitly to generate answers. We assigned weights to the references used in the Google and Perplexity hits on the GAI platform. The references used by them are classified as below.

Journals and Scientific Reports- 1.0 score
Scientific News/Science Magazines – 0.6 score
Blogs/News Bulletins- 0.4 score
Social media -0.2 score

How the GAI platforms rate their generated content is an interesting observation we adopted. We used ChatGPT to get the credibility scores as below. We employed five scales, such as *Source Credibility*, Reputability and transparency of sources cited, *Content Accuracy,* factual correctness, scientific integrity, *Completeness*, depth, breadth, and inclusion of key findings or mechanisms, *Objectivity,* neutral tone, avoidance of speculation or marketing language and *Verifiability,* availability and traceability of supporting references or data.

## 5. Workflow

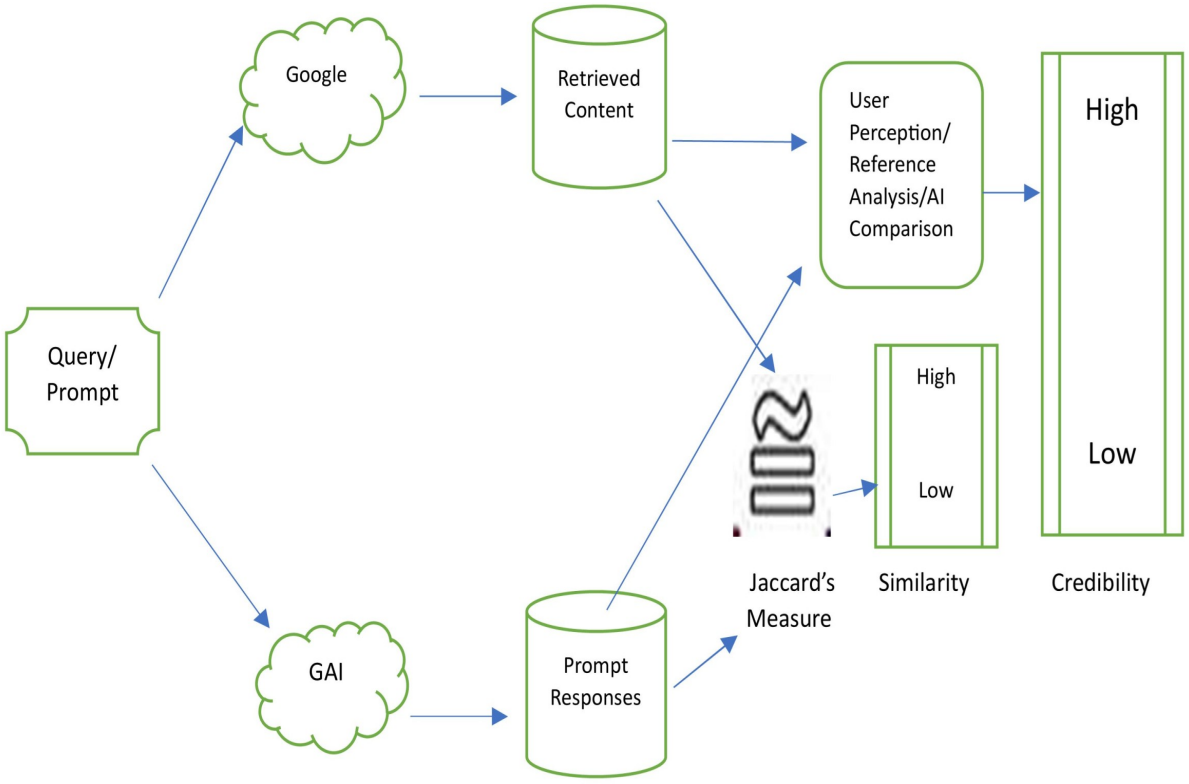



Figure 1. Workflow of the study

We identified ten queries/prompts, mainly focused on scientific concepts, and used them as basic inputs in both Google and AI platforms. The retrieved responses from Google search and AI platforms are stored as text files, resulting in voluminous content for Google searches and summarised output for AI platforms. As we ignore the volume of the information generated, this length has no impact on analysis. The content similarity is measured using Jaccard's measure, and credibility is calculated using users' perception, reference analysis (only for Google and Perplexity), and AI-generated comparison. The results are discussed in detail before we draw inferences.

# 6. Experimentation

## 6.1 Similarity

The primary goal of this work is to find and explore the relationship and credibility between Google search and Generative models. To start with, we feed ten randomly framed queries into Google and prompts in the Generative Platforms, ChatGPT, Qwen, Perplexity and Poe. The resulting hits in Google that appear at the top of the retrieval page are unified in a single file, which is typically lengthy. The same query, when fed into the generative models, resulted in a brief summary-type output that ranges from 300 to 1000 words. However, words do make sense here as we do not attach any significance to response length.

The content relatedness is assessed through the calculation of Jaccard's Similarity Index between the five platforms (one Google and four GAI). Each source file and each other file is matched to find content similarity. The Jaccard values are presented in Table 1. The similarity values are expressed as a percentage of the overlap between Google search-retrieved content and AI-generated content. A clear observation is that the content of the Google retrieval and Generative Platforms is not related, as the retrieval resulted in varying content. However, AI-generated content between four GAI shows considerable similarity.

| Q No. | Prompt/Query | G Vs Qwen | G Vs ChatGPT | G Vs Perplexity | G Vs Poe | ChatGPT Vs Qwen | Poe Vs Per | ChatGPT Vs Per | Qwen Vs Per | Poe Vs Qwen | Poe Vs ChatGPT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Q1 | Why does Science become less disruptive? | 41.60% | 36.80% | 36.50% | 29.70% | 63.70% | 48.60% | 53.40% | 56.50% | 59.70% | 68.60% |
| Q2 | Climate risks for younger generation | 46.00% | 46.30% | 46.00% | 41.90% | 65.60% | 59.50% | 58.70% | 57.60% | 57.90% | 63.20% |
| Q3 | Obesity drugs slash some cancer risks | 30.70% | 36.10% | 31.40% | 30.90% | 59.20% | 58.70% | 55.10% | 57.20% | 64.30% | 57.90% |
| Q4 | Side effects of weight loss drugs | 37.50% | 33.40% | 34.10% | 29.60% | 65.80% | 59.70% | 54.80% | 55.20% | 53.80% | 54.70% |
| Q5 | Whether a microbe with a tiny genome may evolve into a virus | 25.00% | 23.10% | 20.50% | 23.60% | 45.10% | 59.90% | 50.80% | 55.10% | 62.30% | 60.00% |
| Q6 | How tunnelling photons challenge the interpretation of quantum mechanics? | 44.00% | 42.70% | 38.60% | 36.20% | 64.40% | 55.70% | 57.30% | 55.60% | 63.40% | 59.10% |
| Q7 | Do brain scans spot the signs of ageing? | 37.20% | 33.80% | 28.80% | 28.80% | 65.50% | 54.80% | 50.80% | 49.10% | 59.30% | 58.30% |
| Q8 | Does a bile acid link calorie restriction to longevity? | 29.60% | 26.80% | 30.90% | 24.10% | 65.70% | 53.70% | 54.70% | 52.40% | 64.20% | 64.60% |
| Q9 | Compare the credibility of information from Google vs. AI-generated models | 30.80% | 25.10% | 28.10% | 23.80% | 61.70% | 57.80% | 54.70% | 56.00% | 56.60% | 59.20% |
| Q10 | Can the integer partitions detect the prime numbers? | 33.40% | 30.40% | 39.00% | 30.80% | 52.60% | 47.60% | 36.60% | 40.90% | 48.90% | 44.40% |
| Mean similarity values | | 35.58% | 33.45% | 33.39% | 29.94% | 60.93% | 55.60% | 52.69% | 53.56% | 59.04% | 59.00% |

Table 1. Jaccard Similarity Index

Most of the Google-based similarity metrics (e.g., "G Vs Qwen", "G Vs Chatgpt", "G Vs Perplexity", "G Vs Poe") are highly correlated with each other, suggesting they tend to agree in their similarity assessments. "Chatgpt vs. Qwen" and other non-Google comparisons show a moderate correlation with Google-based metrics. Some pairs, such as "G Vs Perplexity" and "Poe V Per", show negative or weak correlations, indicating differing patterns in similarity judgments.

The GAI platforms exhibit a higher similarity, with a minimum mean value of 52.69% and a maximum of 60.93%. They align with other AI platforms and tend to be more homogeneous.

| Comparison | Mean (%) | Standard Deviation |
|---|---|---|
| Google Vs Qwen | 35.58 | 6.86 |
| Google Vs ChatGPT | 33.45 | 7.34 |
| Google Vs Perplexity | 33.39 | 7.38 |
| Google Vs Poe | 29.94 | 5.98 |
| ChatGPT Vs Qwen | 60.93 | 6.37 |
| Poe Vs Per | 55.6 | 4.54 |
| ChatGPT Vs Per | 52.69 | 6.01 |
| Qwen Vs Per | 53.56 | 4.96 |
| Poe Vs Qwen | 59.04 | 4.84 |
| Poe Vs ChatGPT | 59 | 7.07 |

Table 2. Mean and Standard Deviation Values

Most of the mean values are acceptable, especially for comparisons involving ChatGPT, Qwen, Poe, and Perplexity, where the standard deviations are relatively low. However, for comparisons involving Google (e.g., Google vs. ChatGPT or Google vs. Perplexity), the higher standard deviations suggest less consistent performance, so the mean should be interpreted with caution.

## 6.2 Heatmap showing the similarity values between Google and GAI in shades

The heatmap below provides a comprehensive visual overview of these correlations. Darker colors represent stronger correlations (positive or negative), while lighter colors indicate weaker relationships.

This map helps quickly identify which similarity measures tend to move together and which diverge, offering insights into the consistency and divergence among the evaluated models' responses.

**Consistency Among Google-Based Metrics:** A clear cluster of strong agreement exists among metrics that utilise Google as a reference, reflecting similar evaluation standards or data sources.

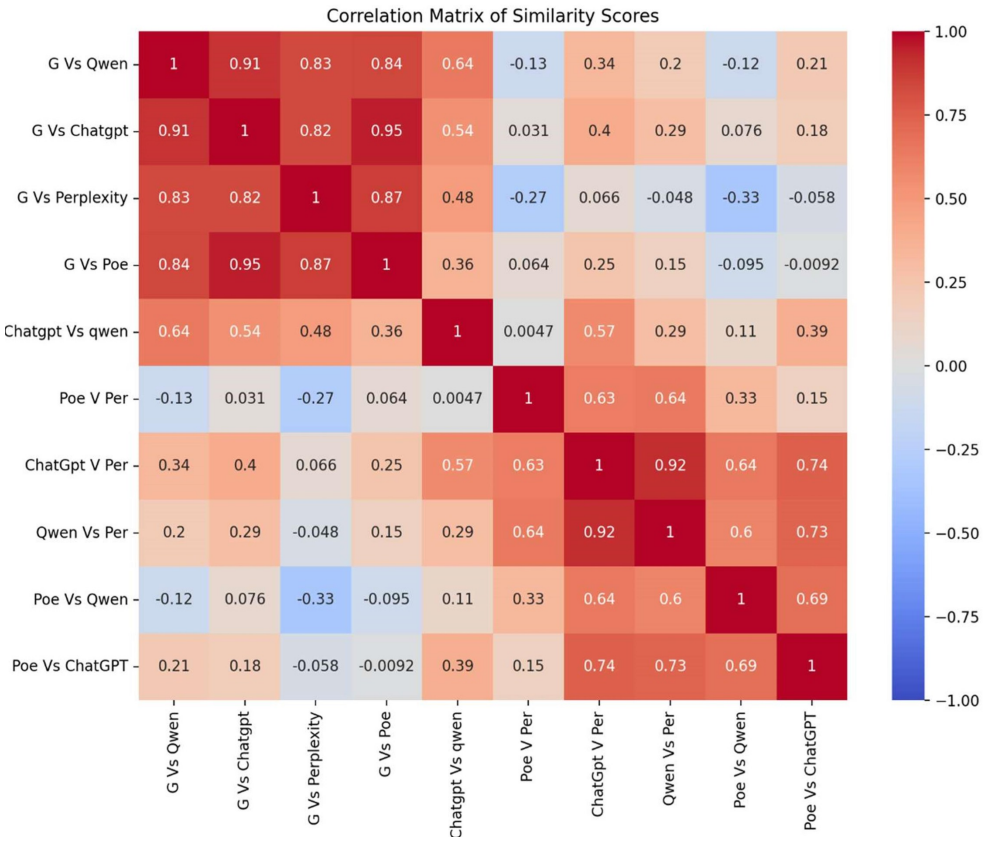



Figure 2. The heatmap of the correlation matrix of the similarity scores

**Divergence in Non-Google Pairings:** Metrics comparing Poe, Perplexity, and other non-Google models often diverge from the Google-based cluster, revealing different perspectives or criteria in their similarity judgments.

Some metrics, such as “Chatgpt Vs qwen”, serve as a bridge between the Google-based metrics and other model pairings, showing moderate correlations with both groups.

## Description

The strongest correlations are found among the Google-based similarity metrics (“G Vs Qwen”, “G Vs Chatgpt”, “G Vs Perplexity”, “G Vs Poe”). These values consistently show correlations above 0.8 with each other, indicating that these metrics tend to assess similarity in a very consistent manner and often agree in their judgments.

Comparisons involving ChatGPT and Qwen (e.g., “ChatGPT vs. Qwen” with “G Vs Qwen” or “G Vs ChatGPT”) also show moderate to high positive correlations, suggesting these models share similar perspectives or response patterns on the evaluated queries. As expected, all diagonal values in the correlation matrix are 1, indicating perfect correlation between each metric and itself.

The metric “Poe V Per” often shows weak or even negative correlations with several Google-based metrics

(e.g., “G Vs Perplexity”, “G Vs Poe”). This suggests that the way Poe and Perplexity compare with each other is notably different from how Google-based metrics evaluate similarity, indicating a divergence in assessment criteria or model behaviour.

Other non-Google pairings, such as “Poe Vs Qwen” and “Poe Vs ChatGPT”, tend to have lower correlations with the Google-based metrics, highlighting that these model pairings may capture unique or less-aligned aspects of similarity.

| Metric Pair | Correlation Value | Pattern |
|---|---|---|
| G Vs Qwen & G Vs ChatGPT | 0.91 | Highest (Strong) |
| G Vs Qwen & G Vs Perplexity | 0.83 | Highest (Strong) |
| G Vs Poe & Poe V Per. | 0.36 | Lowest (Weak) |
| G Vs Perplexity & Poe V Per. | (Negative/Low) | Lowest (Divergent) |
| ChatGPT vs. Qwen & G Vs Qwen | 0.64 | Moderate (Bridging) |

Table 3. Pattern of Correlation Values

These patterns highlight both the clusters of agreement and areas of divergence among the evaluated similarity metrics, offering insight into how different models and reference points align or differ in their assessments.

## Inference

· All “Google vs ...” similarity scores are strongly and positively correlated with each other (e” 0.83), meaning the Google answer tends to align at the same level of similarity with every other model. To express the Google results explicitly, each AI model is similar and consistent, whereas the AI clusters demonstrate consistency and homogeneity among them.

We present below the full scatter matrix of every numeric similarity column, allowing us to understand and visually inspect how each pair of models co-moves.

· Each off-diagonal pane is a scatter plot of one metric versus another. Tight, upward-tilting point clouds (e.g. any “G vs ...” against another “G vs ...”) confirm the strong correlations we quantified.

· Near-horizontal or diffuse blobs (e.g. “Poe V Per” vs most others) signal weak or negative relationships.

· The diagonal shows one-dimensional histograms, revealing that most similarity scores cluster around 0.3–0.7, with “Poe V Per” skewing a bit lower.

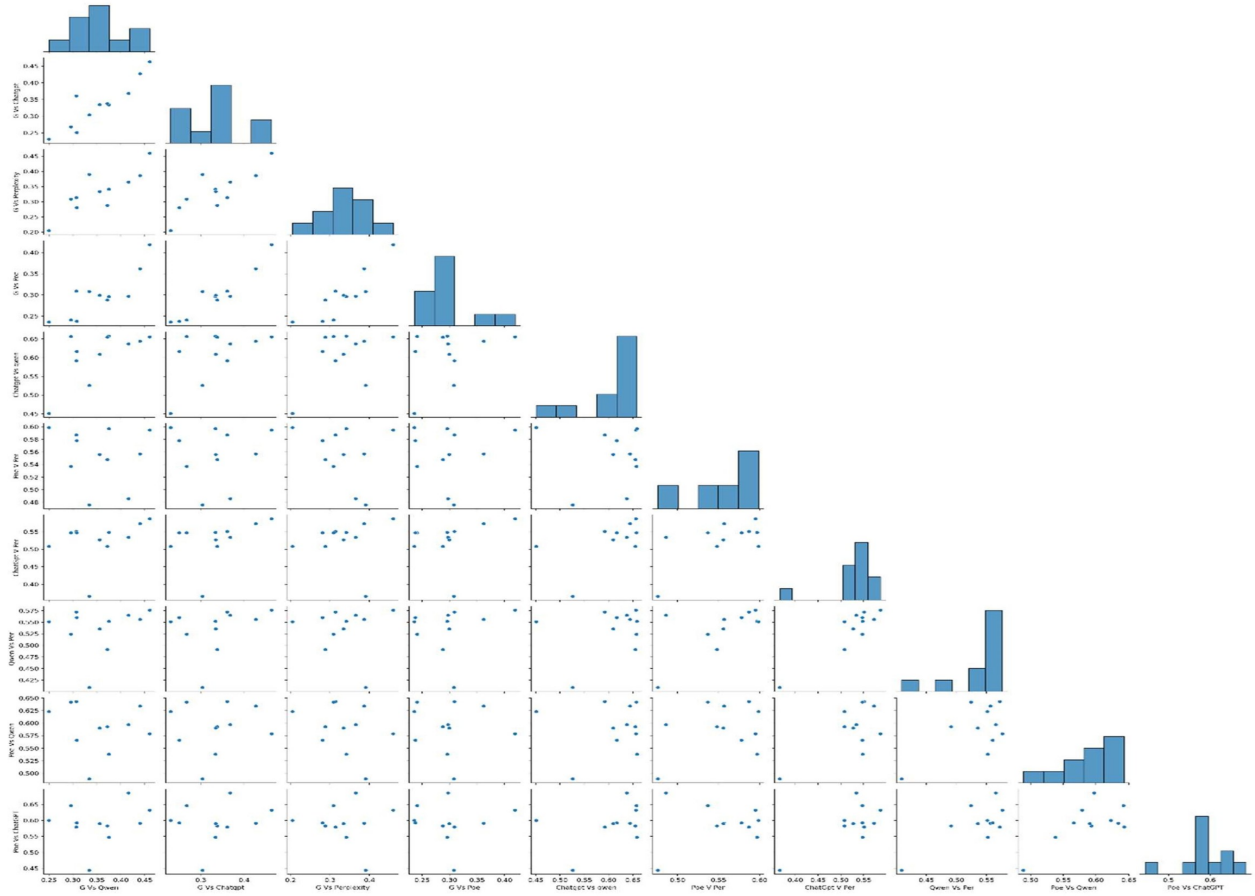


Figure 3. Complete scatter matrix of pairs

Drill into specific outliers (points far from the main cloud) to see which questions generated especially divergent answers.

## Outliers

Outliers are unique observations that can cause distortion in group values. When analysing values, outlying observations may cause problems because they can strongly influence the outcome.

## Divergent scores - Inference

We converted every similarity column to *zz*-scores, summed their squared values, and took the square root. The resulting "*divergence_score*" reflects how far each question's pattern of pairwise overlaps departs from the overall centre of the data cloud (akin to a Mahalanobis distance).

## Top five stand-outs

1. "Can the integer partitions detect the prime numbers?" – by far the most unusual, sitting more than 5.8 standard-deviation units from the centroid.

2. "Climate risks for the younger generation" and

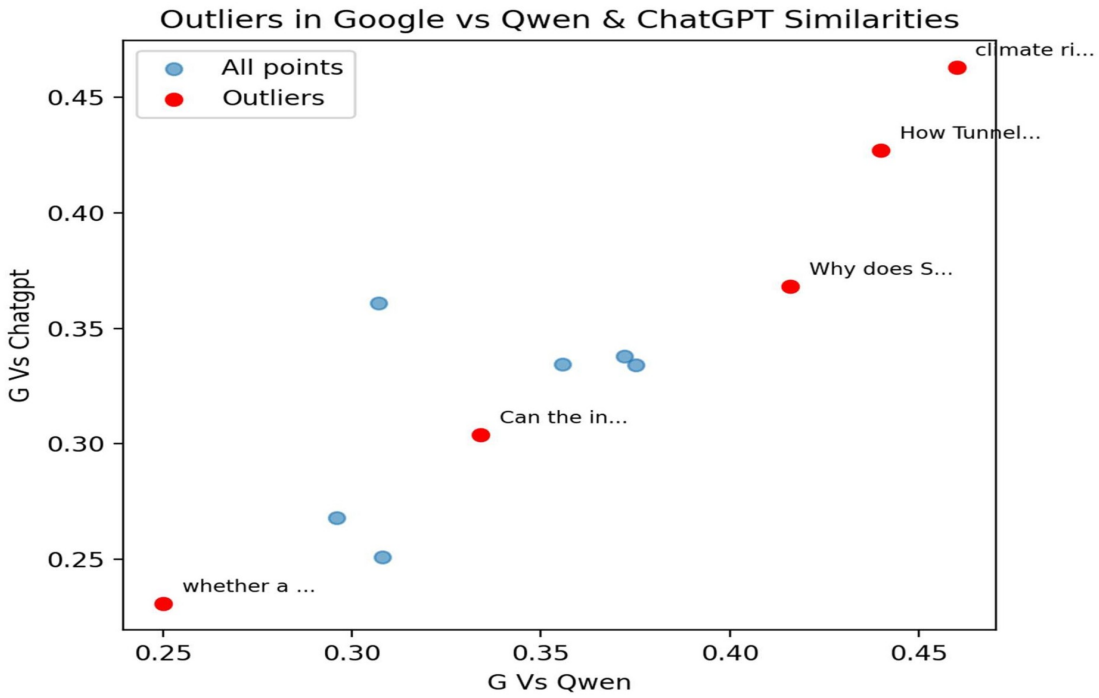



Figure 4. Outlier values of the selected pairs

3. “Whether a microbe with a tiny genome may evolve into a virus” – both around 4.4 units. 4-5. Two physics/ science questions round out the list with scores near 2.8.

***Visual cue***

The scatter plot highlights the five in red against the main mass of points (Google vs. Qwen on the *x*-axis, Google vs. ChatGPT on the *y*-axis). Their isolation indicates that they do not follow the typical tight linear band of most queries.

**Why do they diverge?**

· **Integer-partition / prime-number query:** highly specialised; different models offer very different phrasing or partial answers, so pairwise Jaccard overlaps drop unevenly.

· **Climate-risk and microbe-virus questions:** one or two models produce markedly shorter or broader answers, suppressing overlap for some pairs but not others.

· **The two physics items involve nuanced quantum concepts;** stylistic or explanatory choices vary sharply across models, yielding a jagged similarity profile.

We open the original answers for these five items to inspect which models disagree the most.

1. For each of the five most “divergent” questions, we located the pair of models whose Jaccard overlap is lowest (i.e., where the wording differs the most).

2. The table shows those weakest-overlap pairs and their similarity score.

| Prompts | Lowest Pair | Similarity |
|---|---|---|
| Can the integer partitions detect the prime numbers? | G Vs ChatGPT | 0.304 |
| Climate risks for the younger generation | G Vs Poe | 0.419 |
| Whether a microbe with a tiny genome may evolve into a virus | G Vs Perplexity | 0.205 |
| How do Tunnelling photons challenge the interpretation of quantum mechanics? | G Vs Poe | 0.362 |
| Why does Science become less disruptive? | G Vs Poe | 0.297 |

Table 4. Low similarity pairs

**Inference**

· "Can the integer partitions detect the prime numbers?" → Google vs. ChatGPT is the most disagreeing pair (0.304).

· "Climate risks for younger generation" → Google vs. Poe has the weakest overlap (0.419).

· "Whether a microbe with a tiny genome may evolve into a virus" → Google vs. Perplexity bottoms out at 0.205—very low, meaning totally different phrasing or content.

· The two physics-related prompts ("Tunnelling photons …", "Why does Science become less disruptive?") are also driven by disagreements between Google and Poe.

**Next Logical Step**

To inspect the exact wording differences, we would need the raw answers each model produced for these prompts. They are not included in the spreadsheet—only the pairwise Jaccard scores were provided. When we have a separate document or export containing the full text responses, we can upload it and highlight precisely where the wording diverges.

· Each dot corresponds to one model-pair (for example, G vs Qwen, ChatGPT vs Qwen, etc.).

· Because we only have one averaged value per pair, the scatter simply visualises their heights; there is no "cloud" to analyse correlation.

· The highest average overlaps are between ChatGPT & Qwen and Poe & Perplexity (≈ 0.59).

· The lowest is Google (G) versus Poe (≈0.30)

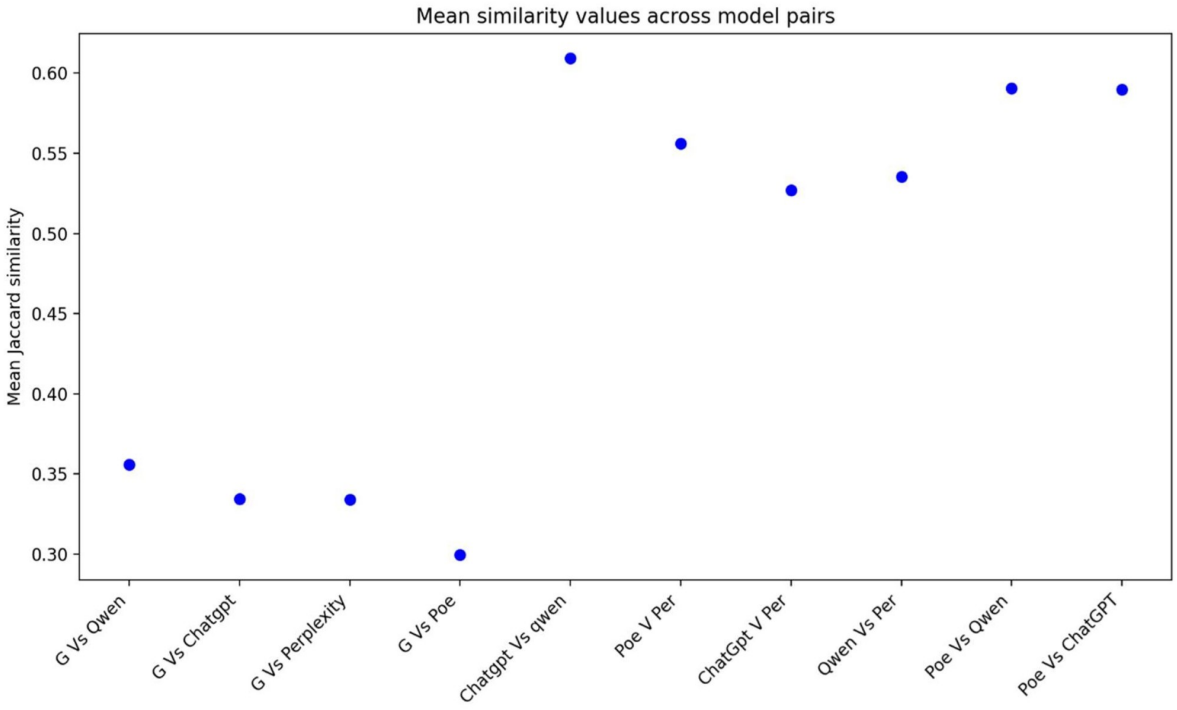
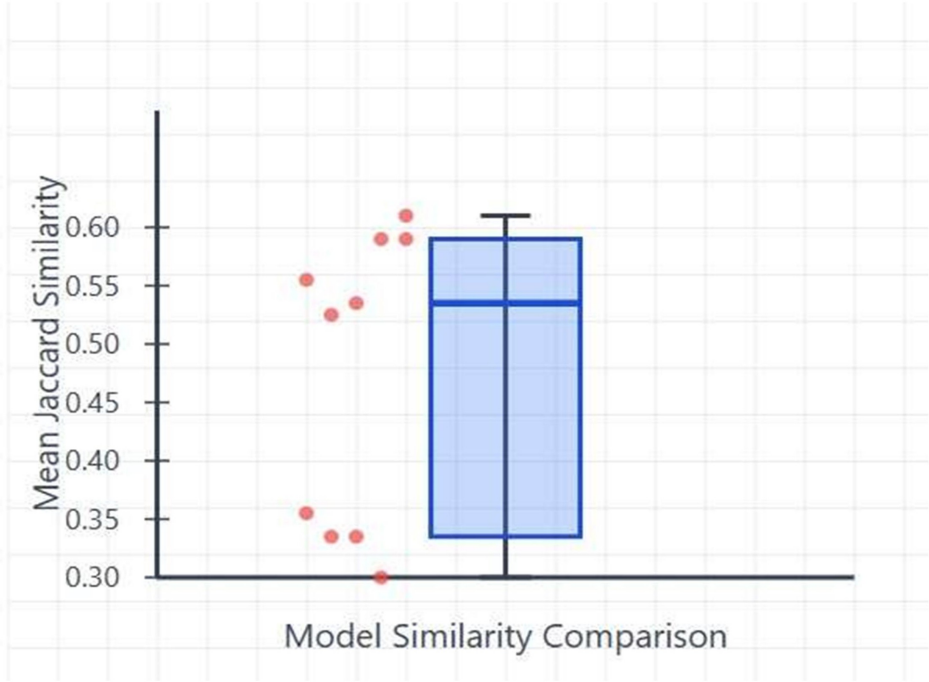



Figure 5. Mean similarity values across model pairs



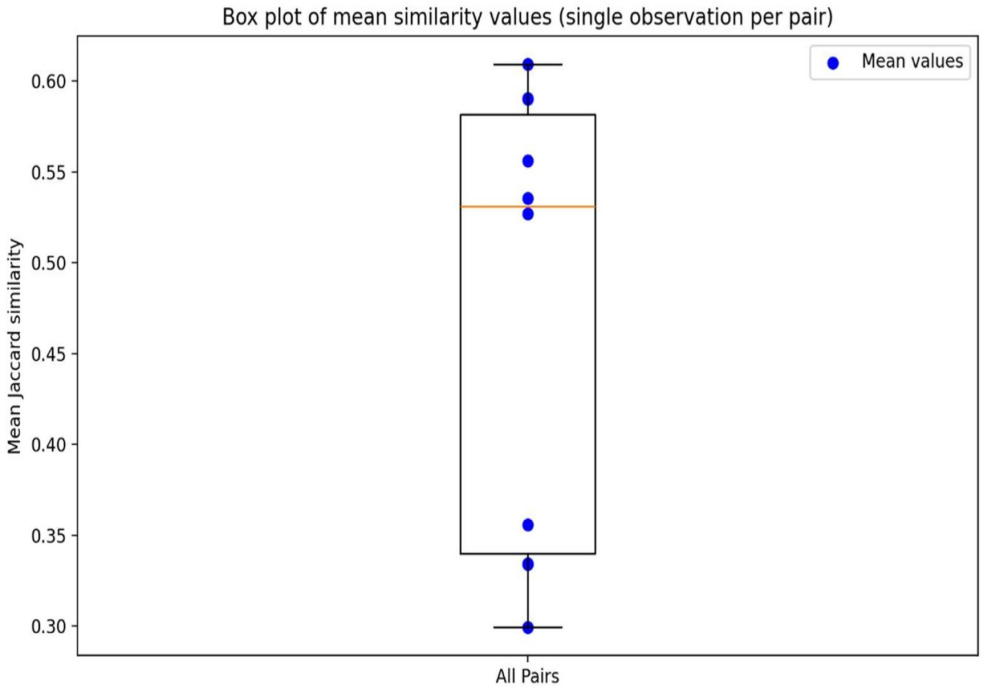



Figure 6a and 6b. Model similarity comparison

· The solitary horizontal line is the "box" produced when there is only one number—its min, Q1, median, Q3, and max are all identical.

· The blue dots are the ten mean-similarity scores themselves; we overlaid them so we can still see where each value sits.

## 6.3 Discussion on Summarised Similarity Values

The above analyses presented a correlation analysis comparing Google search results and outputs from generative AI (GAI) models — ChatGPT, Qwen, Perplexity, and Poe — using the Jaccard Similarity Index. The study assesses the similarity between content from Google (G) and AI-generated responses, as well as the consistency of the GAI models with each other. Results show that Google-based comparisons (e.g., Google vs Qwen, Google vs ChatGPT) are highly correlated, indicating consistent alignment between Google and these models. However, GAI platforms show even higher internal similarity (mean scores 52.69–60.93%), suggesting homogeneity among them. Metrics like "Poe Vs Perplexity" often diverge from Google-based assessments. Some queries, such as those involving math or subtle science topics, show significant divergence, indicating varied model interpretations. Outliers and low-similarity pairs (e.g., Google vs. Poe at 29.7%) highlight discrepancies in content overlap. Overall, the analysis reveals strong internal consistency among GAI models but less alignment with Google-based metrics.

### 6.4 Credibility Scores

The credibility scores are determined using user evaluation, reference analysis, and GAI evaluation. The credibility and value are operationalised by the parameters as outlined below.

### 6.4.1 User Evaluation

Twelve users are recruited to assess the responses from Google and GAI platforms, which provide a scoring based on assessments. The user perceptions of the output evaluate several AI platforms—ChatGPT, Qwen, Poe, Per, and Google—across multiple qualitative dimensions. The assessment criteria we used include:

***Completeness:*** Refers to how fully the AI addresses the query or task.

***Depth and Intensity:*** The level of detail and analytical rigour in the response.

***Scientific Integrity:*** Adherence to scientific principles, logical consistency, and accuracy.

***Evidence Support:*** The extent to which responses are backed by credible evidence or references.

***Reliability:*** Consistency and trustworthiness of the information provided.

The responses generated by Google Search and GAI platforms are presented to the users who assess using the above scale. We used 7-point Likert items to provide a more accurate measure of a participant's true evaluation [26]. The mean values of all the scores are arrived and presented in Table 5.

| Criteria | ChatGPT | Qwen | Poe | Per | Google |
|---|---|---|---|---|---|
| Completeness | 3.9 | 3.25 | 3.1 | 3.2 | 5.9 |
| Depth and Intensity | 3.1 | 3.6 | 2.7 | 3.8 | 6.3 |
| Scientific Integrity | 2.4 | 3.8 | 1.7 | 5.4 | 6.8 |
| Evidence Support | 1.85 | 3.9 | 1.6 | 5.9 | 6.9 |
| Reliability | 2.8 | 4.2 | 2.2 | 5.3 | 6.8 |
| Mean | 2.81 | 3.75 | 2.26 | 4.72 | 6.54 |

Table 5. Mean scores of all five platforms based on user scales

Google has the highest mean score (6.54), indicating superior overall performance across all criteria, followed by Perplexity. Qwen performs moderately well across most criteria, particularly in Reliability and Scientific Integrity.

ChatGPT exhibits middling performance, with the highest scores in Completeness but notably lower scores in Scientific Integrity and Evidence Support.

Poe scores lowest overall, especially in Scientific Integrity and Evidence Support. The retrieval pages from

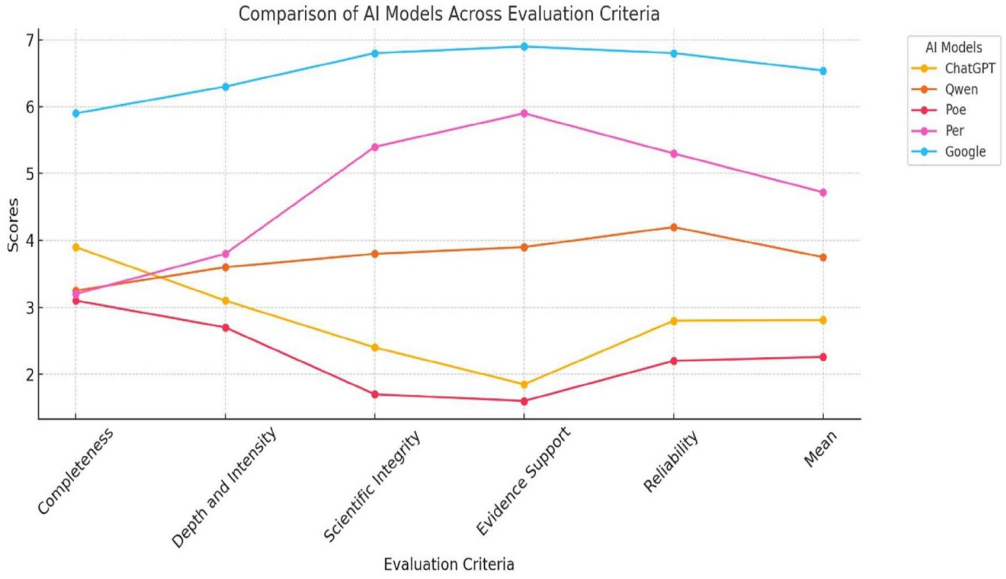



Figure 7. Comparison of Google and other AI models in user evaluation

Google win decisively on the three criteria - content quality, value, and relevance. They provide a comprehensive, data-rich, and multi-perspective examination of the question. GAI platforms offer a useful conceptual overview but lack the depth, evidence base, and connection to current research that can be retrieved through Google search results.

Poe scores lowest overall, especially in Scientific Integrity and Evidence Support. The characteristics of the content from Google retrieval results are summarised.

***Content Quality -*** Clear, organised, and conceptually strong. - Data-rich, methodologically rigorous, multi-perspective.

***Value-*** Good for quick understanding. -High value for academic and policy research.

***Relevance -*** Relevant, but more general. - Directly tied to current empirical studies and debates.

While GAI platforms offer a helpful summary with good accessibility, Google Responses provide significantly greater depth, empirical support, and scholarly context, making them the top choice for anyone seeking comprehensive insight into the issues.

### 6.4.2 Reference Analysis

As outlined in the section, we calculated the credibility scores based on the sources mentioned. Since only Perplexity platforms support references, the reference analysis is limited to this GAI only.

Google has higher average credibility (0.826 vs. 0.585) and exhibits more consistent performance, as shown

| | Google | Perplexity |
|---|---|---|
| | 0.87 | 0.68 |
| | 0.87 | 0.6 |
| | 0.87 | 0.56 |
| | 0.71 | 0.6 |
| | 0.8 | 0.6 |
| | 0.85 | 0.8 |
| | 0.86 | 0.7 |
| | 0.9 | 0.42 |
| | 0.8 | 0.36 |
| | 0.73 | 0.63 |
| **Mean scores** | 0.82 | 0.59 |

Table 6. Credibility analysis using references used in Google and Perplexity

by a lower standard deviation (0.058 vs. 0.134).

Google platform maintains credibility above 0.7 in all queries, whereas perplexity has a Lower average credibility (0.585 vs. Google's 0.826). It also exhibits higher variability in performance (standard deviation of 0.134), scoring below 0.5 in 5 out of 10 queries, which indicates lower perceived credibility in those cases.

Google is perceived as more credible than Perplexity across all queries, and Perplexity exhibits inconsistency, particularly in queries such as Q8 and Q9, where it scored significantly lower. The reasons observed in this work are the accuracy of responses, Clarity and depth of information, reliable citations and the tone and confidence of the response. Google-based responses may be more reliable and consistent for factual and complex queries. Perplexity, while sometimes providing good responses (e.g., Q6), appears to be less consistent and less trusted overall. Users may prefer Google for high-stakes or critical information, while Perplexity may be better for exploratory or general-purpose use.

### 6.4.3 Content Credibility

The way GAI platforms evaluate their content is an interesting observation. We asked ChatGPT about the possible framework for assessing content, and it listed five primary criteria, which are explained below.

### Evaluation Criteria Definitions

**1. Source Credibility**: Reputability and transparency of sources cited.

**2. Content Accuracy**: Factual correctness, scientific integrity.

**3. Completeness**: Depth, breadth, and inclusion of key findings or mechanisms.

**4. Objectivity**: Neutral tone, avoidance of speculation or marketing language.

**5. Verifiability**: Availability and traceability of supporting references or data.

| Query | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Google** | 92 | 94 | 96 | 98 | 98 | 98 | 99 | 92 | 92 | 98 | 95.7 |
| **Perplexity** | 85 | 89 | 89 | 90 | 86 | 88 | 90 | 68 | 82 | 88 | 85.5 |
| **ChatGPT** | 73 | 82 | 86 | 84 | 80 | 80 | 87 | 80 | 76 | 82 | 81 |
| **Qwen** | 78 | 84 | 78 | 74 | 82 | 78 | 82 | 74 | 72 | 78 | 78 |
| **Poe** | 64 | 71 | 74 | 66 | 66 | 68 | 76 | 64 | 62 | 68 | 67.9 |

Table 7. Content Credibility Scores

For each of the extracted responses from Google and the four GAI platforms, we loaded the files and fed the following prompt. "***Produce a composite, explainable Credibility Score of the five attached files using Source Credibility, Content Accuracy, Completeness, Objectivity and Verifiability***".

ChatGPT, while generating the scores for each GAI and Google, has listed the reasons behind the scores. *Although these scores do not reflect perfection, the overall results lead us to accept the assessment as a whole*.

## 7. Summary

We compared the similarity of content between the Google and GAI platforms, excluding the relevance analysis of these two models. There are differences between the two models in responses, specifically in distinguishing between cognitive and affective dimensions. Its high scores across all five criteria reflect its reliability as a source for empirical insights into public perceptions on issues**.**

GAI produce a short excerpt and may be useful for conceptual understanding. They produced useful summaries, but they should be cross-verified with primary sources.

Even Perplexity lists the references it used to generate answers, not all of which are really cited in the answers, and it mentions studies without specific citations. We found that the references are general experimental findings without traceability. The kind of extraction from an advanced platform like Perplexity is still a hallucination experience, which many users fail to realise. AI platforms are found to hallucinate' false outputs and

groundless answers. [27, 28, 29]. The most effective way to detect hallucinations is by subjecting the training data to validation and inconsistency detection. Generative AI mechanisms are not explicit about the training data, and integrity remains an issue.

## 8. Limitations

For Google search results, we used only the responses from the top hits. When we use lower-ranked search results, the content results may reflect a different outcome.

We acknowledge the use of a limited number of queries/prompts for analysis. This is due to the extensive analysis and study of feature similarity and credibility in a single exercise. The experts' views are limited in number and volume, and are under-described. The credibility metrics we have used are not comprehensive, as the empirical validation method is complex and not entirely clear. Credibility metrics can be improved by using a large user population of experts for evaluation. We hope to overcome these limitations in future research. The similarity and credibility assessment of the AI models are likely to undergo changes as these platforms refine and improve their models.

## 9. Conclusion and Future Directions

AI models face challenges with veracity, attribution, explanation, transparency and inference procedures. [30]. Generative platforms have made communication faster and easier; however, they have also led to users becoming dependent on synthetic content. These tools nurtured distrust for scholarly resources and hindered the process of knowledge foraging. The challenges in today's information world force users to seek easier access to knowledge without fully comprehending it. We need to establish mechanisms for transparency in the content processing process.

AI models continue to evolve, incorporating new features and trying to create enhanced content; however, their success depends on how these models mimic human expertise.

Natural Language Processing, particularly in-depth text analysis, may yield meaningful insights to inform comprehensive outcomes. We can observe that NLP techniques are more adept at detecting the synthetic nature of AI models. We intend to analyse the volume and length of answers generated on Google and GAI platforms, which is a measure of content value.

**Declaration:** AI tools are used for a few statistical derivations and not used for writing.

**Funding:** This work has not received any funding.

**Conflict of Interest:** The paper has no conflicts of interest.

## References

[1] Lindsey Witmer Collins. *How AI is undermining online authenticity*. https://www.fastcompany.com/91364892/how-ai-is-undermining-online-authenticity

[2] Bozkurt, A., Sharma, R. C. (2024). Trust, credibility and transparency in human–AI interaction: Why we need explainable and trustworthy AI and why we need it now. *Asian Journal of Distance Education, 19*(2). http://www.asianjde.com/ojs/index.php/AsianJDE/article/view/819

[3] Luna, Javier Canales. (2023). ChatGPT vs Google Bard: A comparative guide to AI chatbots. *DataCamp Blog*. https://www.datacamp.com/blog/bard-vs-chat-gpt

[4] Borji, Ali. (2023, April 3). A categorical archive of ChatGPT failures (arXiv:2302.03494v8) [Preprint]. arXiv. https://arxiv.org/abs/2302.03494v8

[5] Guo, Biyang., Zhang, Xin., Wang, Ziyuan., Jiang, Minqi., Nie, Jinran., Ding, Yuxuan., Yue, J., Wu, Yupeng. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection (arXiv:2301.07597v1). https://arxiv.org/abs/2301.07597v1

[6] Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., Hovy, E., Ji, H., Menczer, F., Miguez, R., Nakov, P., Scheufele, D., Sharma, S., Zagni, G. (2024). Factuality challenges in the era of large language models and opportunities for fact checking. *Nature Machine Intelligence, 6*(8), 852–863.

[7] Zhang, Peng., Boulos, M. (2023). Generative AI in medicine and healthcare: Promises, opportunities and challenges. *Future Internet, 2023*.

[8] Li, Alice., Sinnamon, L. (2024). Generative AI search engines as arbiters of public knowledge: An audit of bias and authorit. https://arxiv.org/abs/2405.14034

[9] Pichappan, P., Krishnamurthy, M., Vijayakumar, P. (2023). Analysis of ChatGPT as a question-answering tool. *Journal of Digital Information Management, 21*(2), 50–60.

[10] Ou, M., Zheng, H., Zeng, Y., Hansen, P. (2024). Trust it or not: understanding users' motivations and strategies for assessing the credibility of AI generated information. *New Media & Society,* November.

[11] Zhang, He., Xie, Jingyi., Wu, Chuhao., Cai, Jie., Kim, Chanmin., Carroll, John. M. (2024). The future of learning: Large language models through the lens of students, *Association for Computing Machinery*, New York, NY, USA.

[12] Huschens, Martin., Briesch, Martin., Sobania, Dominik., Rothlauf, Franz. (2023, September 5). Do you trust ChatGPT? Perceived credibility of human and AI generated content (arXiv:2309.02524v1)[cs.HC].

[13] Knight, Samsun., Bart, Yakov. (2023). Generative AI and user-generated content: Evidence from online reviews. Social Science Research Network.

[14] Zhou,Tao., Lu, Hailin. (2025). The effect of trust on user adoption of AI generated content. *The Electronic Library, 43*(1), 61–76.

[15] Cheng, S., Tsai, S., Bai, Y., Ko, C., Hsu, C., Yang, F., et al. (2023, December 25). Comparisons of quality,

correctness, and similarity between ChatGPT generated and human written abstracts for basic research: Cross sectional study. *Journal of Med Internet Res, 25*, e51229.

[16] Kim, H. J., Yang, J. H., Chang, D., Lenke, L. G., Pizones, J., Castelein, R., et. al. (2024, June 26). Assessing the reproducibility of structured abstracts generated by ChatGPT and Bard compared to human written abstracts in spine surgery: Comparative analysis. *Journal of Med Internet Res, 26*, e52001.

[17] Jang., Lukasiewicz. (2023). Consistency analysis of ChatGPT (arXiv:2303.06273) [Preprint]. arXiv. 2303.06273

[18] Shen, X., Chen, Z., Backes, M., Zhang, Y. (2023). In ChatGPT we trust? Measuring and characterising the reliability of ChatGPT, arXiv [Preprint]. arXiv: 2304.08979

[19] Xu, Danni., Fan, Shaojing., Kankanhalli, Shaojing. (2023, October 29–November 3). Combating misinformation in the era of generative AI models. In: *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)* (pp. 1–8). Ottawa, ON, Canada: ACM. https://doi.org/10.1145/3581783.3612704

[20] Wahid, Risqo M., Mero, J., Ritala, P. (2023). Editorial: Written by ChatGPT, illustrated by Midjourney: Generative AI for content marketing. *Asia Pacific Journal of Marketing and Logistics*.

[21] Tai, Yuehong Cassandra., Patni, Khushi., Hemauer, N. D., Desmarais, B. A., Lin, Y.-R. (2025). GenAI vs. human fact checkers: Accurate ratings, flawed rationales. In *Proceedings of the Web Science Conference* (pp. 1–12).

[22] Hanimann, Anina., Heimann, Andri., Hellmueller, Lea., Trilling, D. (2023). Believing in credibility measures: Reviewing credibility measures in media research from 1951 to 2018. *International Journal of Communication, 17*(2023), 214–235.

[23] Guo, Biyang., Zhang, Xin., Wang, Ziyuan., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection (arXiv:2301.07597v1).

[24] Travieso, Gonzalo., Benatti, Alexandre., Costa, Luciano da F. (2024). An analytical approach to the Jaccard similarity index (arXiv:2410.16436v1)

[25] Talaver, Oleh V., Vakaliuk, Tetiana A. (2025, May 14). A model for improving the accuracy of educational content created by generative AI. *In: 7th International Workshop on Augmented Reality in Education (AREdu 2024)* (pp. 149–158). Kryvyi Rih, Ukraine.

[26] Finstad, Kraig. (2010). Response interpolation and scale sensitivity: Evidence against 5 point scales. *Journal of User Experience, 5*(3), 104–110.

[27] Farquhar, S., Kossen, J., Kuhn, L., et. al. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*, 625–630.

[28] Xiao, Y., Wang, W. Y. (2021). On hallucination and predictive uncertainty in conditional language generation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 2734–2744). PA: Association for Computational Linguistics.

[29] Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., Saenko, K. (2018). Object hallucination in image captioning. In Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (Eds.), *In:* Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4035–4045). Association for Computational Linguistics.

[30] Blau, Wolfgang., Cerf, Vinton. G., Enriquez, Juan., Francisco, Joseph. S., Gasser, Urs., Gray, Mary. L., Greaves, Mark., Grosz, Barbara J., Jamieson, Kathleen Hall., Haugk, Gerald H., Hennessy, John L., Horvitz, Eric., Kaiser, David I., Londono, A. J., Lovell-Badge, R., McNutt, M. K., Minow, M., Mitchell, T. M., Ness, Susan ., Parthasarathy, Shobita., Perlmutter, Saul., Press, William H., Wing, Jeannette M., Witherell, Michael. (2024). Protecting scientific integrity in an age of generative AI. *Proceedings of the National Academy of Sciences of the United States of America, 121*(22), e2407886121. https://doi.org/10.1073/pnas.2407886121, pp 1-3