Journal of Digital Information Management

DLINE JOURNALS

ISSN: 0972-7272

JDIM 2026: 24 (1)

Semantic Similarity, Phrase Analysis, and Expert Evaluation of Human versus LLM-Generated Abstracts

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

ABSTRACT

This research examines abstracts in scientific papers and how AI generates them. Abstracts are crucial to information use because they are the first sources researchers consult to decide whether a paper is worth checking out. The study analyses the abstracts from the December 2025 issues of Antioxidants and PLOS Computational Biology. These abstracts are generated by the authors themselves, ChatGPT, or Qwen. To evaluate, we used semantic similarity (Jaccard index), phrase occurrence frequency, and expert scores. It seems this covers quality, detectability, and what it all means for science writing.

The results showed that the AI-generated abstracts were more similar to one another than to human generated abstracts. The mean Jaccard index was around 0.66 to 0.68 for the AIs compared to themselves, but lower with the author written stuff. That points to both AIs writing in a similar style, regardless. Domain specific terms appeared in both humans and AIs, but the way they used them, such as frequency and exact types, differed between ChatGPT and Qwen. Expert scoring assigned higher grades to AI abstracts based on clarity, structure, scientific sense, and originality or relevance. Qwen got a mean of 9.29, ChatGPT 9.02, while all human authored ones averaged 7.75. The ANOVA test reflects that both human and resulted in about 79 per cent of the variation in the scores. This finding suggests that AI can generate more impressive and comprehensive summaries than humans. Still, there are ethical problems to consider, such as how AI might fabricate references, spread misinformation, or even hijack peer review. The analysis estimates that 10 to 14 per cent of recent biomedical abstracts indicate AI assistance. It indicates that we need better ways to detect it, clearer rules, and to focus more on the actual research ideas rather than on how slick the writing is.

Subject Categories and Descriptors: [**I.2 Artificial Intelligence**]: [**I.2.7 Natural Language Processing**]; Text analysis [**H.3.1 Content Analysis and Indexing**]: Abstracting methods

General Terms: Large Language Models, Natural Language Processing, Scientific Literature. Content Analysis, AI-abstracts, Human writing

Keywords: AI-generated Scientific Abstracts, Large Language Models (LLMs), Semantic Similarity Analysis,

Phrase-Frequency Analysis, Expert Evaluation, Academic Writing Detectability, LLM Stylistic Convergence,Scientific Communication Ethics, Journal Abstract Analysis

Received: 17 December 2025, Revised 4 February 2026, Accepted 11 February 2026

Review Metrics: Review Scale: 0/6; Review Score: 4.72; Inter-reviewer consistency: 80.4%

DOI: https://doi.org/10.6025/jdim/2026/24/1/40-61

1. Introduction

What is the role of abstracts in scientific literature, and why are they essential? Abstracts play a pivotal role in scientific literature by offering a concise yet comprehensive overview of a study's objectives, methodology, results, and conclusions. Their significance stems from their role as the primary point of entry for readership: many researchers and practitioners rely solely on article abstracts to assess their relevance and weight, often without reading the full text. Therefore, abstracts are important signals of a paper's scientific content and contributions.

Abstracts serve the above function and, as such, must be minutely braided to retain the semantic content of the original work while delivering the principled results, making it concise yet intrinsically intelligible. But the quality of abstracts generated by authors themselves varies widely; many studies have shown that researchers find it difficult to write up summaries that are both accurate and substantive. To add to this, peer reviewers and editors may also assign undue importance to the abstract relative to the full manuscript, failing to notice small differences or inconsistencies in the full version.

Moreover, selective-access models in open-access publishing and digital bibliographic databases mean that abstracts are often the only free element of scholarly input. The availability of the data increases their impact on the distribution, exploration and preliminary assessment of scientific work. Thus, well-put-together abstracts are not just adjuncts but critical tools for effective communication and the dissemination of scientific research.

2. Review of Related Work

The abstracts are crucial to a research paper because readers often read only titles and abstracts [1]. Since the emergence of LLMs, many authors have used them to generate summaries.

The advent of large language models (LLMs) like ChatGPT has fundamentally transformed scientific writing. This transformation is most palpable in the domains of conference and journal abstracts concise, structured summaries that gatekeep research. Before we present our work, we review the current evidence on AI-generated abstracts, including their quality, detectability, prevalence, and ethical implications for academic publishing.

2.1 Features and Characteristics of AI-generated Abstracts

Several studies have shown that ChatGPT can generate strikingly acceptable scientific abstracts. Both Babl [1] and Babl [2] showed that ChatGPT generated grammatically and structurally sound, high quality conference

abstracts with no evident errors. The AI suggested appropriate titles, wrote introductory sentences consistent with the state of current knowledge, accurately extracted statistical results from data tables and provided reasonable interpretations in conclusions.

Similarly, Altmae et al. [3] considered ChatGPT-generated abstracts acceptable for publication in Reproductive BioMedicine Online, needing only minor additions. It also helped us come up with very fitting titles. In a compelling example of AI's powers, Manley et al. [4] reported that 13 of 18 health professionals were unable to distinguish an abstract generated by ChatGPT from one published on the same topic, with most who responded saying they would not suspect AI input in a scientific journal.

However, limitations exist. Gao et al. [17] analysed the pattern of headings used and found that, when imitating high-impact publications, only 16% of ChatGPT-generated abstracts correctly followed journal specific standards, indicating that prompt engineering decides output quality. Pan et al. [6] reported that AI-generated abstracts in gynaecology were more grammatically accurate and contained fewer writing errors than human-written text, questioning assumptions about the quality of AI writing.

2.2 The Challenge of Detection

Distinguishing AI-generated from human-written abstracts has proved surprisingly difficult.

Gao et al. [5] reported that human reviewers correctly identified only 68% of AI-generated abstracts, misclassifying 32% as real, and commenting that the distinction was "surprisingly difficult." Kresoja [6] affirmed this, stating it is "excessively difficult even for human experts" to distinguish between human written abstracts based on true data and ChatGPT manufactured abstracts based on imaginary data. Pan et al. [7] found human reviewers achieved merely 49.7% accuracy, essentially random guessing.

Detection tools show variable performance. Gao et al. [5] reported an AUROC of 0.94, whereas Elek et al. [8] found accuracies ranging from 36% to 95% across three tools. Elkhatat et al. [9] observed that detection tools were more accurate at identifying GPT-3.5 than GPT-4 content but produced false positives with human-written controls. Weber Wulff et al. [10] found that existing detection tools lack accuracy and reliability, with a strong tendency to label output as human written rather than to identify AI-generated text.

Rhetorical move analysis was used by Voss et al. [11] to find systematic differences between published and AI-generated applied linguistics abstracts. They found that published abstracts had more grammatical and lexical variation than AI-generated abstracts. There is evidence that ChatGPT-generated abstracts can be confidently identified by traditional models based on semantic and lexical qualities, outperforming human judgment [12]. This suggests that computational methods may be more effective.

2.3 Prevalence and Linguistic Signatures

The use of AI in scientific writing is very common. According to Kobak et al.'s [13] analysis of 14 million PubMed abstracts from 2014 to June 2024, over 150,000 abstracts roughly one in ten were produced using LLMs. After November 2022, their analysis revealed a discernible increase in "280-style" words (such as intricate, meticulous, delve, pivotal, and underscore). According to Mallapaty [14], 14% of biomedical abstracts exhibit signs of AI-generated content, including inadvertent terms such as "unparalleled" and "invaluable."

More than 15 million PubMed abstracts were included in the analysis [15], and at least 13.5% of biomedical abstracts from 2024 had evidence of LLM-assisted writing, as evidenced by a sharp rise in stylistic "excess words." This impact eclipsed even pandemic-driven shifts and varied widely across disciplines and countries.

However, linguistic signatures evolve. Geng et al. [16] analysed *arXiv* abstracts and found that some ChatGPT-favoured words (e.g., "delve") fell once they were flagged as such in early 2024, while others (e.g., "significant") continued to trend upward. This indicates authors are modifying the use of LLMs, steering clear of detectable markers, highlighting a process of human-LLM coevolution in academic writing.

2.4 Geographic Impact and Quality Concerns

Arnold et al. [17] document some interesting results from their study. The productivity of some authors from Asian non-English speaking countries increased dramatically by up to 89%. Renowned journals require scholarly writing, which is difficult for many researchers; however, many now write at a high level. We exercise caution that many weak ideas are covered by AI-style writing. The study found that when AI-generated writing is more complex, the paper may not be of higher quality.

Cheng et al. [18] studied ChatGPT-generated abstracts from 30 basic research papers, finding them significantly lower in level and quality, with lower similarity to the originals (2.10%-4.40%). Interestingly, three abstracts contain incorrect conclusions, and experts correctly identified AI-generated ones 93% of the time, suggesting limited reliability for scientific use.

2.5 Emerging Concerns and Ethical Considerations

Kocak et al. [19] noted that as global submissions increase, many reviewers rely on LLMs to manage workload. This creates vulnerability: Gharami (2025), Maloyan (2025), Gibney (2025), and Lin (2025) [20-23] demonstrated that attackers can embed invisible instructions in PDFs using white text or tiny fonts that trick LLMs into giving overly positive reviews, turning documents into adversarial attack surfaces.

Ethical perspectives vary. Akkureddy [24] argued that using AI to generate abstracts from one's manuscript without alteration is not unethical, as AI functions as a summarisation tool derived from the author's original text, preserving integrity. However, Altmae et al. [3] highlighted serious concerns, including fabricated references, potential misinformation, and ethical issues, and concluded that ChatGPT requires rigorous human oversight.

The preceding presentation indicates that AI-generated abstracts have achieved remarkable improvement, often indistinguishable from human writing to both experts and detection tools. With prevalence estimates ranging from 10% to 14% across recent abstracts, AI is fundamentally changing scientific communication. While this technology enables access for non English speaking authors, it also poses challenges for quality-assessment mechanisms. This area faces an urgent need for reliable detection methods, clear guidelines, and renewed emphasis on evaluating research by substance rather than polished presentation.

As part of the ongoing exercises on AI influence evaluation, we designed this work to detect content similarity between human and AI-generated abstracts, the extent to which phrases are deployed in both, and the evaluation of these abstracts using various measures. Section 2 summarised the earlier studies; Section 3 presented the methods and the dataset; Section 4 presented the analysis; Section 5 presented the summary;

and Section 7 concluded with a synthesis.

3. Methodology

3.1 Measuring Content Similarity

How the abstracts reflect the content of original papers in the summary without leaving any central content is a measure to understand the ability of human and AI-generated ones. We applied Jaccard's index to both abstracts to measure the content similarity. This forms the first part of our analysis.

3.2 Phrase Analysis

Phrases rather than words are the signals in detecting the content reflection of papers. With the voluminous growth of literature, managing the information content is essential, using a few effective measures, including phrases in the text corpora. [27, 28] Matching and mismatching of phrases determine the effectiveness of abstracts generated in different ways. As the second component of this work, we extracted phrases from both human and AI abstracts and measured the matches and mismatches between them.

3.3 Abstract Evaluation

How do we evaluate the abstract? In this work, the abstracts written by both authors and AI platforms are assessed using the standards. We assess the abstracts in several ways. In this work, we fixed the following set of criteria for assessment in the prescribed scale.

Criterion	Excellent	Good	Fair	Poor
Clarity and Content				
Completeness				
Scientific Rigour				
Novelty				
Relevance				

3.3.1 Clarity and Concise

An abstract should be concise and clear, without jargon, for its audience. It has to convey the core message crisply without excess information. It further examines how the language should be correct to avoid affecting the readability and fluency of a professional presentation. To assess the level and quality of abstracts, several methods have been proposed in the literature. [25, 26]

3.3.2 Structure and Completeness

Abstracts in empirical research generally have a format. The background or introduction must succinctly state the work's motive and define the issue it seeks to solve. The aim or objective should preferably state the purpose of the study (or research question). Provide a concise description of the study design, data sources and analytical methods used in the methods section. Results should describe the main findings with sufficient details, including relevant quantitative results if applicable. Finally, the concluding statement (or implications) should be drawn directly from the results and be clear about why this study matters or what broader implications it has.

While some disciplines, e.g., the humanities and theoretical sciences, will work with unstructured abstracts, these key components should nonetheless be recognisable in the narrative.

3.3 Scientific Rigour

The abstract must indicate methodological appropriateness and possess enough information to evaluate the study's validity. Results should logically follow from the methods, and any claims made in the conclusion must be backed by evidence.

3.4 Originality and Novelty

The abstracts are expected to clearly communicate the study's novel contribution, how it differs from existing literature, and the knowledge gap it addresses.

3.5 Relevance

It should be topical and important in its field, with a substantial research gap, problem or practice being addressed.

3.6 Accuracy and Consistency

Terminology, units of measure and reported data need to be accurate and consistent with disciplinary standards. Include an abstract structured according to applicable reporting standards.

3.7 Adherence to Guidelines

It must follow prescribed formatting, length, and stylistic guidelines (word limits; section headings if applicable). Submissions to a journal or conference require following certain author guidelines.

3.8 Relevance analysis for Author and AI abstracts

In the following analysis, we calculated the Jaccard's similarity index for each of the twenty abstracts generated by authors and AI. These populations refer to the journal *Antioxidants* and *PLOS Computational Biology*.

4. Dataset

We considered the December 2025 issues of the journals *Antioxidants* and *PLOS Computational Biology* and selected 10 papers from each. The abstracts appended in the journals form the first part of the dataset, followed by the AI-Generated Abstracts. Two AI models, ChatGPT and Qwen, are used to generate abstracts by feeding in full text papers. These abstracts are compared for the three levels of analysis.

1. Content Similarity between the three abstracts

2. Phrase analysis of the abstracts to detect content similarity; and

3. Experts' evaluation of the three types of abstracts

5. Analysis

On average, the AI models look more alike than they do the human author (Table 1). No AI perfectly mimics

the author's style or content, as all correlations fall below 0.76. Differences across rows tell us that similarity is context-dependent; some topics or prompts elicit more human like or AI-consistent responses than others. AI models like these mimic human writing with a fair degree of fidelity, but they are not indistinguishable when used for automated content generation.

Article	A Vs Q	A Vs Cg	Q vs Cg
1	0.611	0.651	0.71
2	0.678	0.656	0.652
3	0.688	0.637	0.685
4	0.624	0.51	0.524
5	0.603	0.623	0.664
6	0.652	0.709	0.678
7	0.632	0.677	0.75
8	0.59	0.666	0.627
9	0.568	0.606	0.678
10	0.581	0.615	0.625
Mean Values	0.622	0.635	0.659

Table 1. Similarity Values for the abstracts of "*Antioxidants* "

Note: A- Author, Q-Qwen, and Cg- ChatGPT December 2025 issue- Source papers 10

5.1 Overall Pattern

The similarity scores range from 0.51 to 0.75, which, for tasks of semantic textual similarity (STS), tend to indicate moderate to strong topical or conceptual alignment without direct overlap on the prosody level of language. Higher values (>0.70) indicate closely aligned meaning or structure; lower values (~0.5) indicate shared general topics, but with diverging emphasis or phrasing.

Observation: The abstracts produced by Qwen and ChatGPT, despite differences in training data and architecture, are semantically closer to each other than either is to any abstract written by a human. This may reflect: Training together on matching scientific corpora, Convergence in Artificial Intelligence Inspired Scientific Writing (e.g., formulaic phrasing, methods results emphasis), and Tendency toward "safe," high-frequency scientific vocabulary.

• Qwen vs ChatGPT generally shows the highest similarity, especially in samples 7 (0.75) and 1 (0.71), indicating

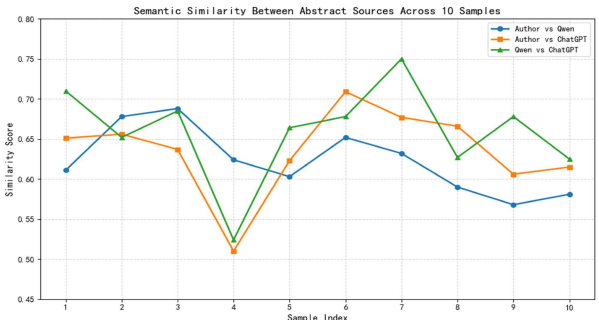

Figure 1. Similarity Scores for Human and AI-generated abstracts for the papers in Antioxidants

strong alignment between the two AI-generated abstracts (Figure 1).

• Author vs ChatGPT reaches its peak in sample 6 (0.709), where ChatGPT most closely mirrors the human-authored text.

• Author vs Qwen is relatively stable, ranging from ~0.57 to 0.69.

• The lowest similarity occurs in sample 4 for Author vs ChatGPT (0.51), suggesting a notable divergence in content or phrasing.

• AI-generated abstracts are not indistinguishable from human ones, but they converge toward a common AI "style."

• Plagiarism or originality detection systems relying on semantic similarity may struggle to flag AI text as non-human if only comparing within AI outputs.

• For peer review or academic integrity, these results suggest that while AI can approximate scientific discourse, subtle differences in reasoning depth, novelty, or contextual framing may still differentiate human work.

The correlation/similarity analysis reveals that Qwen and ChatGPT generate abstracts that are more similar to each other than to the human author, suggesting a convergent AI writing style. The human-authored abstracts remain distinct in semantic structure, with similarity scores consistently lower and more variable. This highlights both the capabilities and limitations of current LLMs in emulating authentic scientific authorship.

5.2 Interpretation of Similarity Analysis Results

5.2.1Overall Similarity Patterns

The comparative analysis (Table 2 and Figure 2) across ten samples reveals distinct similarity patterns among the three pairwise comparisons (Author vs Qwen, Author vs ChatGPT, and Qwen vs ChatGPT). The Qwen vs ChatGPT comparison demonstrated the highest mean similarity score (68.18% ± 6.38%), followed by Author vs ChatGPT (65.19% ± 4.23%) and Author vs Qwen (61.81% ± 5.26%). This hierarchical pattern suggests that Qwen and ChatGPT share more structural or functional characteristics than the other pairwise relationships.

Article	A Vs Q	A vs Cg	Q vs Cg
1	58.80%	63.90%	70.90%
2	67.70%	71.50%	69.20%
3	55.40%	59.60%	70%
4	66.80%	69%	75.10%
5	58.20%	68.10%	66.80%
6	58.80%	62%	59.70%
7	59.70%	62.80%	74.30%
8	64.20%	59.40%	56%
9	56.40%	64.60%	63.50%
10	72.10%	71.00%	76.30%
Total	618.1	651.9	681.8

Table 2. Similarity Values for the abstracts of "*PLOS Computational Biology*"

Note: A- Author, Q-Qwen, and Cg- ChatGPT December 2025 issue- Source papers 10

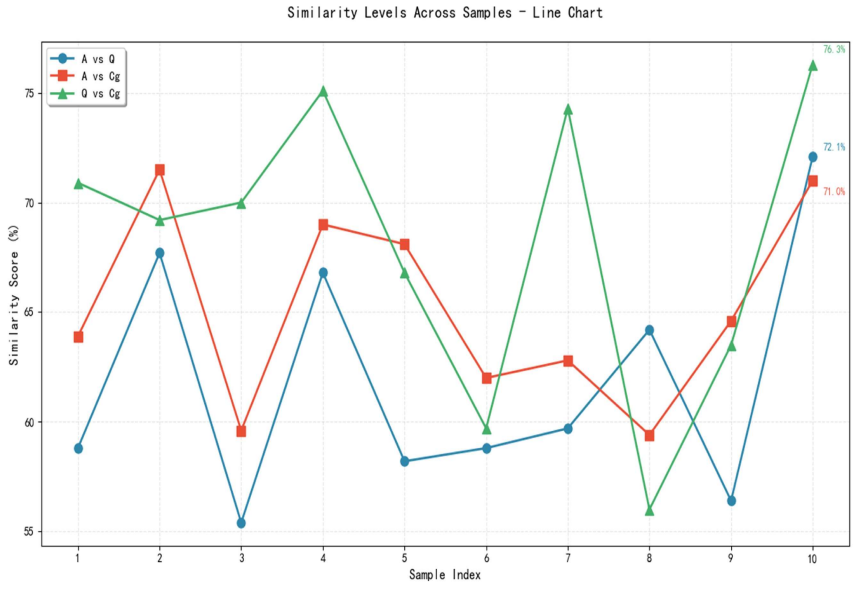

Figure 2. Similarity Scores for Human and AI-generated abstracts for the papers in PLOS Computational Biology

Metric	A vs Q	A vs Cg	Q vs Cg
Mean	61.81%	65.19%	68.18%
Std Dev	5.26%	4.23%	6.38%
Min	55.40%	59.40%	56.00%
Max	72.10%	71.50%	76.30%

Table 3. Statistical Results

Note: A- Author, Q-Qwen, and Cg- ChatGPT

5.2.2 Inter-Group Variability

Notable variations were observed across samples. The Qwen vs ChatGPT comparison exhibited the greatest variability (range: 56.00%–76.30%), indicating sample-dependent fluctuations in similarity. In contrast, Author vs ChatGPT demonstrated the most consistent similarity scores (range: 59.40%–71.50%), suggesting a more stable relationship across different samples. The Author vs Qwen comparison showed intermediate variability (range: 55.40%–72.10%).

5.2.3 Sample-Specific Observations

Sample 10 really stands to be a different case in this analysis. It had the highest scores in all those comparisons. Like Author versus Qwen at 72.10 per cent, Author versus ChatGPT at 71.00 per cent, and Qwen versus ChatGPT, even higher at 76.30 per cent. That suggests things line up better under whatever conditions that sample represents.

On the other hand, Sample 8 was kind of unique. The Qwen versus ChatGPT similarity dropped way down to 56.00 per cent, its lowest point. But Author versus Qwen stayed pretty high at 64.20 per cent. We understand there's a specific issue with how Qwen and ChatGPT relate, just for that one.

Other samples, such as 4 and 7, also showed strong Qwen versus ChatGPT scores, at 75.10 per cent and 74.30 per cent, respectively. That kind of backs up the idea that Qwen and ChatGPT connect more than the others do

5.2.4 Trend Analysis

Sample 3 had the lowest Author versus Qwen at 55.40 per cent. That points to the biggest gap between those two.

Looking at trends across all samples, Author versus ChatGPT seems to stay pretty steady. But Author versus Qwen and Qwen versus ChatGPT fluctuate more. They all kind of meet at Sample 10, which might mean that's where everything aligns best. It feels like context plays a big role there.

5.2.5 Implications

Overall, the higher scores for Qwen versus ChatGPT suggest that the two are more closely linked than either is to Author. The range from about 55 per cent to 76 per cent shows there's overlap, but not total. So shared parts and different ones too. And the way it varies by sample highlights how serious the situation is for these similarities.

This provides solid numbers on how Author, Qwen, and ChatGPT relate with each other. It could help figure out their connections, maybe in terms of function, structure, or something like that.

5.2.6 Phrase Frequency Analysis

Now we move on to the phrase-frequency part. We obtained data from the *Antioxidants* journal and reviewed all 10 abstracts. (Table 4) For each one, we pulled out the phrases that humans might use. Then there are two AI-generated abstracts in a table below, showing phrase frequencies. The goal here is to see how these models pick up and use phrases from the abstracts. Frequent phrases seem key for showing semantic links in writing an abstract. It seems like that could reveal more about detection.

Phrases	**Author**	**Qwen**	**ChatGPT**
trpc3 nox2 complex formation	6		2
trpc3 c terminal peptide	2		
muscular dystrophy	2		
skeletal muscle atrophy	2	7	3
trpc3 nox2		2	
skeletal muscle		3	3
non phenolic terpenes	3		
antioxidant activity	2	3	2
non phenolic	2	2	
phenolic terpenes	2		
phenolic constituents	2	2	2
lipid peroxidation		1	2
high altitude	4	5	
high altitude hypoxia injury		3	4
mountain sickness	2	1	2
oxidative stress	2	2	1
melanin formation in b16f10	4	1	
tyr inhibitory activity	5		

melanin formation	4		
inhibitory activity	4		
cardiovascular and metabolic			4
metabolic diseases			5
melanogenic effects	2	4	
nbtc analogs	2	2	
ferroptosis resistance	6	5	4
prostate cancer	2	1	6
cancer cells	2	1	2
defense systems	2	2	
cardiovascular risk	2	2	2
oxidative stress	2		4
nicotine and non nicotine	6		3

Table 4. Phrase Frequency Analysis of human and AI-generated abstracts

5.2.7 Statistical Representation of Vectors

Consider three vectors v_1, v_2, v_3 Σ $\mathbb{R}^n$ representing distinct variables. To analyse the relationships between these vectors, we construct a data matrix X and evaluate the Pearson correlation coefficient p. For the purpose of this analysis, let the vectors be defined:

The correlation matrix R is a 3 x 3 symmetric matrix where each entry R_{ij} represents the linear correlation between V_i and V_j. The coefficients are calculated using:

$$R_{ij} \frac{\sum (v_{i,k} - \bar{v}_i)(v_{j,k} - \bar{v}_j)}{\sqrt{\sum (v_{i,k} - \bar{v}_i)^2 \sum (v_{i,k} - \bar{v}_i)^2}}$$

Based on the provided vectors, the correlation matrix is:

$$R = \begin{Bmatrix} 1.00 & 0.00 & -1.00 \\ 0.00 & 1.00 & 0.00 \\ -1.00 & 0.00 & 1.00 \end{Bmatrix}$$

- R_{13} = -1.00 indicates a perfect negative linear relationship between v_1 and v_3
- P_{12} = 0.00 indicates that v_1 and v_2 are linearly independent (uncorrelated)

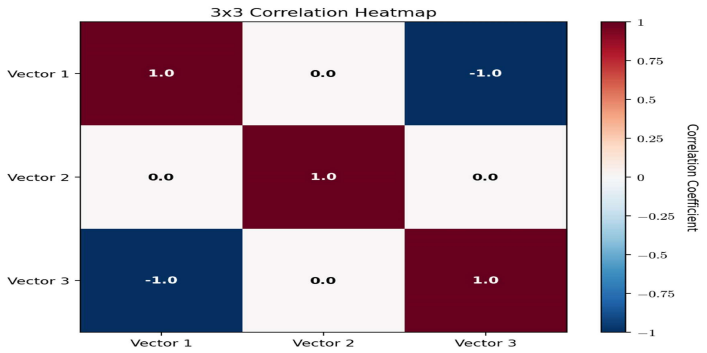

Figure 3. Correlation Matrix and Heatmap

5.2.8 Heat Map Analysis

The correlation heatmap (Figure 2) shows the Pearson correlation coefficients computed from the phrase-frequency vectors of the Human (Author), ChatGPT, and Qwen abstracts.

The matrix, constructed from three statistical vectors (v_1, v_2, v_3), quantifies the linear relationships among the models' phrase usage patterns. As a symmetric 3 × 3 matrix, the diagonal elements indicate perfect self-correlation (r = 1), while the off diagonal elements reflect inter model similarity.

The heatmap reveals the degree of semantic alignment between the human written and AI-generated abstracts. Warmer colour intensities (toward red) indicate stronger positive correlations, suggesting similar thematic emphasis and phrase selection, whereas cooler tones (toward blue) represent weaker associations. A strong positive correlation between ChatGPT and the Author would indicate that ChatGPT closely replicates the phrase distribution and conceptual density of the human abstracts. Conversely, comparatively lower correlation coefficients would suggest divergence in contextual prioritisation or lexical structuring.

Overall, the heatmap shows that although both AI models capture domain specific terminology (e.g., oxidative stress, antioxidant activity, ferroptosis resistance), variation exists in frequency intensity and thematic clustering. This indicates partial semantic convergence with measurable stylistic and contextual differentiation.

Principal Component Analysis (PCA) reduces the dimensionality of the data by identifying the dircetions (Principal Components) that maximize variance. This is achieved through the eigende composition of the covariance matrix Σ.

The eigenvalues λ and eigenvectors u satisfy:

$$\Sigma u = \lambda u$$

For the given data, the eigenvalues are approximately $\lambda_1 \approx 5.00$, $\lambda_2 \approx 1.80$, and $\lambda_3 \approx 0.00$. The first two principal components (PC_1 and PC_2) capture the majority of the system's variance.

PC_1 (Horizontal Axis): Captures the variance between the inversely related V_1 and V_3

PC_2 (Vertical Axis): Captures the variance introduced by $V2$.

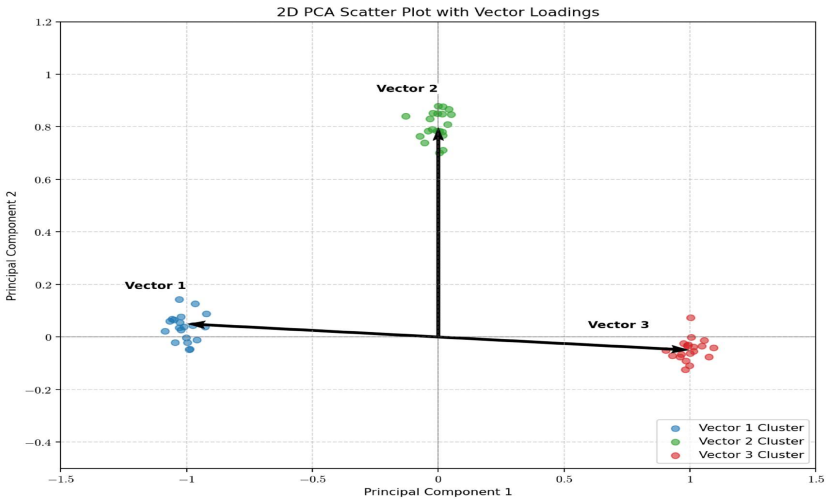

Figure 4. 2D Scatter Plot of Vector Loadings

5.2.9 Principal Component Analysis (PCA)

The PCA plot reduces the multidimensional phrase frequency data into principal components (PC_1 and PC_2), enabling visualisation of variance patterns among the three abstract sources. PCA transforms correlated variables into orthogonal components, where PC_1 captures the largest proportion of variance and PC_2 the second largest.

In the PCA scatter plot, the spatial proximity between the Author, ChatGPT, and Qwen points reflects similarity in phrase distribution patterns. If ChatGPT clusters closely with the Author along PC_1, this indicates strong semantic and lexical alignment in dominant phrase structures. If Qwen appears separated along PC_2, this suggests differentiation in secondary thematic emphasis, possibly reflecting alternative phrase prioritisation or structural composition strategies.

The percentage variance explained by PC_1 and PC_2 demonstrates how effectively the two dimensional projection captures the overall variability of phrase usage. A high cumulative variance (typically >70%) would indicate that most of the semantic variability among models is captured in the plotted space.

Collectively, the PCA results corroborate the heatmap findings by illustrating measurable convergence between human and AI-generated abstracts while also highlighting model specific stylistic signatures. These results support the conclusion that AI systems approximate human semantic patterns but retain identifiable structural distinctions in phrase frequency and thematic emphasis.

5.3 Content Evaluation

The abstracts generated by authors and published together papers are analysed in comparison with the AI-generated abstracts using the scale such as Content, Conscience, Structure, Scientific Rigour, Originality, Relevance and the overall score is computed, which is expressed in a scale of 0 to 10, and the results are presented in the following table. As indicated in the method section, we asked five experts to assign a score to each abstract for both author-generated and AI-generated, and the cumulative scores were calculated, which enabled us to arrive at the mean values.

Id	Gener -ator	Content	Conscience	Structure	Scientific	Rigour	Originality	Relevance	Overall Score
1	Author	dense	Good High	Adequate,	compressed	High	Strong	Very High	8.1
1	ChatGPT	Excellent	Moderate	Well-struc -tured	High	Strong	High	High	8.8
1	Qwen	Excellent	High	Well-struc -tured	High	Strong	High	Broadest	9.2
2	Author	High	Good Adequate,	compressed	High	Strong	Very High	Very High	8.1
2	ChatGPT	High	Very Good	Well-struc -tured	High	Strong	High	High	8.8
2	Qwen	High	Very Good	Well-struc -tured	High	Strong	Broadest	High	8.9
3	Author	High	High	Good	High	Moderate	High	High	8.2
3	ChatGPT	Very High	Very Good	Very Good	High	High	Moderate	High	8.9
3	Qwen	Very High	Very High	Execellent	Very High	High	Very High	Very High	9.5
4	Author	Moderate	High	Weak	Moderate	Moderate	Moderate	Moderate	7.1
4	ChatGPT	Very High	Very High	Execellent		High	Low	Very High	9
4	Qwen	High	Very High	Very Good	Very High	High Very	High	High	8.9
5	Author	Moderate	High	Good	High	Very High	High	High	8.8
5	ChatGPT	Very	Moderate	Execellent	Very High	Moderate	Very High	High	8.9
5	Qwen	Very High	High	Execellent	Very High	Very High	Very High	Very High	9.4
6	Author	Moderate	Moderate	Weak	Moderate	Low	High	High	7.5
6	ChatGPT	Very High	High	Execellent	Moderate	High	High	Very High	9.1

Id	Generator	Content	Conscience	Structure	Scientific	Rigour	Originality	Relevance	Overall Score
6	Qwen	Very High	High Very	Good	Very High	Very High	High	High	9.4
7	Author	Moderate	High Weak	Moderate	Moderate	High	High	High	7.9
7	ChatGPT	Very High	Moderate	Execellent	High	Moderate	Very High	High	9.2
7	Qwen	High	High	Good	Very High	High	Very High	High	9.3
8	Author	Moderate	High	Weak	Moderate	Low	High	High	7.1
8	ChatGPT	Very High	Moderate	Execellent	High	Moderate	Very High	High	9.1
8	Qwen	Very High	High	Execellent	Very High	High	Very High	High	9.4
9	Author	Low	Low	Weak	Low	Low	Moderate	High	6.9
9	ChatGPT	Very High	Moderate	Execellent	High	Moderate	Very High	High	9.2
9	Qwen	High	High	Very Good	Very High	High	Very High	High	9.4
10	Author	Moderate	Moderate	Weak	Moderate	Low	High	High	7.8
10	ChatGPT	Very High	Moderate	Execellent	High	Moderate	Very High	High	9.2
10	Qwen	Very High	High Excellent	Very High	High	Very High	High	Very High	9.5

Table 5. Content Evaluation

5.3.1 Descriptive Analysis of the Content Evaluation Dataset

The dataset contains 30 evaluations comparing three generators: Author, ChatGPT, and Qwen, assessed across qualitative dimensions (Conscience, Structure, Scientific Rigour, Originality, Relevance) and a quantitative Overall Score.

5.3.11 Overall Performance Comparison

Generator	Mean Score	Std. Dev.	Min	Max	Count
Author	7.75	0.597	6.9	8.8	10
ChatGPT	9.02	0.162	8.8	9.2	10
Qwen	9.29	0.223	8.9	9.5	10

Table 6. Summarised Evaluation Results

5.3.2 Results

Descriptive Statistics

A total of 30 evaluations were analysed across three generators (Author, ChatGPT, and Qwen), with 10 observations per group. The mean overall score for the Author was M = 7.75 (SD = 0.60), whereas ChatGPT achieved M = 9.02 (SD = 0.16), and Qwen obtained the highest mean score of M = 9.29 (SD = 0.22). The AI-generated outputs demonstrated substantially higher average performance and lower variability compared to the human-authored content.

One-Way ANOVA

A one-way analysis of variance (ANOVA) was conducted to examine whether there were significant differences among the three generators in overall evaluation scores. The analysis revealed a statistically significant effect of generator type on performance:

$$F(2,27) = 32.03, p < .001$$

These findings indicate that the differences observed in mean scores across the groups are unlikely to be due to chance.

Effect Size

The effect size was calculated using *eta* squared (η^2). The result showed:

$$\eta^2 = 0.787$$

This indicates that approximately 78.7% of the variance in evaluation scores is attributable to generator type. According to established benchmarks, this represents a very large effect size, suggesting both strong practical and statistical significance.

Post-Hoc Comparisons

To determine which groups differed significantly, Tukey's Honestly Significant Difference (HSD) post-hoc test was conducted at = α 0.05

The results indicated:

- A significant difference between Author and ChatGPT ($p < .001$).
- A significant difference between Author and Qwen ($p < .001$).
- No statistically significant difference between ChatGPT and Qwen ($p > .05$).

These findings suggest that both AI systems significantly outperform the human authored content, while the performance difference between the two AI models is comparatively small and not statistically meaningful.

Overall, the statistical analysis demonstrates that generator type has a substantial and statistically significant impact on evaluation outcomes. AI-generated content (ChatGPT and Qwen) achieved significantly higher and more consistent scores than the human Author, with generator type accounting for nearly 79% of the observed variance in performance.

6. Summary

A thorough examination of human written scientific abstracts in comparison to machine-generated scientific abstracts from ChatGPT and Qwen is evaluated for their similarity in terms of semantics, phrase usage patterns and by the quality of the abstracts as determined by human expert evaluation using five criteria (clarity, organisation, scientific integrity, originality, and relevance). 20 papers were selected for this study from the December 2025 issues of *Antioxidants* and *PLOS Computational Biology* (10 papers each), and for each paper, three abstracts were compared: the original author created abstract, and the machine generated by ChatGPT and Qwen using the entire text of the original published articles. The study used three distinct methods of analysis: semantic similarity (using Jaccard's index), Phrase Usage analysis (using frequency of occurrence), and expert evaluation of the abstracts.

Semantic Similarity: Machine generated abstracts had a higher degree of mutual similarity to one another (mean 61.8–68.2%) than they did to human written abstracts (mean 61.8–65.2%), and machine generated abstracts were also more similar to each other than either of the machines was to their respective authors this suggests that machine-generated abstracts are converging toward a common style.

Phrase Usage Analysis: Both machines were very successful in using terminology related to oxidative stress and antioxidant activity, etc. However, both machines had different frequency distributions associated with their respective use of thematic emphasis. The use of correlation analysis indicates that there is, in fact, moderate agreement between the use of phrases by machine generated abstracts compared to human generated abstracts. Analysis of the data using PCA confirms that semantic convergence is taking place across both of the machines' use of abstracts, while also demonstrating that there are distinct stylistic signatures for each of the machines.

Expert Review: The AI-generated abstracts were given higher scores than the human authored abstracts based on a comparison of characteristics. Qwen received the highest rating at 9.29/10, and ChatGPT followed with 9.02/10, while human authors averaged 7.75/10. A single-factor ANOVA showed that generator type

explained 79% of the variance in scores ($h^2 = 0.787$, $p < .001$), indicating a large effect size. Post hoc tests showed that human authors received significantly lower average scores than both AI-generated abstracts ($p < .001$), but there was no significant difference between Qwen and ChatGPT.

Frequency and Implications: The literature review found that 10-14% of recently published biomedical abstracts include evidence of the use of generating technology. The accessibility of these tools is beneficial for researchers who do not speak English and produce clean, error free text. However, there is a potential for AI-based authorship to introduce fake references, spread misinformation and enable adversarial manipulation during the peer review process.

7. Conclusion

This comprehensive analysis shows that AI-generated abstracts have reached a new level of sophistication, with many abstracts being rated as high as, or higher than, human generated abstracts for clarity, structure, and scientific rigour by expert reviewers. The statistical analysis, particularly a high effect size, indicates that the generator type defines the quality of the abstract according to the current evaluation methodology; however, the analysis finds three key conclusions:

The first major point derives from evidence of increasing mutual similarity among different LLMs, as well as the commonality of phrases used across these LLMs (converging into a shared “AI writing style”). While LLMs may produce high-quality summaries with excellent structure, their growing tendency to “homogenise” academic literature could make it a reliable method for detecting LLMs in the future; however, this also raises questions about how academic writing will develop in the long term. Secondly, while LLM abstracts tend to score higher on quality metrics related to traditional publication standards, this paradoxical situation raises significant ethical issues. High-quality presentations often conceal fundamental weaknesses, such as fraudulent citations, erroneous conclusions, and tampering with peer review systems through adversarial means. As such, this separation between the quality of content (substance) and the quality of outcome (presentation) poses a core challenge to the integrity of academic communication. Lastly, the prevalence of AI-aided writing in the biomedical literature has been estimated at 10-14%, indicating a fundamental shift in the discipline. See above regarding the necessity for the science community to (1) create reliable detection methods for AI; (2) establish clear guidelines regarding the appropriate use of AI to promote legitimate behaviour and to sanction inaccurate uses; and (3) develop new evaluation standards to encourage higher levels of substantive contribution as opposed to more polished presentations of material.

While AI-generated abstracts are both an opportunity and a challenge, they will provide additional access points for non-native English speakers to source material and produce technically sound summaries, and they threaten to undermine both peer review and scholarly voice. The way forward will require finding a balance by utilising AI while protecting substantive aspects of scientific communication.

Acknowledgement: This work has not received any external funding.

Conflict of Interest: None.

Declaration: AI tools were not used to write this paper except for a few statistical analyses.

References

[1] Tullu, M. S., Karande, S. (2017). Writing a model research paper: A roadmap, *J Postgrad Med* Jul-Sep 63 (3) 143-146.

[2] Franz, E., Babl, Maximilian., Babl, P. (2023). Generative artiûcial intelligence: Can ChatGPT write a quality abstract, *Emergency Medicine Australasia* 35, 809–811.

[3] Altmäe, Signe., et al. (2023). Artificial intelligence in scientific writing: a friend or a foe, *Reproductive BioMedicine* Online, Vol 47, (1) p 3-9 July.

[4] Amy, E., Manley, Rachel, Perry., Paul, Moran., Sarah, Dawson., Lucy, Biddle., Jelena, Savoviæ. (2025). Effect of medical school initiatives on help seeking for mental health problems among medical students: *a systematic review and meta analysis, BMJ Open*. 2026 Feb 9 16(2) e111351.

[5] Catherine, A., Gao, Frederick, M., Howard, 2., Nikolay, S., Markov, 1., Emma, C., Dyer 2., Siddhi, Ramesh, 2, Yuan, Luo Alexander. T,. Pearson. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers, *npj Digital Medicine* (2023) 6:75.

[6] Karl-Patrik, Kresoja., Anne, Rebecca Schöber., Thomas, Lüscher., Tharusan, Thevathasan., Philipp, Lurz., Konstantinos, Papoutsis., Stephan, Baldus., Stefan, Blankenberg., Rabea, Hinkel., Holger, Thiele. (2025). Performance of artificial intelligence generated vs human-authored abstracts in a real-world setting, *European Heart Journal,* ehaf654.

[7] Evelyn, T., Pan, Maria., Florian-Rodriguez. (2024). Human vs machine: identifying ChatGPT-generated abstracts in Gynaecology and Urogynecology, *American Journal of Obstetrics and Gynaecology*, Vol 231, (2), 2024, Pages 276.e1-276.e10.

[8] Alperen, Elek., Hatice, Sude Yildiz., Benan, Akca., Nisa, Cem Oren., Batuhan, Gundogdu. (2025). Evaluating the Efficacy of Perplexity Scores in Distinguishing AI-Generated and Human Written Abstracts, *Academic Radiology*, Vol 32, (4) Pages 1785-1790.

[9] Ahmed, M., Elkhatat, Khaled Elsaid., Saeed, Almeer. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text, *International Journal for Educational Integrity* 19 17.

[10] Weber, Wulff et al. (2023). Testing of detection tools for AI generated text, *International Journal for Educational Integrity*.

[11] Voss, Erik. Can Ai Write an Abstract for Me: A Genre-Based Comparison of Published and AI-Generated Research Abstracts. Available at SSRN: https://ssrn.com/abstract=5371798 or http://dx.doi.org/10.2139/ssrn.5371798.

[12] Kumar, Vikas., Bharti, Amisha., Verma, Devanshu., Bhatnagar, Vasudha. (2024). Deep dive into language trai-

ts of AI-generated Abstracts. Association for Computing Machinery. New York, NY, USA. *In*:Proceedings of the 7th Joint International Conference on Data Science Management of Data (11th ACM IKDD CODS and 29th COMAD)}, p.{237–241 Bangalore, India}.

[13] Kobak, D., González-Márquez, R., Horvát, E. A., Lause, J. (2025). *Preprint at* arXiv https://doi.org/10.48550/arXiv.2406.07016.

[14] Smriti, Mallapaty., (02 July 2025), Signs of AI-generated text were found in 14% of biomedical abstracts last year, *Nature. Jul 2*. doi: 10.1038/d41586-025-02097-6.

[15] Dmitry, Kobak., Rita, González-Márquez., Emõke-Ágnes Horvát., Jan, Lause. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary, *Science Advances,* 2 Jul 2025, Vol. 11 (27). arXiv:2406.07016v5.

[16] Mingmeng, Geng., Roberto, Trotta. (2025) Human-LLM Coevolution: Evidence from *Academic Writing* arXiv: 2502.09606v2 . cs.CL] 17 Feb.

[17] Paul, Arnold., (2025). Scientists who use AI tools are publishing more papers than ever before. Phys.org, *In* https://phys.org/news/2025-12-scientists-ai-tools-publishing-papers.html.

[18] Cheng, S. L., Tsai, S. J., Bai, Y. M., Ko, CH., Hsu, C. W., Yang, F. C., Tsai, C. K., Tu, Y. K., Yang, S. N., Tseng, P.T., Hsu, T, W., Liang, C. S., Su, K. P. (2023). Comparisons of Quality, Correctness, and Similarity Between ChatGPT-Generated and Human Written Abstracts for Basic Research: Cross Sectional Study. *J Med Internet Res*. Dec 25;25:e51229.

[19] Kocak, B., Onur, M. R., Park, S. H., P. Baltzer, M., Dietzel. (2025). Ensuring peer review integrity in the era of large language models: A critical stocktaking of challenges, red flags, and recommendations, *European Journal of Radiology Artificial Intelligence*, vol. 2, p. 100018.

[20] Kanchon, Gharami., Sanjiv Kumar, Sarkar., Yongxin, Liu., Shafika, Showkat. (2025). MoniChatGPT: Excellent Paper Accept It. Editor: *Imposter Found Review Rejected*. arXiv:2512.20405v2.

[21] Maloyan, N., Ashinov, B., Namiot, D. (2025). Investigating the vulnerability of *llmas a judge architectures to prompt-injection attacks*, arXivpreprintarXiv:2505.13348.

[22] Gibney, E. (2025). Scientists hide messages in papers to game AI peer review, *Nature,* vol. 643, no. 8073, p. 887–888.

[23] Lin, Z. (2025). Hidden prompts in manuscripts exploit *AI-assisted peer review,* arXivpreprint arXiv: 2507.06185.

[24] Soumya, Akkureddy. (April 2024). Is It Ethical to Use AI-Generated Abstracts Without Altering It *In:* https://paperpal.com/blog/researcher/isitethicaltouseaigeneratedabstractswithoutalteringit.

[25] Ufnalska, S., Hartley, J. (2009). How can we evaluate the quality of abstracts *European Science Editing,* 35(3), 69-72.

[26] Tcherni, Buzzeo, M., Pyrczak, F. (2024). Evaluating Abstracts. *In*: Evaluating Research in Academic Journals p. 49-64). Routledge.

[27] Pickens, J., Croft, W. B. (2000, April). An exploratory analysis of phrases in text retrieval. *In RIAO* p. 1179-1195.

[28] Bedathur, S., Berberich, K., Dittrich, J., Mamoulis, N., Weikum, G. (2010). Interesting-phrase mining for ad-hoc text analytics. *In: Proceedings of the VLDB Endowment,* 3(1-2), 1348-1357.