International Journal of Scientific Research and Modern Education (IJSRME)
International Peer Reviewed - Refereed Research Journal, Website: www.crystalpen.in
Impact Factor: 7.137, ISSN (Online): 2455 - 5630, Volume 11, Issue 2, July - December, 2026

SCIENTIFIC EVIDENCE CHARACTERISTICS AND THEIR ASSOCIATION WITH SUPPORTIVE VERSUS REFUTING ATTRIBUTION IN AUTOMATED CLAIM VERIFICATION

Divya Vishwanath* & Paul Johnson Asamoah**
* Janardhan Rai Nagar Rajasthan Vidyapeeth, Udaipur, Rajasthan, India
** Power Life University College, Accra, Ghana, West Africa

Cite This Article: Divya Vishwanath & Paul Johnson Asamoah, "Scientific Evidence Characteristics and Their Association with Supportive Versus Refuting Attribution in Automated Claim Verification", International Journal of Scientific Research and Modern Education, Volume 11, Issue 2, July - December, Page Number 14-22, 2026.
Copy Right: © Crystal Pen Publication, 2026 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Type of Review: Peer Reviewed as per |C|O|P|E| Guidance.
Disclaimer: The scholarly papers reviewed and published by Crystal Pen Publication, India, reflect the views and opinions of their respective authors and do not necessarily represent the views or opinions of Crystal Pen Publication. The publisher disclaims any responsibility for any harm, loss, or damage resulting from the use of the published content by any party.
DOI: https://doi.org/10.5281/zenodo.24050612

Abstract:

This analysis examines whether scientific abstract characteristics distinguish supporting from refuting evidence, addressing the interpretation of document presentation in scientific verification. Using the 2020 SciFact benchmark, it analyses 773 annotated claim document pairs covering 693 claims and 437 abstracts through descriptive comparisons, document weighting and partition sensitivity checks. Refuting pairs contain abstracts averaging 32.46 words longer than supporting pairs. The refuting share is 14.1 percentage points higher among structured abstracts, while mean relative evidence position differs by only 0.007. Equal document weighting preserves these pooled directions, but length and structure differences narrow substantially in the development partition. Furthermore, 179 abstracts support some claims and refute others despite unchanged document characteristics. These findings distinguish document presentation from the semantic relationship between evidence and a specific claim. The contribution is a descriptive account of how pooled presentation patterns coexist with partition sensitivity and claim dependent stance. Evidence services and benchmark developers should preserve claim specific provenance and disclose document overlap and weighting choices when reporting aggregate comparisons.

Key Words: Scientific Claim Verification; Evidence Attribution; Abstract Structure; Descriptive Analysis; Benchmark Sensitivity

1. Introduction:

Scientific evidence supports decisions only when the relationship between a claim and its source is interpreted correctly. An abstract can support one proposition while contradicting another, making evidence attribution a relational task rather than a judgement based on document appearance. For researchers, evidence services and institutions developing automated verification systems, this distinction matters because observable presentation characteristics may coincide with annotated outcomes without explaining their substantive meaning. Guo et al., 2022, position automated fact checking around the interaction between claims, evidence and verdicts. This analysis examines that interaction through the document characteristics accompanying supporting and refuting scientific evidence. Scientific claim verification provides a setting in which evidence relationships can be examined directly. Wadden et al., 2020, introduced SciFact, connecting scientific claims with abstracts, stance annotations and rationale sentences. Its contribution was to make evidence retrieval, rationale identification and claim verification jointly observable within a benchmark. These annotations also permit a complementary question: how do the characteristics of evidence documents differ between supporting and refuting claim relationships? Answering this question helps distinguish properties of benchmark composition from the semantic reasoning required to interpret evidence.

The closest verification research emphasises coordination between evidence selection and stance prediction. Zhang et al., 2021, developed a joint framework linking abstract retrieval, rationale selection and stance classification to address information loss and error propagation across separate tasks. Wadden et al., 2022, subsequently incorporated document context into joint rationale and label prediction. These approaches establish the importance of representing evidence in context. Their modelling contributions do not, by themselves, establish whether abstract length, rationale position or structured presentation consistently distinguish supporting from refuting observations. The present analysis addresses this descriptive question rather than evaluating another verification model. The conceptual logic separates textual extent, evidence placement and document organisation. Longer abstracts may provide qualifications that clarify whether a claim follows from reported findings, but they may also contain additional background unrelated to that judgement. Evidence position indicates where an annotated rationale occurs, yet early and concluding statements can each support or contradict a proposition. Structured abstracts organise information into identifiable components, potentially making its function more visible. However, Hartley, 2014, distinguishes the informational advantages of structured presentation from continuing questions about accuracy and omissions. Consequently, none of these characteristics supplies a defensible directional rule for stance without examining the observed claim relationships.

The measurement problem is therefore as important as the substantive comparison. Abstract length and structure belong to the document, while stance belongs to the relationship between that document and a particular claim. Relative evidence position occupies an intermediate analytical level because it combines document length with the location of an annotated rationale. Treating these measures as interchangeable indicators of evidence quality would obscure their different meanings. This analysis preserves those distinctions by retaining separate observed variables and examining their distributions across annotated stance categories. The reviewed evidence does not establish whether these presentation differences persist when repeated documents receive equal total weight or when benchmark partitions are examined separately. This is a bounded empirical gap concerning the stability of descriptive relationships. Repeated abstracts can contribute several observations, while differences in partition

composition can alter pooled summaries. Examining both features allows the analysis to determine whether an apparent document pattern remains visible under alternative representations of the same evidence collection.

This study aims to characterise how abstract word count, relative evidence position and structured presentation differ between supporting and refuting scientific claim document pairs. The central research question asks whether these observable characteristics distinguish the two stance categories and how sensitive the comparisons are to document weighting and source partition. The specific objectives are to compare abstract length across stance categories, compare the relative location of the first annotated rationale, and compare refuting proportions between structured and unstructured abstracts. These objectives require descriptive comparisons rather than directional causal hypotheses. The empirical setting comprises 773 annotated claim document pairs covering 693 claims and 437 abstracts from the training and development partitions of SciFact, released in 2020. The analysis uses a cross section of benchmark evidence rather than annual observations. It compares distributions, central tendency and proportions, then assesses sensitivity through equal total weighting of documents and separate partition summaries. Source identifiers and rationale sentences preserve the connection between each observation and its evidence. This design enables an auditable description of benchmark relationships without attributing differences to unmeasured research quality, institutional practices or verification performance.

The completed analysis contributes by showing that pooled presentation patterns and their stability across samples are distinct empirical questions. Refuting pairs contain longer abstracts and a higher share of structured observations in the combined sample, while evidence position provides little separation. The length and structure differences persist under equal document weighting but become substantially smaller in the development partition. Moreover, 179 abstracts support some claims and refute others despite retaining the same length and format. Together, these findings qualify the interpretation of document characteristics as general indicators of stance and demonstrate the importance of preserving the claim as part of the analytical unit. The practical relevance extends to organisations constructing scientific evidence repositories and verification benchmarks. Reporting document characteristics alongside weighting and partition comparisons can reveal when pooled patterns depend on how evidence is represented. Retaining claim identifiers, source sentences and stance annotations together allows users to inspect those relationships. The transferable contribution is this approach to evidence auditing, while the observed magnitudes remain conditional on the SciFact sample.

2. Theory, Conceptual Framework and Research Questions Development:

Scientific evidence attribution requires an inference about the relationship between a claim and the findings reported in a source. An abstract becomes supporting or refuting evidence through its relevance to a particular proposition, rather than through its scientific status alone. Wadden et al., 2020, operationalised this relationship through scientific claims, evidence labels and annotated rationale sentences. This analysis adopts that relational foundation and examines whether observable abstract characteristics differ between supporting and refuting claim document pairs. The conceptual framework distinguishes evidence content from evidence presentation. Content determines whether the reported findings support or contradict a claim, while presentation determines how much information appears and where relevant statements occur. Abstract word count, relative evidence position and abstract structure represent observed presentation characteristics. Refuting evidence attribution represents the annotated semantic relationship, coded one for contradiction and zero for support. These measures permit an examination of associations without assuming that document presentation determines scientific truth. Abstract word count captures the amount of text available for interpreting a claim. Additional text can supply methodological qualifications, population restrictions and contextual information that narrow the meaning of a reported finding. Conversely, additional background can increase the distance between a claim and its relevant evidence. The competing mechanisms concern contextual completeness and information dispersion, so abstract length alone does not establish an expected direction of association with refutation.

Context can influence how individual statements are interpreted. Wadden et al., 2022, developed a verification approach that jointly identifies rationales and assigns evidence labels using the surrounding document context. Their results establish the relevance of contextual representation for verification systems. They do not establish that longer abstracts contain a greater proportion of refuting evidence. This analysis therefore treats word count as a descriptive document characteristic rather than a measure of evidential strength. The first research question is: How does abstract word count differ between supporting and refuting claim document pairs? The available word counts and binary evidence labels resolve this question through comparisons of distributions, central tendency and dispersion. Any observed difference describes the composition of the annotated evidence rather than an effect of length on scientific conclusions. Relative evidence position captures where the first annotated rationale occurs within an abstract. It equals the position of the earliest evidence sentence, counted from one, divided by the total number of abstract sentences. This normalisation distinguishes location from document length. Evidence appearing near the beginning may state an established relationship, whereas later evidence may report results, qualifications or conclusions.

These possibilities do not make sentence position a direct indicator of support or contradiction. A concluding statement can affirm or reject a proposition, depending on the claim being assessed. Zhang et al., 2021, showed the importance of coordinating abstract retrieval, rationale selection and stance prediction. Their framework supports treating rationale location and evidence stance as related components of verification while preserving their conceptual distinction. The second research question is: How does the relative position of the first evidence sentence differ between supporting and refuting claim document pairs? Because this measure uses annotated rationales, the comparison describes evidence placement after identification. It does not measure the ability of a system to locate evidence without annotations. Abstract structure captures whether the source uses an explicitly organised abstract format. Its theoretical relevance lies in information architecture: separating background, methods, results and conclusions can make the function of individual statements more visible. Hartley, 2014, reviewed evidence that structured abstracts can improve information presentation, while also identifying continuing concerns about omissions and accuracy. Structure therefore describes organisation rather than a guarantee of evidential reliability.

Reporting guidance provides a complementary institutional mechanism. Beller et al., 2013, specified information that systematic review abstracts should communicate so readers can interpret objectives, methods, findings and limitations. Such conventions can shape the visibility of evidence without determining whether it supports a particular claim. The binary structure

indicator records the presence of organisation; it does not measure compliance with a reporting checklist or the quality of the underlying research. The third research question is: How does the proportion of refuting evidence differ between structured and unstructured abstracts? The available structure indicator and evidence labels permit a direct comparison of proportions. This question concerns the distribution of annotated stance across document formats, without attributing differences to reporting requirements that were not recorded. The resulting conceptual framework connects three observed abstract characteristics with one observed evidence outcome through descriptive associations. Word count represents textual extent, relative evidence position represents rationale placement, and structure represents document organisation. These characteristics remain distinct measures rather than components of an assumed latent index. Interpretation remains conditional on the annotated scientific evidence sample and recognises that claims and documents recur across observations. The framework therefore supports descriptive answers to the three research questions without converting document characteristics into causal explanations.

3. Data:

This analysis establishes source provenance, documents sample construction and defines the observed measures used to describe scientific evidence attribution. It preserves the connection between each claim, its annotated stance and the source sentences supporting that annotation.

3.1 Data Sources, Population, and Sample Construction:

The analysis uses SciFact, released by the Allen Institute for AI in 2020. The source comprises a scientific abstract corpus and claims annotated with supporting or contradicting evidence. Wadden et al., 2020, introduced the benchmark to connect scientific claims with evidence documents and explanatory rationales. The analytical files are corpus.jsonl, claims train.jsonl and claims dev.jsonl, obtained through the public allenai/scifact repository and its linked release archive. The repository licences claims and annotations under Creative Commons Attribution 4.0 and corpus abstracts under Open Data Commons Attribution 1.0.
The target population comprises scientific claim and abstract relationships with identifiable supporting or refuting evidence. The accessible sampling frame is narrower: the claims and abstracts distributed in SciFact. Selection includes every training or development claim with at least one annotated evidence document. Claims without annotated evidence are excluded because they cannot supply the binary outcome and rationale position required by the analysis. The public test partition is excluded because its evidence labels are unavailable. Alternative cross validation partitions are not appended because they redistribute existing records.
The unit of analysis is an annotated claim document pair. The data form a cross section drawn from a benchmark released in 2020, within the declared 2010-2025 scope. They do not provide annual observations across that interval. Article publication years and geographical identifiers are absent from the analytical source fields, so publication period and country coverage cannot be established for individual documents. The empirical setting is scientific literature, with no country-based sampling restriction imposed during construction. The source corpus contains 5,183 unique abstracts. The claim files contain 809 training claims, 300 development claims and 300 test claims. Among the training and development records, 693 claims have annotated evidence. Expanding their evidence mappings produces 773 claim document pairs linked to 437 distinct abstracts. Training contributes 564 pairs from 505 claims, while development contributes 209 pairs from 188 claims.

Table 1: Sample construction

Stage	Units	Observations	Reason or Rule
Distributed claim frame	1,409 claims	1,409 claim records	Training, development and public test partitions
Excluded public test partition	300 claims	300 claim records removed	Evidence annotations unavailable
Accessible annotated partitions	1,109 claims	1,109 claim records	Training and development retained for evidence screening
Excluded claims without annotated evidence	416 claims	416 claim records removed	No usable stance and rationale combination
Claims retained for evidence expansion	693 claims	693 claim records	At least one annotated evidence document
Expanded and integrated sample	693 claims and 437 abstracts	773 claim document pairs	One observation per annotated claim and document combination
Final descriptive sample	693 claims and 437 abstracts	773 claim document pairs	Complete values for all four analytical measures

The release archive is identified by its SHA256 checksum in the workbook provenance. This identifies the downloaded content more precisely than the mutable latest release address. The extraction date is also recorded in that provenance record and is distinct from the benchmark release year.

3.2 Data Integration, Cleaning, and Quality Assurance:

Integration uses the document identifier linking each claim's evidence mapping to the abstract corpus. Each claim can reference several evidence documents, while one abstract can contribute evidence to several claims. The intended relationship therefore permits repeated claim and document identifiers while requiring a unique combination of source partition, claim identifier and document identifier. All 773 retained pairs have distinct record identifiers and matching source abstracts. When a claim document pair contains several annotated rationales, construction combines their sentence indices into a sorted union. Repeated indices are retained once. The pair contributes one outcome label because its rationale annotations share the same stance. The original claim text, stance, document identifier, sentence indices and evidence text remain available for inspection. This treatment follows the source schema, which distinguishes evidence annotations from documents merely cited in the originating citation context.

Quality checks reconstructed abstract word counts, abstract sentence counts, relative evidence positions and binary outcomes from the workbook's source sentence records. These checks identified no discrepancies in the reconstructed measures,

no unmatched document identifiers and no evidence indices outside their corresponding abstracts. The evidence sheet contains 4,166 original sentences across the 437 retained documents. All four analytical variables have 773 valid observations. Construction uses neither imputation nor interpolation. A zero outcome denotes supporting evidence, and a zero-structure indicator denotes an unstructured abstract. Neither value represents missing information. Records lacking evidence annotations are excluded rather than assigned a supporting label. Abstract word counts range from 68 to 1,065, and these observed values remain unchanged. Relative positions are greater than zero and no greater than one. The binary indicators contain only zero and one. Selection and transformations were established during workbook construction, before calculation of the descriptive summaries presented here. Repeated sources require care in interpreting the sample. A total of 110 abstracts occur in both the training and development contributions. These are genuine repeated documents across claim relationships, rather than duplicate analytical records. Accordingly, source partition labels do not establish independence between the two contributions.

3.3 Variable Measurement, Construction, and Validation:

The framework uses three observed document characteristics and one annotated evidence outcome. Abstract word count measures textual extent. Relative evidence position measures the location of the first annotated rationale. Abstract structure records the corpus indicator for a structured abstract. These measures remain separate and are not combined into a composite index. Word count is calculated after joining the original abstract sentences with spaces and splitting the resulting text at whitespace. It counts textual tokens under this rule rather than tokens generated by a language model. Relative evidence position converts the earliest source sentence index from zero based indexing to ordinary sentence order and divides that position by the abstract sentence count. Equation 1 defines relative evidence position.

$$R_i = \frac{F_i}{N_i}$$

Here, R_i denotes relative evidence position for pair i, F_i denotes the first annotated evidence sentence counted from one, and N_i denotes the number of sentences in its abstract. A value of one place the first rationale in the final sentence. Smaller values indicate earlier placement. The dependent variable equals one when the source label is CONTRADICT and zero when it is SUPPORT. It records an abstract's relationship to a particular claim, rather than the accuracy of the abstract itself. Zhang et al., 2021, distinguish rationale selection from stance prediction, supporting the conceptual separation between evidence location and evidence orientation used here.

Table 2: Variable dictionary

Variable or Indicator	Analytical Role	Definition and Construction	Source and Identifier	Unit and Frequency	Transformation and Direction	Missing Value Rule	Coverage and Limitation
Refuting evidence attribution	Outcome	One for CONTRADICT and zero for SUPPORT	SciFact evidence label	Binary value per claim document pair	Higher value denotes refutation	Require an annotated label	773 pairs; measures annotated stance
Abstract word count	Document characteristic	Whitespace token count across joined abstract sentences	SciFact corpus abstract	Word count per abstract, repeated across eligible pairs	Higher value denotes greater textual extent	Require matching abstract text	773 pairs; does not measure semantic complexity
Relative evidence position	Document characteristic	First evidence sentence counted from one divided by abstract sentence count	SciFact rationale indices and corpus abstract	Ratio per pair	Higher value denotes later evidence placement	Require valid rationale indices and a positive sentence count	773 pairs; depends on existing rationale annotations
Structured abstract indicator	Document characteristic	Source structured indicator converted to one or zero	SciFact corpus structured field	Binary value per abstract	One denotes structured format	Require a recorded source indicator	773 pairs; does not measure reporting quality

Structure is interpreted as a formatting characteristic. Hartley, 2014, distinguishes the informational advantages of structured abstracts from unresolved questions about their accuracy. Similarly, Beller et al., 2013, identify reporting elements relevant to interpreting systematic review abstracts. The available binary indicator does not record compliance with those elements, so it cannot serve as a reporting completeness score.

Table 3: Descriptive statistics

Variable	Valid Observations	Mean	Standard Deviation	Median	Minimum	Maximum
Abstract word count	773	234.344	114.234	197.000	68.000	1,065.000
Relative evidence position	773	0.602	0.222	0.625	0.042	1.000
Structured abstract indicator	773	0.335	0.472	0.000	0.000	1.000
Refuting evidence attribution	773	0.343	0.475	0.000	0.000	1.000

The sample contains 265 refuting pairs and 508 supporting pairs. Structured abstracts account for 259 pair observations, equivalent to 33.5 percent. Because abstracts recur across claims, these percentages describe pair observations rather than the distribution of unique documents.

3.4 Analytical Framework, Identification, and Reproducibility:

The analytical framework is descriptive. Its objects of interest are the distributions of abstract length and evidence position across stance categories, together with the proportion of refuting pairs across abstract formats. The baseline summaries comprise valid counts, means, sample standard deviations, medians and observed ranges. No regression estimates or hypothesis test results are asserted in this data section. Interpretation remains conditional on the benchmark's annotated evidence relationships. Document presentation is not randomly assigned, claims differ in content, and repeated abstracts contribute several observations. Consequently, differences between stance categories cannot identify causal effects of length, location or structure. Pair weighting also gives abstracts linked to more retained claims greater representation than abstracts linked to fewer claims.

Relative evidence position is available only after rationale annotation. It therefore supports retrospective description of identified evidence. Wadden et al., 2022, jointly model evidence labels and rationales using document context, illustrating why

annotation dependent descriptions must remain distinct from evaluations of systems that identify evidence themselves. Reproducibility rests on preserved source identifiers, original evidence sentences, explicit variable definitions and the recorded archive checksum. The workbook contains the analytical pairs, source sentences and construction documentation. Its relative position and binary outcome columns retain calculation formulas. Independent reconstruction from the source sentence sheet reproduced the stored analytical values without discrepancy. The resulting data support comparisons of observed scientific evidence characteristics under consistent definitions and sample membership. Their contribution is an auditable description of annotated evidence relationships, bounded by the accessible benchmark and its available metadata.

4. Methods:

4.1 Research Design and Analytical Objectives:

This analysis uses a descriptive cross sectional design to examine how scientific abstract characteristics differ between supporting and refuting claim document pairs. The analytical unit is an annotated relationship between a claim and an evidence document. The comparisons address abstract length, relative evidence position and abstract structure, using the sample construction rules in Table 1 and the measurement definitions in Table 2. The analytical targets are differences in document characteristic distributions across stance categories and differences in refuting proportions across abstract formats. These comparisons describe the retained benchmark observations. They do not identify causal effects or evaluate predictive performance. Scientific claim verification distinguishes evidence selection from the assessment of whether that evidence supports or contradicts a proposition, as formalised by Wadden et al., 2020. The design preserves this distinction by treating stance as a property of the claim document relationship. Useful variation arises from differences in abstract characteristics and annotated stance across retained pairs. Because claims and documents recur, observations are not assumed to represent independent draws from a wider scientific population. The analysis therefore reports descriptive magnitudes without significance tests or population confidence intervals. Its interpretation remains conditional on the annotated sample.

4.2 Comparative Specification and Measurement Alignment:

The baseline comparisons give each claim document pair equal weight. Abstract word count and relative evidence position are summarised separately for supporting and refuting observations using valid counts, means, sample standard deviations, medians, quartiles and observed ranges. Equation 1 defines relative evidence position. Its denominator standardises evidence location by abstract sentence count without treating location as a measure of evidential quality. Equation 2 defines the difference in group means.

$$\Delta_Z = \bar{Z}_R - \bar{Z}_S$$

Here, Δ_Z is the descriptive difference for characteristic Z, $\bar{Z}_R$ is its mean among refuting pairs, and $\bar{Z}_S$ is its mean among supporting pairs. Characteristic \(Z\) denotes either abstract word count or relative evidence position. Positive values indicate a higher mean among refuting observations. The relative word count difference is additionally expressed as a percentage of the supporting mean. For abstract structure, the comparison concerns the proportion of observations labelled refuting within each format. Equation 3 defines this proportion.

$$p_g = \frac{r_g}{n_g}$$

Here, p_g is the refuting proportion in format group g, r_g is the number of refuting pairs in that group, and n_g is its total number of pairs. Group g denotes structured or unstructured abstracts. The contrast subtracts the unstructured proportion from the structured proportion and multiplies the result by 100 to express the difference in percentage points. The four analytical measures are observed or directly calculated variables. They are not combined into a latent construct or composite index. Length and structure remain document characteristics, while stance and rationale location depend on the particular claim relationship. This separation is consistent with the distinction between rationale selection and stance prediction developed by Zhang et al., 2021. The full retained sample contributes to each pooled comparison. Group counts change only through the categories defining the relevant contrast. No additional observations are removed for analytical transformations. Statements concerning structured abstracts refer to the recorded format indicator rather than compliance with reporting standards or methodological quality.

4.3 Computation, Processing and Construction Validation:

Computations use Python 3.12.14, NumPy 2.3.5 and lxml 6.1.1. The workbook is read through its spreadsheet XML records, resolving stored strings and numeric cell values. Analytical measures are converted to numeric arrays before calculation. Means, medians, minima and maxima are calculated directly from the retained observations. Standard deviations use the sample denominator, and quartiles use NumPy's default linear interpolation method. Processing begins with the cleaned pair records and their original stance labels. The procedure separates supporting and refuting observations, calculates the continuous variable summaries, and constructs the abstract format by stance counts. Differences are calculated from unrounded values. Rounding is applied only when presenting results.

All analytical variables are complete in the retained sample. No imputation, interpolation or replacement of unavailable values with zero is performed. No observations are trimmed or winsorised. Valid extreme values remain in the calculations. The processing rules therefore preserve the sample described in Table 1 rather than introducing further exclusions after examining group differences. Construction validation independently reconstructs the analytical measures from the workbook's source sentence records. Abstract sentences are ordered by their original indices, joined with spaces and split at whitespace to reproduce word counts. Sentence counts are recalculated from those records. The earliest annotated evidence index is converted to sentence order and divided by the reconstructed sentence count to reproduce relative position. Binary stance values are checked against the original SUPPORT and CONTRADICT labels. Additional checks assess record identifier uniqueness, document matches, admissible binary values and evidence index bounds. These procedures verify internal consistency between source records and calculated measures. They do not constitute independent reassessment of the source annotations. The distinction matters because agreement with an annotated rationale and the validity of an explanation are separate evaluation questions, as demonstrated by DeYoung et al., 2020.

4.4 Sensitivity Analysis, Dependence and Reproducibility:

Two exploratory sensitivity analyses follow the baseline comparisons. The first addresses unequal representation of documents. Each pair receives a weight equal to the reciprocal of its document's total number of retained pairs, so every document contributes a total weight of one across the complete sample. Equation 4 defines the resulting weighted group mean.

$$\bar{Z}_{h,w} = \frac{\sum_{i \in h} w_i Z_i}{\sum_{i \in h} w_i}, \qquad w_i = \frac{1}{m_{d_i}}$$

Here, $\bar{Z}_{h,w}$ is the weighted mean of characteristic Z in group h, i indexes claim document pairs, Z_i is the observed characteristic, w_i is its weight, d_i identifies its document, and m_{d_i} is that document's total number of retained pairs. Group h denotes a stance category for continuous comparisons or an abstract format for the weighted refuting proportion. For the latter calculation, Z_i is the binary refuting indicator. The same pooled observations enter the baseline and weighted comparisons. Differences therefore reflect weighting rather than sample deletion. Weighting reduces the influence of frequently represented documents but does not remove dependence between related claims or establish representative sampling. The second sensitivity analysis repeats the unweighted comparisons separately within the original training and development partitions. It examines whether pooled differences retain similar directions and magnitudes across those contributions. Because documents overlap between partitions, this comparison is interpreted as a check on sample composition rather than independent replication. Kapoor and Narayanan, 2023, identify information overlap as a consequential threat to claims of generalisation, supporting this interpretive boundary.

The weighting and partition comparisons are identified as exploratory because they were introduced after inspection of the pooled summaries. Their evaluation concerns changes in direction and magnitude, rather than continued statistical significance. The descriptive weighting and partition outputs are reported together in Table 6. Replication follows the retained source identifiers, sentence indices, original labels and explicit construction rules. The workbook contains the analytical records, source sentences and provenance documentation, including the release archive checksum. Its position and binary outcome columns retain spreadsheet formulas. The computations are deterministic and require no random seed or numerical optimisation. Together, these procedures answer the research questions through reproducible descriptive comparisons while preserving the distinction between document presentation and annotated evidence stance.

5. Findings:

This analysis examines differences in abstract length, evidence position and abstract structure between supporting and refuting evidence. It answers the three research questions through descriptive comparisons of annotated claim document pairs. Additional comparisons assess sensitivity to repeated documents and differences between the source partitions.

5.1 Sample Composition and Measurement:

The analytical sample contains 773 claim document pairs covering 693 claims and 437 scientific abstracts. Supporting evidence accounts for 508 observations, or 65.7 percent, while refuting evidence accounts for 265 observations, or 34.3 percent. All observations contain complete values for abstract word count, relative evidence position, abstract structure and annotated stance. The sample represents a cross section of the SciFact benchmark released in 2020. Its outcome follows the distinction between supporting and refuting scientific evidence established by Wadden et al., 2020. The comparisons weight each annotated pair equally unless stated otherwise. Abstract word count measures whitespace separated tokens. Relative evidence position divides the first annotated evidence sentence, counted from one, by the abstract sentence count. Abstract structure distinguishes structured from unstructured abstracts. Refuting evidence is coded one and supporting evidence zero. These measures describe document presentation and annotated stance rather than verification accuracy.

Table 4: Abstract characteristics by evidence stance

Measure	Supporting Evidence	Refuting Evidence
Claim document pairs	508	265
Mean abstract word count	223.21	255.68
Standard deviation of word count	109.64	119.90
Median abstract word count	180.00	225.00
Word count at the first quartile	149.00	156.00
Word count at the third quartile	283.00	327.00
Minimum word count	68	82
Maximum word count	1,065	1,065
Mean relative evidence position	0.599	0.606
Standard deviation of relative position	0.223	0.220
Median relative evidence position	0.615	0.636

5.2 Abstract Length and Evidence Attribution:

Refuting pairs contain longer abstracts in the pooled sample. Their mean word count exceeds that of supporting pairs by 32.46 words, equivalent to 14.5 percent of the supporting mean. The median difference is 45 words. Agreement between the mean and median comparisons indicates that the pooled difference is not confined to the largest observations. Nevertheless, the distributions overlap substantially. The central half of supporting observations spans 149 to 283 words, compared with 156 to 327 words for refuting observations. Both categories reach the same maximum of 1,065 words. Abstract length therefore distinguishes the group summaries without separating the two stance categories. The first research question is resolved descriptively: refuting pairs are associated with longer abstracts in the combined sample. This finding concerns textual extent rather than the strength of

evidence. Wadden et al., 2022, demonstrated the relevance of document context to scientific verification, but that result does not imply that additional words necessarily indicate refutation.

5.3 Evidence Position and Annotated Stance:

Evidence position provides less separation between the categories. Mean relative position is 0.599 for supporting pairs and 0.606 for refuting pairs, a difference of 0.007. Median positions differ by 0.021. These differences are small relative to the within category standard deviations of approximately 0.22. The quartile comparisons reinforce the overlap. Supporting evidence has a first quartile of 0.467 and a third quartile of 0.750. The corresponding values for refuting evidence are 0.500 and 0.765. Both categories include evidence near the beginning of abstracts and evidence first appearing in the final sentence. The second research question is therefore answered by a limited pooled difference: the first annotated rationale appears slightly later among refuting pairs, but its location offers little descriptive separation. This result preserves the distinction between locating a rationale and determining its stance. Zhang et al., 2021, treat those tasks as connected but distinct components of scientific verification.

5.4 Abstract Structure and Refuting Evidence:

Structured abstracts contribute 259 observations, while unstructured abstracts contribute 514. Refuting evidence occurs in 43.6 percent of structured observations and 29.6 percent of unstructured observations. The difference is 14.1 percentage points, as shown in Table 5.

Table 5: Evidence stance by abstract structure

Abstract Format	Supporting Pairs	Refuting Pairs	Total Pairs	Refuting Share
Unstructured	362	152	514	29.6 percent
Structured	146	113	259	43.6 percent
Combined	508	265	773	34.3 percent

The third research question is resolved through a higher pooled refuting share among structured observations. Supporting evidence nevertheless remains the larger category within both formats. Structure consequently identifies a difference in sample composition rather than a rule determining stance. This distinction matters because abstract organisation and abstract accuracy are separate properties. Hartley, 2014, identifies informational advantages of structured abstracts alongside continuing concerns about accuracy and omissions. The observed structure difference does not establish superior evidential quality. Likewise, the structure indicator does not measure the reporting elements specified by Beller et al., 2013. The findings cannot therefore attribute the higher refuting share to reporting completeness.

5.5 Sensitivity to Document Representation and Source Partition:

Two exploratory comparisons assess whether the pooled patterns depend on sample composition. The first assigns each pair the reciprocal of its document's total number of retained pairs. Each abstract consequently contributes a total weight of one across its observations. The second calculates the comparisons separately within the original training and development partitions.

Table 6: Sensitivity of descriptive differences

Comparison	Refuting Minus Supporting Mean Word Count	Refuting Minus Supporting Mean Relative Position	Structured Minus Unstructured Refuting Share
Pooled pair weighting	32.46 words	0.007	14.1 percentage points
Equal total weight per document	38.55 words	0.011	16.8 percentage points
Training partition	43.42 words	0.012	18.8 percentage points
Development partition	2.74 words	−0.006	1.2 percentage points

Giving each document equal total weight preserves the direction of the pooled length and structure differences. The word count difference increases to 38.55 words, while the structure difference increases to 16.8 percentage points. Unequal document repetition therefore does not explain away these pooled descriptive patterns under this weighting scheme. Partition comparisons reveal a more substantial qualification. The word count difference is 43.42 words in training but only 2.74 words in development. The structure difference similarly falls from 18.8 to 1.2 percentage points. Evidence position remains weakly separated and changes direction. These comparisons show that the pooled patterns are concentrated more strongly in the training contribution. Because 110 abstracts appear in both partitions, these results assess composition sensitivity rather than independent replication.

Document reuse also clarifies the meaning of stance. Of the 437 abstracts, 179 supply supporting evidence for some claims and refuting evidence for others. These documents account for 469 observations, or 60.7 percent of all pairs. Their word count and structure remain fixed while their annotated stance varies across claims. Evidence orientation therefore depends on the claim being evaluated and cannot be assigned to document format alone. The findings answer the three research questions at the descriptive level. Refuting pairs have longer abstracts and occur more frequently among structured observations in the pooled sample, while evidence position differs little. Equal document weighting preserves these pooled directions, but partition comparisons reveal substantial variation in their magnitudes. The evidence supports sample specific descriptive differences rather than uniform relationships across benchmark partitions.

6. Discussion:

Document form can distinguish groups of scientific evidence without determining whether an individual claim is supported. Tables 4 and 5 show that refuting pairs contain longer abstracts and occur more frequently among structured observations, while evidence position provides little separation. The contribution of this analysis lies in examining these presentation characteristics alongside their sensitivity to document representation and partition composition. Guo et al., 2022, organise automated verification around claims, evidence and verdicts. The present findings sharpen that distinction by showing that document characteristics describe the distribution of annotated stance without constituting evidence of its semantic direction. The pooled patterns therefore support a relational interpretation of verification: information acquires evidential meaning through

the claim against which it is assessed. The limited difference in evidence position qualifies the role of textual location in explaining stance. Equation 1 normalises the first rationale position by abstract length, allowing Table 4 to distinguish relative placement from document size. Despite this adjustment, supporting and refuting evidence occupy substantially overlapping locations. This pattern suggests that locating an evidentiary statement and interpreting its relationship to a claim are separate analytical tasks. DeYoung et al., 2020, distinguish agreement with annotated rationales from the faithfulness of explanations to model predictions. The present comparison adds a descriptive boundary to that distinction: rationale location alone provides little separation between stance categories in this sample. It consequently directs attention toward the meaning of the selected evidence rather than its position as a substitute for interpretation.

Document reuse provides the clearest evidence that stance cannot be treated as an intrinsic property of an abstract. The 179 abstracts associated with both supporting and refuting labels retain the same length and structure across their claim relationships. Their changing labels therefore expose the limits of interpreting the pooled differences in Tables 4 and 5 as document level determinants. Equal document weighting preserves those differences in Table 6, but preservation under reweighting does not convert them into causal pathways. Jacovi and Goldberg, 2020, argue for distinguishing the criteria used to evaluate interpretations rather than treating explanation quality as a single property. Here, the corresponding theoretical refinement separates observable presentation, annotated evidence orientation and explanatory faithfulness. These concepts cannot be used interchangeably when assessing scientific verification. Partition sensitivity changes the substantive interpretation of the pooled associations. Table 6 shows that the length and structure differences become markedly smaller in the development contribution, while the small position difference changes direction. The combined sample therefore conceals variation that matters for transferring descriptive relationships between benchmark partitions. Document overlap further means that the partition comparison cannot establish independent replication. Kapoor and Narayanan, 2023, demonstrate how failures to account for information overlap can undermine claims of generalisation in machine learning research. The practical implication is specific: benchmark reports should disclose document overlap, retain partition specific summaries and distinguish pair weighting from document weighting. This analysis shows why stable pooled directions and consistent performance across partitions are different evidential requirements.

The international relevance of these findings concerns evidence verification infrastructure rather than differences between national economies. Schlichtkrull et al., 2023, construct AVeriTeC around claims assessed by multiple organisations, with evidence and justifications connecting source material to verdicts. That setting differs from the present scientific abstract benchmark, so its results cannot establish that the same presentation patterns operate across domains. The comparison nevertheless identifies a shared design requirement: verification systems must preserve the relationship between a specific claim and its evidence. Tables 4-6 indicate that length, structure and location should remain descriptive audit characteristics rather than standalone grounds for accepting or rejecting claims. For research organisations and evidence services, the actionable contribution is to retain claim identifiers, source sentences and stance annotations together, while reporting how descriptive relationships vary across samples. This approach makes the limits of presentation-based interpretation visible without overstating what the evidence establishes.

7. Conclusion and Implications:

Reliable scientific verification depends on preserving the relationship between a claim and its evidence, rather than treating document presentation as a verdict. This analysis shows that observable presentation patterns distinguish pooled evidence categories unevenly, while their magnitude depends on sample composition. The contribution is an empirical distinction between descriptive separation and consistency across samples: patterns that persist after balancing document representation can still weaken substantially across benchmark partitions. Evidence attribution consequently requires an analytical framework that separates document properties from the meaning of particular claim relationships. The same source can support one proposition and refute another, placing relational interpretation at the centre of verification. For managers of evidence services, these findings favour retaining traceable links between claims, source sentences and annotated judgements. Benchmark developers should report partition composition and document overlap alongside pooled summaries, allowing users to assess where apparent regularities hold. Research governance policies can support this practice by requiring transparent documentation of sampling, annotation and aggregation decisions. Operationally, presenting evidence in context provides a defensible basis for review without substituting presentation cues for substantive assessment. The wider contribution is a basis for auditing scientific evidence systems across institutions, while keeping conclusions proportionate to the relationships actually observed and examined.

The benchmark setting provides a foundation for extensions across scientific domains, languages and evidence repositories. Document reuse and partition overlap motivate validation designs that separate sources and related claims. Additional annotations could distinguish reporting completeness, methodological quality and semantic complexity from presentation characteristics. Longitudinal collections could examine whether descriptive relationships change as evidence accumulates. Controlled changes to presentation, combined with independent assessments of claim interpretation, could test mechanisms that observational comparisons cannot identify. Such extensions would establish which patterns generalise across technological and institutional settings while preserving the distinction between evidence organisation, annotated stance and verification performance.

References:

1. Beller, E. M., Glasziou, P. P., Altman, D. G., Hopewell, S., Bastian, H., Chalmers, I., Gøtzsche, P. C., Lasserson, T., Tovey, D., & PRISMA for Abstracts Group. 2013. PRISMA for abstracts: Reporting systematic reviews in journal and conference abstracts. PLOS Medicine, 10, issue 4, Article e1001419. https://doi.org/10.1371/journal.pmed.1001419
2. DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4443-4458. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.408

3. Guo, Z., Schlichtkrull, M., & Vlachos, A. 2022. A survey on automated fact checking. Transactions of the Association for Computational Linguistics, 10, 178-206. https://doi.org/10.1162/tacl_a_00454
4. Hartley, J. 2014. Current findings from research on structured abstracts: An update. Journal of the Medical Library Association, 102, issue 3, 146-148. https://doi.org/10.3163/1536-5050.102.3.002
5. Jacovi, A., & Goldberg, Y. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4198-4205. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.386
6. Kapoor, S., & Narayanan, A. 2023. Leakage and the reproducibility crisis in machine learning based science. Patterns, 4, issue 9, Article 100804. https://doi.org/10.1016/j.patter.2023.100804
7. Schlichtkrull, M., Guo, Z., & Vlachos, A. 2023. AVeriTeC: A dataset for real world claim verification with evidence from the web. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 65128-65167. Curran Associates. https://doi.org/10.52202/075280-2842
8. Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., &Hajishirzi, H. 2020. Fact or fiction: Verifying scientific claims. In B. Webber, T. Cohn, Y. He, & Y. Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7534-7550. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.609
9. Wadden, D., Lo, K., Wang, L. L., Cohan, A., Beltagy, I., & Hajishirzi, H. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full document context. In M. Carpuat, M. de Marneffe, & I. V. Meza Ruiz, editors, Findings of the Association for Computational Linguistics: NAACL 2022, pages 61-76. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.6
10. Zhang, Z., Li, J., Fukumoto, F., & Ye, Y. 2021. Abstract, rationale, stance: A joint model for scientific claim verification. In M. Moens, X. Huang, L. Specia, & S. W. Yih, editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3580-3586. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.290