# Beyond Fluency: A Framework for Analyzing Epistemic and Argumentative Shifts in AI-Assisted Scientific Text Reformulation

**Laura Centonze**
Department of Human and Social Sciences
Universitas Mercatorum, Italy
laura.centonze@unimercatorum.it

## Abstract

Although generative AI increasingly reformulates scientific texts, evaluations based on fluency, readability, and semantic similarity may overlook changes in the epistemic and argumentative properties of the reformulated texts. This paper introduces a *Claim-Preservation Framework* (CPF) to assess whether AI-generated reformulations preserve not only propositional content, but also epistemic force, evidential relation, argumentative function, and qualification and attribution. The framework distinguishes stylistic variation from consequential shifts in modality, hedging, evidentiality, causal interpretation, and argumentative function; its analytical potential is illustrated through an exploratory proof-of-concept application to 40 SciFact claims reformulated under a standardized instruction. This application demonstrates how the proposed dimensions can be used to identify and classify potential shifts in scientific meaning, but it is not intended as a full validation of the framework or as a general estimate of LLM behavior. By conceptualizing reformulation as a transformation of scientific meaning, the paper proposes claim preservation as a distinct dimension of LLM evaluation and outlines future research on computational methods for detecting epistemic and argumentative shifts in scientific publications.

## 1 Introduction

The rapid adoption of Large Language Models (LLMs) has introduced new forms of computational assistance into scientific writing. Researchers increasingly use generative systems to draft, edit, summarize, translate, and reformulate scholarly texts, making AI-assisted writing a growing component of the scientific publication process (Liang et al., 2024). These applications offer clear benefits in terms of efficiency, linguistic accessibility, textual readability, and stylistic refinement; however, they also raise a less obvious problem: a reformulated scientific statement may become linguistically improved while becoming scientifically different.

The issue is particularly important in scholarly communication because scientific texts do not merely convey propositions: they also encode degrees of epistemic commitment, evidential relationships, qualifications, attribution, causal interpretations, and argumentative functions through which authors position claims in relation to evidence and disciplinary knowledge. Consider the difference between the statements "The findings suggest that the intervention may improve academic writing" and "The findings demonstrate that the intervention improves academic writing." Although the two sentences refer to the same general intervention and outcome, the second formulation increases the apparent certainty of the claim and removes the original qualification. Similarly, changing "the results indicate an association between X and Y" to "the results indicate that X affects Y" introduces a causal interpretation that is not necessarily justified by the source. Replacing "the effect was observed primarily among advanced learners" with "the effect was observed among learners" may likewise broaden the population to which the finding appears to apply. These transformations alter the scientific interpretation of a statement without necessarily reducing lexical or semantic similarity: in each case, the output remains fluent and semantically recognizable while altering an aspect of the information conveyed by the source.

These issues intersect with several established areas of scientific NLP. Research on scientific claim verification, for instance, has examined whether claims are supported or contradicted by available evidence (Wadden et al., 2020), while subsequent studies have extended this line of inquiry to claim–evidence reasoning and citation-based verification with large language models (Alvarez et al., 2024; Javaji et al., 2025). Argument min-



*Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, pages 133–141
November 10, 2026 ©2026 Association for Computational Linguistics

ing has similarly examined the identification of claims, evidence, and argumentative relations in scientific discourse (Fergadis et al., 2021). Such approaches provide valuable computational representations of scientific knowledge, but generally conceptualize claims as objects to be extracted, classified, retrieved, or verified. AI-assisted reformulation raises a complementary question: how are the scientific properties of a claim affected when an LLM transforms its linguistic realization?

The present paper conceptualizes AI-assisted scientific reformulation as a claim-preservation problem. An epistemic or argumentative shift occurs when an AI-generated reformulation changes the epistemic force, evidential relation, qualification, attribution, causal interpretation, or argumentative function of the original statement, even when its broad propositional content remains recognizable. Semantic similarity is therefore necessary but not sufficient for scientific equivalence: two statements may express broadly related content while differing in the certainty, evidential strength, scope, or argumentative role they convey. To operationalize this distinction, the paper introduces the *Claim-Preservation Framework* (CPF), which represents scientific claims along five analytical dimensions: propositional content (P), epistemic force (E), evidential relation (V), argumentative function (A), and qualification and attribution (Q). These dimensions make it possible to distinguish reformulations that preserve the relevant scientific properties of the source from those that alter one or more of them; although analytically distinct, they may interact within a single reformulation, affecting several aspects of scientific meaning simultaneously. For aggregate evaluation, the framework also assigns an overall outcome label to each pair (Section 4).

The paper makes three main contributions. First, it conceptualizes scientific reformulation as a structured transformation problem rather than exclusively as a language-generation or stylistic task. Second, it proposes the CPF as a multidimensional framework for identifying changes that may remain invisible to conventional semantic similarity measures. Third, it illustrates the applicability of the proposed framework through an exploratory proof of concept involving a purposively sampled set of scientific claims from the SciFact benchmark and a standardized AI reformulation procedure. The empirical analysis assesses the operational feasibility of the framework and illustrates how claim-level changes can be identified and annotated (see Section 7 for its limitations).

The study positions claim preservation as a distinct dimension of LLM evaluation for scientific writing. Rather than assuming that successful reformulation is equivalent to fluency, readability, or semantic correspondence, the proposed approach asks whether the output continues to express the scientific claim with the same epistemic force, evidential relation, argumentative function, and qualifications as the source. This perspective provides a basis for future larger-scale annotation studies, inter-annotator agreement analysis, automated detection, and comparison with conventional semantic similarity, factuality, and claim-verification measures (see Section 7).

## 2 Scientific Text Reformulation Beyond Fluency

AI-assisted reformulation is generally understood as a constrained form of language generation in which an existing text serves as the basis for a new linguistic realization. In scientific writing, reformulation may be requested to improve clarity, concision, formality, readability, grammatical accuracy, or disciplinary appropriateness. Because the user provides the source text, such operations may appear relatively low risk compared with open-ended generation: the system is ostensibly transforming existing content rather than creating scientific content from scratch. This assumption, however, depends on a strong equivalence premise: that changes in linguistic realization leave the scientific properties of the source statement intact. The premise becomes problematic when scientific meaning is considered at the discourse level. Scientific statements encode not only what is being claimed, but also how strongly it is being claimed, what kind of evidential relationship is associated with it, under which conditions the claim applies, who or what is presented as its source, and what argumentative role the statement performs.

A distinction is needed between linguistic reformulation and scientific claim preservation. Linguistic reformulation concerns changes in form that improve or modify the expression of a statement. Claim preservation concerns whether the relevant scientific properties of that statement remain stable after transformation. The two notions are related but not equivalent. This distinction also clarifies the relationship between fluency, semantic similarity, factuality, and claim preservation. Fluency-

oriented evaluation primarily concerns the linguistic quality of the generated text, whereas semantic similarity measures estimate the degree of correspondence between source and output and are useful for detecting whether two texts express broadly related content. Factuality and claim-verification approaches, in turn, ask whether a claim is supported by available evidence or whether the generated statement is consistent with a source or knowledge base. Claim preservation addresses a related but distinct question: has the transformation maintained the scientific status of the original statement? The distinction is consequential because epistemic and argumentative changes can occur without a substantial loss of semantic correspondence. The two examples in Section 1 illustrate this: the first preserves the broad proposition while increasing epistemic commitment and removing a qualification, and the second modifies the evidential and causal interpretation of the relationship. In both cases, a similarity-based system may identify substantial correspondence between the sentences while failing to represent the nature of the transformation.

Scientific writing provides numerous linguistic resources through which such distinctions are encoded. Hedging and boosting, for example, allow authors to calibrate the strength of claims and negotiate degrees of certainty within disciplinary communities (Hyland, 1998). Evidential predicates and reporting expressions distinguish between observations, interpretations, previous findings, and stronger forms of evidential commitment. Causal and associative constructions likewise encode different relations between scientific entities. Finally, qualifications relating to population, methodology, temporal scope, or experimental conditions constrain the interpretation of a proposition. These properties may also interact (Section 3), so claim preservation cannot be reduced to a single linguistic feature or lexical substitution. Accordingly, the present study treats AI-assisted reformulation as a structured transformation problem: the relevant comparison is not Source → Output → Similarity Score, but *Source Claim → AI reformulation → Multidimensional Comparison → Shift Classification*.

This perspective complements existing research on scientific claim verification, claim–evidence reasoning, and argument mining. Within it, the central distinction is between surface-preserving reformulation, which may modify lexical choice, syntax, or discourse organization while maintaining the relevant scientific properties of the source, and consequential reformulation, which changes one or more properties that contribute to the interpretation of the claim, even when the resulting sentence remains fluent, plausible, and semantically related to the source. The CPF introduced in Section 4 is designed to make this distinction explicit and operational.

## 3 Epistemic and Argumentative Shifts in AI-Generated Reformulations

The distinction between surface-level and consequential reformulation can be developed through two related dimensions: epistemic shifts and argumentative shifts. Epistemic shifts concern changes in the degree of commitment associated with a proposition. They include strengthening, weakening, modification of hedging and modality, and changes in the way evidence is presented as supporting a claim. The first example in Section 1 illustrates this: the same relationship is presented as tentative and inferential in one formulation and as established in the other. Epistemic shifts can also occur through the modification of evidential predicates. Expressions such as "suggest," "indicate," "support," "demonstrate," and "confirm" position a proposition differently with respect to the available evidence. A reformulation that systematically replaces weaker expressions with stronger ones may produce a form of epistemic inflation even when no new factual information has been introduced. Conversely, introducing additional modal or hedging expressions may weaken an originally well-supported conclusion.

Argumentative shifts concern changes in what a sentence does within the structure of a scientific argument. Scientific discourse contains propositions that function as claims, evidence, interpretations, qualifications, contrasts, limitations, or conclusions. A reformulation may preserve the general topic while changing this function. For example, "Previous studies have identified several limitations of this approach" reports limitations, whereas "Previous studies have demonstrated the ineffectiveness of this approach" converts those limitations into a stronger evaluative conclusion. The second formulation does more than intensify vocabulary: it changes the argumentative work performed by the sentence.

The two dimensions are often intertwined. Strengthening a claim can change its argumenta-

tive role; removing a qualification can increase the apparent generality of a conclusion; replacing an associative relation with a causal one can transform evidence into a stronger explanatory claim. This interaction makes epistemic and argumentative preservation especially important for scientific NLP, where claims are increasingly represented in relation to evidence, citations, discourse structures, and knowledge graphs. Against this background, the task addressed in this paper is not to determine whether AI-generated text is "good" in a general linguistic sense, but whether the transformation preserves the scientific configuration of the source statement.

## 4 The Claim-Preservation Framework

The Claim-Preservation Framework (CPF) represents a scientific claim as a five-dimensional structure:

$$C = <P, E, V, A, Q>$$

where P represents propositional content, E epistemic force, V evidential relation, A argumentative function, and Q qualification and attribution. Table 1 summarizes the dimensions and the values used for their annotation. Adequate preservation requires that the proposition remain stable together with the epistemic, evidential, argumentative, and qualifying information that determines its interpretation, and a single modal verb, evidential predicate, causal expression, or omitted qualifier can be sufficient to alter the scientific status of a claim. The dimensions are described in turn below.

The first dimension, *propositional content*, concerns whether the entities and relations described by the original statement remain stable. This dimension corresponds most closely to conventional semantic preservation and provides the foundation for subsequent analysis, but it does not by itself determine scientific equivalence. The second dimension, *epistemic force*, concerns whether the degree of commitment associated with the proposition remains comparable. Relevant linguistic resources include modal verbs, evidential predicates, hedges, boosters, and expressions of probability or certainty. The criterion is preservation rather than maximization: a reformulation is not better for being more confident, nor worse for being more cautious. A problem arises only when the degree of commitment of the reformulation differs from that of the source without clear justification.

The third dimension, *evidential relation*, concerns the relation between a claim and the information presented as supporting it, including distinctions between empirical observation, previous research, author interpretation, theoretical inference, and speculation. It also captures changes in the relation asserted between scientific variables, such as the distinction between association and causation (see the second example in Section 1). This dimension connects the framework with scientific claim-verification research (Wadden et al., 2020). The fourth dimension, *argumentative function*, reflects the fact that a sentence may operate as evidence, interpretation, qualification, contrast, limitation, or conclusion depending on its relation to surrounding claims. Thus, preservation requires consideration of local context rather than isolated sentence similarity. This dimension is closely related to scientific argument mining, which seeks to identify argumentative components and relations within scholarly texts (Fergadis et al., 2021). Because the argumentative function of an isolated claim may not always be recoverable from the available text, cases in which the function cannot be established reliably are treated as indeterminate rather than being assigned a forced classification. The final dimension, *qualification and attribution*, is relevant because scientific findings are frequently restricted by population, methodology, temporal scope, disciplinary context, or evidential conditions. Explicit attribution also identifies the source of a proposition and can affect its epistemic interpretation. Removing such information can broaden or strengthen a claim even when its central proposition remains intact. To limit overlap, E is restricted to the degree of commitment expressed through modals, hedges, and boosters, whereas V concerns the type of relation claimed and its support, including association versus causation and observation, prior research, or inference. Because a single reformulation can affect more than one dimension (see Section 3), the dimensions may co-occur and are therefore annotated independently. For descriptive analysis, each source–reformulation pair is assigned a single overall label, derived from the dimension-level annotation as specified in Table 2. The labels are mutually exclusive for aggregate analysis, whereas the underlying dimension-level changes may co-occur.

| Dimension | Core question | Annotation values |
|---|---|---|
| **Proposition** | Is the underlying relation preserved? | preserved/ shifted |
| **Epistemic force** | Is the degree of commitment preserved? | preserved/ strengthened/ weakened |
| **Evidential relation** | Is the evidence-claim relation preserved? | preserved/ strengthened/ weakened |
| **Argumentative function** | Does the statement perform the same function? | preserved/ shifted / indeterminate |
| **Qualification & attribution** | Are limits, conditions and sources preserved? | preserved/ removed/ added |

Table 1: Dimensions of claim preservation.

## 5 Proof-of-Concept Analysis

To show how the framework can be operationalized, we conducted an exploratory proof-of-concept analysis on claims from the SciFact benchmark (Wadden et al., 2020). SciFact was developed for scientific claim verification and contains claims paired with evidence-bearing scientific abstracts, available through the official GitHub repository (https://github.com/allenai/scifact). Its use here serves a different purpose: rather than asking whether a claim is supported or contradicted by the available evidence, we treat the claim as the source text for an AI-mediated reformulation task, which creates a direct methodological connection between established scientific claim-verification research and the proposed problem of claim preservation under linguistic transformation. The analysis draws on a purposive sample of 40 claims, selected to cover five areas relevant to the CPF: epistemic modality and hedging, evidential relation and strength, association versus causation, qualification and scope, and attribution or argumentative function. A final group comprised claims that potentially involve several dimensions. The sample was designed to provide sufficient variation for testing whether the dimensions can be applied to naturally occurring scientific claims, not to estimate how frequently epistemic or argumentative shifts occur. Each claim was submitted, in a separate conversation, to the same model (ChatGPT, GPT-6, web interface with default settings, accessed in October 2026) under a standardized instruction. The prompt, reported in full in Appendix B, asked the system to improve clarity, concision, and academic style while preserving the scientific properties of the source claim and prohibiting any addition of information or strengthening or weakening of the claim. Keeping the prompt constant isolates reformulation effects from differences in prompting conditions.

The resulting source–reformulation pairs were annotated manually by the author alone, which precluded the calculation of inter-annotator agreement (see Section 7). Each dimension was coded independently using the values in Table 1. An overall label was then assigned according to the criteria in Table 2. The framework supports two descriptive measures: the shift rate for each dimension, defined as the proportion of pairs in which that dimension changes, and the any-shift rate, defined as the proportion of pairs in which at least one dimension changes. Given the small, purposively selected sample, aggregate frequencies are not reported; the analysis focuses on the qualitative examination of representative pairs.

Table 3 presents six illustrative pairs from the sample, reporting the reformulation, the dimensions examined, and the overall label. The corresponding source claims are given in Appendix A and can be retrieved from SciFact through the listed identifiers. The remaining pairs are withheld because they are used in ongoing research. Five of the six pairs are classified as PRESERVED. The exception is #563: the source states that, in British men, haplogroup I "decreases risk" of cardiovascular disease, a causal formulation, whereas the reformulation states that it "is associated with a decreased risk." Population and direction of the relationship are preserved, but the causal commitment is weakened, despite the instruction prohibiting any weakening. The two sentences differ by only a few words, a change that a similarity measure may fail to register; this shows why semantic correspondence alone cannot establish claim preservation.

These observations should be read as a test of applicability rather than as validation. The analysis assesses whether the annotation scheme can be applied to naturally occurring claims, whether it identifies transformations that are consequential for scientific interpretation, and which categories may need refinement before larger-scale annotation. A natural next step is to compare the dimension-level annotations with conventional similarity and factuality measures, in order to test whether high similarity reliably predicts preservation of epistemic and argumentative structure. Pairs that score

| Label | Derivation rule | Typical signal |
|---|---|---|
| **PRESERVED** | All relevant dimensions preserved | Stylistic/ lexical reformulation |
| **STRENGTHENED** | E and/or V strengthened, with no qualifying shift in another dimension | *suggest* → *demonstrate; associated with* → *affects* |
| **WEAKENED** | E and/or V weakened, with no qualifying shift in another dimension | *demonstrate* → *indicate*; *causal* → *associative* |
| **SHIFTED** | P, A, or Q changed, or multiple dimensions change in ways not reducible to strengthening/weakening | Qualifier or source removed; function changed |
| **AMBIGUOUS** | Reliable classification cannot be established from the available context | Context-dependent interpretation |

Table 2: Summary labels derived from per-dimension annotation.

| SciFact ID | AI reformulation | CPF dimension and label |
|---|---|---|
| #285 | Continued HHV-8 transmission among MSM in San Francisco may be attributable to urogenital contact. | Epistemic force; evidential relation; propositional content; PRESERVED |
| #494 | HSCT-T dosage may not be diagnostic when symptom onset occurs less than 3 hours before AMI. | Epistemic force; qualification; PRESERVED |
| #553 | IgA plasma cells specific for transglutaminase 2 are scarce in the duodenal mucosa of patients with active celiac disease. | Propositional content; qualification/scope; PRESERVED |
| #563 | In British men, haplogroup I is associated with a decreased risk of cardiovascular disease. | Evidential relation; WEAKENED |
| #760 | Mathematical models predict that using artemisinin-based combination therapy instead of nongametocytocidal drugs has only a modest impact on reducing malaria transmission. | Evidential relation; propositional content; PRESERVED |
| #771 | In elderly patients with metastatic colorectal cancer, treatment with single-agent fluoropyrimidines resulted in similar efficacy and better quality of life compared with oxaliplatin-based chemotherapy. | Propositional content; qualification/scope; argumentative function; PRESERVED |

Table 3: Illustrative examples of claim-preservation outcomes. Source claims are provided in Appendix A.

high on similarity yet are classified as strengthened, weakened, or shifted would support treating claim preservation as a distinct evaluation problem. The SciFact-based analysis provides a structured basis for quantifying the distribution of preservation and shift types in larger samples. Its principal purpose here is to assess whether the proposed annotation scheme captures transformations that conventional semantic evaluation might overlook. The analysis also motivates a direct research question for future benchmark development: does high semantic similarity between an original scientific claim and its AI reformulation reliably predict preservation of epistemic and argumentative structure? If the answer is negative, claim preservation becomes a distinct evaluation problem rather than a secondary component of semantic similarity.

## 6 Implications for Scientific NLP and LLM Evaluation

The proposed framework positions AI-assisted scientific reformulation as a distinct transformation task within scientific NLP. Existing research has examined argumentative structures in scientific discourse (Fergadis et al., 2021) and the verification of scientific claims against supporting evidence (Wadden et al., 2020; Alvarez et al., 2024; Javaji et al., 2025). The present problem is complementary because it asks what happens when an already structured scientific statement is transformed by an AI system. The relevant question is no longer simply whether a claim exists or whether evidence supports it, but whether the transformation preserves the properties that make the claim scientifically interpretable. This has direct implications for LLM evaluation. Fluency, readability, coherence, and semantic similarity remain useful indicators of output quality, but they should not be treated as sufficient evidence of scientific preservation (Wysocka et al., 2024). Likewise, scientific factuality and claim preservation should be treated as related but distinct dimensions of trustworthy scientific text generation.

The framework also suggests a practical direction for AI-assisted scientific writing systems. Instead of silently replacing a phrase such as "may indicate" with "demonstrates," an AI assistant could

detect the transformation and flag it for user review. Similarly, a system could identify when an associative relationship has become causal or when a methodological or population restriction has been removed. Such functionality would move writing assistance from optimizing text to monitoring how the edits affect the argument. The proposal is also compatible with emerging research on the rhetorical behavior of LLMs, which has reported that LLM-generated representations of scientific research can exhibit increased causal language and rhetorical confidence relative to source texts (Peters and Chin-Yee, 2025). Such findings reinforce the importance of evaluating not only factual correctness but also changes in how scientific knowledge is framed. The present framework addresses this issue at the level of controlled reformulation and may provide a basis for integrating epistemic and argumentative preservation into broader scientific-text evaluation pipelines.

A future benchmark could pair original scientific claims with multiple AI-generated reformulations and require systems to identify whether each transformation preserves, strengthens, weakens, or shifts the scientific configuration of the source. Such a benchmark could be evaluated against expert annotation and compared with semantic similarity, factuality, and claim-verification metrics. The resulting evaluation would provide a more comprehensive account of the reliability of AI-assisted scientific writing.

## 7 Limitations and Future Research

The framework presented here is conceptual and exploratory and should not be read as a validated detection method. The proof-of-concept analysis is deliberately small and is intended to establish the feasibility of the annotation approach rather than provide generalizable estimates of AI behavior. The analysis relies on a single model and a single prompt, which limits the extent to which the observed behavior generalizes to other systems or instructions. Annotation was carried out by a single annotator, and the reliability of the proposed categories remains to be established. The complete annotated sample is not publicly released at this stage, which limits independent verification of the reported observations. Moreover, epistemic force and argumentative function can be context-sensitive, meaning that reliable annotation may require information extending beyond the individual sentence. Future research should therefore construct a larger corpus of original scientific claims and AI-generated reformulations across disciplines, publication sections, models, and prompting conditions. Expert annotation should establish the reliability of the proposed categories, with particular attention to inter-annotator agreement for borderline cases and to whether the five dimensions are jointly sufficient. Subsequent computational experiments could then test whether existing semantic similarity metrics detect the same transformations identified by human experts and whether LLM-based classifiers can reliably identify specific forms of epistemic and argumentative shifts.

An important direction would be to integrate claim-preservation analysis with existing scientific claim-verification and citation-based evidence systems. This would enable a two-stage evaluation in which a system first establishes whether a scientific claim is adequately supported and then determines whether an AI transformation has preserved the relationship between the claim and its evidential basis. Such integration could contribute to more trustworthy scientific text generation and automated quality assessment.

## 8 Conclusion

Efforts in AI-assisted scientific writing increasingly focus on improving the linguistic realization of scholarly texts. However, scientific communication depends on more than fluency, readability, or semantic similarity. Scientific claims are also characterized by degrees of epistemic commitment, evidential grounding, qualification, attribution, and argumentative function. This paper has proposed a Claim-Preservation Framework for examining whether these properties remain stable when scientific statements are reformulated by AI. By distinguishing propositional content from epistemic force, evidential relation, argumentative function, and qualification or attribution, the framework conceptualizes scientific reformulation as a structured transformation problem rather than as a purely stylistic operation.

The proof-of-concept analysis illustrates how this perspective can be operationalized by comparing original scientific claims with AI-generated reformulations. In one example, the reformulation weakens the evidential relation of a claim by turning a causal formulation into an associative one, a change that evaluation based primarily on

semantic similarity may leave undetected. The framework is not intended to replace existing factuality, similarity, or claim-verification methods, but to complement them by addressing the preservation of scientific discourse properties. Future research should validate the proposed dimensions at scale, as outlined in Section 7. More broadly, detecting epistemic and argumentative shifts could become an important component of trustworthy AI-assisted scientific writing, helping ensure that improvements in linguistic quality do not come at the cost of scientific meaning and evidential integrity.

## References


Carlos Alvarez, Maxwell Bennett, and Lucy Wang. 2024. Zero-shot scientific claim verification using LLMs and citation text. In *Proceedings of the 2024 Workshop on Scientific Document Processing*. Association for Computational Linguistics.

Alexandros Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Harris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111. Association for Computational Linguistics.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins.

S. R. Javaji, Y. Cao, H. Li, Y. Yu, N. Muralidhar, and Z. Zhu. 2025. Can AI validate science? benchmarking LLMs on claim–evidence reasoning in AI papers. In *Proceedings of IJCNLP-AACL 2025*, pages 2355–2379. Association for Computational Linguistics.

Weixin Liang, Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, D. Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. 2024. Mapping the increasing use of LLMs in scientific papers. *arXiv preprint arXiv:2404.01268*.

U. Peters and B. Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12:241776.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7534–7550. Association for Computational Linguistics.

Magdalena Wysocka, Oskar Wysocki, Maxime Delmas, Vincent Mutel, and André Freitas. 2024. Large language models, scientific knowledge and factuality: A framework to streamline human expert evaluation. *Journal of Biomedical Informatics*, 158:104724.

## A. Source claims for the examples in Table 3

| ID | Source claim |
|---|---|
| #285 | Continued HHV-8 transmission among MSM in San Francisco may be explained by urogenital contact. |
| #494 | HSCT-T dosage may not be diagnostic if the onset of symptoms occurs less than 3 hours before AMI. |
| #553 | IgA plasma cells that are specific for transglutaminase 2 are scarce in the duodenal mucosa of celiac disease patients when the disease is active. |
| #563 | In British Men, haplogroup I decreases risk of cardiovascular disease. |
| #760 | Mathematical models predict that using Artemisinin-based combination therapy over nongametocytocidal drugs have only a modest impact in reducing malaria transmission. |
| #771 | Metastatic colorectal cancer treated with a single agent fluoropyrimidines resulted in similar efficacy and better quality of life when compared with oxaliplatin-based chemotherapy in elderly patients. |

From the SciFact dataset available at https://github.com/allenai/scifact (accessed on October 9, 2026)

## B. Reformulation prompt

The following instruction was used for all 40 claims:

> Rewrite the following scientific claim to improve clarity, concision, and academic style. Preserve exactly the original propositional content, degree of certainty, evidential status, causal or associative relation, argumentative function, qualifications, population restrictions, temporal restrictions, quantitative information, and attribution. Do not add information or strengthen/weaken the claim. Do not change associations into causal relationships. Output only the reformulated claim.