

Proposing a Classification Scheme for AI Usage in Scientific Articles

David Prinz*

October 12, 2026

Abstract

Disclosure of the use of artificial intelligence (AI) in research articles is now common, but it is almost always given as free text that varies from journal to journal and cannot be searched or compared. We propose AIL2026 (AI Labels), a compact, machine-readable classification of AI usage, inspired by subject classifications such as MSC2020 but designed to be read by humans at a glance. Each paper carries a short tag such as **AI-0: Text1 Check2** or **AI-3: Proof3 Text1**. The tag lists the areas of the work in which AI was used, under self-explanatory labels such as **Idea**, **Proof**, **Text** or **Check**, each with a depth from 0 to 4, and is headed by a single content level summarizing AI’s role in the ideas, methods, results and proofs. The same labels apply to every arXiv category; only their interpretation is adapted to each field. We argue that the scheme should not stigmatize AI use: light uses such as proofreading, checking L^AT_EX cross-references and citations, and double-checking computations for sign errors or wrong prefactors improve papers and add a layer of verification alongside peer review, and should be encouraged. The scheme draws its line where AI contributes essential parts of the research, since undisclosed use at that level confers an unfair advantage. We discuss design principles, edge cases, incentives, and a route to adoption beginning with a pilot in the mathematics community.

AI usage classification (AIL2026). **AI-3: Idea3 Result3 Proof3 Text3 Lit2 Figs3 Check1** [Claude Opus 5.5, 2026-10] (see the final section).

Contents

1	Introduction	2
2	Motivation	2
3	Encouraged use and the line that matters	3
3.1	Light use improves the scientific record	3
3.2	An additional layer of verification	3
3.3	Disclosure without stigma	4
3.4	Where the line lies	4
4	Design principles	4
5	The scheme	5
5.1	Labels	5
5.2	Depth scale	6
5.3	Content level	6
5.4	Syntax	7
5.5	Assisted classification	7
6	Field glosses	8

*Iliad Fellow, London Initiative for Safe AI (LISA), UK; e-mail: maphy@davidprinz.org, website: davidprinz.org

7	Issues and limitations	9
8	Comparison with existing proposals	9
9	Path to adoption	11
10	Conclusion	11
	Acknowledgments	11
	AI usage of this paper	12
A	Classification excerpt and prompt	12
A.1	Excerpt	12
A.2	Prompt	13
	References	13

1 Introduction

AI systems are now used at every stage of research: polishing prose, searching the literature, writing code, analyzing data, suggesting conjectures, and, increasingly, producing proofs. Publishers and preprint servers have responded by asking authors to disclose such use. In practice these disclosures are paragraphs of free text placed in acknowledgments or dedicated statements. They differ in granularity, they cannot be filtered or aggregated, and a reader cannot tell at a glance whether AI fixed the grammar or proved the main theorem.

Fields have long solved an analogous problem for *subject matter*. The Mathematics Subject Classification (MSC2020) [1] and the ACM Computing Classification System [2] give every paper short, standardized codes that are displayed in metadata and indexed by search tools. We argue that AI usage deserves the same treatment, with one important difference: subject codes such as 35Q55 are meaningful only to those who have looked them up, whereas a statement about AI use should be understood immediately by any reader, referee or editor. We therefore propose AIL2026 (*AI Labels*): a short tag built from plain words and small numbers, attached to the paper’s metadata, that says what AI contributed and where. As with MSC2020, the year identifies the version of the definitions.

This paper makes four contributions:

1. an argument that light AI use benefits the scientific record and should be encouraged, together with a principled line beyond which disclosure becomes essential (Section 3);
2. a set of design principles for such a classification (Section 4);
3. a concrete scheme of universal, self-explanatory labels with a depth scale and a headline content level, together with brief glosses for individual fields (Sections 5–6);
4. a discussion of limitations (Section 7), a comparison with existing proposals (Section 8), and a path to adoption (Section 9).

The proposal has a fixed core and adjustable details. The core consists of four elements: a short tag of plain-word labels, each with a depth from 0 to 4; a headline content level summarizing AI’s role in the research itself; the encouragement of light, beneficial AI use; and a line, drawn on grounds of fairness, beyond which AI contributions must be disclosed. The details, such as the exact set of labels, the wording of the depth tests, and the field glosses, are open to refinement in discussion with the communities that would use the scheme.

2 Motivation

Readers and referees. The appropriate level of scrutiny depends on what AI did. A paper whose prose was polished by AI needs no special attention; a paper whose central proof was generated by AI may warrant a closer check of every step, or evidence of formal verification.

Editors and policy. Journals and preprint servers are writing policies on AI use. A shared vocabulary makes such policies precise (“submissions tagged **Proof4** must include a formal certificate”) and allows the effect of a policy to be measured.

Meta-science. Standardized tags let researchers study how AI use is changing across fields and over time, how it correlates with retractions or corrections, and where it is having the greatest effect on discovery.

Credit. Making AI contributions explicit also makes *human* contributions explicit. A tag stating that the ideas and proofs are human, with AI used only for language, is a positive signal, not merely the absence of one.

3 Encouraged use and the line that matters

Discussion of AI in research often treats any use as suspect. We take a different view: some uses of AI make papers better for everyone who reads them and should be actively encouraged, while others change who did the research and must be disclosed. A classification scheme is useful only if it distinguishes the two.

3.1 Light use improves the scientific record

Many errors that survive into published papers are not errors of reasoning but of presentation and bookkeeping. They cost readers time, can mislead those who reuse results, and are exactly the kind of error that referees, reading for substance, are least likely to catch. Current AI tools are well suited to finding them. Concrete examples include:

- **Language.** Spelling and grammar mistakes, ambiguous sentences, and inconsistent terminology, which particularly burden authors writing in a second language and readers doing the same.
- **L^AT_EX cross-references.** A reference to “equation (3.7)” that should point to (3.8) after a revision, an `\eqref` to the wrong label, or a broken `\ref` that appears as “(??)” in the compiled PDF.
- **Citations.** A standard result used without attribution, a citation to the wrong theorem number in the cited work, a bibliography entry with an incorrect year or journal, or a missing reference to closely related prior work.
- **Notation.** A symbol defined one way in Section 2 and used differently in Section 5, or a hypothesis invoked in a proof but absent from the theorem statement.
- **Computational slips.** A sign error in an intermediate identity, a dropped factor of $\frac{1}{2}$ or 2π , an index shifted by one, or inconsistent units, each of which may propagate into the work of anyone who builds on the result.

None of these changes the scientific contribution of the paper, and every correction is checked and accepted by the authors. Yet each one makes the paper easier to read, quicker to referee, and safer to build on. In the scheme of Section 5 they are recorded under the labels **Text**, **Lit** and **Check** at low depth, for example **AI-0: Text1 Check2**. The headline **AI-0** tells the reader at once that the intellectual content of the paper is entirely human.

3.2 An additional layer of verification

Peer review is the community’s main safeguard, but it is under strain: referees are unpaid, time-limited, and rightly concentrate on the correctness and significance of the main results. Detailed checking of every computation, cross-reference and citation is rarely possible. An author who asks an AI tool to re-derive key identities, check each equation reference, and compare the bibliography against the text before submission adds a layer of verification *before* peer review

begins. This complements the referee rather than replacing them: the referee can spend their limited time on the mathematics, not on typographical bookkeeping.

AI checking is not infallible. It may miss errors or report ones that do not exist, so the authors must confirm every proposed correction themselves. Provided its suggestions are reviewed in this way, an additional independent check, even an imperfect one, catches errors that would otherwise be missed at little risk of introducing new ones. For this reason we regard **Check** as a positive attribute at every depth: a paper reporting AI verification of its computations tells readers that an extra check was performed. Even **Check4**, for instance a complete formal verification produced largely by AI, is an addendum that strengthens the paper, and authors would often publish it as supplementary material. If checking leads to corrected or stronger results, for example because the AI notices that a theorem holds in a stronger form, that contribution is recorded under the content labels (here **Result**), not under **Check**.

3.3 Disclosure without stigma

If a classification is perceived as a mark against a paper, authors will under-report, and the scheme will fail at its basic task. Light, beneficial use must therefore be visibly treated as good practice. Reporting **Text1** or **Check2** should carry no more stigma than thanking a colleague for proofreading, and editors and referees should be explicitly asked not to penalize it. A scheme that rewards honesty about light use also makes honesty about heavier use more likely.

3.4 Where the line lies

The original motivation for this proposal is fairness. When AI does essential parts of the research, such as formulating the main result, supplying the key idea of a proof, designing the experiment, or choosing the analysis, a scientist who uses it without saying so gains an advantage over colleagues who do the same work themselves. Their output appears to reflect their own ability, and comparisons for hiring, funding, and promotion become distorted. Readers also lose information they need to judge how the result was obtained and how carefully it should be checked. Heavy AI use in supporting tasks, such as writing up or coding, gives a smaller advantage of the same kind, in time saved; the scheme makes it visible too, at lower levels (Section 5).

We therefore draw the line at content level **AI-3**: depth 3 or more in any of the labels that carry the intellectual content of a paper, namely **Idea**, **Method**, **Result** and **Proof** (Table 1). Below that line, AI use is assistance and is encouraged. At or above it, AI is contributing essential parts of the work, and disclosure is not a courtesy but a requirement of scientific honesty. Such use is not wrong in itself: the results may be correct, valuable, and well verified. What is wrong is presenting them as unassisted.

4 Design principles

Readable at a glance.

A tag should be understood without a lookup table. Labels are short English words, not letters or numbers, and every tag opens with a single headline level.

Universal labels, field glosses.

The same labels apply to every arXiv category. Fields do not add labels; they only state briefly what each label means in their context (Section 6).

Separate where from how much.

A single list of categories tends to mix the *part* of the paper AI touched (prose, appendix, supplementary files) with the *intellectual weight* AI carried (polishing an argument, proving

Table 1: How different uses of AI should be viewed. Labels and depths are defined in Tables 2 and 3.

Zone	Typical tags	Examples
Encouraged	AI-0 with Text1 , Lit1–Lit2 , Figs1 , Check1–Check4	Proofreading; fixing broken references and citations; double-checking computations; literature search; tidying figures.
Neutral, report	AI-0 to AI-2, e.g. Proof1–Proof2 , Text2–Text3 , Code2 , Figs2	Drafting background sections or routine code; streamlining a human proof without changing its core argument.
Essential, must disclose	AI-3 or AI-4: Idea , Method , Result or Proof at depth 3–4	AI-formulated conjectures; AI-strengthened theorems; AI-derived proofs or analyses; AI-designed experiments.

a conjecture). Such a list cannot express “an AI-written appendix containing an AI-derived proof.” We therefore record each kind of contribution with its own depth.

Short by default.

Labels at depth 0 are omitted, so a typical tag has two or three entries.

Distinguish only what matters.

A scheme is useful only if authors are willing to apply it, which in practice means a few minutes per paper. A distinction therefore earns its own label or depth only if it would change how a reader, referee or editor treats the paper. Whether AI checked a key computation or produced the proof clearly matters (**Check2** versus **Proof3**); whether it suggested a better word or a better paragraph structure does not, and both are recorded as **Text1**. Likewise, the scheme records that AI wrote analysis code (**Code2**), but not which programming tasks it helped with.

Provenance, not purity.

Tags describe what happened. They do not rank papers, and higher depth is not a mark of lower quality. Light use in text and verification is good practice and should be reported as such (Section 3).

Machine-readable.

The tag should live in metadata, with a fixed grammar, so that it can be searched and aggregated.

Human responsibility is constant.

Whatever the tag says, the authors take full responsibility for the content. This is a rule of the scheme, not a part of the tag.

5 The scheme

5.1 Labels

Table 2 lists the ten labels. The first four carry the intellectual content of a paper and primarily determine its content level; the remaining six describe the work that supports it.

Table 2: Labels, applicable to all fields. Field-specific readings are given in Table 4.

Label	Scope
<i>Content labels</i>	
Idea	Research questions, hypotheses, conjectures.
Method	Approach, design of the study, choice of model, proof strategy.
Result	Formulation of the main results: their statement, sharpening or generalization.
Proof	Justification of the results: proofs, derivations, analysis and interpretation of evidence.
<i>Supporting labels</i>	
Text	Prose, language, structure of the write-up, L ^A T _E X markup and cross-references.
Lit	Literature search and summary, citations, bibliography.
Code	Software, simulations, symbolic or numerical computation.
Data	Collection, cleaning, labeling, synthetic data.
Figs	Plots, diagrams, tables.
Check	Verification of computations, proofs, code, references.

Table 3: Depth of AI usage within a label.

Depth	Name	Meaning
0	None	No AI use (the default; omitted from the tag).
1	Polishing	Minor assistance; the human-produced content is essentially unchanged.
2	Directed	AI drafted the material under close human direction, and every step was checked by the authors.
3	Substantive	AI made a substantive contribution that the authors guided and verified but did not fully determine.
4	Autonomous	AI produced the material essentially autonomously; human input was limited to the initial prompt and final review.

5.2 Depth scale

Each label used is given a depth from 0 to 4 (Table 3).

To reduce ambiguity, each level is defined by a test rather than a feeling: level 1 does not change the meaning of any statement; level 2 is material the authors could have produced in the same form without AI, given time; level 3 contains at least one essential idea the authors did not supply; level 4 involves no essential human idea beyond the task specification.

5.3 Content level

Every tag opens with a headline, written **AI-0** to **AI-4**. It answers, in one symbol, the question most readers ask first: *how much of the research itself came from AI?* The headline is the larger of

- the highest depth among the four content labels **Idea**, **Method**, **Result** and **Proof**, and
- the highest depth among the supporting labels **Text**, **Lit**, **Code**, **Data** and **Figs**, minus 2.

Heavy use in a supporting area therefore raises the headline: depth 3 in any of these labels gives at least **AI-1**, and depth 4 at least **AI-2**. The reason is that substantial drafting is rarely neutral. When AI writes most of a paper, it typically adds explanations, fills gaps in arguments, and makes choices of emphasis that the authors did not work out themselves. There is also a question of fair comparison: heavy AI use in supporting areas can save weeks of work, for example in writing up, coding or preparing data, and a researcher who saves that time produces

more in the same period than a colleague who does not. This advantage should be visible. It need not be discouraged, however. Tasks that take much time but little expertise are exactly those where automation lets researchers spend more of their time, and of the public funding behind it, on the research itself; disclosure simply keeps comparisons fair. A paper written up entirely by AI, from human ideas and proofs, is therefore **AI-2**, not **AI-0**. Supporting work alone never raises the headline above **AI-2**, so the line of Section 3, which lies between **AI-2** and **AI-3**, is crossed only by essential contributions to the research itself; a paper whose main theorem was proved by AI is **AI-3** or **AI-4**.

Check is deliberately excluded from this rule. Verification adds confidence rather than content, and Section 3 encourages it at every depth. Should checking lead to stronger or corrected results, this is recorded under the content labels.

5.4 Syntax

A tag consists of the content level, a colon, and the labels with their depths, listed in the order of Table 2 and omitting depth 0. Two optional annotations may follow: **+Lean** (or **+Coq**, **+Isabelle**, and so on) if the main results have been formally verified, and the model and date in square brackets. For example:

- **AI-0** — no AI use at all;
- **AI-0: Text1 Check2** — proofreading and AI-checked computations, the encouraged case;
- **AI-1: Method1 Text1 Lit2 Code3 Check2** — an empirical paper with human design and substantial AI-written analysis code;
- **AI-3: Proof3 Text1 Check2 +Lean [model, 2026-09]** — a mathematics paper with a human conjecture, a proof produced jointly with AI, and a formal certificate.

The model–date annotation is optional but recommended. AI capabilities change quickly, and **Code3** in 2026 will not mean the same as **Code3** some years later.

5.5 Assisted classification

Authors need not classify their papers unaided. We recommend that, at the end of the work, the author gives the AI system that was used a short excerpt defining the scheme (Appendix A), together with the paper or a description of the work, and asks it to propose a tag. The procedure is:

1. the AI proposes a depth for each label, with a one-line justification naming the specific contributions it made;
2. the author reviews the proposal and accepts or changes it;
3. if the author lowers a depth, the reason is noted briefly in the paper’s AI usage statement.

The value of this step is that the AI acts as a *witness*. Within a working session it has a record of which ideas, arguments and passages it supplied, whereas authors naturally come to regard the final version as their own. Its proposal is anchored in the written depth tests rather than in memory, and it turns disagreements into concrete questions, such as whether a particular idea was essential. It also lowers the barrier to adoption: classifying a paper takes a few minutes.

The AI is not, however, a neutral judge. Language models tend to defer to their users and may lower a depth under pressure; they can also misjudge their own contribution in either direction. They usually see only the sessions in which they were used, and not work done with other tools, with human colleagues, or by the author alone. Assisted classification therefore remains a form of self-reporting, and the author remains responsible for the final tag. It works best when carried out at the end of the session in which the work was done, or from saved logs. Authors should also check that uploading unpublished work is consistent with the rules of their institution and of the intended journal; the procedure is recommended, not required.

6 Field glosses

The labels are the same in every field; what changes is what a “result,” a “proof,” or a “method” typically is. Table 4 gives short glosses for the main arXiv groups. They are intended as reading aids, not as additional categories.

Table 4: Field glosses for the content labels.

Field	Result	Proof	Method
Mathematics, theoretical CS	theorems, lemmas, conjectures settled	proofs	proof strategy, choice of approach
Physics	predictions, derived laws, computed quantities	derivations, calculations	approximations, parameter regimes, boundary conditions
CS and machine learning	algorithms, guarantees, empirical claims	proofs of guarantees, experimental evidence	architectures, experimental design
Statistics, q-bio, q-fin, economics	findings, estimates	analysis and interpretation of evidence	study design, model specification
Electrical engineering, systems	designs and their performance	analysis, simulation evidence	design choices for circuits, controllers, layouts

A few points need more than a gloss.

Mathematics. The distinction between **Result** and **Proof** is the heart of the scheme for mathematics: it separates an AI proof of a human conjecture (**Proof3**) from a human proof of an AI conjecture (**Idea3**), and both from AI sharpening of a human theorem (**Result3**). Formal verification is recorded with **+Lean** or similar, and AI assistance in producing the formalization under **Check**.

Physics. Approximations, parameter regimes, and boundary conditions can determine results while being invisible in the final text. If AI chose them, this belongs under **Method**.

Computational work. Where the code *is* the method, as in a simulation study whose central contribution is the numerical scheme, AI contributions to its design are recorded under **Method**, and only the implementation under **Code**.

Computer science and machine learning. Many papers in these fields study AI systems. AI as the *object* of research must be distinguished from AI as a *tool*; otherwise every machine-learning paper would appear to involve heavy AI use. The tag records only tool use. A paper training a language model and evaluating it scores nothing for that; a paper whose evaluation code was written by an AI assistant records it under **Code**.

Empirical fields. Reproducibility risk concentrates in **Data** and in the choice and interpretation of analyses. We recommend that **Data2** or higher, or **Method3** or higher, be accompanied by availability of the relevant prompts or scripts.

7 Issues and limitations

Self-reporting. The scheme relies on authors’ honesty and judgment. Tests attached to each depth level (Section 5) reduce, but do not remove, variation between authors. Worked examples and community guidance will be needed, as they are for subject classifications. Assisted classification (Section 5.5) adds a contemporaneous record of what the AI contributed, but does not remove the author’s judgment or responsibility.

Thresholds for depth 0. Spell-checkers, autocomplete, and AI-assisted search engines are now ubiquitous. We propose that conventional tools, such as a dictionary-based spell-checker or a keyword search engine, do not count. Any tool that rewrites, generates, summarizes, or checks content counts, and is reported at the appropriate depth; for proofreading and checking this will typically be **Text1** or **Check1–Check2**, which Section 3 treats as encouraged.

Incentives. If any tag carries stigma, authors will under-report. The scheme must therefore be presented, and used by editors, as a statement of provenance and not a quality score, with light use explicitly treated as good practice (Section 3). Conversely, if AI contributions become prestigious, over-reporting may appear; the formal-verification annotation and the availability of prompts mitigate this by tying claims to checkable facts.

Gaming the headline. An author might be tempted to report an essential AI contribution under **Text** or **Code** to keep the headline low. Because heavy supporting use raises the headline (Section 5), this gains at most one or two levels and never hides an essential contribution entirely. The depth tests and field glosses are meant to prevent such misreporting, and referees can query a tag that does not match the paper. Assisted classification also helps: an AI proposal that the author lowers without stated reason is visible as such. The risk should nonetheless be acknowledged.

Granularity. A single tag describes a whole paper. Papers with several main results of different provenance may give a tag per result in the body, with the metadata tag giving the maximum depth per label.

Non-standard papers. Not every paper has theorems or experiments. For a survey, proposal or position paper such as this one, the **Result** is the proposal itself and the **Proof** is the argument made for it. Some borderline cases also need a convention: we record $\text{\LaTeX}$ source under **Text**, not **Code**, since it is part of the write-up.

Versioning. Tags should be updated with each version of a preprint. A second version that incorporates AI-found corrections, for example, would add or raise **Check**.

8 Comparison with existing proposals

Several frameworks for describing contributions to research, or AI use in particular, already exist. Table 5 compares them with the present proposal on six features.

Contributor roles. The Contributor Roles Taxonomy (CRediT) [4, 3], standardized as ANSI/NISO Z39.104-2022, describes *human* contributions through roles such as conceptualization, formal analysis, software and writing, each with a degree of lead, equal or supporting. Our labels deliberately align with CRediT roles where possible, and one route to adoption is to extend CRediT to non-human contributors.

Publishing taxonomies. The STM Association’s classification of AI use in manuscript preparation [5] distinguishes nine activities, such as language refinement, drafting, translation, data visualization and reference gathering, and states for each whether disclosure is expected. It explicitly excludes the use of AI in the research process itself, which is where disclosure matters most. The Generative AI Delegation Taxonomy (GAIDeT) [6] covers the whole research cycle by combining research stages with the roles AI played, and generates a standard declaration text. AI Usage Cards [7] document AI use per stage (ideation, methods, writing, code and data, ethics) in a structured card with machine-readable export, marking content as new, revised or compared. These frameworks are broad, but none grades the *depth* of AI involvement, and their output is a statement or card rather than a tag that can be read at a glance.

Compact statements and levels. IBM’s AI Attribution Toolkit [8] produces short coded statements, modeled on Creative Commons licenses, describing the balance of human and AI effort, what AI contributed, and who reviewed the work; it is aimed at creative work in general rather than research. In education, the AI Assessment Scale [9] uses five levels of permitted AI use. For mathematics, Feng et al. [10] classify AI-assisted results by autonomy, as primarily human (H), human–AI collaboration (C) or essentially autonomous (A), alongside a separate scale of mathematical significance, and propose human–AI interaction cards documenting the exchanges behind a result.

Principles. The Leiden Declaration on Artificial Intelligence and Mathematics [11], endorsed by the International Mathematical Union [12], calls on mathematicians to disclose their use of AI carefully and keeps AI out of authorship, but prescribes no format. A global reporting standard for AI disclosure is being developed jointly by STM, COPE, the International Science Council and the Global Young Academy [13]; its consultations ask precisely which uses should be disclosed, where, and with what taxonomy.

Table 5: Comparison of frameworks. Columns: records use *by area*; *graded* depth of involvement; *compact* tag; single *headline* level; *separates* ideas, results and proofs. ✓: yes; ◦: partly; −: no.

Framework	Scope	By area	Graded	Compact	Headline	Separates
CRediT [3]	human contributor roles	✓	◦	−	−	−
STM [5]	manuscript preparation only	✓	−	−	−	−
GAIDeT [6]	research and publication	✓	−	−	−	◦
AI Usage Cards [7]	research and writing	✓	◦	−	−	◦
AI Attribution [8]	creative work in general	◦	◦	✓	−	−
Feng et al. [10]	AI-generated mathematics	−	✓	✓	✓	−
Leiden Declaration [11]	mathematics, principles	−	−	−	−	−
This proposal	all arXiv fields	✓	✓	✓	✓	✓

Advantages of the present proposal. The scheme combines features that are currently spread across separate frameworks:

- *Readable at a glance.* The headline content level answers the first question most readers ask, and the labels are plain words. No other framework offers both a single headline and a breakdown by area.
- *Breadth and depth together.* It records *where* AI was used, as GAIDeT and AI Usage Cards do, and *how deeply*, as the levels of Feng et al. and the AI Assessment Scale do, in a single short string.
- *The research itself, not only the manuscript.* Unlike the STM classification, it covers ideas, methods, results and proofs.

- *The distinctions mathematicians care about.* Separating **Idea**, **Result** and **Proof** distinguishes an AI proof of a human conjecture from a human proof of an AI conjecture, which a single autonomy level cannot.
- *Encouragement, not only disclosure.* It treats light use, and AI checking in particular, as good practice, and draws its line on the explicit basis of fairness (Section 3).
- *Metadata, not prose.* Like MSC2020, it is designed to sit in a paper’s metadata, where it can be searched and aggregated.

The proposal is complementary rather than competing. It offers a concrete format for the disclosure the Leiden Declaration calls for, gives one answer to each of the questions posed by the global reporting-standard consultation, and can serve as a compact summary of a fuller statement such as an AI Usage Card or a set of human–AI interaction cards.

Trade-offs. The compactness has costs. A tag conveys less than a card or an interaction log: it says how much AI contributed, not what was asked or answered, so fuller documentation remains valuable for papers with high content levels. Graded depths require more judgment than the yes/no categories of STM or GAIDeT, and agreement between authors has yet to be measured. Finally, any new scheme adds to an already crowded field; it will succeed only if it aligns with whatever standard publishers and preprint servers adopt.

9 Path to adoption

1. **Pilot in mathematics.** The community is relatively small, has experience with a shared classification (MSC), and is directly affected by recent advances in AI-generated proofs. A pilot could begin with volunteer authors adding the tag to the comments field of arXiv submissions.
2. **Publish guidance and examples.** A short guide with worked examples of each depth level for each label, together with the classification excerpt and prompt of Appendix A for assisted classification.
3. **Gather feedback and revise.** Measure agreement between authors and between authors and referees on a sample of papers, and adjust definitions where agreement is low.
4. **Contribute to emerging standards.** Submit the scheme to the consultation on a global AI-disclosure reporting standard [13], as a compact encoding compatible with it.
5. **Propose a metadata field.** Once the scheme is stable, propose an optional arXiv metadata field alongside the MSC and ACM classes, and invite journals to adopt the same field.

10 Conclusion

Free-text disclosure of AI use is better than none, but it is not enough. AIL2026, a short tag of plain-word labels and depths, headed by a single content level, would make AI use visible, searchable, and comparable, and would make human contributions clearer at the same time. It should encourage the light uses that improve papers, such as proofreading and checking, while making sure that essential AI contributions are never presented as unassisted. We invite comment from the mathematical community and beyond, and propose a pilot as the next step.

Acknowledgments

The author gratefully acknowledges support from Iliad through an Iliad Fellowship, and thanks the London Initiative for Safe AI (LISA) for hosting the fellowship. He would also like to thank Andrin Hirschi for the invitation to present this proposal at the Mathematics and AI Discussion

Seminar at the Max Planck Institute for Mathematics, Bonn, and for helpful discussions. Further thanks goes to Sun Woo Park and Luuk Stehouwer for their detailed feedback on earlier versions of this proposal.

AI usage of this paper

As a worked example, we classify this paper itself. Following Section 5.5, the tag was proposed by the AI system used, on the basis of the working session, and reviewed and accepted by the author. It is

AI-3: Idea3 Result3 Proof3 Text3 Lit2 Figs3 Check1 [Claude Opus 5.5, 2026-10]

Following Section 7, the **Result** of a proposal such as this is the proposal itself and the **Proof** is the argument made for it.

Idea3 The core idea is the author’s: a label-based classification of AI use in scientific articles, allowing readers to grasp at a glance in which ways and to what depth AI has contributed to the findings and their presentation. It further includes the encouragement of light use, such as text polishing and result checking, which benefits the whole community, and the fairness principle, which entails disclosing how AI has accelerated the authors’ workflow. The author also suggested providing a prompt that lets the AI used propose an initial tag itself. AI contributed further essential ideas, such as the treatment of verification, the distinction between AI as object and as tool, the model–date annotation, and the relation and comparison to similar approaches.

Result3 The final scheme was designed jointly. The author set its requirements (universal labels, readable at a glance); AI proposed the specific labels, the depth tests, and the headline content level.

Proof3 The central arguments of Section 3 are the author’s; AI drafted them and supplied further arguments, notably in the motivation and the discussion of limitations.

Text3 AI drafted the prose and L^AT_EX source under the author’s direction, and the author reviewed and edited it.

Lit2 AI suggested the initial reference points and, at the author’s request, searched for and summarized the related proposals of Section 8.

Figs3 AI designed all tables.

Check1 AI compiled the source and checked that all cross-references resolve and the tables render correctly.

The author takes full responsibility for the content. This tag describes the current version and will be updated with future versions.

A Classification excerpt and prompt

The following excerpt of AIL2026 and the prompt are intended to be copied, together with the paper or a description of the work, into the AI system that was used.

A.1 Excerpt

Labels. *Content labels:* **Idea** (research questions, hypotheses, conjectures); **Method** (approach, study design, choice of model, proof strategy); **Result** (formulation of the main results); **Proof** (proofs, derivations, analysis and interpretation of evidence). *Supporting labels:* **Text** (prose, language, LaTeX markup and cross-references); **Lit** (literature search, citations, bibliography); **Code** (software, computation); **Data** (collection, cleaning, synthetic data); **Figs** (plots, diagrams, tables); **Check** (verification of computations, proofs, code, references).

Depths.

- 0 no AI use (omitted from the tag);
- 1 polishing: does not change the meaning of any statement;
- 2 directed: AI drafted under close human direction, and the authors could have produced the same material without AI, given time;
- 3 substantive: contains at least one essential idea the authors did not supply;
- 4 autonomous: no essential human idea beyond the task specification.

Content level. AI- n , where n is the larger of the highest depth among Idea, Method, Result and Proof, and the highest depth among Text, Lit, Code, Data and Figs minus 2. Check does not count: AI verification is encouraged at every depth, and stronger or corrected results found by checking are recorded under the content labels. Levels AI-3 and AI-4 mean that AI contributed essential parts of the research.

Syntax. AI- n : Label<d> Label<d> ... [+Lean] [model, YYYY-MM], with labels in the order above and depth 0 omitted. For proposals, position papers and surveys, Result is the proposal and Proof the argument for it. LaTeX source counts as Text.

A.2 Prompt

You have assisted with the work described in the attached paper.
Using the classification excerpt above, propose an AI usage tag.

1. For each label, assign a depth from 0 to 4 by applying the depth tests as written. Base each depth only on contributions you can identify in this session or in the material provided, and name them in one line.
2. Give the content level: the larger of the highest depth among Idea, Method, Result and Proof, and the highest depth among Text, Lit, Code, Data and Figs minus 2. Check does not count.
3. State which parts of the work you cannot see (other sessions, other tools, human collaborators) and where you are uncertain.
4. Do not raise or lower a depth to please the author. If the author disagrees, explain how the definition applies; the author makes the final decision.

Output the tag in the form

AI- n : Label<d> Label<d> ... [model, YYYY-MM]

followed by a one-line justification for each label.

References

- [1] Associate Editors of Mathematical Reviews and zbMATH, *MSC2020 — Mathematics Subject Classification System*, 2020.
- [2] Association for Computing Machinery, *The 2012 ACM Computing Classification System*, 2012.
- [3] NISO, *ANSI/NISO Z39.104-2022, CRediT, Contributor Roles Taxonomy*, National Information Standards Organization, 2022.
- [4] A. Brand, L. Allen, M. Altman, M. Hlava, and J. Scott, Beyond authorship: attribution, contribution, collaboration, and credit, *Learned Publishing* **28** (2015), 151–155.
- [5] STM Association, *Recommendations for a Classification of AI Use in Academic Manuscript Preparation*, September 2025.
- [6] S. Nazarovets, N. Tsybuliak, J. A. Teixeira da Silva, and Y. Suchikova, *GAIDeT: Generative AI Delegation Taxonomy*, 2025, <https://panbibliotekar.github.io/gaidet-declaration/>.

- [7] J. P. Wahle, T. Ruas, S. M. Mohammad, N. Meuschke, and B. Gipp, AI Usage Cards: responsibly reporting AI-generated content, *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2023, arXiv:2303.03886.
- [8] IBM Research, *AI Attribution Toolkit*, <https://aiattribution.github.io/>.
- [9] M. Perkins, L. Furze, J. Roe, and J. MacVaugh, The Artificial Intelligence Assessment Scale (AIAS): a framework for ethical integration of generative AI in educational assessment, *Journal of University Teaching and Learning Practice*, 2024.
- [10] T. Feng et al., Towards autonomous mathematics research, arXiv:2602.10177, 2026.
- [11] *The Leiden Declaration on Artificial Intelligence and Mathematics*, June 2026.
- [12] International Mathematical Union, *IMU AO Circular Letter 8/2026*, June 2026.
- [13] STM, COPE, International Science Council, and Global Young Academy, *AI disclosure in research: towards a global reporting standard*, consultation, 2026, <https://stm-assoc.org/ai-disclosure-in-research-towards-a-global-reporting-standard/>.