Journal of Information Literacy
Volume 20 Issue 1
ISSN 1750-5968

Project Report

Adding a simple log to in-text citations to strengthen information literacy and argumentation while blocking casual misuse of generative AI

http://dx.doi.org/10.11645/20.1.847

Peter Tamas
Lecturer, Wageningen University. Email: peter.tamas@wur.nl. ORCID: 0000-0002-5409-1273.

Leonie Kamminga
Information Specialist, Wageningen University. Email: leonie.kamminga@wur.nl. ORCID: 0000-0002-0508-7796.

Abstract

This report describes an intervention that required students to submit a structured log demonstrating engagement with the literature alongside a required report to strengthen information literacy, reinforce argumentation skills and limit the ease of misuse of generative AI. The intervention was deployed in a class of 120 interdisciplinary MSc students in the environmental sciences. It required students to (1) examine how citations function in argument, (2) practice source evaluation using a stepwise protocol, and (3) submit a supplemental citation log. Supporting instructional materials addressed the conceptual foundations of citation as argument infrastructure, introduced a tiered model of justification for research decisions, and provided classroom exercises to build relevant skills. All student groups produced citation logs that demonstrated engagement with source material, though most submissions contained minor formatting and file-handling inconsistencies that impeded automated validation. Instructors observed that this requirement did not pose unusual difficulty and open student evaluation comments did not single out the new expectations. The pilot suggests that these additions are workable, that they make students' engagement with the literature more transparent and that they may constrain casual misuse of generative AI. Looking forward, clearer instructions and in-class guidance will be needed to reduce avoidable submission errors and support efficient assessment.

Keywords
academic integrity; citations; generative AI; information literacy; source evaluation

Tamas & Kamminga. 2026. Adding a simple log to in-text citations to strengthen information literacy and argumentation while blocking casual misuse of generative AI. Journal of Information Literacy, 20(1), pp. 208–236. http://dx.doi.org/10.11645/20.1.847

1. Introduction

As part of a larger project, we asked students to undertake one aspect of active citation (Moravcsik, 2010): enabling readers' access to the data that supports the claim backed through citation. This request was motivated by an interest in improving both their information literacy (IL) and argumentation skills. Our objectives for IL were that students see papers not as separable knowledge entities but as high dimensional nodes in a complex network of claims, that they know how citation functions within this network, that they care that what they write is properly woven into this network, and that they have the tools required to test the integrity of the network of claims within which their own efforts are housed. Turning now to argumentation, we expect our students to be able to position themselves within this web through careful consideration of an appropriate diversity of plausible options. The forms of IL and argumentation that we are working with are reasonable to expect of our students today. What this project adds to these reasonable expectations is a slight formalisation of both coupled with a reporting format, a log, that enable students to demonstrate their capacity to engage with the literature properly while blocking causal abuse of generative AI.

The aspect of IL required by our project is source evaluation. More specifically, students must be able to determine the extent to which a claim made in their text is supported by the content of the cited source. We found many examples in the literature that describe how researchers have studied citation practices, including how citation error rates are measured. However, we did not find guidance aimed at students that matched the needs of this project. In particular, existing materials rarely focus on the practical task of checking whether specific citing claims in one text are warranted by claims in cited texts. The instructional material developed for this course, therefore, serves several purposes. It introduces the functions of citation within the academic record and outlines the current state of citation practice. It then specifies how citation is used and specifies which types of citation are to be checked. Finally, it introduces a stacking 'how to' protocol with examples (appendix 1).

Turning now to argumentation, all students we have worked with have been taught argument structure at some point in their education. However, we have not seen much in the way of evidence of this training in the reports they have submitted. When we reviewed our own assignments, we also recognised a practical constraint: there are far more decisions required in our assignments than would reasonably fit in the time required for each to receive what we recognise as full justification. Made formal, full justification requires identification of plausible alternatives, definition of assessment criteria, specifying a measurement strategy, carrying out measurements of each alternative, assessing the results, and carrying the limitations found for the option chosen forward through the rest of their work. We, therefore, provided students with material that described several levels of argumentation and clearly stated our expectations for which level was required at what point in our assignment description.

2. Intervention

Our intervention was located in the context of a semester long building assignment worth half the final grade in which groups prepared a research proposal. This assignment required them at several points to reference the literature to assemble arguments in support of design decisions. In this class we have historically stated and made clear in our rubrics that our interest is in the transparency and the quality of argument. Despite this, we noted that many of our groups

seemed to spend the time they had on substantive research in the hopes of finding the 'correct' decision rather than on transparently documenting necessarily imperfect decisions. For this cycle we decided to clarify our expectations with respect to both argument and citation checking. These are discussed separately before their integration is presented.

For argumentation, students were provided with a set of power-point slides (appendix 4) that presented a version of Toulmin's approach to argument analysis that was simpler on one dimension and more complex on another. Starting with the simplification, these slides only discuss 'claim', 'data' and 'warrant.' They discuss none of the other components, such as 'qualifier' and 'backing', that are frequently introduced in classes on argument analysis. The function rendered by these components was adequately met by telling students that the data and the warranting required for one argument may be claims arising from another and that all warrants, just like all scientific models, are specific to a determinate set of circumstances. The requirement that we be able to represent features of argument other than that of a basic syllogism were, therefore, met by recursion and specification.

With that foundation in place, we then set very clear expectations for argument in our assignment. Our baseline requirement was that each decision made be supported by at least some justification. At the very minimum they were required to use logic/argumentation to demonstrate careful consideration. For those cases where students could not reasonably be expected to do the research required to provide a substantive rationale for a decision, we invited them to invent a patently ridiculous citation, such as (Astley, 1985), so that we would know that they knew that research was required but practically impossible. The second level of justification we specified was 'partial justification'. For partial justification students were permitted to cherry-pick a few sources to back up the choices they made which would, for full points, require they demonstrate sensitivity to limitations. The most serious consideration we specified, 'full justification', required students to draw on at least four scientific sources, to decide on (and report) which criteria best fit their circumstances when choosing between options, to determine how to measure the candidate sources for these criteria, to measure those sources, to analyse the results of the measurement and, based on all that, to decide which option was least bad while carefully carrying forward the bad parts of the option chosen as limitations.

For guidance on checking citations, the students received a text prepared by the instructor sketching a simplified history of and discussion of the role of citation in enabling the extended conversation that is science (appendix 1). This text introduced a framework for understanding why authors cite and, for those citations that bring forward specific components to support the argument the citing author is making, it proposes a four-stage approach to validation. The stages are:

1. Consolidation: There are four distinct components to this first step. First, testing if the in-text citation matches an entry in the bibliography. Second, checking that all bibliography entries have matching in-text citations. Third, determining if the bibliography item matches a record that does or has existed. And fourth, determining if that record is possible to retrieve.
2. Textual entailment: Examining surface features of the cited work for content that either supports or contradicts the citing claim.
3. Validity: Determining whether the entailing cited content is empirically warranted within the context of the record that it is found (e.g. while results may be claimed, the study out of which they arose may be empirically unsound).

4. Transferability: Determining the extent to which claims which may be appropriate in the context of the cited work are also appropriate in the context of the citing work.

In addition to this text students received a supplement which contained an early draft of a stacking set of exercises that may be useful in walking through the steps of citation in a manner that demonstrates the fragility of argument in the academic record, improves understanding of the pipelines that constitute chatbots and reveals the instability of the results they produce (appendix 4).

Our intervention required students to submit a log that demonstrated textual entailment. This log is a zip file containing a spreadsheet with columns “citing quotation from report”, “source file name”, “supporting statement from source” and all of the files listed. Validation of this supplemental material is done automatically using a script. An example of this submission format may be found in the supplemental material.

Assessment for argument and appropriate citation were handled separately and differently. For argument, we inserted rows in our grading rubric in each section where it was required and, for each, we assessed against the expectation set for that section. For example, in a section where we state that we expect at most arguments supporting the assertion that an arbitrary choice fits, we do not reward consideration and rejection of alternatives. Our assessment of the literature engagement logs submitted by students were considered on the same model as plagiarism: students were required to submit a complete log and, if analysis of a log found either that a supporting sentence presented in the spreadsheet did not exist or that there was directly contradicting information in the cited source, the report was returned to the students to repair and resubmit.

The student reading material on citation (appendix 1), the material on argumentation (appendix 2) explicit instructions (appendix 3) and the initial draft of exercises (appendix 4) were placed in the course guide and assignment instructions were inserted into two cycles of a highly diverse interdisciplinary MSc class in the environmental sciences (YRM20306, n=163) course at Wageningen University in the Netherlands.

3. Ethical review

When queried, the university ethics committee responded that our efforts were a normal curricular update, we are not collecting any novel data from students and all reporting was handled by the university’s fully anonymised course review system, so it did not require any additional consideration.

4. Assessment

Assessment of the innovation from the student perspective was undertaken through analysis of relevant comments made by students in response to the open items ‘what were the strengths of the course’ and ‘what could be improved’ on the standard fully anonymous end of class evaluation form. Assessment of the innovation from the perspective of the instructors was undertaken through notes taken and confirmed by participants during the normal debriefing session we hold after each cycle of the course.

5. Results

For this assignment, the class of 120 was divided into 32 groups. Every group submitted a log file that demonstrated substantial engagement with the literature. Manual inspection of the logs submitted found that every group was able to create a table in which they placed quotes from their report, quotes from the cited source and an identifier of that source. Their submissions, however, contained many minor errors that made machine interpretation of their submissions difficult. For example, 12/32 named their files correctly, 14/32 submitted a CSV, 15/32 named their folders correctly,17/32 had the correct folder structure and 23/32 converted all sources to PDF, 4/32 groups included URLs in their tables rather than attached file and many PDF files submitted were not readable. In addition, variations in the formatting of the submitted files, inspection of the tables provided by students found frequent reference to features of the sources that are not compatible with checking by a simple script. For example, students reference figures and tables as supporting their claims neither of which are reliably recognised and properly handled by scripts.

During the course students asked no more questions about the added expectations for citation and argument than they did for other expectations that we had. When students did raise comments in the normal one-on-one discussion that happens in an interactive class, they usually contained some combination of frustration with the extra work, acknowledgement that this was part of transparency, and, looking specifically at the citation reporting requirement, that the supplemental information made things more fair because it blocked students from using generative AI to produce work that looked as good or better than their manual efforts.

Instructors involved in the courses reported no more difficulty with this component than encountered with other ‘annoying’ requirements such as properly formatted bibliographies. No students mentioned the reporting requirement in their evaluation of the course first course. At the time of submission, evaluation for the second cycle of courses has not been submitted.

6. Conclusion

The students in our class accepted transparent reporting of their engagement with the literature as a part of an assignment but they demonstrated a remarkable ability to misinterpret expectations. The implications of this initial result are that it is reasonable to expect students to provide citation logs but some combination of improved instructions, increased assistance during class and automated submission validation will be necessary to support low-effort inspection of this supplemental information.

Declarations

Ethics approval

Ethical approval was not considered necessary in alignment with the Wageningen University's guidance on the conduct of ethical research.

Funding

Not applicable.

AI-generated content

None of the content is AI generated. Prior drafts were improved by AI pointing out, in no uncertain terms, where I was missing content that would normally be expected.

References

Akhtar, M., Cocarascu, O., & Simperl, E. (2023). Reading and reasoning over chart images for evidence-based automated fact-checking. In A. Vlachos & I. Augenstein (Eds.), *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 399–414). Association for Computational Linguistics.

Al-Benna, S., Rajgarhia, P., Ahmed, S., & Sheikh, Z. (2009). Accuracy of references in burns journals. *Burns*, *35*(5), 677–680.

Armstrong, M. F., Conduff, J. H., Fenton, J. E., & Coelho, D. H. (2018). Reference errors in otolaryngology-head and neck surgery literature. *Otolaryngol Head Neck Surgery*, *159*(2), 249–253.

Brown, A. (2025, January 27-30). Why most published research findings are still false, and why you should care. *2025 Pan Pacific Strategic Electronics Symposium*.

Camp, N. T., Bengtson, J. A., & Sandstrom, J. C. (2025). The citation catastrophe: Propagation of AI-generated counterfeit citations in scholarship. *The Journal of Academic Librarianship*, *51*(4), 103065.

Curlewis, K., Leung, B., Sinclair, L., Ricketts, D., & Rogers, B. (2023). Quotation errors related to the wound management of open lower limb fractures (WOLLF) randomized clinical trial. *European Journal of Orthopaedic Surgery and Traumatology*, *33*(4), 701–707.

Davids, J. R., Weigl, D. M., Edmonds, J. P., & Blackhurst, D. W. (2010). Reference accuracy in peer-reviewed pediatric orthopaedic literature. *Journal of Bone and Joint Surgery - American Volume*, *92A*(5), 1155–1161.

Deemter, K. van. (2024). The pitfalls of defining hallucination. *Computational Linguistics*, *50*(2), 807–816.

Fleck, L. (1935). *Entstehung und entwicklung einer wissenschaftlichen tatsache: Einführung in die lehre vom denkstil und denkkollectiv*. B. Schwabe.

Gazendam, A., Cohen, D., Morgan, S., Ekhtiari, S., & Ghert, M. (2021). Quotation errors in high-impact-factor orthopaedic and sports medicine journals. *JBJS OPEN ACCESS*, *6*(3).

Gosling, C. M., Cameron, M., & Gibbons, P. F. (2004). Referencing and quotation accuracy in four manual therapy journals. *Manual Therapy*, *9*(1), 36–40.

Henry, S. (1982). Citation context analysis. *Progress in Communication Sciences 3*, *0*, 287–310.

Horbach, S. P. J. M., Aagaard, K., & Schneider, J. W. (2021). *Meta-Research: How problematic citing practices distort science*. OSF.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine*, *2*(8), 0696–0701.

Ioannidis, J. P. A. (2018). Massive citations to misleading methods and research tools: Matthew effect, quotation error and citation copying. *European Journal of Epidemiology*, *33*(11), 1021–1023.

Jin, Q., Chen, F., Zhou, Y., Xu, Z., Cheung, J. M., Chen, R., Summers, R. M., Rousseau, J. F., Ni, P., Landsman, M. J., & others. (2024). Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *Npj Digital Medicine*, *7*(1), 190.

Luo, M., Li, C. C., Molina, D., Andersen, C. R., & Panchbhavi, V. K. (2013). Accuracy of citation and quotation in foot and ankle surgery journals. *Foot & Ankle International*, *34*(7), 949–955.

Maes, S. (2024). *Fixing reference hallucinations of LLMs*. Center for Open Science.

Mattiazzi, A., & Vila-Petroff, M. (2024). Unveiling the ethical void: Bias in reference citations and its academic ramifications. *Current Research in Physiology*, *7*.

Mogull, S. A. (2017). Accuracy of cited “facts” in medical research articles: A review of study methodology and recalculation of quotation error rate. *PLOS ONE*, *12*(9).

Moher, D., Tugwell, P., & Jadad, A. R. (1996). Assessing the quality of randomized controlled trials: Current issues and future directions. *International Journal of Technology Assessment in Health Care*, *12*(2), 195–208.

Moravcsik, A. (2010). Active citation: A precondition for replicable qualitative research. *PS: Political Science & Politics*, *43*(1), 29–35.

Nigel Gilbert, G. (1977). Referencing as persuasion. In *Social studies of science* (Vol. 7, Issue 1, pp. 113–122). Sage Publications Sage CA: Thousand Oaks, CA.

Reddy, M. S., Srinivas, S., Sabanayagam, N., & Balasubramanian, S. P. (2008). Accuracy of references in general surgical journals - An old problem revisited. *Surgeon - Journal of the Royal Colleges of Surgeons of Edinburgh and Ireland*, *6*(2), 71–75.

Shigarov, A. (2023). Table understanding: Problem overview. *WIREs Data Mining and Knowledge Discovery*, *13*(1), e1482.

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, *99*(4), 178–182.

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. In *Scientometrics, 121,* 1635–1684.

Tfelt-Hansen, P. (2015). The qualitative problem of major quotation errors, as illustrated by 10 different examples in the headache literature. *Headache*, *55*(3), 419–426.

Toulmin, S. (1958). *The uses of argument*. Cambridge University.

Wager, E., & Middleton, P. (2007). Technical editing of research reports in biomedical journals. *Cochrane Database of Systematic Reviews*, *2*.

Weinstein, M. (1990). Towards an account of argumentation in science. *Argumentation*, *4*, 269–298.

Wright, M., & Scott Armstrong, J. (2008). The Ombudsman: Verification of citations: Fawlty towers of knowledge? *Interfaces, 38*(2), 125–139.

Xu, Z., Jain, S., & Kankanhalli, M. (2025). Hallucination is inevitable: An innate limitation of large language models. In *Arxiv*.

Zhao, D., & Strotmann, A. (2015). Re-citation analysis: A promising method for improving citation analysis for research evaluation, knowledge network analysis, knowledge representation. *ISSI*.

Appendices

Appendix 1: classroom reading on citation checking

Title: No, really, check the damn citations

TL;DR:

Every citation is a promise:

1. the source exists,
2. the source supports,
3. the source is valid, and
4. the source is relevant.

Sometimes broken:

1. sources do not exist,
2. sources do not support,
3. sources are not valid, and
4. sources are not relevant.

Your work is only as good as their word.

How one 'do not' became 14,000 'did it'

The Newcastle–Ottawa scale (NOS) is a simplified scale to appraise the quality of non-randomised studies. It has been used in thousands of meta-analyses. Stang published a commentary in 2010 that heavily criticised the NOS. The most heavily cited reference to the NOS is not the original description, but the heavy criticism of it by Stang. The title of the commentary is such that someone who has not read it may believe that this is a critical presentation of an important tool that everyone should use. Thousands of articles (June 2025: 14.5k citations) have defended using the NOS by citing a paper that concludes that the tool should not be used (Ioannidis, 2018).

Introduction

Every citation is a promise: that the cited source exists, that it supports, that it is valid and that it is relevant. Citation, the very mechanism by which science carries out the extended conversation on which our progress depends, is in trouble. Studies have long found startling citation error rates and now AI fabricates content and references at scale. Citation is, also, easy to neglect. Checking citations is not rewarded, it is time consuming and it is boring. This paper lays out where citations come from, argues for why they matter, summarises where we are today, and provides step by step instructions that break down what it means to check citations.

Where citations come from

Science emerged from a tradition called natural philosophy, so it is no surprise that empirical results arise from a form of argument (Weinstein, 1990). Argument in its simplest classical form (Toulmin, 1958) has three parts. A claim, some data, and rules which, if followed, make it so that the data support the claim. In empirical research, where claims are called results, data support a result when they are related (this is our model/theory/conceptual framework) and when they are produced correctly (and here we have our methods). We've known for at least a century that things are not quite so simple. How scientists see the world (our models) and what we recognise as acceptable ways to study it (our methods) change over time and across disciplines (Fleck, 1935). That said, the same basic argument structure can be found in every

corner of empirical science. Taking the next step, as well captured by the metaphor ‘standing on the shoulders of giants’, science accumulates models, methods and evidence by building on the work of those who came before. Citation is how we, together, stand taller.

Why citation matters

Citation is the backbone of scholarly communication. Each reference not only signals the origin of a claim but also serves as a conduit for trust. If I assert in an article “the moon is made of green cheese”, you will be sceptical of my article. If I, however, assert “the moon is made of green cheese (Astley, 1987)”, the problem will either be with my attribution (...no way Astley claimed that) or with Astley, 1987. In the first case, you tested the argument in my article. In the second, you test the integrity of my citation and then the cited source. Stripping a claim of a citation shifts the burden of proof from the record cited to the text being read. This matters. If, like many readers, you happen to trust peer review (and you should not (Smith, 2006)), you will be more trusting when reading an implausible claim backed by citation. Citation is how we create the web of knowledge on which our progress is founded.

Things are not good

As important as they are, citations are often wrong. The larger category of error is something called a citation error rate, which is the number of error containing citations divided by the total number of citations. While definitions vary, this category tends to include everything starting from typos in bibliographies to misrepresentations of the cited content. Quotation errors pick up the portion of citation error that pertains to the accuracy of a quotation with major quotation errors indicating a citing claim that is not supported by the relevant cited text to minor quotation errors that do not disrupt understanding. Historically, research on both citation and quotation error rates have produced troubling results. In studies done before the emergence of large language models, citation error rates averaged between 20% to 30% with about half of those being errors in the actual quotations (Al-Benna et al., 2009; Armstrong et al., 2018; Curlewis et al., 2023; Davids et al., 2010; Gazendam et al., 2021; Gosling et al., 2004; Luo et al., 2013; Mogull, 2017; Reddy et al., 2008; Wager & Middleton, 2007) . Since the arrival of generative AI things have, if the conclusions of an article subtly titled the citation catastrophe is any indicator, gotten much worse (Camp et al., 2025). Citation error is argued to hinder accurate dissemination of what we have learned by worsening the signal to noise ratio in scientific communication, to further marginalise underrepresented voices by mixing known discriminatory social factors, like popularity, accessibility and eminence, into empirical relevance, and to slow scientific progress by introducing errors into subsequent research (Horbach et al., 2021; Mattiazzi & Vila-Petroff, 2024; Tfelt-Hansen, 2015).

A step-by-step guide for checking citations

Given that citation is important and that things are bad, it would be easy to argue that all citations should be checked. This is not necessary. Researchers have long argued (Nigel Gilbert, 1977) and continue to demonstrate (Tahamtan & Bornmann, 2019) that scientists use citations for many purposes. For example, a well cited but hard to retrieve description proposes that authors use citation to:

1. refute some aspect of the cited work,
2. note or make perfunctory mention of that work,
3. compare or review that work,
4. use or apply what they cite in their own work, or

5. argue that some aspect of the cited work is substantiated by their own study (Henry, 1982).

Which citations we should check depends on our purpose. Since this essay is about making mistakes copying things forward, the protocol that follows for checking citations is built for the fourth purpose: use or application of claims found in the cited work. While parts of this protocol are appropriate for the other purposes, each would require its own adaptations.

How to check citations

When an author uses a citation to back a claim they promise the following:

1. This citation matches a source that exists
2. That source contains a claim that entails the citing claim
3. That cited claim is valid there
4. That cited claim is relevant here

At the very least, the first two steps in testing should be taken for every citation backed claim in an article which, were it wrong, would destroy a claim in the citing article that you find interesting. The paragraphs below introduce each of these steps, discuss their importance and suggest how most efficiently to do those that are possible.

This citation matches a source that exists

There are two parts to this first step. First, the relationship between in-text citations and bibliography items, and second, the relationship between bibliography items and the records they reference.

In text citations match bibliography items

From the perspective of a reader looking at a claim, the only thing that matters for this first step is that there is a bibliography entry for an in-text citation. While it is not immediately relevant to the citation you are looking at in a text, it also matters that all bibliography items are used. There may be extra items in a bibliography. Writing is messy, mistakes happen, and formatting bibliographies is boring with different requirements across journals. That said, authors are rewarded for stuffing their bibliographies. Journals, for example, want to look good. They are ranked according to something called a journal impact factor. This rating goes up when articles from their journal are cited. Authors and editorial boards know this. Sometimes authors will stuff their bibliography with items from a journal to which they are submitting to make themselves look more appealing (Ebrahimy & Osareh, 2015), sometimes editorial boards will ask authors to include reference to articles in their journal (Wilhite & Fong, 2012) and, worst, on occasion journals will just place bibliography items in a paper without telling the author (Besançon et al., 2024). While none of these forms of bibliography stuffing bears on the citation you are looking at right now, they do raise red flags on the quality of the science you are reading.

How to: For each in text citation chosen, check if there is a corresponding bibliography item. For each bibliography item, check that there is a corresponding in text citation.

Bibliography items point to records that exist

Bibliography items may not point to a record that exists. Apart from honest error, these days the most likely source of false records is large language models. Authors tend not to report use of LLMs, and these models do confabulate references (Maes, 2024). Hallucination is hard wired into how LLMs work (Xu et al., 2025). Whether or not a bibliography item describes a source that seems plausible, it must be checked to make sure it exists.

How to: Drop information from the bibliography item into an online consolidation service. For articles and other things that have DOIs, both OpenAlex and Crossref have open web interfaces and public APIs for this purpose. For books and manuscripts try WorldCat. If those don't work, assume that either you or the author made a typo. The user should go back and check and expand to try other sources. If still stuck, consult a librarian. It is certainly easier to check things that have DOIs which are in English but the assumption that 'all relevant knowledge is in English and has a DOI' is both harmful and wrong.

Cited record contains a claim that entails the citing claim

Unlike the previous steps which could be done quite quickly, testing whether a cited record supports a citing claim requires the reader to retrieve, to find all of the information in the cited record that is relevant to the citing statement and then to decide on the nature and adequacy of the relationship (Henry, 1982; Zhao & Strotmann, 2015) . There have been startling advances in the ability of generative AI to provide summaries, but we should not use models that generate text to try and find evidence which might be absent (Akhtar et al., 2023; Deemter, 2024; Jin et al., 2024; Shigarov, 2023) . No matter what approach is used, not finding something does not mean it is not there. The best we may claim is 'I tried hard, here is how, and I did not find anything'. While it is not possible to argue that absence is evidence, it is entirely possible to report when and if a cited source contradicts the citing claim. For example, a citing article may fail to mention that the results reported were not significant or that other studies reported in the cited record concluded differently.

How to: Read the cited text for relevant information, perhaps with the help of models that are built only for information retrieval. Since all models make errors, particularly when images, graphs, figures, tables and equations are mixed in, use a combination of approaches that are known to fail differently and always have a human review top candidates to make the final decision.

Cited claims are valid there

The steps taken so far in testing citations have rested entirely on the words found. This next step tests whether the words found in the cited record are warranted. This is a critical step as even twenty years ago there was a convincing argument put forward that most published findings are false (Ioannidis, 2005) and, apparently, things have not improved (Brown, 2025), which makes 'wrong' a good starting assumption.

Testing whether claims are valid in their home context is a form of quality assessment...something that has been discussed within systematic review since its invention more than a generation ago (Moher et al., 1996) . There is an incredible diversity of ways in which the quality of a study is understood and a confusing array of standards and tools used for their measurement. Those worth taking seriously all require a deeply competent assessor, they often demand access to underlying data and they most certainly take more time than most of us have to execute.

It may be far beyond our capacity to fully understand the overall quality of a study, but it is possible to ask discriminating questions that are simple and useful. These tests may look for selective omission of inconvenient truths: a study may ignore something that is broadly known to be relevant in the field, such as the impact of systematic non-response on the generalisability of questionnaires (Wright & Scott Armstrong, 2008), or they may examine the internal

coherence of a study by asking ‘if this, then that’ questions. For example, in a study of systematic reviews of research on human responses to climate change I wanted to know if authors were over-stating their claims. I decided that those reviews that did not test the quality of the studies they examined should say ‘has been reported’ rather than ‘it is so’. Given this expectation, I was able quickly to find evidence of quality assessment and to find and classify the claims made. No matter what path is followed, these quick and dirty tests are not comprehensive. All these tests permit is description of a single quality of the report examined. It indicates nothing about other dimensions of quality. The quality you describe is not relevant for readers who have different interests. For example, a reader might find that the primary conclusion of a study is false, and record this as a quality. This finding is irrelevant for readers who want to know what models were used.

How to: Neither full quality assessment nor the quick checks suggested can be automated. Further, even trivial quality assessment takes time and skills we may not have…which is why this perspective suggested that only the first two steps of citation checking always be done. For this third step the reader must define (and provide justifications for) expectations of the cited record that are both reasonable to have and matter for the citing claim. Once stated, the reader should note down what they will look for in the cited record that is relevant to these expectations and why these indicators are both reasonable and unequivocally related to the expectation. The reader should then search through the cited record for relevant information.

Cited claims are relevant here

What is valid in the cited paper is only relevant to a citing paper if inference is justified. For example, in the Netherlands studies report that most commercial cows produce about 30 litres of milk a day. While these studies may be valid for the Netherlands, things are very different in other countries. As such, a paper discussing watering requirements for dairy herds in Kenya would be suspect if it quoted a Dutch study on watering requirements without talking about the differences in daily milk production. This same sort of failure may occur for every component that goes into an empirical argument: just because a model, a method, an analytic choice or result fits in the context of the cited author does not mean that it also fits in the context of the citing author.

It takes a few steps to figure out if a valid claim imported by citation fits. The first step is to identify those dimensions that limit the generalisability of the claim imported by citation. The second is to describe differences between cited and citing on those dimensions. The third and final step is to identify how much differences found matter. While that three-step process was rather easy to state and hopefully seems sensible, it is very difficult to execute. In practice, authors only rarely report what limits the generalisability of their results and, if they do, there is no guarantee that they will anticipate those constraining factors that are relevant to the sort of inference required by a citation. Going back to our study of Dutch cows, it would be surprising to find ‘these findings do not work with Kenyan cows’ in the limitations section of the cited text. The reader would have to know about cows in the Netherlands and in Kenya and how the differences between these two condition the relevance of Dutch findings to the Kenyan context…and then we would look for evidence in the citing article that the author was sensitive to this difference.

Comprehensive assessments of fit are likely beyond what we are able to do. It, again, may be possible to set tests that are both low effort and useful. For example, in my own field, the social

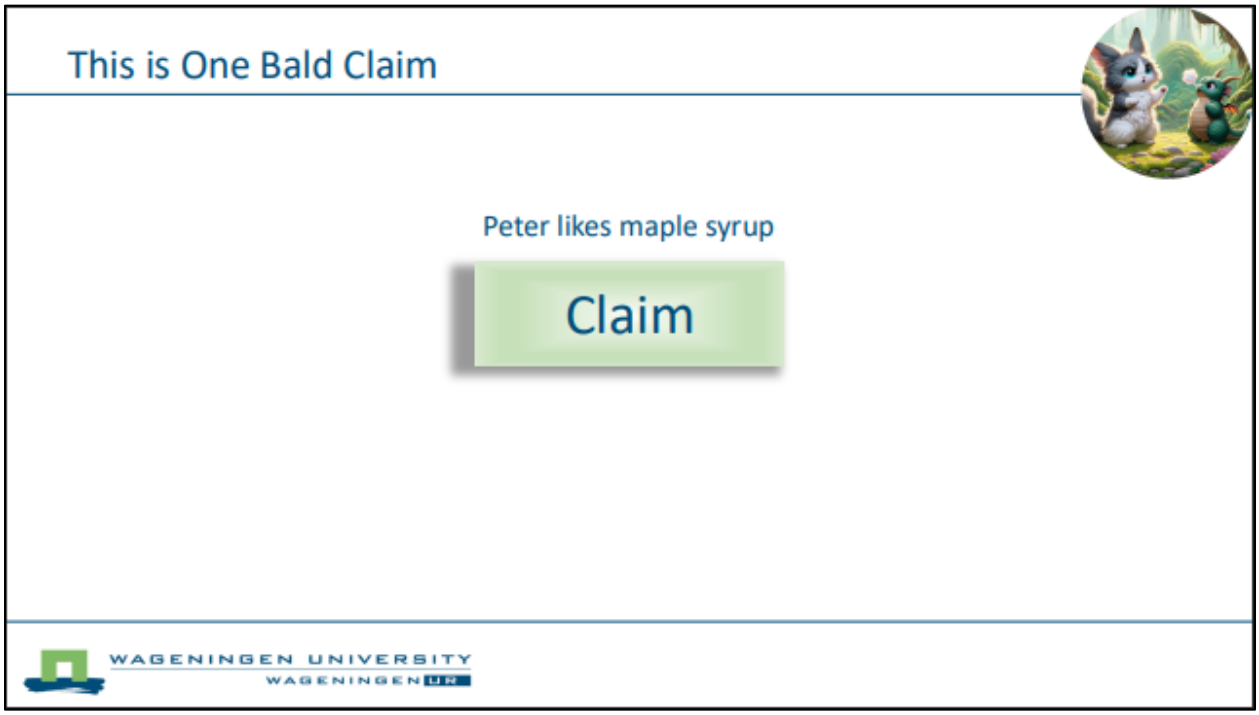
sciences, studies will often import questionnaires. In principle, we should use others' work and questionnaire design is scientific work so we should use others' questionnaires. This is how science builds knowledge. Different people, however, are well known to respond to the same questions quite differently. This means that there should at least be some discussion in the importing article of the appropriateness of a questionnaire and it is quite easy to test if such discussion exists in the citing article. As in the case of quality assessment, these quick and dirty tests do not support comprehensive claims about fit. All they support, and all that may be needed at the moment, is a description of fit on a very limited number of dimensions.

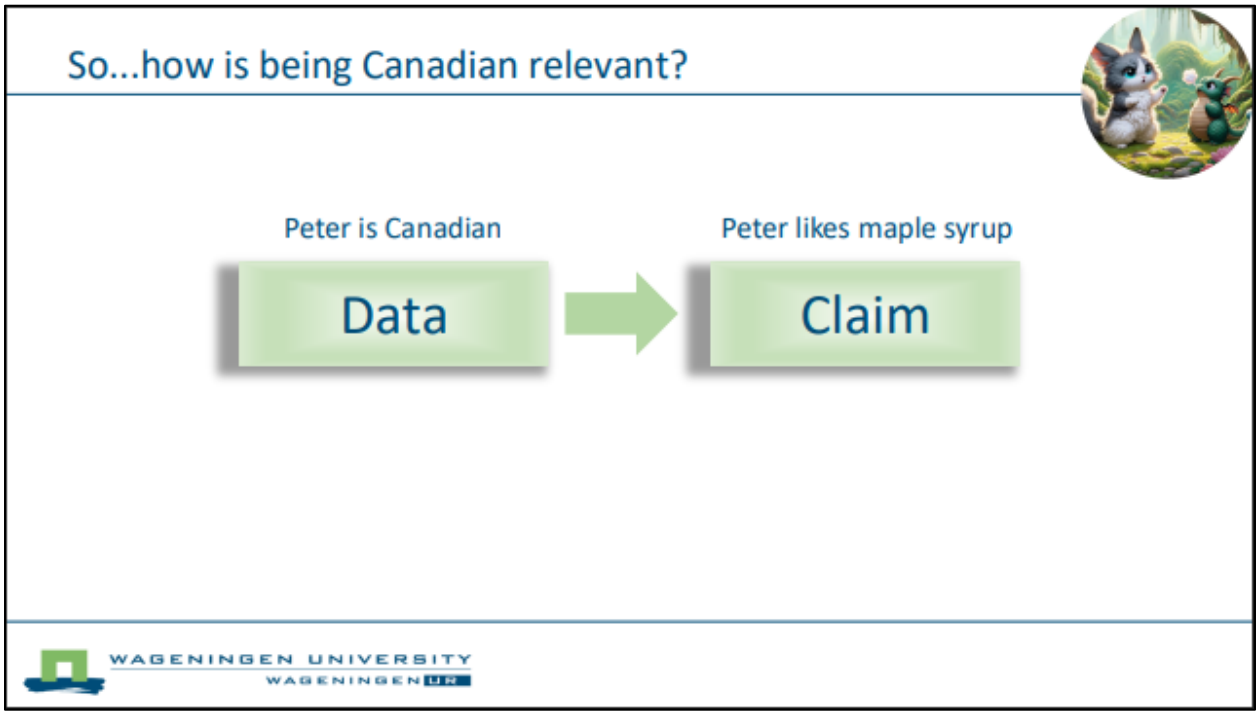
How to: Again, this step cannot be automated. For this step the reader must define (and provide justifications for) dimensions that condition the relevance of the claims found in the cited record to the context in which the citing claim is made that are both reasonable to expect discussion of and that matter. Once stated, the reader should note down what they will look for in either the cited or the citing record that is relevant to these expectations and why these indicators are both reasonable and unequivocally related to their expectations. The reader should then search through both records, as appropriate, for relevant information using the method they find most efficient.

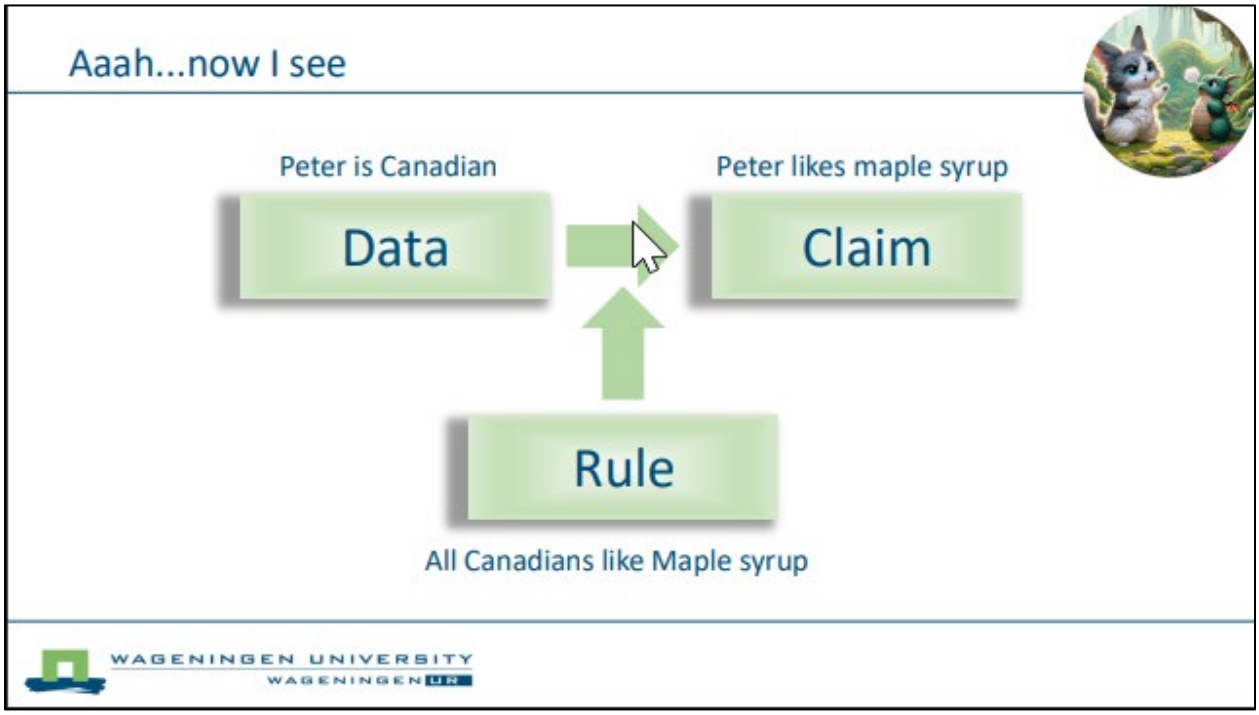
Conclusion

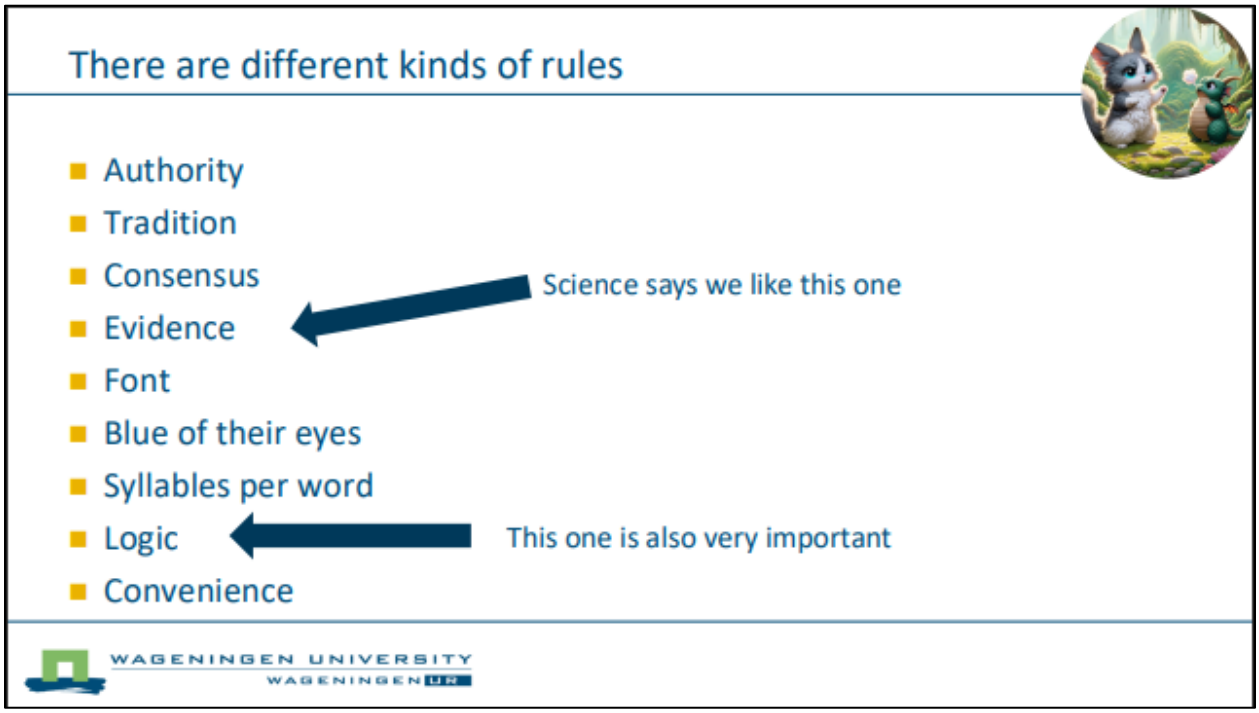
In an ideal world we would not need to check citations. We do not live in an ideal world. We make mistakes, we are short on time, it is easier to publish when we make loud claims and we are promoted more for publishing more. When the stakes are low and cautious humility is rewarded, perhaps trust is good enough. When what we do matters, when the amount of information compounds exponentially every day and when scientists are motivated by outdated incentive systems, we must verify. This verification requires, at the very least, that we test those citations that, if wrong, would destroy the features of the article that are critical for our own purposes.

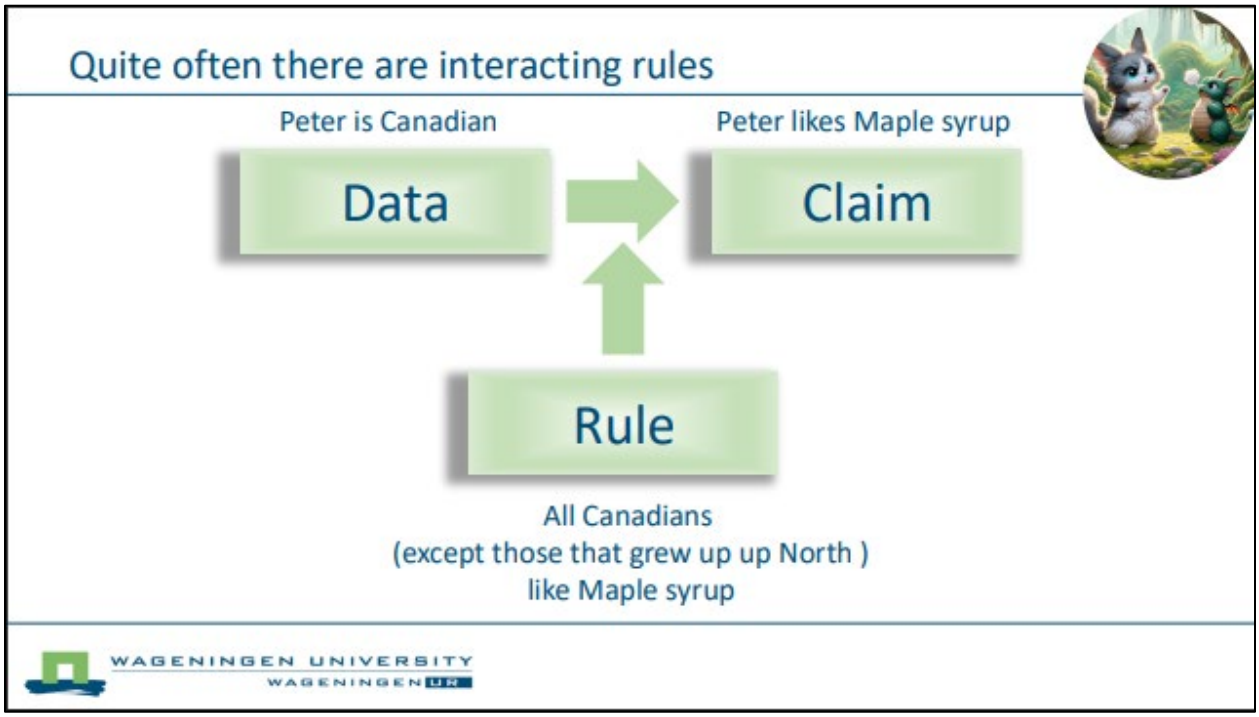
Appendix 2: slides on argumentation

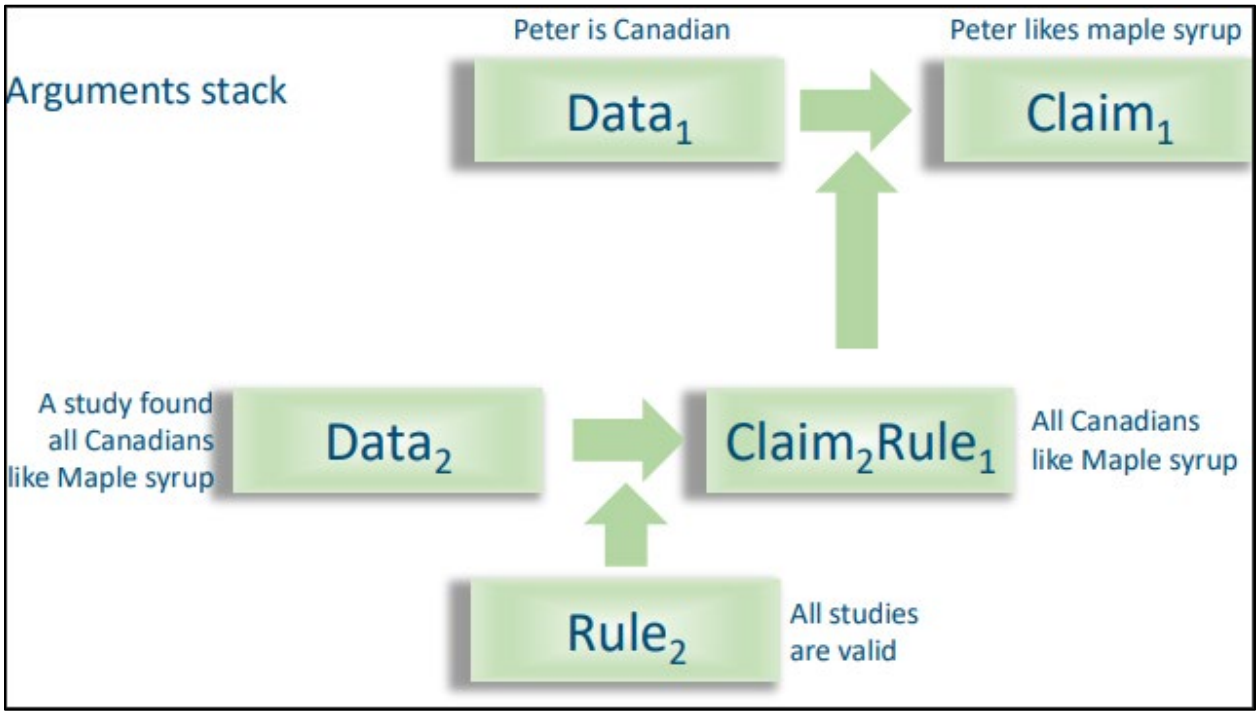


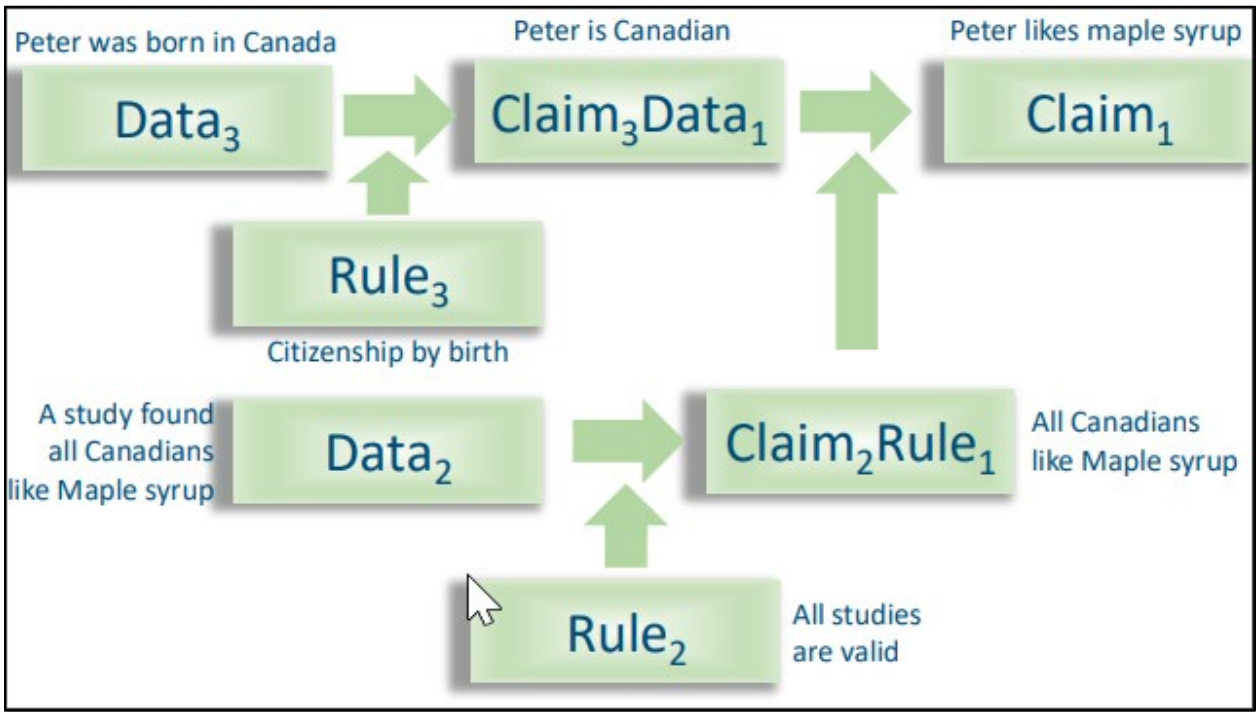

So...how is being Canadian relevant?
Peter is Canadian
Peter likes maple syrup
Data
Claim
WAGENINGEN UNIVERSITY
WAGENINGEN UR

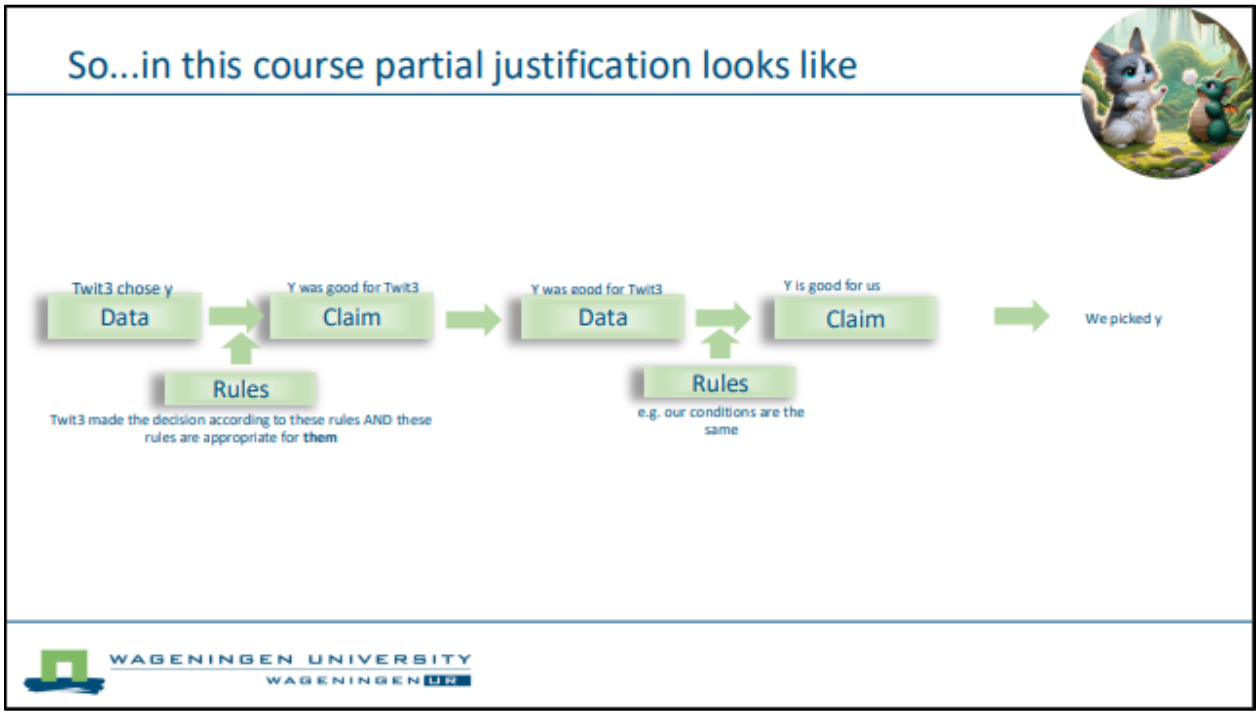

Aaah...now I see
Peter is Canadian
Peter likes maple syrup
Data
Claim
Rule
All Canadians like Maple syrup
WAGENINGEN UNIVERSITY
WAGENINGEN UR

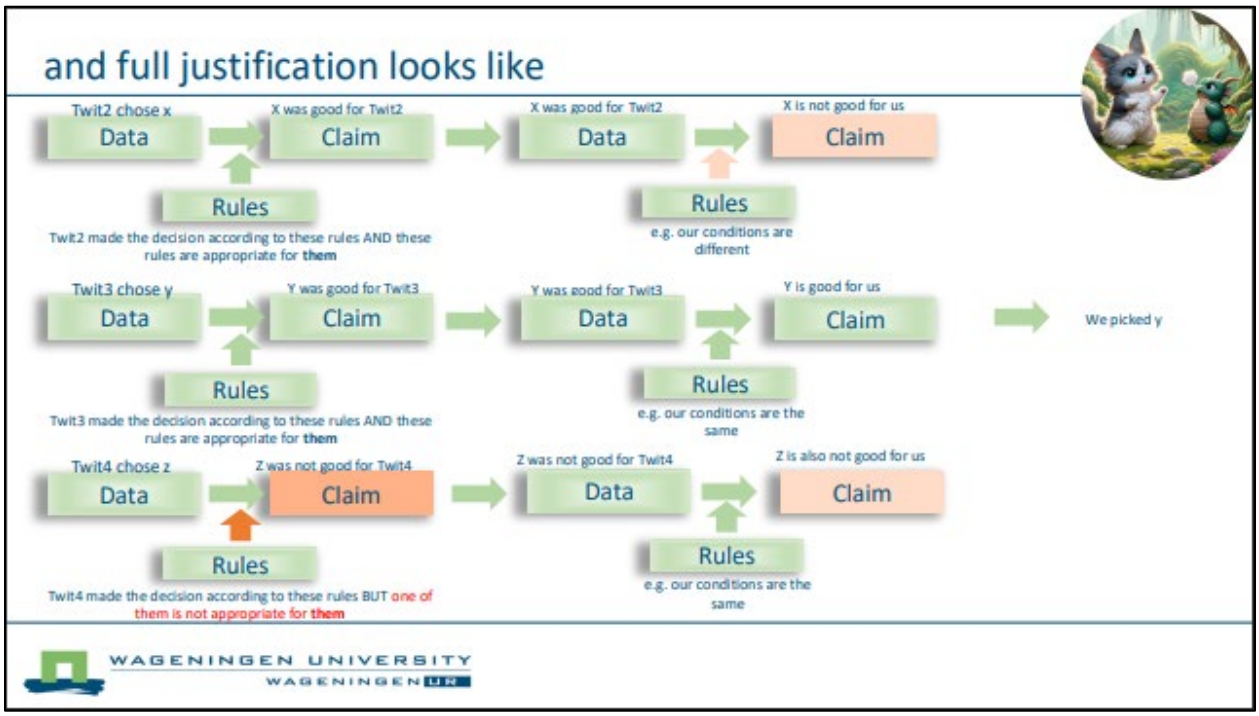

There are different kinds of rules
Authority
Tradition
Consensus
Evidence
Font
Blue of their eyes
Syllables per word
Logic
Convenience
Science says we like this one
This one is also very important
WAGENINGEN UNIVERSITY
WAGENINGEN UR


Quite often there are interacting rules
Peter is Canadian
Peter likes Maple syrup
Data
Claim
Rule
All Canadians
(except those that grew up up North)
like Maple syrup
WAGENINGEN UNIVERSITY
WAGENINGEN UR


Arguments stack
Peter is Canadian
Peter likes maple syrup
Data1
Claim1
A study found all Canadians like Maple syrup
Data2
Claim2Rule1
All Canadians like Maple syrup
Rule2
All studies are valid


Peter was born in Canada
Peter is Canadian
Peter likes maple syrup
Data3
Claim3Data1
Claim1
Rule3
Citizenship by birth
A study found all Canadians like Maple syrup
Data2
Claim2Rule1
All Canadians like Maple syrup
Rule2
All studies are valid


So...in this course partial justification looks like
Twit3 chose y
Data
Y was good for Twit3
Claim
Rules
Twit3 made the decision according to these rules AND these rules are appropriate for them
Y was good for Twit3
Data
Y is good for us
Claim
Rules
e.g. our conditions are the same
We picked y
WAGENINGEN UNIVERSITY
WAGENINGEN UR


and full justification looks like
Twit2 chose x
Data
X was good for Twit2
Claim
Rules
Twit2 made the decision according to these rules AND these rules are appropriate for them
X was good for Twit2
Data
X is not good for us
Claim
Rules
e.g. our conditions are different
Twit3 chose y
Data
Y was good for Twit3
Claim
Rules
Twit3 made the decision according to these rules AND these rules are appropriate for them
Y was good for Twit3
Data
Y is good for us
Claim
Rules
e.g. our conditions are the same
We picked y
Twit4 chose z
Data
Z was not good for Twit4
Claim
Rules
Twit4 made the decision according to these rules BUT one of them is not appropriate for them
Z was not good for Twit4
Data
Z is also not good for us
Claim
Rules
e.g. our conditions are the same
WAGENINGEN UNIVERSITY
WAGENINGEN UR

Appendix 3: Instructions for students

(This text is adapted from our course description)

What we look for when grading your report
We want to see that you are using the tools provided in this course to carefully think through each step of your report. What matters most from this class is not the content of the report you write but the skills required to prepare a similar report in the future. The content you put in this report, therefore, really does not matter. What does matter is the clarity and the quality of your documentation of the process by which you arrived at the decisions you made in writing it.
For most decisions that you make you are expected to present a justification. This means that you use will use the Claim Data Rule structure shown in the provided slides. It takes a lot of work to do a full justification for a choice. There is not enough time in this course to fully justify all the decisions you must make. We, therefore, tell you in this outline how much effort we expect at each point.

For every decision, you'll need to reason, i.e. use logic/argumentation to demonstrate careful consideration in making your decision. (replace following with appropriate examples) If you want to look only at households within 5 km of a river, why does that make sense for your research? If you want to operationalise biodiversity by looking at three particular indicators, why is that sensible for your research purposes, population and context?

For some decisions (indicated in the assignment instructions), you need to explicitly draw on scientific literature to support your argumentation. For 'partial justification', you may cherry-pick a few sources to back up your choices, showing sensitivity to limitations without having to fully consider alternative options or the criteria for evaluating trade-offs and choosing one over the other option. In the places where we require 'full justification', you need to draw on at least four scientific sources and transparently consider and evaluate alternative options. This means that you must decide (and report) which criteria matter when choosing between options, determine how to measure the candidate sources for these criteria, measure those sources, analyse the results of your measurement and, based on all that, decide which option to follow.

Because this is a class early in your studies, in most cases we support 'invention' and procedural logic when your choices rely on information that is not readily available. For example, (insert your own example) if it is plausible but not certain that a particular list of sampling units exists, you could mention that you will, for now, assume it exists and base further choices on that assumption. If you realise that a particular instrumentation decision relies on information you'll still need to get from the field, you may discuss why that information matters or how you'd like to retrieve it.

NOTE: you do not need to provide citations/references to the course material. Mentioning something that we introduce and why it matters is enough. You only need to provide citations for external material.

NOTE 2: Please indicate clearly what you are doing where. If we ask for one fully justified decision in a bigger section, clearly highlight where you provide this.

Sample text framing expectations for a report section
Since there is no time to carry out an extensive literature study, it is wise to select a topic with which you are already somewhat familiar (partial justification).

Description of the supplemental information to be submitted by the student
For this class, in addition to the report, you are required to submit a zip folder containing supplemental information. This zip folder (named ‘group x.zip’) will contain a spreadsheet file and the PDFs cited in your report. Please download the example that is in the same folder on (LCMS name) as this assignment description.

Appendix 4

Classroom exercises for checking citations
The following exercises require students to practice each step of citation checking. For each there is a version that does not require software which meets our learning objectives. Once the software we are developing is complete, the version that includes use of software will allow us to achieve learning objectives that combine IL with machine learning and AI. The software is a full patch-board environments (on the model of KNIME or Orange3) that students use to explore and learn about how models actually handle text and how sensitive results are to changes in pipelines.

None of these exercises have yet been tested. They were created to signal to students what might be involved in learning about citation. If they are at all interesting, to ensure felt relevance, these exercises should be structured around a short article that students bring with them that is central to a project that they are working on at the time. Some or all of these may also be integrated into expectations for a report that requires review of the literature.

Testing how sensitive an article's conclusions are to citation error

Overview
The conclusions of an article depend, in part, on claims that are brought forward by citation from the academic record. Citation errors may, therefore, undercut conclusions. In this exercise students describe and assess the vulnerability of the conclusions put forward in an article to citation error.

Task
- Read your article and note down the main conclusions.
- Read it again and highlight all statements that are backed by a citation.
- For each highlighted statement, look at the main conclusions and ask: "What would happen to this conclusion if this citation was bad?"
    - If it would have no effect on the conclusion, mark it 0.
    - If it would kill the conclusion, mark it 10 and provide a justification
    - Intermediate values estimate partial impact and provide a justification
- Rank-order the conclusions by their average risk across citations.
- Have a peer redo one of your assessments to see if they reach the same conclusion

Assessment
- Completion: to what extent is the assignment completed
- Justification: to what extent does the student provide an adequate justification for their decisions (factors both for and against a conclusion are fully considered)
- Humility: to what extent does the student condition their conclusion by, for example, reporting uncertainty

Learning goals
Understand
Cognitive/skills
- See academic texts as complex recursive arguments that span records
- Detect and describe how citations support conclusions

- Recognise that claims often rely on implicit or unstated assumptions that may be specific to one discipline (complicating inter-disciplinarity)

Affective

- Demystify the authority of articles
- Accept that conclusions are contingent
- Value transparency

Compare

Cognitive/skills

- Compare the relative importance of different citations for overall conclusions
- Identify patterns: some conclusions may rest on one fragile citation, others on many low-risk ones
- Notice differences in what citations do for an argument (e.g. essential empirical backing vs. perfunctory acknowledgement)

Affective

- Appreciate that not all citations matter equally
- Recognise that the fragility of an argument isn't apparent till it is examined

Assess

Cognitive/skills

- Quantify the "risk" of a conclusion failing if a citation is invalid.
- Assess robustness of conclusions by average risk scores.
- Reflect on how citation intent (perfunctory vs essential) shapes fragility

Affective

- Understand how and why humility is necessary: conclusions are provisional
- Accept/normalise recognition that science does not deal in Truths but more often in shaky arguments.
- Acknowledge that reduced expectations are a hallmark of sound science

Apply

Cognitive/skills

- Build a simple risk ranking of your article's conclusions.
- Document which citations carried the most weight and why.
- Justify your ranking in a short written assessment.
- Produce a reproducible record (others can check your coding of risk values).

Affective

- Value accountability by linking risk judgments to transparent evidence.
- Invite critique: have peers replicate your ranking to see if they agree
- Prefer clarity about fragility over the comfort of false certitude

Information/AI Literacy Integration

Cognitive/skills

- Connect risk analysis of citations to broader inclination and skills for critical evaluation of information

Affective

- Adopt a stewardship mindset: weigh citations responsibly before relying on them.

- Recognise the scientific record as a network of fragile claims that requires care and scrutiny.

Check the match between in text citations and bibliography items (undeveloped)

Overview
Citation stuffing by authors, genAI and journals happens. Mismatches between in-text citations and bibliography items are the sort of error that is produced by careless researchers and lazy frauds.

Manual task:
Find your selected in-text citation(s) in the bibliography.

Software assisted task:
Use the provided patch board to build a workflow that

1. Converts your article into a structured format
2. Extracts both in-text and bibliography items
3. Matches extracted items.
4. Reports mis-matches

Learning goals
To recognise and work with structure within an academic text. To learn about perverse incentives in citation.

Check whether cited records exist (incomplete draft)

Manual task:
Paste the selected citations into webUIs for OpenAlex and WorldCat and record whether or not the cited record exists in their databases

Software-assisted task:
Use the patch board environment to create a workflow that submits information to OpenAlex and WorldCat APIs which reports whether or not the cited record exists in their databases.

Learning goals
To use open-science services. To learn about open global efforts to better capture and make meta-data available.

Retrieve cited records (undeveloped)
Manual task:
Identify citations that are critical for those aspects of an article that interest you. Use the library's retrieval tools to download your chosen cited records. Load those retrieved records into a program that supports their organisation and annotation (e.g. EndNote, Zotero)

Patch-board task (requires fully automated retrieval pipelines at the host library):
Identify citations that are critical for those aspects of an article that interest you. Use the patch board environment to use the library's retrieval API to download, process and organise your chosen cited records to support further analysis.

Learning goals
Learn how to retrieve and manage records.

Check cited records for entailing statements
Manual task:
Read the downloaded records for information that is relevant to the matching citing statement(s). Classify whether that relevant information either supports or contradicts the citing claim.

Software assisted task:
Use the patch-board environment to:

1. Convert citing statements into sets of independent claims (e.g. warmth and humidity encourage bacterial growth -> warmth encourages bacterial growth, humidity encourages bacterial growth.)
2. Identify matching cited record
3. Filter matching cited records to surface possibly relevant statements
4. Present the most likely candidates to the user to identify which, if any, statements in the cited documents either support or contradict the citing claim
5. Allow the user to classify responses

Learning goals
Develop and apply a targeted reading strategy...either manual or machine assisted. Learn about error types (the difference between including things that are not relevant and missing things that are relevant). Discuss the limits of machine learning/LLMs in both working with tables/figures/graphs and establishing that something does not exist.

Test supporting statements in cited record (Are they good there?)
Task
Given two statements in cited articles:

- Identify reasonable questions to ask to determine the extent to which each assertion is defensible in the context of the cited article.
- Describe the extent to which the information required to determine if the claims in the cited article are valid is present in that article.
- Determine the extent to which those statements are justified.

Learning goals
Understand
Cognitive/skills

- Determine what makes an assertion defensible within its own study
- Recognise the difference between internal validity (good for that article) and transferability (good beyond it)
- Learn to ask context-appropriate questions about measurement, sample, setting, and assumptions

Affective

- Appreciate the effort needed to design and report a rigorous study
- Experience frustration at incomplete reporting
- Value transparency in reporting

Compare

Cognitive/skills

- Notice variation in what counts as adequate evidence between studies
- Compare how differently articles report what's needed to check validity

Affective

- Accept that no two studies will provide exactly the same kinds of information
- Respect that complexity makes full reporting impossible, leaving room for trust and for abuse.

Assess

Cognitive/skills

- Develop simple rubrics to test whether the cited claim is adequately supported in its own article
- Identify missing information and reflect on how its absence limits your confidence
- Estimate the seriousness of weaknesses for overall defensibility

Affective

- Show humility: absence of evidence ≠ evidence of absence
- Accept that judgments of validity are always partial
- Recognise the risks of overstating what the article shows

Apply

Cognitive/skills

- Apply your rubrics to two cited statements
- Document what questions you asked, what you found, and what was missing
- Write a short justification of how seriously you take each claim, given the evidence
- Produce a reproducible record of your checks

Affective

- Value accountability by linking judgments directly to evidence
- Invite peer critique to see if others reach the same assessment
- Prefer transparency over speed or clarity alone

Information Literacy Integration

Cognitive/skills

- Understand the difference between "claim present in the article" and "claim adequately supported by evidence in the article."
- Connect citation-checking to broader goals of transparency, reproducibility, and critical evaluation.

Affective

- Adopt a stewardship mindset: treat cited claims as shared resources that require care
- Cultivate suspicion of black-box claims where evidence isn't visible.

Test transferability of cited statement to citing record (Are they good here?)

Task:
Given three statements in your citing article that are backed by a citation on which key claims in the citing article depend, for each:

- Look back and forth between the citing and cited records and identify factors that shape whether the claim can be defensibly copied forward. For example, a measurement strategy may have been validated in the cited study, but circumstances in the citing study may differ.
- Make a list of the information you would need before deciding if the transfer is defensible.
- Examine both texts for discussion of that information.
- Repeat for two more citation-dependent claims in the citing paper.
- Have peers in the class reproduce one of your analyses to check if they get the same result

Write an assessment justifying how seriously you take the citing paper's claims, based on what you found.

Learning goals

Understand

Cognitive/skills

- Identify factors that condition whether a cited finding generalises to a citing study.
- Recognise that generalisability is limited and conditional.
- Distinguish between validity within the cited study and appropriateness of transfer to the citing study.

Affective

- Appreciate the level of effort required to design a scientific study.
- Understand the fragility of scientific knowledge Given frustration with incomplete reporting,
- Value transparency in reporting.
- Develop a healthy suspicion towards unremarked citations

Compare

Cognitive/skills

- Understand that the criteria that shape transferability may vary between cases.
- Demonstrate sensitivity by respecting how differences in setting, methods, or measures could shift transferability.

Affective

- Appreciate that the same rules may not fit all circumstances
- Respect that the complexity of science makes it difficult fully to report so we must, to some extent, trust which makes science vulnerable to abuse

Assess

Cognitive/skills

- Develop context appropriate rubrics to test how strongly citing claims are supported.
- Understand factors that condition transferability
- Reflect on how missing information undermines defensibility

Affective

- Show humility: absence of evidence ≠ evidence of absence

- Own the risk of over-claiming when citations don't support transfer

Apply
Cognitive/skills
- Use appropriate rubrics to assess the transferability of claims
- Transparently document design considerations, choices and results
- Justify your assessment of the citing paper's claims
- Produce a reproducible record of your checks (so others can audit your reasoning)

Affective
- Value accountability: link your judgments to evidence
- Look for critique from peers
- Understand why transparency and auditability are more important than quick clear judgments

Information Literacy Integration
Cognitive/skills
- Understand the difference between "appears in the literature" and "is valid for my case."
- Connect citation-checking to broader literacy goals: transparency, reproducibility, critical evaluation

Affective
- Adopt a stewardship mindset toward the scientific record: treat citations as a shared responsibility
- Cultivate suspicion toward "black box" practices that hide conditions and limits
- Recognise the scientific record as an easily poisoned commons on which we, utterly, depend

Compare performance of AI workflows
Task
Given a patch-board working environment and a meaningful task, build a set of candidate workflows (including black box LLM), assess their performance and then select one that best fits the student's needs (steps, models, parameters).

Learning goals
Understand
Cognitive/skills
- Identify the stages of the workflow (parsing, chunking, retrieval, comparison, NLI).
- Explain how each stage can change outcomes (e.g., window size, coref).
- Recognise provenance (models, parameters, data sources) and traceability needs.

Affective
- Value transparency and reproducibility; express why provenance matters for public trust. (Reflexivity)
- Acknowledge uncertainty as normal; state limits of one's own understanding without defensiveness. (Anticipation/Reflexivity)
- Commit to documenting choices so others can audit and learn. (Responsiveness)

Compare
Cognitive/skills

- Run alternative workflows in the patch board (e.g., 2-sentence vs 3-paragraph windows).
- Compare outputs with at least two metrics and understand how they differ.
- Describe sensitivity: show how parameter changes alter results.

Affective

- Adopt an evidence-first stance over preference for a method or model.
- Demonstrate curiosity about failure cases and edge conditions
- Respect diverse sources/languages; consider multilingual and domain biases when comparing setups.

Assess

Cognitive/skills

- Assess reliability/robustness of results.
- Identify known limits and estimate their impact (e.g. figures/tables, missing context, ambiguous phrasing, languages).
- Reflect on bias introduced by choices that are black-boxed in most AI.

Affective

- Own potential harms from over-claiming.
- Intellectual humility (e.g., absence of evidence is not evidence of absence).
- Prioritise fairness: voice concerns when results systematically disadvantage topics, languages, or groups.

Apply

Cognitive/skills

- Assemble candidate workflows for a defined task (select steps, models, parameters).
- Consider and record rejected alternatives with reasons.
- Justify design decisions; state known limitations of design chosen.
- Produce clear, reproducible documentation (what/why/how) that others can run.

Affective

- Demonstrate accountability for choices
- Seek feedback from peers/stakeholders; integrate critiques into revisions.
- Prefer solutions that improve auditability and future reuse, not just short-term performance.

Information Literacy integration

Cognitive/skills

- Articulate the role of AI as decision support, not authority.
- Connect outputs to literacy goals: transparency, reproducibility, critical evaluation.

Affective

- Adopt a stewardship mindset toward the scientific record.
- Develop suspicion towards black box solutions.