



Talk, Check, Decide: A Skill-Based Workflow for Writing and Verifying Geological Excursion Reports with an AI Assistant

Stefan Mohr¹

¹Independent Researcher, Hamburg, Germany

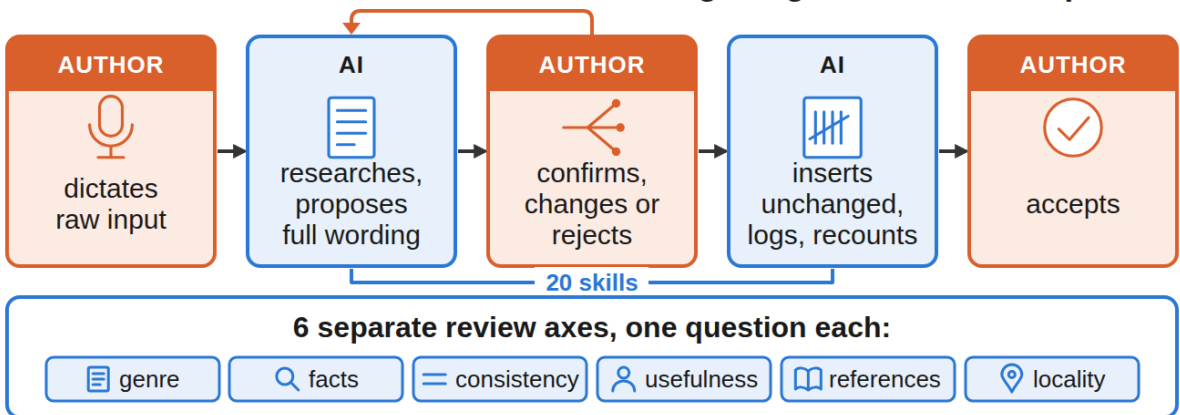
5 Correspondence to: Stefan Mohr (stefan@familie-mohr.com)

Abstract. Large language models write fluent text but invent references and blur the line between what was seen and what was read. Geological excursion reports are sensitive to both failures, because every statement must be bound to a place and a source. This paper describes a workflow in which an AI assistant (Claude) works under a family of 20 written procedures, called skills, and two templates for its projects, that separate research, writing, checking and decision, and in which the author, by his own account, works almost entirely by voice. The workflow is shown on two reports: a post-trip report on 20 localities in Germany and a pre-trip field guide on 21 localities in southern Norway and Denmark. Six review axes are kept apart: fact support, internal consistency, reference accuracy, locality evidence, genre and usefulness. In the Norway report, seven consistency runs produced 18, 4, 6, 4, 0, 2 and 71 findings, the last after the author had changed the target format from a planning report to a dossier; in the Germany report, a full reference check found 13 errors in the 211 entries it held on 18 September 2026. In three documented cases, an error was caught by one axis only. The results suggest that the value of AI assistance lies above all in systematic, counted checking, with a human in the loop: the assistant proposes the complete wording, the author confirms it, and only then is it inserted unchanged, logged and verified.

Keywords: excursion report; field guide; large language model; fact checking; reference verification; geological localities; dictation



Talk, Check, Decide: AI-assisted scientific geological excursion reports



Graphical abstract. The workflow in brief: the author dictates raw input; the assistant, working under 20 written skills, researches and proposes the full wording; the author confirms, changes or rejects it; the assistant inserts it unchanged, logs the change and recounts; the author accepts the result. Six separate review axes each ask one question: genre, facts, consistency, usefulness, references and locality. Orange marks the author, blue the assistant.

1 Introduction

Excursion reports and field guides are a working genre of geology. A good one lets a reader find the outcrop, see what the author saw, and check what the author claims. Guidance for field trip guidebooks therefore asks for enough information to repeat the trip, including coordinates for each stop (Geoscience Information Society, 2018), and field manuals insist on recording observations separately from their interpretation (Compton, 1985; Coe, 2010; Lisle, Brabham and Barnes, 2011).

Large language models are now widely used in scientific writing (Kobak et al., 2025; Liang et al., 2025). They are also known to produce plausible but false statements (Ji et al., 2023; Huang et al., 2025) and to fabricate references (Walters and Wilder, 2023). For an excursion report the risk is specific: a model readily attaches a regional statement to a single quarry, or a correct date to the wrong structure, and the result reads as well as a true sentence.

This paper asks three questions. RQ1: Where does an AI assistant help in writing geological field guides before a trip and excursion reports after it, and where does it fall short? RQ2: Which errors does AI assistance introduce into such guides and reports, and which checks catch them before publication? RQ3: To what extent does an iterative approach, in which skills are derived from analysed practice and revised after each review round of the report, support the building of skills and improve the quality of the reports produced with them? It describes the workflow, reports what the check records of two reports show, and discusses what failed. Errors are reported openly, because the rules of the workflow were derived from them.



2 Related work

Fabricated and wrong references. Walters and Wilder (2023) found 55 % of citations produced by GPT-3.5 and 18 % of those produced by GPT-4 to be fabricated, and substantive errors in many of the real ones. Later models fabricate less, but still often, and they differ widely. In radiology, Claude 3.5 Sonnet fabricated 3.1 % of its references and Claude 3 Opus 18.3 %, against 27.7 to 60.6 % for four models from OpenAI and Google, and only 59.6 % of the references from Claude 3 Opus were fully accurate (Güneş, Cesur and Çamur, 2026). A comparison of 26 models found 55.4 % of all responses fabricated; Claude Opus 4.6 fabricated 18.5 % and gave fully correct references in 54.6 %, while Claude Opus 4.8 fabricated 10.2 % but declined about half of the requests (Topaz et al., 2026a, preprint). Earlier studies in medicine report similar or higher rates (Bhattacharyya et al., 2023; Chelli et al., 2024), and the problem has been noted in geography (Day, 2023). Access to sources does not remove it: even with retrieval, about 30 % of the statements made by GPT-4o were not supported by the sources it cited (Gao et al., 2023; Wu et al., 2025), and commercial deep research agents still produced hallucinated URLs for 3 to 13 % of their citations (Rao, Wong and Callison-Burch, 2026, preprint). Verification steps reduce hallucination (Manakul, Liusie and Gales, 2023; Dhuliawala et al., 2024). The errors now reach the published literature: among the open-access papers in PubMed Central, the share with at least one fabricated reference rose from one in 2,828 in 2023 to one in 277 in the first seven weeks of 2026 (Topaz et al., 2026b). Human-written literature is not free of such errors either: in marine biology, only 75.8 % of cited assertions were clearly supported by the cited source (Todd et al., 2010; see also Jergas and Baethge, 2015; Mogull, 2017).

Responsibility and disclosure. Publishers and ethics bodies agree that an AI tool cannot be an author, that the author remains responsible for all content, and that the use of AI must be disclosed (Nature, 2023; COPE Council, 2024; Elsevier, 2026; ICMJE, n.d.). How this is implemented varies widely (Hosseini, Resnik and Holmes, 2023; Ganjavi et al., 2024). Commentators warn that AI must not take over the normative and interpretive parts of science (Messeri and Crockett, 2024; Binz et al., 2025).

Language models in the geosciences. Geoscience-specific models (Deng et al., 2024) and reviews of generative AI in the field (Foroumandi et al., 2023; Hadid, Chakraborty and Busby, 2024) show growing use, alongside concern about bias (Hatvani, Mudelsee and Kern, 2024). The closest work to the present one used a language model to write geosite descriptions and found that 44.8 % of the Polish and 1.3 % of the English texts needed editing after manual and field checks (Wolniewicz, 2024). A GPT-based geology system reached a score of at least 5 out of 10 on 79.6 % of 152 questions (Baucon and Neto de Carvalho, 2024).

Locality data. Open databases are central to locality work but carry spatial uncertainty and biases (Peters and McClennen, 2016; Ralph et al., 2025). Georeferencing practice treats a locality as a point with an uncertainty radius (Wieczorek, Guo and Hijmans, 2004; Chapman and Wieczorek, 2020), and occurrence records need cleaning before use (Zizka et al., 2019).

Speech and structured elicitation. Speech input is about three times faster than typing on a phone (Ruan et al., 2018). In survey experiments, spoken answers contained about three times as many words and more themes than written ones (Höhne,



75 Gavras and Claassen, 2024; Lenzner, Höhne and Gavras, 2024), although plain dictation into text gave less detail than
recorded speech (Revilla et al., 2020; Revilla and Couper, 2026). Dictated clinical notes contain errors that review removes
(Zhou et al., 2018; Blackley et al., 2019). Speech recognition has been used to capture geological field data (Zhang et al.,
2020). Knowledge elicitation research describes structured interviews as a method to draw out expert knowledge (Cooke,
1994; Hoffman et al., 1995).

Documentation practice. Software engineering records design decisions explicitly because their reasons are otherwise lost
80 (Jansen and Bosch, 2005; Tyree and Akerman, 2005). Version control and a record of every change support reproducible
research (Ram, 2013; Sandve et al., 2013; Wilson et al., 2017), and simple checklists reduce errors in routine work (Haynes
et al., 2009).

Gap. A search of 24 queries in English and German, by web search and a partial Crossref search, without Scopus or Web of
Science, found no study on using a language model to write or check geological excursion reports, or on dictation as a way
85 of writing them.

3 Materials and methods

3.1 Cases

The workflow was applied to two reports written in English by the author (Table 1). The Germany report describes a tour
that took place; its evidence includes the author's own observations, finds and photographs. The Norway report plans a tour
90 for 2027; it rests on literature and databases only, and its photo places and field notes were removed when it became a
dossier; they pass to a separate results report after the tour. It was a planning report until 26 September 2026 and has since
been a pre-excursion dossier (version v100); in both forms it is the field guide before the tour, and this paper calls it the
Norway report in both.

Table 1. The two cases.

	Germany (post-trip)	Norway (pre-trip)
Tour	8 days, 2 to 9 August 2026, from Hamburg	10 days, summer 2027, from Hamburg via Hirtshals and Kristiansand
Localities	20: quarries, dumps, show mines, track sites, a geotope, museums	21: pegmatite quarries, silver mines, fossil sites, peridotite sites, museums, a moler pit in Denmark
Size	about 55,000 words, 226 references in version v15	about 42,000 words, 7 appendices
Project files	report, appendices, register of repeated statements, work plan, base data, locality table	27 files, among them a decision log with 130 entries, a log with 243 entries and a work plan with 132 steps in 21 blocks
Evidence	observation, finds, literature, databases	literature, databases

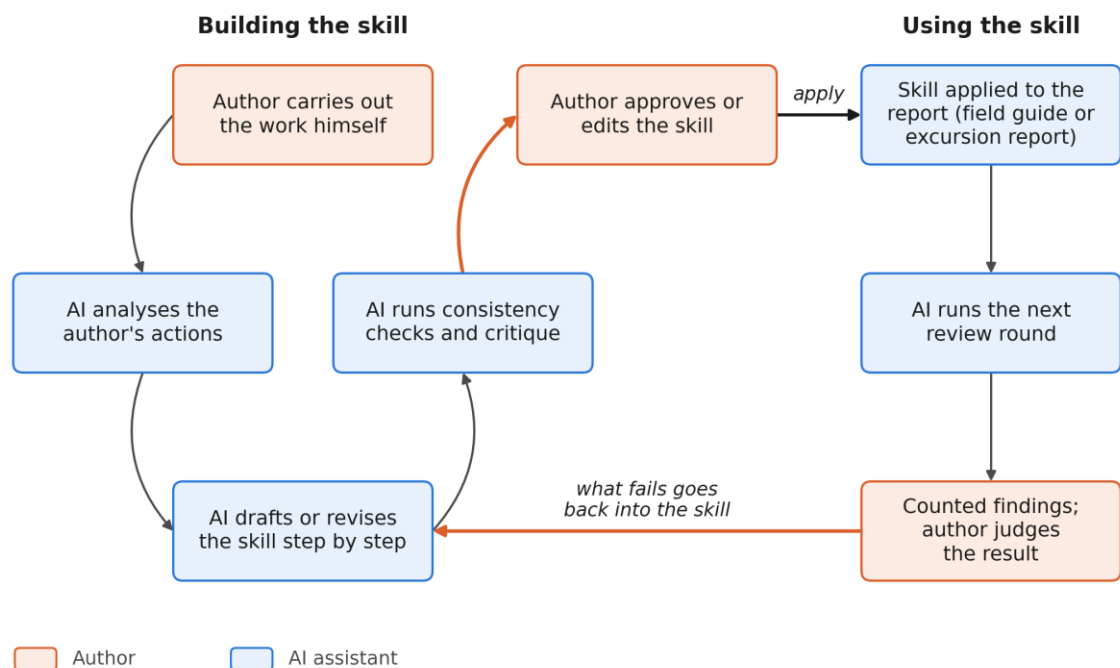


3.2 Working cycle and interaction

The skills were not designed in advance. They were distilled from work already done in four steps: try a task on a real case; formalise what worked into a procedure; write the procedure as a work plan with blocks, decision points and acceptance criteria before executing it; and write back into a skill the rules that proved sound. For this, the working records of both projects were read and 268 working patterns were written down and clustered into 12 groups; the gaps between groups became new skills. In the author's practice, the same analysis is now also applied to single working sessions: at the end of a session, the assistant analyses it and writes the procedure it followed into a method document, which the author checks and which then becomes the basis of a work plan or a skill. In this sense, AI was used to set up the use of AI: the assistant that did the work also analysed its records and revised the skills that steer it, and the author approved or edited them.

Feedback loops. Skill building is iterative. The assistant analyses the actions the author carries out himself and drafts a skill from them step by step; the skill is then put through consistency checks and a deliberately harsh critique, and it is judged by the results of the next review round of the actual report. What fails there goes back into the skill. The loop does not close: every report is a practical test of the skills, and every change to a skill passes the author's control (Figure 1).

Templates. Once the working patterns had settled, two framework skills were written as templates: one for document projects such as reports (template-dokumentprojekt) and one for skill projects (template-skill-projekt). Each for its kind of project, they lay down a file set, naming, versions, identifiers, logs, a chat record, wording proposals and the check after every write, so that every new project starts with all conventions learned so far and the way of working can be reproduced, as recommended for research projects in general (Wilson et al., 2017). Each of their rules names the incident it came from; the document template holds 31 such rules, the skill template 12. In one case, 10 files of the Norway project carried six different counts from 15 to 38; this became the rule that all project files share one version, starting at the highest count plus one (v39). The present paper was itself set up with the document template.



Structural analogy to reinforcement learning from human feedback: human judgement of the results steers the improvement. No reward model is trained and no parameters change; the assistant revises a written procedure, and the author approves or edits it.

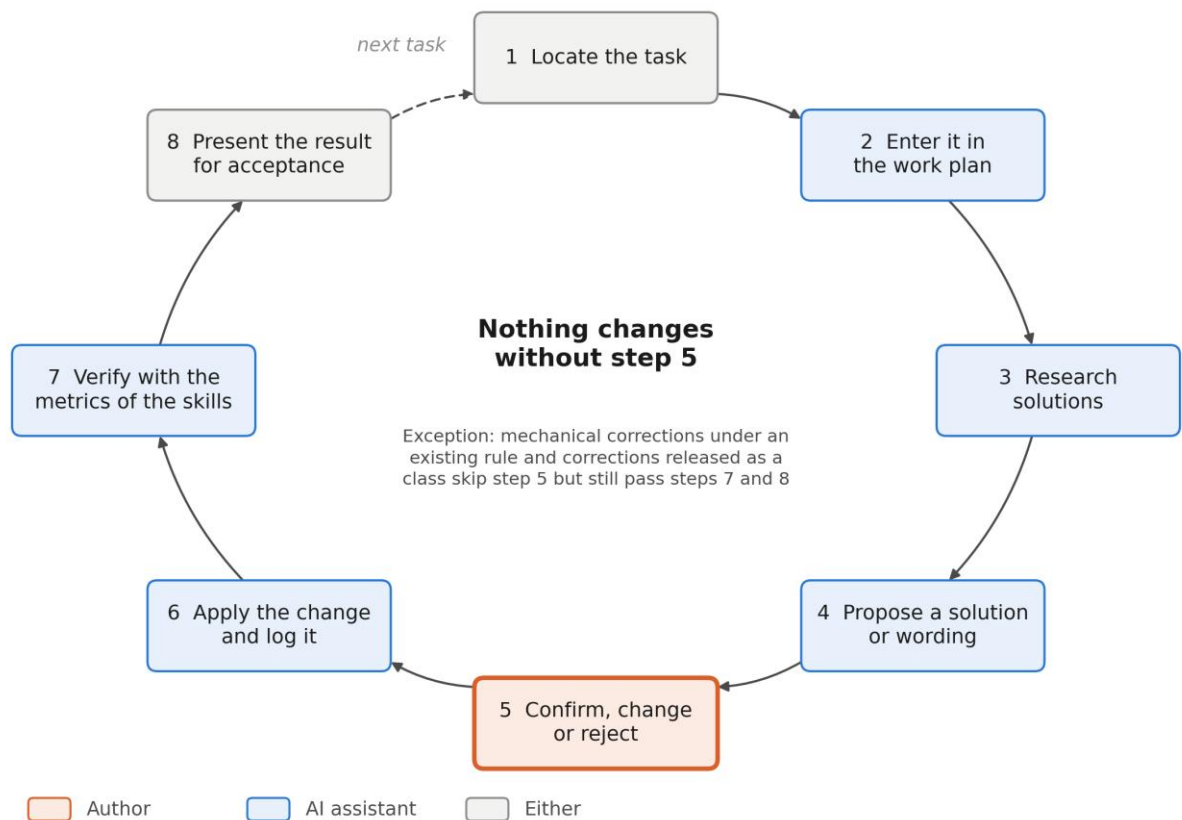
Figure 1. Iterative skill development. In building the skill, the assistant analyses the author's own actions, drafts or revises a skill and puts it through consistency checks and critique, and the author approves or edits it; in using the skill, it is applied, the assistant runs the next review round of the report, and the author judges the result. What fails feeds back into the skill, which closes the loop. Orange marks the author, blue the assistant. The analogy to reinforcement learning from human feedback is structural.

Human in the loop. Every task followed the same eight steps (Figure 2): locate the task; enter it in the work plan; research solutions; propose a solution or wording; have the author confirm, change or reject it; apply the change and log it; verify it against the metrics of the skills (the change is at the intended places, the number of changes matches the proposal, the old wording no longer occurs); and present the result for acceptance. Nothing was changed without step 5, except mechanical corrections under an existing rule and corrections the author had released as a class. After the work was done, a second check compared the result once more with the chat and the work plan, to find what had been asked for but not done or not recorded. When the work on this paper moved to a new chat on 27 September 2026, this check found that the log handed over lacked one entry and the chat record two inputs, and that seven statements in the working files contradicted other files or the paper; all of them were corrected. Findings outside the task at hand were not fixed on the way but set out to the author and entered in the work plan as steps of their own. While this paper was revised, for example, a check showed that the rule on the running chat described in section 3.3 was not yet carried by any skill; it was entered as a step and written into the document template only after the author had decided so. The assistant also reported its own errors as soon as it noticed them: when one



proposal for this paper left out a count that the register of repeated statements required to change with it, the omission was reported before the next proposal and corrected by a proposal of its own.

Wording proposals. The central procedure of the workflow is the wording proposal. Since this procedure was introduced on 26 September 2026, no sentence entered this paper as an intention or a summary; the author also used it for the two reports from that date, where this was not counted. The assistant first set out the complete wording, numbered, with its place in the text, the old wording where one was replaced, the sources with the state of their verification, and the consequences, such as further passages, new reference entries or numbers that change. The author confirmed, changed or rejected it. Only the confirmed wording was inserted, unchanged; when the author added to a proposal, the extended version was proposed again rather than worked in silently. Each insertion was then logged with the number of the proposal and, from the tenth proposal on, with the number of the confirming input in the chat record, and compared with the proposal in every copy of the text. In this paper, the research questions and all numbered proposals took this route: 75 by 28 September 2026, this one included, of which 68 changed the text, 6 only the figures and 1 only the name of the document. What the author confirmed was therefore the text itself, not an account of it, and every sentence inserted in this way can be traced to its proposal and, from the tenth proposal on, also to the input that released it. The author's answers took four fixed forms, which keep a spoken release unambiguous: "yes" releases only the proposals just set out; "all" releases every proposal listed; "yes, unless ..." asks the assistant first to check the drawbacks and state them; and "list ..." asks for findings without any write.



150 **Figure 2.** Human in the loop: the eight steps every task passed through. Nothing was changed without the author's confirmation (step 5); mechanical corrections under an existing rule and corrections released as a class skipped it but still passed steps 7 and 8. The colours show who carries each step: the author confirms, changes or rejects; the assistant enters, researches, proposes, applies and verifies; locating a

155 By his own account, the author worked almost entirely by voice and dictated nearly every input in the working chats; the share was not measured. Instructions, decisions and field observations were dictated; the system asked one question at a time, stated why it asked, and wrote the result into the files. Decisions were prepared so that a one-word spoken answer sufficed, and the system kept the written record (Figure 3). Speaking to the system has to be learned and practised. Slips of the tongue, misrecognised words, corrections several sentences later, pauses and even revoking a statement at the end and

160 restating it did no harm in the author's experience; the system was treated as a conversation partner, not as a dictation device (Figure 4).

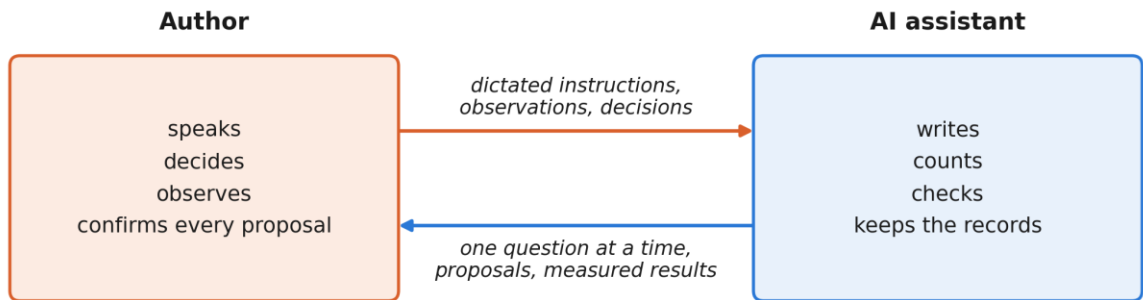


Figure 3. Division of labour between author and AI assistant. The author speaks, decides, observes and confirms every proposal; the assistant writes, counts, checks and keeps the records. The author passes dictated instructions, observations and decisions to the assistant; the assistant returns one question at a time, proposals and measured results. Orange marks the author, blue the assistant.

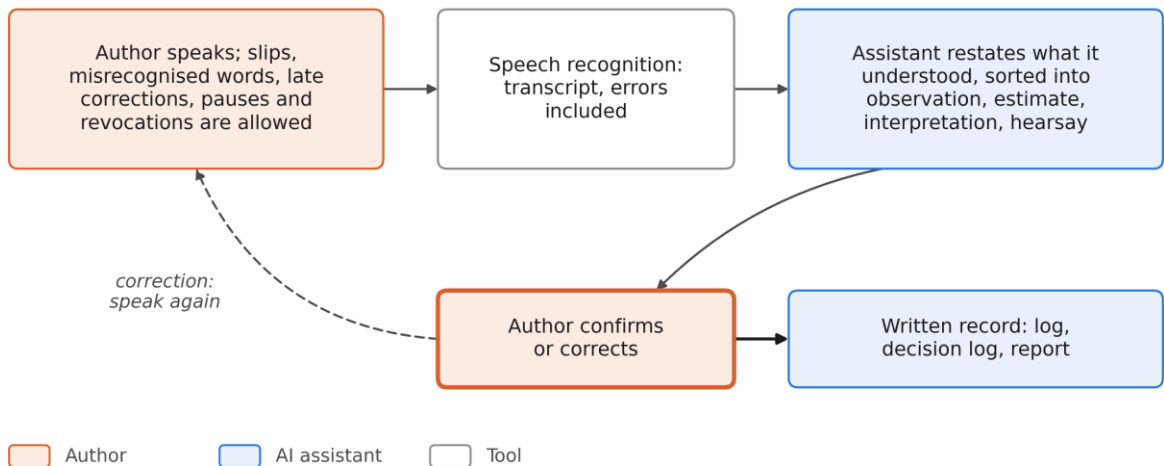


Figure 4. Speaking with read-back: slips, corrections and revocations are caught because the assistant restates what it understood and the author confirms or corrects it. Orange marks the author, blue the assistant, the white box a tool.

3.3 Skills

A skill is a written procedure that the assistant loads before a task. The family used here has 20 skills in two groups: `geo-` skills for geological reports and localities and `science-` skills that work on any document, plus two framework skills, templates for document projects and for skill projects (Appendix A, Table A1). The 20 skills cover research, elicitation of the author's observations, writing, rules, clarification, checking, steering and maintenance by change propagation; the two templates form the framework. Each skill names the tasks it hands to others. Skill names are the identifiers of the skill files, most of them German; Table A1 gives the task of each in English.

Eleven rules recur across the skills: count instead of assuring; never take a reference, coordinate or name from memory; leave gaps visible with the markers OPEN (nothing found; OFFEN in the skills), UNCLEAR (uncertain; UNKLAR in the skills) and --- (does not exist there); never transfer a regional statement to a single point; keep observation, interpretation



and literature apart; change nothing without approval; leave decisions with the author; freeze findings and track status separately; keep working notes out of the deliverable; check every write by recounting; and keep everything that matters in files, not only in the chat. Since 28 September 2026, the document template extends the rule against memory to the running chat: a fact is looked up again in its source or in the project files instead of being taken from what was said earlier in the chat, because a long chat is summarised and its earlier wording is lost (section 5).

3.4 Review axes

Six checks were kept separate because an axis can find errors the others miss (Table 2). Section 4 reports a counted result for each of them. The skills prescribe the order: genre first, then fact support and consistency, then usefulness, then a new consistency run, because every improvement shifts numbers and cross-references; in the Germany report, the genre analysis came only after the first two consistency runs. Reference accuracy and locality evidence can run at any time and run at the latest before submission. A finding that belongs to another axis is handed on, not fixed in passing.

Table 2. Review axes.

Axis	Question	Skill
Genre	Measured against which type of report is it complete?	geo-exkursionsbericht-typenanalyse
Fact support	Does the cited source carry the statement?	science-faktenbeleg-pruefung
Internal consistency	Does the document agree with itself?	science-interne-konsistenz-pruefung
Reference accuracy	Does each entry describe the work it claims, and does its DOI lead there?	science-quellenapparat-pruefung
Locality evidence	Is the statement bound to the site, not the region?	geo-fundstellen-evidenz-pruefung
Usefulness	Does it serve the reader it is written for?	science-dokumentkritik-drei-sichten

Decisions were separated from writing. A factual decision has a better answer and receives a recommendation; a purpose decision has two correct answers and receives counted consequences but no recommendation (science-entscheidungspunkt-klaeren). Each rule of lasting effect went into the decision log.

3.5 Data

The results are taken as quoted in the project records of both projects (logs, work plans, critiques, the Germany report and the skills, check reports as quoted in the working chats, a summary document compiled from the chats, and the chats themselves), as produced up to 28 September 2026; the Norway records begin on 10 September. The 268 working patterns rest on the chats, and 32 of the 117 statements about the two projects checked for this paper rest on the summary document only (section 5, "Records outside the chat"). Counts are those stated in the reports, logs and work plans; for the Norway report they refer to version v100 of 26 September 2026, for the Germany report to its version v15, so numbered since 21 September 2026, unless a date is given. The localities of the Germany tour were selected before it with ChatGPT and



Claude; for writing and checking, the assistant was Claude, used through its chat and agent interfaces, first as Claude Opus 5 and later as Claude Opus 5.5.

205 3.6 Use of AI in the preparation of this paper

This paper was drafted with Claude from the author's dictated instructions, the skills listed in Appendix A and the project records of both reports. The literature was searched and every reference verified against a publisher, society or institutional page, a DOI landing page or a Crossref record. The author is responsible for all content.

4 Results

210 4.1 Consistency runs

In the Norway report, findings fell over the first six runs and rose only after large changes. The steepest rise, to 71 findings, came in the seventh run, after the author had changed the target format from a planning report to a dossier: the run measured a text written largely for the earlier format against the new one. It was also made in two separate parts and then merged (Figure 5). Documented non-findings, that is suspicions checked and found correct, ranged from 6 to 12 in the first six runs and reached 42 in the seventh, and they stopped later runs from raising them again. In the Germany report, the first two runs over the whole report, on 13 and 15 September, produced 29 and then 8 findings, and three later runs, on 19 and 21 September, produced 3, 5 and 4; three runs over chapter 1 produced 10, 9 and 7 findings; the first of them also recorded 16 non-findings.

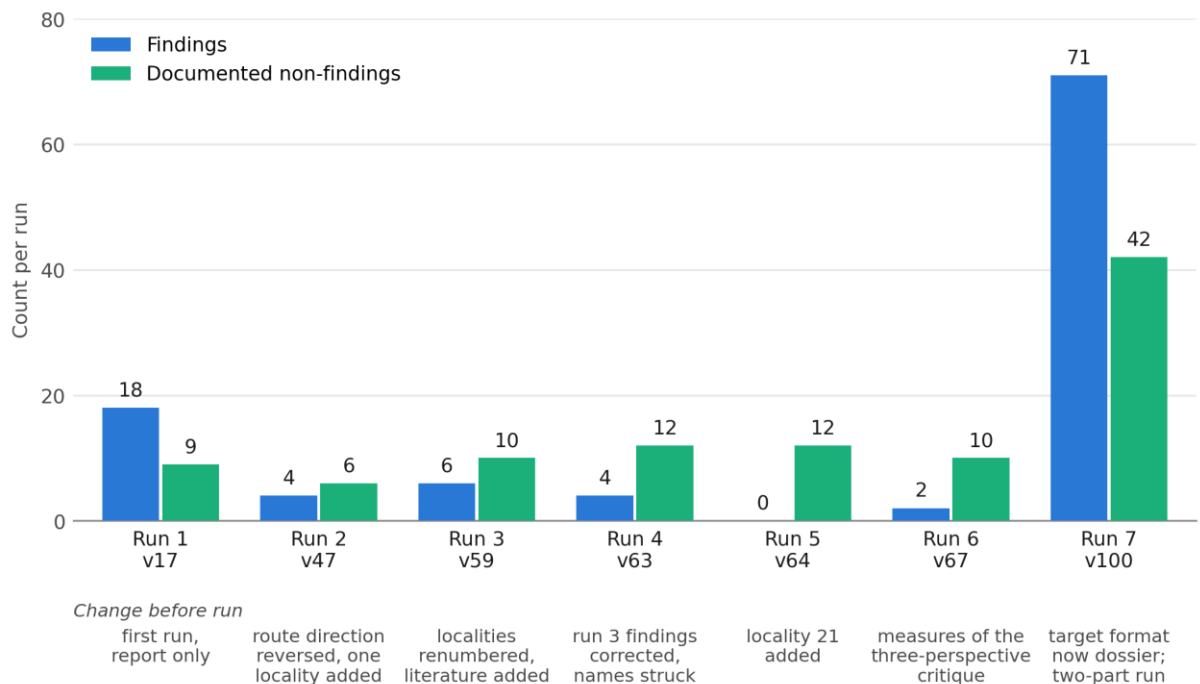


Figure 5. Findings and documented non-findings in the seven consistency runs of the Norway report, with the change that preceded each run. Blue bars count findings, green bars documented non-findings, that is suspicions checked and found correct; below each run stand the version of the report it checked and the change made before it. Findings rose only after large changes, most steeply to 71 in the seventh run, after the target format had changed from a planning report to a dossier.

In the author's view, findings clustered where new text met old; this was not counted. Examples from Germany: one section named two track horizons and another six, both citing the same paper; the age of the Ries impact was attached grammatically to the Steinheim crater, which cannot be dated isotopically; and the abstract announced 5 discussion themes for a chapter that by then had 10.

4.2 Fact support

In the final check of chapter 1 of the Germany report on 24 September 2026, 36 statements carried a source: 18 were verified, 14 partly verified, 3 could not be checked because the sources are printed books without accessible text, and 1 was misattributed. A check of statements about museums, 20 statements in 19 rows, found 12 rows verified, 4 partly verified and 3 misattributed; its date and the version it checked were not recorded. The three misattributions of the museum check were of three different kinds: a paper by two named authors listed as anonymous, taken over from an unchecked bibliography; a correction made in one section but not in another; and a conclusion of the writer placed inside a citation, where it read as a statement of the source.



4.3 Reference accuracy

A full check of the Germany reference list covered the 211 entries it held on 18 September 2026 and found 13 errors, about 6 %. Five were found by checks that needed no search, among them comparisons of an entry with the document's own practice. Four of the six errors among entries with a DOI sat among the 11 most cited works. Three DOIs pointed to a real but different paper, one differing by a single digit from its neighbour in the same volume. About one entry in eight could only be checked in part. Requiring either the publisher or two independent sources before correcting prevented five changes that would have introduced new errors.

4.4 Locality evidence

In the Norway report, all 55 citations of Mindat were checked, in version v67, for whether the cited page supports the statement for that very site; each was listed with its sentence and its reference entry. A first pass produced 26 proposals for change. A second pass by separate instances of the assistant, which had not seen the results of the first pass, confirmed 21 of them, 4 in part and 1 with a condition; all 26 were carried out, in 22 changes to the text.

4.5 Genre and usefulness

Measured against four candidate genres in its version v13 of 19 September 2026, the Germany report met 9 of 11 obligations of an excursion report; 53 of its 58 figures were still placeholders. The four candidates were chosen from the 15 types of the genre selection skill in a discussion in which the assistant acted as a discussion partner: it reflected the author's reasoning critically, which made him reconsider it at two or three points, and the four were kept that best served the aim of the report. The author chose as target genre the research excursion written up as a project report, one of the four candidates; in this paper, it counts as an excursion report after the trip. The first three-perspective critique of the Norway report (a practical, a collecting and a scientific view), made on a version before v67 whose date and number were not recorded, produced 45 measures, of which the author struck 9; a second critique on 26 September produced 45 more. The first critique found gaps that no correctness check could: the daily plan existed only in the route drafts, and radioactive minerals in the Iveland pegmatites were mentioned nowhere. Both led to new content, among them a field identification table of 20 minerals with a radioactivity column.

4.6 Failures and the rules they produced

Every rule of the two templates names the incident it came from, 43 in all; Figure 6 shows six failures and the rules they produced; two of these rules are carried by the document template, three by the reference check and one by the work plan skill. Two of them concern error classes named in the answer to RQ2 (section 5): an entry with the right title and DOI but the wrong authors, caught only by comparing the first author on the DOI landing page, and an over-wide replacement anchor that deleted about 800 words, whose first reported restoration was itself unchecked and missed a second error.

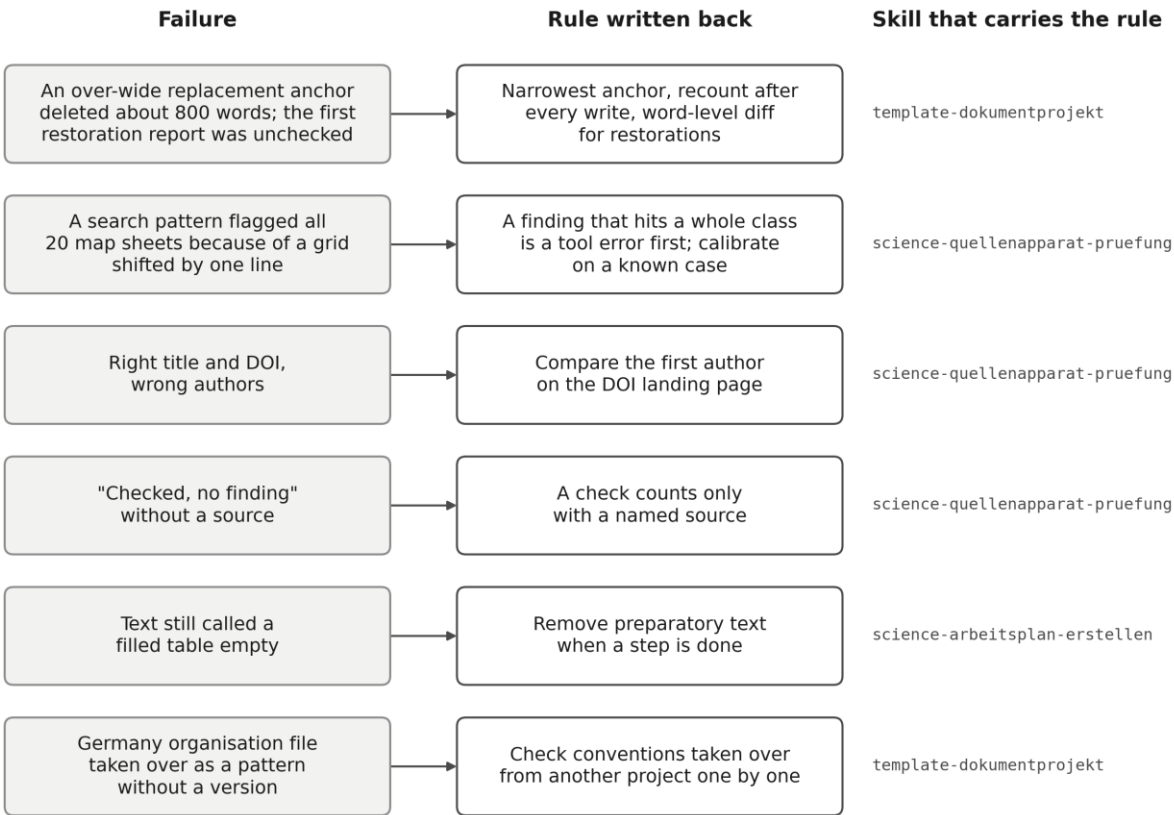


Figure 6. Failures and the rules they produced, with the skill that now carries each rule. Each row reads from left to right: a failure that occurred during the work (grey box), the rule written back in response (white box) and the skill that now carries it. The six rows are examples; two rules are carried by the document template, three by the reference check and one by the work plan skill.

270 **5 Discussion**

Answers to the research questions. RQ1: The assistant helped in research, provided that every reference it proposed was verified; in drawing out the author's observations one question at a time; and above all in counted checking along separate axes. It fell short where only the author could contribute, in field observation, and where sources were out of reach: about one reference in eight in the Germany list could be verified only in part. RQ2: The errors that arose in AI-assisted writing, not all of them introduced by the assistant, were misattributions, misplaced attributions, wrong DOIs or authors, regional statements transferred to single sites, stale numbers and cross-references, and editing accidents; misattributions are the target of fact support, misplaced attributions and stale numbers and cross-references of internal consistency, and wrong DOIs or authors of reference accuracy, although the classes overlap: the museum check found a wrong author list and a correction not carried through as misattributions (section 4.2); editing accidents led to the check after every write (section 4.6), and regional statements transferred to single sites are the target of the locality evidence axis (section 4.4). In three documented cases, an



error was caught by one check only, as the paragraph on separate axes below shows. RQ3: Building skills in loops supported the building of skills, as far as the two templates show: all 43 of their rules name the incident they came from, and they carry these rules into every new project; the rules of the other skills were not counted. That it also improved the reports is indicated by findings that fell over successive runs between large changes, but not shown: with two cases and no comparison group, better skills cannot be separated from reports that simply matured.

Separate axes. The central result is that in three documented cases an error was found by one axis only; how often each type occurred or was missed was not counted. The misplaced age was invisible to the fact check, because the value and its citation were correct; only the consistency check found that it was attached to the wrong structure. The writer's own conclusion inside a citation was invisible to the consistency check; only reading the source clause by clause found it. The DOI pointing to a neighbouring paper was invisible to both. This matches the finding that retrieval alone leaves citation support incomplete (Gao et al., 2023) and supports keeping verification as a separate step (Dhuliawala et al., 2024).

Counting instead of assuring. Every finding carried a count or a location, and success was reported as measured values. This made the checks falsifiable and turned the assistant's tendency to confirm into something the author could test. The comparison with a checklist is deliberate (Haynes et al., 2009): the rules are simple, and their value lies in being applied every time.

Region and point. Database records and regional literature were, in the author's view, a frequent source of statements that looked like evidence for a site without being so; their number was not counted. Treating every value as carrying the level it came from follows georeferencing practice (Wieczorek, Guo and Hijmans, 2004) and addresses the biases documented for open locality data (Ralph et al., 2025).

Speaking. In the author's experience, speaking was faster than typing and led to a real dialogue with the system. The speed is supported by measurement (Ruan et al., 2018). The richer content is supported for recorded speech in surveys (Lenzner, Höhne and Gavras, 2024) but not for plain dictation (Revilla et al., 2020). What differed here is that the system asked back, one question at a time, and sorted the answer into observation, estimate, interpretation and hearsay. Whether this closes the gap between dictation and recorded speech was not measured. Self-repair is a regular part of spontaneous speech (Schegloff, Jefferson and Sacks, 1977; Levelt, 1983). Language models, however, handle revoked or restated content less reliably than other disfluencies (Rohanian et al., 2023), which is a further reason why the system restates what it understood and the author confirms it.

Language. The author is not a native speaker of English. He worked with the assistant in German and, by his own account, almost entirely by voice, while the reports and this paper were written in English; the assistant carried the content from one language into the other and kept the English consistent across files. It also opened literature in other languages: more than a third of the 226 references of the Germany report have German titles and one is in French, and the Norway report draws on Norwegian acts and survey publications. Non-native speakers spend considerably more time reading and writing scientific English (Amano et al., 2023), and literature in languages other than English is often left out of international work (Amano, González-Varo and Sutherland, 2016); AI language tools have been proposed as a way to lower both barriers (Berdejo-



315 Espinola and Amano, 2023). This makes one rule more important: a statement taken from a source in another language has to be checked against the original wording, not against an English rendering of it.

Human responsibility. The workflow keeps a human in the loop for every change, with the two exceptions named in section 3.2 (Mosqueira-Rey et al., 2023): the assistant proposes, the author decides, and purpose decisions receive no recommendation. An explicit confirmation step matters because automation bias is not prevented by training or instructions alone (Parasuraman and Manzey, 2010), and it follows guidance to combine high automation with high human control (Amershi et al., 2019; Shneiderman, 2020). This is consistent with the positions of publishers and ethics bodies (COPE Council, 2024; ICMJE, n.d.) and with warnings against handing interpretation to AI (Messerli and Crockett, 2024).

Iterative skill development. In the author's view, the permanent feedback between skill, report and review mirrors the principle of reinforcement learning from human feedback, in which human judgements of outputs steer the improvement of a model (Christiano et al., 2017; Ouyang et al., 2022). The analogy is structural, not formal: no reward model is trained and no parameters change; what is revised is a written procedure, revised by the assistant and approved or edited by the author after seeing the results (Schön, 1983; Sutton and Barto, 2018). Loops of self-critique improve model output even without a human (Madaan et al., 2023), and human feedback is itself inconsistent and partial (Casper et al., 2023). The human role here is therefore control and the test against real use, not the engine of the loop. In the author's practice, the loop still needs a human prompt: the author explicitly asks for recursive improvement, that is, for a finished result to be critiqued and improved once more, as with the repeated consistency runs of the Norway report (Figure 5).

Documents or files. The two reports were kept as Markdown files; this paper was written in a shared online document, Claude Docs, with a Markdown copy. The document gave the author a live view, comments and a revision history, and the assistant could change single passages by exact wording, which suits the rule of the narrowest anchor. The files gave what the checks need: every change can be counted, compared word by word and versioned in the file name (Ram, 2013; Wilson et al., 2017). Keeping both doubled the work: each of the 19 rounds of change to this paper logged between 19:30 and the end of 26 September was made in the document and in the file, and the two were compared in 11 separate checks. The comparison itself had to be adapted, because the document splits words at formatting boundaries and a word count reported differences where the text was the same; only a character comparison without spaces gave a clean result. Direct links to Google Docs or similar editors offer the same live view to co-authors. For work built on counted checks, the file is the better master copy, and the online document the place where the author reads, comments and confirms. This paper itself left that setup on 27 September: to end the double work, only the online document has been kept since then, and the Markdown copy was frozen. Figures followed a fixed order of routes: if an editable source exists, change the model and export it with the original tool; otherwise change or rebuild the script that drew the figure; otherwise add without deleting; otherwise make a minimal raster edit in a narrow, uniform area; failing all of these, ask the author for the source file. The six figures of this paper took the second route: each was redrawn from a script, and the scripts are kept with the figures in the project files.

Records outside the chat. In the author's view, the most general lesson of this work is that everything that matters has to leave the chat and be kept in documents that can be read, checked and exported. The working chat is where the work



happens, but it is not a record: it grows long, it is summarised when it grows too long, and it cannot be cited. This paper shows the cost. The 268 working patterns could only be recovered by reading the chats afterwards; 32 of the 117 statements about the two projects checked for this paper could be traced only to a summary document compiled from the chats, not to a saved check report; and of the first 30 inputs in the chat in which this paper was revised, only their order survived, because the chat had been summarised. Logs, work plans and decision logs kept the rest traceable. The workflow therefore now adds a chat record, laid down in both templates: at the end of each working session, and before a chat is summarised, every input of the author is written into a document with its result and a pointer to the log entry, decision or file it led to. The record keeps the author's words and the log keeps the measured values, so that no number is held twice. The same records serve the assistant: when the work on this paper moved to a new chat on 27 September 2026, the new chat resumed from a handover document and the project files, not from the old conversation, and the second check described in section 3.2 compared them with the state of the work. This follows the reason why software engineering records design decisions (Jansen and Bosch, 2005; Tyree and Akerman, 2005) and why reproducible research keeps a record of every step (Sandve et al., 2013).

Limits. The evidence comes from two reports by one author and, for writing and checking, one assistant, and the counts are taken from the author's own records. About one reference in eight in the Germany list could not be fully verified. Two skills, for genre selection and genre analysis, were each tested once or not at all. The effort was large: 243 log entries for a pre-trip field guide. A lighter version would keep the register of repeated statements, one consistency run per chapter and one fact check per chapter, and drop most of the project files.

6 Conclusions and outlook

An AI assistant can support the writing of field guides and excursion reports, but its main value in these two cases lay in systematic, counted checking along separate axes, with the author speaking, deciding, observing and confirming the complete wording of every proposed change before it was inserted, and the system writing, counting, checking and keeping the records. The rules that made this work were learned from incidents and written back into skills; the two templates record 43 of them with their incident. Building the skills in loops supported their building, as far as the two templates show, since every one of these rules came from an incident; whether it also improves the reports is indicated but remains to be shown.

The Norway tour in 2027 will test the workflow in the other direction: a separate results report will follow the dossier, and its field notes will be filled by dictation on site. Two questions should be measured then: whether guided dictation yields as much content as recorded speech, and how the share of follow-up corrections changes with the register of repeated statements. Beyond that tour, it remains to be tested whether the skills transfer to other authors and other field sciences. A further question is whether a skill can improve itself: whether it could end each task by proposing, from the incidents of that task, the rule it lacked, so that the loop described in section 5 would run with the author's confirmation as its only human step.



380 Appendix A. The skills

Table A1. The 20 skills and the two framework skills, by role.

Role	Skill	Task
Research	science-recherchebericht-mit-quellen	multi-step web research with numbered references
Research	science-deep-research-prompt-generator	structured brief for a deep research tool
Research	geo-fundort-datenbank-recherche	locality links in Mindat, Mineralienatlas and the Paleobiology Database; exact matches only
Research	geo-mindat-datenbank-recherche-mit-api	the same with the Mindat API, including mineral lists
Elicitation	science-beobachtung-erfassen-befragung	draws out the author's observations one question at a time and grades them as observation, estimate, interpretation or hearsay, or marks the point as not observed
Writing	geo-exkursionsbericht-erstellen	chooses the genre among 15 types before writing
Writing	geo-excursion-report	builds the report with itinerary table, route map, georeferenced overlay and stratigraphic chart
Writing	geo-fundstellen-bibliographie	per-locality bibliography in four categories; every DOI resolved
Writing	geo-fundstueckregister	specimen register with inventory numbers, open nomenclature and labels
Rules	geo-fundstellen-evidenzregeln	locality identity, no transfer from region to point, specimen-bound records, markers
Clarification	geo-frage-belegbasiert-klaeren	one geological question answered with source, evidence level, scope and confidence
Clarification	science-entscheidungspunkt-klaeren	one decision point turned into an applicable rule after counting the cost of each option
Checking	science-faktenbeleg-pruefung	fact table, check in the original source, peer-reviewed replacements, change report
Checking	science-interne-konsistenz-pruefung	the document against itself in six dimensions
Checking	science-quellenapparat-pruefung	every reference entry, DOI and URL, and citations in both directions
Checking	geo-fundstellen-evidenz-pruefung	a locality document against the evidence rules
Checking	geo-exkursionsbericht-typenanalyse	a finished report against every candidate genre
Checking	science-dokumentkritik-drei-sichten	usefulness from three reader perspectives, with 15 measures each
Steering	science-arbeitsplan-erstellen	checkable work plan before execution
Maintenance	science-aenderungsfortpflanzung	register of repeated statements, consulted before every change
Framework	template-dokumentprojekt	file set, naming, shared versions, identifiers, decision log, log, chat record, wording proposals and the check after every write
Framework	template-skill-projekt	file set, naming, versions, identifiers, log,



chat record, wording proposals and the
check after every write for skill projects

Three of the 22 entries carry a version number in their header: `science-arbeitsplan-erstellen` (5.0.0),
`template-skill-projekt` (1.5.0) and `template-dokumentprojekt` (1.6.0); version 1.5.0 of 27 September
385 2026 added the chat record and the wording proposals to both templates, and version 1.6.0 of 28 September 2026 added the
rule on the running chat to the document template; the others are identified by name, in their state of 27 September 2026.
The skills and templates listed here are published at <https://github.com/nordie69/Talk-Check-Decide> under the MIT licence.

Data availability

The skills and templates on which the workflow rests are available at <https://github.com/nordie69/Talk-Check-Decide> under
390 the MIT licence (Appendix A). The project records of the two reports, namely logs, work plans, check reports and chat
records, and the scripts that drew the figures are not published. The two reports themselves are not yet published; the author
intends to make them publicly available on a suitable platform.

Author contributions

SM conceived the study, developed the workflow and the skills, carried out the fieldwork and both excursion reports, curated
395 the project records, and wrote, reviewed and edited the manuscript. The use of an AI assistant is described in section 3.6.

Competing interests

The author declares that he has no conflict of interest.

Financial support

This research received no external funding.

400 References

- Amano, T., González-Varo, J.P. and Sutherland, W.J. (2016) 'Languages are still a major barrier to global science', *PLOS*
Biology, 14(12), e2000933. <https://doi.org/10.1371/journal.pbio.2000933>
Amano, T., Ramírez-Castañeda, V., Berdejo-Espinola, V., Borokini, I., Chowdhury, S., Golivets, M., González-Trujillo,
J.D., Montaña-Centellas, F., Paudel, K., White, R.L. and Veríssimo, D. (2023) 'The manifold costs of being a non-native
405 English speaker in science', *PLOS Biology*, 21(7), e3002184. <https://doi.org/10.1371/journal.pbio.3002184>



- Amershi, S., Weld, D., Vorvoreanu, M. et al. (2019) 'Guidelines for human-AI interaction', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York: ACM, pp. 1-13. <https://doi.org/10.1145/3290605.3300233>
- Baucon, A. and Neto de Carvalho, C. (2024) 'Can AI get a degree in geoscience? Performance analysis of a GPT-based artificial intelligence system trained for Earth science (GeologyOracle)', *Geoheritage*, 16(4), 121.
- 410 <https://doi.org/10.1007/s12371-024-01011-2>
- Berdejo-Espinola, V. and Amano, T. (2023) 'AI tools can improve equity in science', *Science*, 379(6636), p. 991. <https://doi.org/10.1126/science.adg9714>
- Bhattacharyya, M., Miller, V.M., Bhattacharyya, D. and Miller, L.E. (2023) 'High rates of fabricated and inaccurate references in ChatGPT-generated medical content', *Cureus*, 15(5), e39238. <https://doi.org/10.7759/cureus.39238>
- 415 Binz, M., Alaniz, S., Roskies, A. et al. (2025) 'How should the advancement of large language models affect the practice of science?', *Proceedings of the National Academy of Sciences of the United States of America*, 122(5), e2401227121. <https://doi.org/10.1073/pnas.2401227121>
- Blackley, S.V., Huynh, J., Wang, L., Korach, Z. and Zhou, L. (2019) 'Speech recognition for clinical documentation from 1990 to 2018: a systematic review', *Journal of the American Medical Informatics Association*, 26(4), pp. 324-338.
- 420 <https://doi.org/10.1093/jamia/ocy179>
- Casper, S., Davies, X., Shi, C. et al. (2023) 'Open problems and fundamental limitations of reinforcement learning from human feedback', *Transactions on Machine Learning Research*. Available at: <https://openreview.net/forum?id=bx24KpJ4Eb> (Accessed: 26 September 2026).
- Chapman, A.D. and Wiczorek, J.R. (2020) *Georeferencing Best Practices*. Copenhagen: GBIF Secretariat.
- 425 <https://doi.org/10.15468/doc-gg7h-s853>
- Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P. and Ruetsch-Chelli, C. (2024) 'Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis', *Journal of Medical Internet Research*, 26, e53164. <https://doi.org/10.2196/53164>
- Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S. and Amodei, D. (2017) 'Deep reinforcement learning from human preferences', in *Advances in Neural Information Processing Systems 30*. Available at: <https://papers.nips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html> (Accessed: 26 September 2026).
- 430 Coe, A.L. (ed.) (2010) *Geological Field Techniques*. Chichester: Wiley-Blackwell. ISBN 9781444330625.
- Compton, R.R. (1985) *Geology in the Field*. New York: John Wiley and Sons. ISBN 0471829021.
- Cooke, N.J. (1994) 'Varieties of knowledge elicitation techniques', *International Journal of Human-Computer Studies*, 41(6), pp. 801-849. <https://doi.org/10.1006/ijhc.1994.1083>
- 435 COPE Council (2024) *COPE position: Authorship and AI tools*. Version 1. Committee on Publication Ethics. <https://doi.org/10.24318/cCVRZBms>
- Day, T. (2023) 'A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT', *The Professional Geographer*, 75(6), pp. 1024-1027. <https://doi.org/10.1080/00330124.2023.2190373>



- 440 Deng, C., Zhang, T., He, Z., Chen, Q., Shi, Y., Xu, Y., Fu, L., Zhang, W., Wang, X., Zhou, C., Lin, Z. and He, J. (2024) 'K2: A foundation language model for geoscience knowledge understanding and utilization', in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24)*. New York: ACM, pp. 161-170.
<https://doi.org/10.1145/3616855.3635772>
- Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A. and Weston, J. (2024) 'Chain-of-Verification reduces hallucination in large language models', in *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563-3578. <https://doi.org/10.18653/v1/2024.findings-acl.212>
- Elsevier (2026) *Generative AI policies for journals*. Available at: <https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals> (Accessed: 26 September 2026).
- Foroumandi, E., Moradkhani, H., Sanchez-Vila, X., Singha, K., Castelletti, A. and Destouni, G. (2023) 'ChatGPT in hydrology and earth sciences: opportunities, prospects, and concerns', *Water Resources Research*, 59(10), e2023WR036288.
<https://doi.org/10.1029/2023WR036288>
- 445 Ganjavi, C., Eppler, M.B., Pekcan, A., Biedermann, B., Abreu, A., Collins, G.S., Gill, I.S. and Cacciamani, G.E. (2024) 'Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis', *BMJ*, 384, e077192. <https://doi.org/10.1136/bmj-2023-077192>
- 455 Gao, T., Yen, H., Yu, J. and Chen, D. (2023) 'Enabling large language models to generate text with citations', in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465-6488.
<https://doi.org/10.18653/v1/2023.emnlp-main.398>
- Geoscience Information Society (2018) *Guidelines for Authors, Editors, and Publishers of Geologic Field Trip Guidebooks*. Revised August 2018. Available at: <https://geoinfo.org/wp-content/uploads/2020/01/guidebook-guidelines.pdf> (Accessed: 26 September 2026).
- 460 Güneş, Y.C., Cesur, T. and Çamur, E. (2026) 'Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties', *Diagnostic and Interventional Radiology*, 32(2), pp. 173-181.
<https://doi.org/10.4274/dir.2025.253101>
- Hadid, A., Chakraborty, T. and Busby, D. (2024) 'When geoscience meets generative AI and large language models: foundations, trends, and future challenges', *Expert Systems*, 41(10), e13654. <https://doi.org/10.1111/exsy.13654>
- 465 Hatvani, I.G., Mudelsee, M. and Kern, Z. (2024) 'Essential questions in earth and geosciences according to large language models', *Open Geosciences*, 16(1), 20220677. <https://doi.org/10.1515/geo-2022-0677>
- Haynes, A.B., Weiser, T.G., Berry, W.R. et al. (2009) 'A surgical safety checklist to reduce morbidity and mortality in a global population', *New England Journal of Medicine*, 360(5), pp. 491-499. <https://doi.org/10.1056/NEJMsa0810119>
- 470 Hoffman, R.R., Shadbolt, N.R., Burton, A.M. and Klein, G. (1995) 'Eliciting knowledge from experts: a methodological analysis', *Organizational Behavior and Human Decision Processes*, 62(2), pp. 129-158.
<https://doi.org/10.1006/obhd.1995.1039>



- Höhne, J.K., Gavras, K. and Claassen, J. (2024) 'Typing or speaking? Comparing text and voice answers to open questions on sensitive topics in smartphone surveys', *Social Science Computer Review*, 42(4), pp. 1066-1085.
- 475 <https://doi.org/10.1177/08944393231160961>
- Hosseini, M., Resnik, D.B. and Holmes, K. (2023) 'The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts', *Research Ethics*, 19(4), pp. 449-465. <https://doi.org/10.1177/17470161231180449>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. and Liu, T. (2025) 'A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions', *ACM Transactions*
- 480 *on Information Systems*, 43(2), pp. 1-55. <https://doi.org/10.1145/3703155>
- ICMJE (n.d.) *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: Use of AI by authors*. Available at: <https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html> (Accessed: 26 September 2026).
- Jansen, A. and Bosch, J. (2005) 'Software architecture as a set of architectural design decisions', in *5th Working IEEE/IFIP*
- 485 *Conference on Software Architecture (WICSA'05)*. IEEE, pp. 109-120. <https://doi.org/10.1109/WICSA.2005.61>
- Jergas, H. and Baethge, C. (2015) 'Quotation accuracy in medical journal articles: a systematic review and meta-analysis', *PeerJ*, 3, e1364. <https://doi.org/10.7717/peerj.1364>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A. and Fung, P. (2023) 'Survey of hallucination in natural language generation', *ACM Computing Surveys*, 55(12), pp. 1-38. <https://doi.org/10.1145/3571730>
- 490 Kobak, D., González-Márquez, R., Horvát, E.-Á. and Lause, J. (2025) 'Delving into LLM-assisted writing in biomedical publications through excess vocabulary', *Science Advances*, 11(27), eadt3813. <https://doi.org/10.1126/sciadv.adt3813>
- Lenzner, T., Höhne, J.K. and Gavras, K. (2024) 'Innovating web probing: comparing written and oral answers to open-ended probing questions in a smartphone survey', *Journal of Survey Statistics and Methodology*, 12(5), pp. 1295-1317. <https://doi.org/10.1093/jssam/smae031>
- 495 Levelt, W.J.M. (1983) 'Monitoring and self-repair in speech', *Cognition*, 14(1), pp. 41-104. [https://doi.org/10.1016/0010-0277\(83\)90026-4](https://doi.org/10.1016/0010-0277(83)90026-4)
- Liang, W., Zhang, Y., Wu, Z. et al. (2025) 'Quantifying large language model usage in scientific papers', *Nature Human Behaviour*, 9(12), pp. 2599-2609. <https://doi.org/10.1038/s41562-025-02273-8>
- Lisle, R.J., Brabham, P. and Barnes, J.W. (2011) *Basic Geological Mapping*. 5th edn. Chichester: Wiley-Blackwell. ISBN
- 500 9780470686348.
- Madaan, A., Tandon, N., Gupta, P. et al. (2023) 'Self-Refine: iterative refinement with self-feedback', in *Advances in Neural Information Processing Systems 36*, pp. 46534-46594. <https://doi.org/10.52202/075280-2019>
- Manakul, P., Liusie, A. and Gales, M. (2023) 'SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models', in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp.
- 505 9004-9017. <https://doi.org/10.18653/v1/2023.emnlp-main.557>



- Messeri, L. and Crockett, M.J. (2024) 'Artificial intelligence and illusions of understanding in scientific research', *Nature*, 627(8002), pp. 49-58. <https://doi.org/10.1038/s41586-024-07146-0>
- 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), e0184727. <https://doi.org/10.1371/journal.pone.0184727>
- 510 Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. and Fernández-Leal, Á. (2023) 'Human-in-the-loop machine learning: a state of the art', *Artificial Intelligence Review*, 56(4), pp. 3005-3054. <https://doi.org/10.1007/s10462-022-10246-w>
- Nature (2023) 'Tools such as ChatGPT threaten transparent science; here are our ground rules for their use', *Nature*, 613(7945), p. 612. <https://doi.org/10.1038/d41586-023-00191-1>
- 515 Ouyang, L., Wu, J., Jiang, X. et al. (2022) 'Training language models to follow instructions with human feedback', in *Advances in Neural Information Processing Systems 35*, pp. 27730-27744. Available at: <https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html> (Accessed: 26 September 2026).
- Parasuraman, R. and Manzey, D.H. (2010) 'Complacency and bias in human use of automation: an attentional integration', *Human Factors*, 52(3), pp. 381-410. <https://doi.org/10.1177/0018720810376055>
- 520 Peters, S.E. and McClennen, M. (2016) 'The Paleobiology Database application programming interface', *Paleobiology*, 42(1), pp. 1-7. <https://doi.org/10.1017/pab.2015.39>
- Ralph, J., Von Bargen, D., Martynov, P., Zhang, J., Que, X., Prabhu, A., Morrison, S.M., Li, W., Chen, W. and Ma, X. (2025) 'Mindat.org: The open access mineralogy database to accelerate data-intensive geoscience research', *American Mineralogist*, 110(6), pp. 833-844. <https://doi.org/10.2138/am-2024-9486>
- 525 Ram, K. (2013) 'Git can facilitate greater reproducibility and increased transparency in science', *Source Code for Biology and Medicine*, 8, 7. <https://doi.org/10.1186/1751-0473-8-7>
- Rao, D., Wong, E. and Callison-Burch, C. (2026) 'Detecting and correcting reference hallucinations in commercial LLMs and deep research agents', *arXiv*, 2604.03173. <https://doi.org/10.48550/arXiv.2604.03173>
- 530 Revilla, M. and Couper, M.P. (2026) 'Expanding survey response options: combining dictation and/or voice recording with text to answer narrative open-ended survey questions', *Survey Research Methods*, 20(1), pp. 19-32. <https://doi.org/10.18148/srm/2026.v20i1.8456>
- Revilla, M., Couper, M.P., Bosch, O.J. and Asensio, M. (2020) 'Testing the use of voice input in a smartphone web survey', *Social Science Computer Review*, 38(2), pp. 207-224. <https://doi.org/10.1177/0894439318810715>
- 535 Rohanian, M., Nooralahzadeh, F., Rohanian, O., Clifton, D. and Krauthammer, M. (2023) 'Disfluent cues for enhanced speech understanding in large language models', in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3676-3684. <https://doi.org/10.18653/v1/2023.findings-emnlp.238>



- Ruan, S., Wobbrock, J.O., Liou, K., Ng, A. and Landay, J.A. (2018) 'Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4), 159. <https://doi.org/10.1145/3161187>
- Sandve, G.K., Nekrutenko, A., Taylor, J. and Hovig, E. (2013) 'Ten simple rules for reproducible computational research', *PLOS Computational Biology*, 9(10), e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>
- Schegloff, E.A., Jefferson, G. and Sacks, H. (1977) 'The preference for self-correction in the organization of repair in conversation', *Language*, 53(2), pp. 361-382. <https://doi.org/10.2307/413107>
- Schön, D.A. (1983) *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books. ISBN 046506874X.
- Shneiderman, B. (2020) 'Human-centered artificial intelligence: reliable, safe & trustworthy', *International Journal of Human-Computer Interaction*, 36(6), pp. 495-504. <https://doi.org/10.1080/10447318.2020.1741118>
- Sutton, R.S. and Barto, A.G. (2018) *Reinforcement Learning: An Introduction*. 2nd edn. Cambridge, MA: MIT Press. ISBN 9780262039246.
- Todd, P.A., Guest, J.R., Lu, J. and Chou, L.M. (2010) 'One in four citations in marine biology papers is inappropriate', *Marine Ecology Progress Series*, 408, pp. 299-303. <https://doi.org/10.3354/meps08587>
- Topaz, M., Zhang, Z., Roguin, N., Gupta, P., Li, Z. and Peltonen, L.-M. (2026a) 'Biomedical reference generation remains unreliable across 26 large language models', *arXiv*, 2609.14988. <https://doi.org/10.48550/arXiv.2609.14988>
- Topaz, M., Roguin, N., Gupta, P., Zhang, Z. and Peltonen, L.-M. (2026b) 'Fabricated citations: an audit across 2.5 million biomedical papers', *The Lancet*, 407(10541), pp. 1779-1781. [https://doi.org/10.1016/S0140-6736\(26\)00603-3](https://doi.org/10.1016/S0140-6736(26)00603-3)
- Tyree, J. and Akerman, A. (2005) 'Architecture decisions: demystifying architecture', *IEEE Software*, 22(2), pp. 19-27. <https://doi.org/10.1109/MS.2005.27>
- Walters, W.H. and Wilder, E.I. (2023) 'Fabrication and errors in the bibliographic citations generated by ChatGPT', *Scientific Reports*, 13, 14045. <https://doi.org/10.1038/s41598-023-41032-5>
- Wieczorek, J., Guo, Q. and Hijmans, R. (2004) 'The point-radius method for georeferencing locality descriptions and calculating associated uncertainty', *International Journal of Geographical Information Science*, 18(8), pp. 745-767. <https://doi.org/10.1080/13658810412331280211>
- Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. and Teal, T.K. (2017) 'Good enough practices in scientific computing', *PLOS Computational Biology*, 13(6), e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>
- Wolniewicz, P. (2024) 'The combined use of GIS and generative artificial intelligence in detecting potential geodiversity sites and promoting geoheritage', *Resources*, 13(9), 119. <https://doi.org/10.3390/resources13090119>
- Wu, K., Wu, E., Wei, K., Zhang, A., Casasola, A., Nguyen, T., Riantawan, S., Shi, P., Ho, D. and Zou, J. (2025) 'An automated framework for assessing how well LLMs cite relevant medical references', *Nature Communications*, 16, 3615. <https://doi.org/10.1038/s41467-025-58551-6>



- Zhang, X., Zhao, Y., Xie, J., Li, C. and Hu, Z. (2020) 'Geological big data acquisition based on speech recognition', *Multimedia Tools and Applications*, 79(33), pp. 24413-24428. <https://doi.org/10.1007/s11042-020-09064-5>
- Zhou, L., Blackley, S.V., Kowalski, L., Doan, R., Acker, W.W., Landman, A.B., Kontrient, E., Mack, D., Meteer, M., Bates, D.W. and Goss, F.R. (2018) 'Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists', *JAMA Network Open*, 1(3), e180530. <https://doi.org/10.1001/jamanetworkopen.2018.0530>
- 575 Zizka, A., Silvestro, D., Andermann, T. et al. (2019) 'CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases', *Methods in Ecology and Evolution*, 10(5), pp. 744-751. <https://doi.org/10.1111/2041-210X.13152>