PADERBORN UNIVERSITY

The Responsible AI Literacy (RAIL) Competency Model for Researchers (Validation Version 0.9) – Competency Element Descriptions

Knowledge, Skills, and Responsible Integrity Dispositions for AI-Augmented Research

Stephan Drechsler

June 2026

Disclaimer for the RAIL Competency Element Descriptions

Current Status: Validation Phase (Pre-Publication)

This document presents the **RAIL Competency Model's (Validation Ver. 0.9) Competency Element Descriptions**. Please note that this model is currently undergoing formal validation (e.g., via Nominal Group Technique (NGT)) as part of my ongoing doctoral project. The results and the finalized framework have not yet been published in a peer-reviewed dissertation.

Licensing & Usage Terms: To ensure the integrity of the validation process and to prevent the circulation of divergent versions during this critical phase, this work is currently licensed under: **Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0)**

Note for Validation Participants: While the general public is subject to the 'No Derivatives' (ND) restriction of the CC BY-ND 4.0 license, registered validation experts are expressly permitted and encouraged to propose modifications, deletions, or additions to the model within the scope of the NGT sessions. To register as a validation participant, please go to **Registration: RAIL Model Validation Sessions.**

Future Release Notice: The author is committed to Open Science principles. Upon successful completion of the validation phase and the official publication of the dissertation, the finalized version of the RAIL Researcher Competency Model and its descriptions will be re-released under a more permissive **Creative Commons Attribution 4.0 International (CC BY 4.0)** license to encourage widespread adoption and adaptation within the research community.

Contact & Feedback: If you are interested in using this model for research purposes or wish to provide feedback, please contact me:

If you refer to this document in your work, you can use the following citation format:

Drechsler, S. (2026). *The Responsible AI Literacy (RAIL) Competency Model for Researchers (Validation Version 0.9) – Competency Element Descriptions: Knowledge, Skills, and Responsible Integrity Dispositions for AI-Augmented Research*. Paderborn University, Germany. https://doi.org/10.17619/UNIPB/1-2591

Stephan Drechsler, M.A.
Research Associate
Educational Management and Research in Further Education

Warburger Str. 100
33098 Paderborn
Germany
Office TP6.2.310
Telephone +49 5251 60-5212
E-Mail stephan.drechsler@uni-paderborn.de
Web www.uni-paderborn.de/en/person/26915

Contents

Introduction

This is a comprehensive overview of the concepts and their descriptions of the Responsible AI Literacy (RAIL) competency model for researchers. In the preprint of Drechsler (2026) introducing the RAIL model, the definitions of each competency element were omitted. However, now, to provide clarification of the knowledge, skills, and responsible integrity dispositions identified in the document analysis, you will find each of them and the closest theoretical construct they correspond to. In practice, researchers can use these descriptions to guide decision-making when designing, conducting, or evaluating AI-augmented research, ensuring that their activities align with the expected standards for responsible AI use. The descriptions also serve as a reference for professional development, self-assessment, and team discussions, helping identify individual and group strengths and areas for growth related to responsible AI practices.

It is recommended that this document be read alongside the RAIL Competency Model for Researchers (Validation Version 0.9). The main model paper presents the broader framework, rationale, and competency statements, while this document provides detailed definitions of the individual competency elements needed to interpret and evaluate those statements.

This overview should add value and clarify the concepts and synthesis presented in the RAIL validation version, helping researchers better understand the RAIL competencies, their implications for research practices, and their implications for researchers' professional development. By clarifying these descriptions, the paper aims to help researchers more easily determine whether a competency element is misaligned, missing, or unnecessarily repeated within the competency structure. Ultimately, this comprehensive overview aims to equip researchers with a clear understanding of the RAIL model, enabling more effective and accessible validation within the research community.

The Validation Process

To participate in the validation process, eligible researchers who use artificial intelligence (AI) in their research and have at least 3 years of professional research experience are invited to register using the online form provided with this paper. After registration, participants will receive detailed instructions by email, including an outline of the activities involved.

These typically include reviewing the competency element definitions and providing feedback via a Nominal Group Technique (NGT) session lasting 90 to 120 minutes, depending on group size, with 4 to 9 people per session. There is absolutely no mandatory pre-reading, and experts will not be asked to complete any asynchronous work after the call. The 90-to-120-minute session is entirely self-contained. However, to make things even easier, registered participants will be given access to a dedicated Google NotebookLM instance. This AI-powered briefing assistant lets each expert dynamically query the RAIL model v0.9, generate instant summaries, or listen to audio overviews at each expert's convenience, mitigating the cognitive load of traditional framework review. The NGT session takes place via a lightweight Zoom meeting integrated with a digital whiteboard, with no pre-installation required. All experts need is a stable internet connection, a microphone, and a webcam. This validation method is an excellent way to connect with other interested researchers and to gain insight from them. Compared with other validation methods, NGT has several advantages: While quantitative surveys miss qualitative nuance, open debates are often dominated by one or two vocal participants. NGT uses a structured sequence of silent idea generation, round-robin clarification, and Likert-scale scoring. This ensures that experts' specific expertise is mathematically weighted and captured, even if any view differs from the rest of the group. In total, NGT slots are available in July 2026, after which the results will be analyzed to directly inform the development of model version 1.0. However, additional NGT can be scheduled upon request.

Here, NGT is a structured five-step process that identifies which competency elements are misaligned, missing, or redundant. Here, the NGT sessions are presented in more detail. The core question each NGT group

will address is: Which specific competency elements (Knowledge, Skills, or Responsible Integrity Dispositions) are currently missing, redundant, or misaligned with the actual ethical and methodological demands of AI-supported research in your specific research practices? After this question is introduced as the core challenge of each session, participants will be asked to silently generate ideas, considering what their daily research work demands that the RAIL v0.9 model does not yet address. Then, each participant's ideas will be recorded (one at a time) in a round-robin fashion. After this, the group will clarify the ideas and evaluate whether they can be grouped into broader categories. In the next phase, participants will vote on each idea using a 5-point Likert scale, ranging from 1 = "Not important" to 5 = "Very important". Here, the focus will be on ideas with the highest potential impact on research integrity when using AI. If an SD > 1.0 is found in the scoring results, a targeted discussion will be triggered to understand the source of disagreement among participants, ensuring that dissenting voices are documented. Finally, the prioritized list will be presented, and its role in informing the development of RAIL v.1.0 will be explained.

Every professional status group (e.g., doctoral students, independent researchers, lecturers, directors, professors, and postdocs) is encouraged to participate. Perspectives from all disciplines are welcome, from the humanities to quantum physics to data science, to ensure universal framework applicability and adaptability. To achieve global representation and accommodate the schedules of international experts without disrupting standard working hours, the validation timeline was stratified across three primary geographical corridors: Asia-Pacific (APAC); Europe, Middle East, and Africa (EMEA); and North and South America (The Americas). Sessions were mapped to local time zones, utilizing standardized UTC windows alongside local indicators such as SGT/JST for APAC, CET/SAST for EMEA, and EST/EDT/PST/PDT for the Americas, ensuring equitable access for the global scientific community.

A primary methodological objective in the panel selection process is to ensure that the validation data do not exhibit W.E.I.R.D. (Western, Educated, Industrialized, Rich, and Democratic) population bias. Because perceptions of artificial intelligence, automated decision-making, and digital ethics vary significantly across different socio-political and economic landscapes, relying solely on experts from Western or highly industrialized academic institutions would compromise the universal generalizability of the RAIL model (Rosales, 2024). To deliberately counteract this bias, the validation schedule was structurally balanced across distinct global geographic blocks, explicitly prioritizing equal access for researchers located in the Asia-Pacific (APAC) and African (EMEA) regions alongside Western cohorts. This global stratification ensures that the localized realities, distinct regulatory frameworks, and varied cultural perspectives of non-Western research environments are fundamentally integrated into the consensus-building process, resulting in a robust, globally representative framework for version 1.0.

If you are interested in joining the validation or need further information, please refer to the **contact and registration details** at the beginning of this document. Every researcher who participates in an NGT session will receive a certificate of participation from Paderborn University to verify their role in the model's validation process. The feedback collected through the NGT sessions will be used as data and will be anonymous.

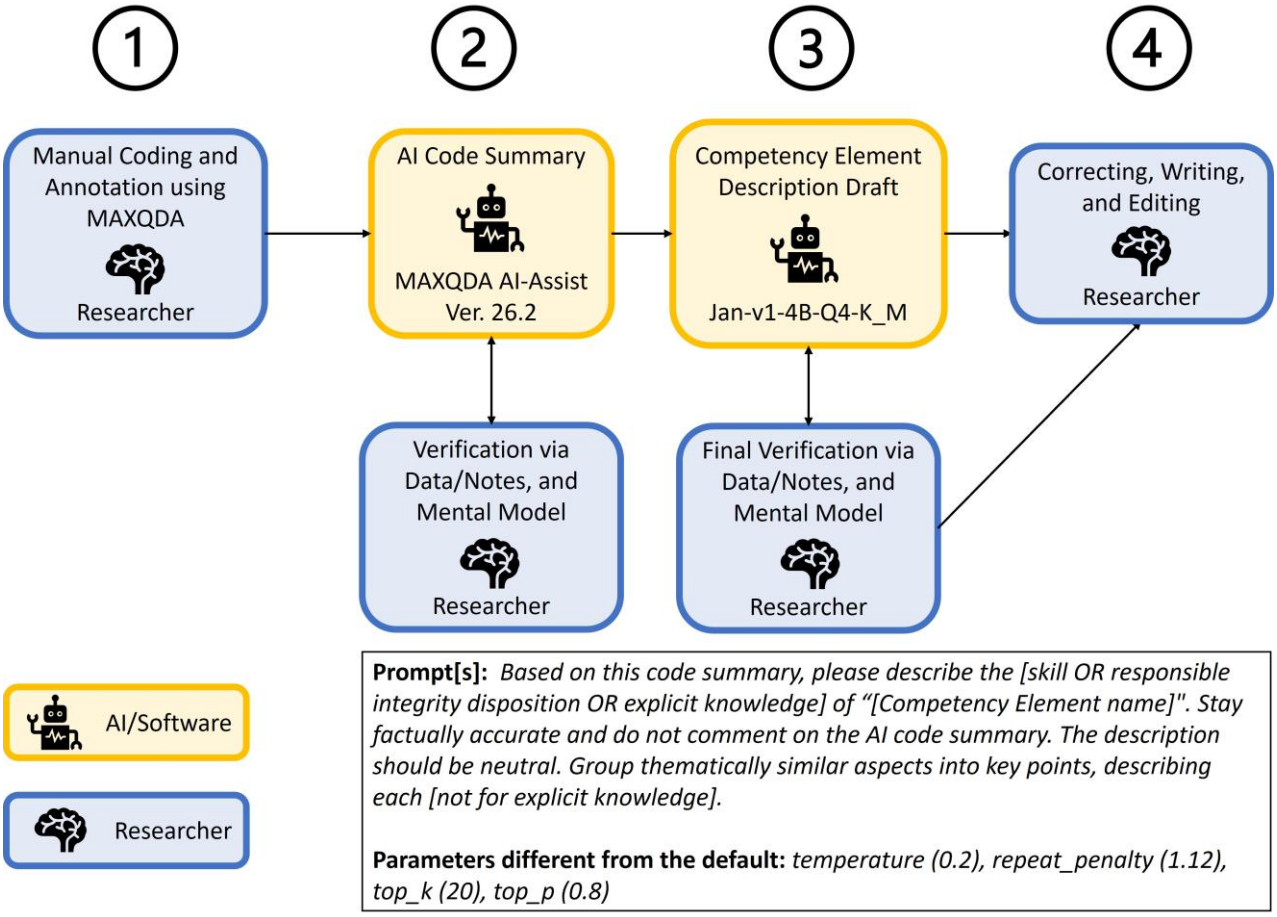
To ensure transparency and enable peer verification and further research, both the validated RAIL model and the anonymized dataset generated during the validation process will be made openly accessible. The dataset will not contain any information that could identify individual participants, and all data will be presented only in aggregate form. Detailed information on how to access the final model and the corresponding dataset will be provided upon completion of the validation process. As outlined in the preprint, the RAIL model version 1.0 will be published under a CC-BY 4.0 license once the validation process is complete.

Competency Elements Explained

The following definitions of the competency elements (1) Responsible Integrity Dispositions (RIDs), (2) Skills, and (3) Knowledge are based on the summaries of the codes identified within the qualitative content analysis (QCA). Please note that the descriptions of the competency elements are based solely on QCA data and

omit contextualization and discussion within the current scientific discourse. To ensure methodological transparency, maintain cognitive autonomy, and systematically prevent self-plagiarism or excessive textual overlap during the subsequent writing and publication process of the doctoral thesis, particularly in relation to prior conceptual preprints (e.g., Drechsler, 2026), the competency element descriptions in this document were generated using a structured human-in-the-loop procedure (see Figure 1):

Figure 1:
Process of Synthesizing the Competency Element Descriptions



To ensure that the competency descriptions adequately represented the data, their development followed a sequential human-in-the-loop procedure. The underlying material was first manually coded and annotated by the researcher in MAXQDA 26. AI support was then used for intermediate summarization and drafting: MAXQDA AI Assist generated code summaries, and a local language model generated initial description drafts. At both stages, the researcher verified the outputs against the coded data, analytic notes, and the developing conceptual understanding of the competency model. Draft descriptions were revised, corrected, and edited manually before inclusion in this paper. Thus, representational adequacy was established through repeated researcher validation of AI-assisted outputs against the empirical material, rather than through reliance on the AI systems themselves. Where necessary, descriptions have been corrected, edited, and contextualized within the RAIL model.

Responsible Integrity Dispositions

Responsible Integrity Dispositions (RIDs) are stable tendencies that influence how researchers approach the ethical use of AI, its reporting, or decision-making in the research process. Indications of those stable tendencies included statements about ethical considerations, caution, transparency, or acknowledgment of limitations when conducting research. As noted in the description of each competency element, the elements are interconnected. At the beginning or end of each section, the supporting references of the document analysis are listed.

Epistemic Stewardship (Formerly: Human Expertise (Decision-making))

Epistemic Stewardship is supported by 25 documents (Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Castillo-Martínez et al., 2024; Charton et al., 2024; Dönmez et al., 2023; Fui-Hoon Nah et al., 2023; Glerean & Silva, 2024; Gupta, 2024; Hill et al., 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. Khan et al., 2024; Lieder & Schäffer, 2024; Magesh et al., 2024; Mudd et al., 2024; Nicholson Thomas et al., 2024; Parker et al., 2025; Perkins & Roe, 2024b; Pretorius & Pretorius, 2025; Rietz & Maedche, 2021; Silva et al., 2024; Tran et al., 2023; Yining et al., 2025; Yongtao et al., 2024). It considers human-in-the-loop principles, ensuring that the researcher safeguards the knowledge-generation process. The researcher takes accountability and responsibility for decisions made throughout the research process, informed by their specific disciplinary knowledge and skills. In this sense, epistemic stewardship is the dispositional commitment to actively exercise and uphold human judgment as the ultimate gatekeeper of knowledge production, particularly in AI-augmented research environments.
It reflects a mindset in which researchers do not delegate epistemological authority to algorithms but instead assume personal responsibility for interpreting outputs, verifying claims, detecting bias or hallucinations (e.g., fabricated or hallucinated outputs), and ensuring that decisions align with ethical principles such as fairness, transparency, and accountability.

- **Human Expertise is the Standard**: Epistemic stewardship recognizes that no amount of training data or model sophistication can replicate human-level critical thinking in complex domains like ethics, domain-specific inference, or qualitative reasoning. Therefore, final decisions must rest with humans who bring contextual understanding, emotional intelligence, and moral judgment.
- **Oversight Over Automation:** Rather than treating AI as an autonomous co-author or decision-maker (which risks responsibility drift[1]), epistemic stewards maintain continuous human supervision throughout the research lifecycle, from prompt design to validation of results. This includes reviewing drafts before submission, auditing outputs for consistency and logic, and actively debugging faulty inferences.
- **Accountability Through Engagement:** It is not enough simply to use AI tools; one must be present during interactions, crafting clear prompts (e.g., methodological prompts), iterating on responses via negotiation between model output and human insight, and accepting full ownership when errors occur due to poor prompting or hallucination.
- **Dynamic Validation Loops:** Because LLMs are prone to generating plausible but false information (e.g., hallucinations), epistemic stewardship requires iterative verification processes involving cross-checking against trusted sources (e.g., literature reviews), peer consultation, or direct data access, processes that reinforce human agency and ensure the integrity of scientific conclusions.

[1] A responsibility drift is a self-contradiction in a philosophical and legal sense, as AI cannot be held accountable. Therefore, it cannot be responsible for anything. However, responsibility drift is understood as a sociological phenomenon because researchers are delegating work and responsibility to AI, although it is not an inherent property of AI.

For instance, in professional practice, an epistemically stewarded researcher might ask an AI to draft a hypothesis but then personally validate its logical structure and relevance using domain knowledge; they may delegate technical writing tasks while retaining editorial control over claims about causality. They protect the scientific record from compromise by automated error or opacity.

In essence, epistemic stewardship is the stable tendency to treat AI-generated content not as authoritative output but as a hypothesis that requires rigorous validation by human expertise.

Ultimately, epistemic stewardship transforms into practice by embedding vigilance, curiosity, and ownership directly into workflow habits, thereby maintaining intellectual sovereignty even in highly AI-integrated laboratories.

AI Coworker Curiosity

AI Coworker Curiosity is supported by 24 documents (Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Glerean & Silva, 2024; Helgesson, 2024; Hill et al., 2024; Ilegbusi, 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. Khan et al., 2024, 2024; Kuiper, 2024; Lieder & Schäffer, 2024; Limongi, 2024; Magesh et al., 2024; Mudd et al., 2024; Nicholson Thomas et al., 2024; Perkins & Roe, 2024b; Picalho et al., 2025; Pretorius & Pretorius, 2025; Rietz & Maedche, 2021; Silva et al., 2024; Skyba et al., 2024; Tran et al., 2023; Yining et al., 2025). It means a stable tendency toward active inquiry into the internal processes, decision logic, and contextual reasoning pathways of one's AI research partner. This disposition manifests through persistent questions about how models arrive at conclusions ("context of discovery"), interest in parameter-level effects ("temperature", "top-p"), desire to challenge assumptions with provocation-style queries, and a belief that human-AI interaction fosters mutual deepening of understanding, treating the AI not as an oracle but as a dynamic collaborator whose learning can be observed and shaped. AI Coworker Curiosity is the dispositional drive to actively investigate, understand, and engage with an AI partner as a dynamic collaborator, not merely as a tool to be used passively. It reflects a mindset rooted in inquiry: researchers deliberately seek insight into how their AI co-worker processes information, generates outputs, adapts to instructions, and interprets data.
This disposition considers the inherent urge to get to know the AI research partner as closely as possible to understand how it functions, why it produces certain outcomes, and how to most effectively combine human and machine capabilities. Moreover, it emphasizes a willingness to try novel AI-based approaches to research problems. As a latent construct, curiosity is inferred from several key behaviors:

- **Exploring the "Black Box":** Recognizing that AI models operate internally without transparency (e.g., no visible training weights or internal states), researchers maintain a persistent interest in probing how decisions are made, what prompts trigger which responses, and how context shapes AI outcomes.
- **Understanding Parameters as Levers of Control:** A curious user doesn't just issue commands; they experiment with settings like temperature, top-p sampling, or max tokens to observe shifts in randomness, coherence, and creativity. These aren't technical afterthoughts but intentional variables for shaping the quality of interaction.
- **Seeking Alternative Interpretations:** Rather than accepting the first plausible answer, researchers deliberately challenge their AI to offer divergent viewpoints, using it not only as a helper but also as a "provocateur" that forces reevaluation of assumptions or biases embedded in human perception.
- **Mutual Learning Dynamics:** The relationship isn't one-way; curiosity fosters reciprocal development. As the researcher gains better control over prompting strategies and AI output patterns, both parties become more effective at navigating complex research problems.

In essence, AI Coworker Curiosity transforms interactions from transactional to co-creation. Instead of treating AI as a tool for their research work, the researcher treats it as an (artificially) intelligent partner whose capabilities vary but can be expanded through continuous exploration and feedback loops.

Ultimately, knowing the AI isn't about memorizing its API; it's about cultivating the willingness to ask questions, test hypotheses, and keep digging until both the researcher and the model are better equipped than before.

Adherence to Scientific Standards

Adherence to Scientific Standards is supported by 23 documents (Burger et al., 2023; Burgui-Burgui, 2023; Chen et al., 2024; Freiesleben, 2023; Glerean & Silva, 2024; Gokul et al., 2024; Helgesson, 2024; Khalifa & Albadawy, 2024, 2024; N. Khan et al., 2024; Lieder & Schäffer, 2024; Magesh et al., 2024; Mohamed et al., 2025; Mudd et al., 2024; Murphy & James, 2024; Nguyen et al., 2024; Nicholson Thomas et al., 2024; Perkins & Roe, 2024b; Salvioni & Almici, 2024; Silva et al., 2024; Tran et al., 2023; Wise et al., 2024; Ye et al., 2025; Yining et al., 2025). The responsible integrity disposition of "Adherence to Scientific Standards", as articulated in the 2026 code summary, is a proactive, iterative commitment by researchers to uphold rigorous scientific practices *through explicit ethical and methodological safeguards*, particularly when integrating AI tools. This disposition relates to seven key dimensions identified in the documents:

- **Proactive Design Rigor**: Researchers must adapt research protocols (e.g., randomized controlled trials) to explicitly meet statistical validity and reproducibility standards, avoiding AI-driven shortcuts that compromise experimental integrity.
- **Integrity Checks**: Systematic use of data integrity validators to detect hallucinations or inconsistencies in AI-generated outputs (e.g., "non-existent quotes" from RAG systems), ensuring all findings align with verifiable evidence.
- **Transparent Ethical Compliance**: Mandatory alignment with AI ethics frameworks, including bias mitigation, algorithmic transparency, and privacy safeguards, to prevent violations like data falsification or plagiarism via uncredited AI outputs.
- **Reproducibility Enforcement**: Implementing Workflow Management Systems (WfMS) to document iterative prompt-testing processes (e.g., RSRP methodology), guaranteeing that AI-assisted analyses can be replicated without ambiguity.
- **Risk-Based Accountability**: Classifying and mitigating AI-related ethical risks using criteria like impact magnitude and likelihood (e.g., avoiding LLMs for critical tasks requiring human judgment).
- **Plagiarism Prevention Mechanisms**: Distinguishing between AI-generated plagiarism and user errors, with tools to detect unauthorized use of others' work in drafting or peer review.
- **Continuous Validation Frameworks**: Adopting tests like the ROBOT[2] (Reliability, Objectivity, Bias, Ownership, Type) assessment to rigorously evaluate AI tools *before* critical deployment, ensuring outputs meet scientific standards without overreliance on uninterpretable models.

This disposition rejects passive compliance; it demands active vigilance through structured practices (e.g., iterative prompt refinement via RSRP (see Lieder & Schäffer, 2024), transparent documentation of AI interactions, and human oversight for high-stakes decisions. Crucially, it recognizes that adherence is not static: as AI tools evolve (e.g., via customized GPTs), researchers must continuously recalibrate their methods to preserve authenticity, reliability, and ethical accountability, ultimately framing scientific standards as a dynamic responsibility rather than a fixed benchmark.

[2] Murphy and James (2024) adapted the test from Hervieux, S. & Wheatley, A. (2020). The Robot Test [evaluation tool], The LibrAIry. https://thelibrairy.wordpress.com/2020/03/11/the-robot-test/

Taking Ownership

Taking Ownership is supported by 23 documents (Barrot, 2025; Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Castillo-Martínez et al., 2024; Cotton et al., 2024; Dönmez et al., 2023; Fui-Hoon Nah et al., 2023; Glerean & Silva, 2024; Gupta, 2024; Helgesson, 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. Khan et al., 2024; Murphy & James, 2024; Nguyen et al., 2024; Perkins & Roe, 2024a; Shukla et al., 2024; Silva et al., 2024; Wise et al., 2024; Ye et al., 2025; Yining et al., 2025; Yongtao et al., 2024).
The responsible integrity disposition of "Taking Ownership" refers to the proactive ethical commitment researchers must exercise to personally bear accountability for all AI-assisted decisions, ensuring that ownership is not an abstract concept but a stable tendency that fosters concrete practices embedded at every stage of research. It entails the following dimensions:

- **Unambiguous Accountability**: Researchers must personally own all outcomes generated via AI tools, acknowledging that AI itself cannot be held legally, ethically, or academically accountable for results (e.g., inaccuracies, biases). This includes taking full responsibility for decisions made during research design (e.g., selecting AI models, tools, and prompt structures).
- **Active Human Oversight**: Researchers must rigorously validate AI-generated content through iterative human review to ensure accuracy, credibility, and ethical integrity. This involves:
  (1) Identifying and correcting errors in AI outputs (e.g., misinterpretations from insufficient prompting).
  (2) Maintaining control over research interpretations and final analyses, never delegating critical judgment to AI.
- **Transparency and Traceability**: Explicitly documenting how AI was used (e.g., local LLM deployment for cybersecurity safeguards) to uphold integrity in the research process. This includes:
  (1) Acknowledging AI's role in methodology without implying co-authorship or equivalence to human work.
  (2) Ensuring journal submissions clearly disclose AI assistance while retaining human authorship as the sole accountable entity.
- **Proactive Risk Mitigation**: Researchers must actively address ethical and methodological risks (e.g., algorithmic bias, data privacy violations) by:
  (1) Using local LLMs or infrastructure provided by the research institution to minimize external data exposure and enhance control over sensitive materials.
  (2) Integrating their own critical analysis to prevent AI-generated content from becoming fraudulent or misleading.
- **Iterative Responsibility**: Ownership is not a one-time act but an ongoing commitment, requiring thorough manuscript revisions, continuous self-evaluation of biases, and alignment with journal guidelines for responsible AI use (e.g., identifying AI-edited content).

Why this disposition ensures integrity: It transforms "taking ownership" from a passive obligation into an active, ethical practice that upholds research validity and trust, without conflating human expertise with AI capabilities. This aligns with the disposition's core principle: *AI cannot be given authorship; human accountability remains irreplaceable*.

Honesty (Formerly: Transparency/Explainability)

Honesty is supported by 29 documents (Bhutta, 2024; Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Charton et al., 2024; Chen et al., 2024; Chopra & Haaland, 2023; Cotton et al., 2024; Freiesleben, 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Helgesson, 2024; Hill et al., 2024; Karakose, 2023; Khalifa & Albadawy, 2024; Lieder & Schäffer, 2024; Limongi, 2024; Magesh et al., 2024;

Mudd et al., 2024; Nicholson Thomas et al., 2024; Perkins & Roe, 2024a; Pretorius & Pretorius, 2025; Rietz & Maedche, 2021; Salvioni & Almici, 2024; Shukla et al., 2024; Silva et al., 2024; Yongtao et al., 2024).
Based on the provided code summary, the responsible integrity disposition of "Honesty" relates to these critical practices and obligations:

- **Transparent Disclosure**: Researchers explicitly disclose *all* aspects of AI usage in their work, including specific AI tools, prompts, parameters (e.g., temperature settings), data sources, and iterative refinement processes, to prevent misrepresentation or unverified claims about AI's role. This ensures honesty by making the full technical context visible to peers and reviewers.
- **Author Accountability**: Authors bear ultimate responsibility for the accuracy and integrity of AI-generated content, even when AI assists with tasks like text editing or literature reviews. This includes verifying outputs against ethical standards (e.g., bias checks) and acknowledging AI's influence on decisions without misattribution as co-authorship.
- **Provenance Documentation**: Maintaining meticulous records of data lifecycles, prompt histories, and model iterations demonstrates honesty by enabling traceability and replication. For instance, sharing exact token limits or prompt revisions ensures that opaque processes don't obscure findings.
- **Bias Mitigation and Verification**: Honesty requires actively identifying and addressing AI-induced biases through diverse user testing, transparency in decision rationales (e.g., "why did the model choose this outcome?"), and iterative refinement to uphold ethical integrity, never treating AI outputs as infallible.
- **Compliance with Publication Standards**: Adhering to journal requirements for explicit AI disclosures (e.g., how tools were applied) reflects honesty by aligning research practices with community expectations for credibility and trustworthiness.

In essence, *honesty* here is the more visible part of the control exercised by taking ownership, where researchers proactively disclose gaps, verify outputs, and accept accountability for AI's influence to preserve scientific integrity. This mitigates the pitfalls of "AI black boxes" by embedding honesty into every step: from prompt design to final publication.

Alignment with Own Knowledge Interest

Alignment is supported by 14 documents (Alfirević et al., 2024; Barrot, 2025; Burger et al., 2023; Butson & Spronken-Smith, 2024; Castillo-Martínez et al., 2024; Cotton et al., 2024; Dönmez et al., 2023; Garg et al., 2024; Helgesson, 2024; Hill et al., 2024; Karakose, 2023; Khalifa & Albadawy, 2024; Shukla et al., 2024; Tran et al., 2023; Wise et al., 2024). It describes a stable tendency to create harmony among research strategy, structure, and the AI system to satisfy one's own research interests, including decisions made to answer the research question. It ensures an AI-generated response, however useful or efficient, is not only factually accurate but *thematically relevant* and *deeply aligned* with the researcher's specific intellectual research goals. It means going beyond surface-level output to ask: Does this information actually feed into what I care about? Instead of accepting any answer as valid simply because it's generated quickly, a responsible researcher evaluates whether the AI has grasped their unique research trajectory, whether it's probing novel mechanisms in neuroscience or exploring ethical trade-offs in policy design. This disposition relates to active researcher agency:

- **Prompt Refinement as Strategy:** Researchers don't just issue generic queries; they tailor prompts to reflect precise interests, so the AI knows where to focus its energy.
- **Critical Selection Over Acceptance:** Even if an output is technically sound, it must resonate with the research question at hand. For instance, a model might generate excellent technical details on

transformer architectures but miss the sociological implications; alignment demands filtering for interest-specific insight, not just completeness.

- **Human Ownership of Relevance:** The AI acts as an ideation engine or brainstormer, but final judgment belongs to the researcher who determines which ideas are worth pursuing. This prevents knowledge drift, in which tools produce content that appears intelligent yet misses core objectives due to misaligned intent.
- **Transparency in Use Cases:** If a piece of generated text contributes meaningfully (e.g., shaping hypotheses or identifying research gaps), its role should be clearly acknowledged in publications, ensuring integrity and avoiding the illusion of sole human involvement when AI was instrumental.

Ultimately, Alignment with one's own knowledge interests protects intellectual sovereignty. It ensures that efficiency doesn't become a substitute for purpose, and that every piece of information retrieved is not random noise but something genuinely worth investigating because it connects back to what matters most in the researcher's work.

Critical Stance

A Critical Stance is supported by 21 documents. It involves the disposition of professional skepticism. It is required, for instance, in auditing, and involves a questioning mind, alertness to potential misstatements (due to error or fraud), and a critical assessment of evidence (Ennar & Damak-Ayadi, 2024). It requires "*trust but verify*" (Brooks, 2023), demanding that auditors not accept one's representations at face value, even if he/she/it has historically been honest.
In the RAIL Model, a Critical Stance toward AI refers to the responsible integrity disposition that researchers must adopt to actively interrogate and validate AI outputs, ensuring they do not become passive extensions of their own cognition but instead serve as tools for enhanced rigor. It embodies a non-negotiable commitment to professional skepticism that prevents over-reliance while preserving intellectual autonomy. This disposition relates to the following key dimensions:

- **Systematic verification protocols**: Researchers must routinely ask foundational questions (e.g., *"Did AI capture the whole picture?"*, *"Does this align with my interpretation?"*, *"Is the outcome reasonable?"*) before accepting AI-generated insights, particularly when outputs involve high-stakes decisions like academic writing or legal analysis.
- **Bias and accuracy audits**: Proactively identifying gaps in AI responses (e.g., hallucinations, omitted context, misformatted citations) through cross-referencing with scholarly databases or domain expertise to ensure outputs reflect truth, not algorithmic limitations.
- **Active role preservation**: Maintaining human oversight as the ultimate gatekeeper for critical judgments, such as verifying that AI-generated summaries of complex topics (e.g., nootropic drug effects) do not oversimplify or misrepresent nuanced realities.
- **Contextual humility**: Acknowledging AI's inherent constraints (e.g., struggles with sarcasm, humor, or deep contextual analysis) to avoid conflating tool outputs with human-level reasoning, thereby preventing "automation bias" in research decisions.

Critically, this disposition rejects *passive compliance* with AI outputs. Instead, it requires researchers to treat every interaction as an opportunity for cognitive accountability: Critical Stance toward AI is not about distrust but disciplined vigilance, transforming AI's potential to assist into a mechanism for deeper intellectual engagement rather than diminished agency.

It directly supports responsible research integrity by preventing over-reliance on AI, safeguarding against misrepresentation, and maintaining the researcher's role as the ultimate arbiter of accuracy and validity in AI-assisted work. Hence, it is foundational to the trustworthy integration of AI across disciplines. In essence,

it is the commitment to keep questioning until confidence can be established through evidence-based validation.

Self-reflectiveness

Self-reflectiveness as a disposition is a stable, personal tendency to actively inspect, evaluate, and learn from one's own thoughts, feelings, and behaviors. It is a proactive, often habitual mindset that fosters self-awareness, personal growth, and emotional regulation, enabling individuals to improve future actions based on insights from past experiences. Grant, Franklin, and Langford (2002) describe it as "the inspection and evaluation of one's thoughts, feelings, and behavior and insight, the clarity of understanding of one's thoughts, feelings, and behavior" (Grant et al., 2002), p. 821).
Self-reflectiveness is supported by 18 documents (Barrot, 2025; Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Castillo-Martínez et al., 2024; Charton et al., 2024; Dönmez et al., 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Karakose, 2023; N. A. Khan et al., 2023; Kuiper, 2024; Lieder & Schäffer, 2024; Magesh et al., 2024; Nguyen et al., 2024; Parker et al., 2025; Pretorius & Pretorius, 2025; Tran et al., 2023; Ye et al., 2025) . In the RAIL model, it refers to the stable tendency to actively monitor and mitigate AI's influence on researchers' cognitive autonomy, critical agency, and intellectual development and habits. It ensures that engagement with AI tools does not erode foundational human capacities but instead strengthens self-awareness through structured reflection. It is not passive awareness, but it fosters an active self-monitoring practice that surfaces risks such as overreliance, declining critical-thinking abilities, and a shift in locus of control from the user to the algorithm. This disposition relates to the following key aspects:

- **Proactive countermeasures for over-reliance**: Researchers must implement deliberate checks (e.g., theory-driven prompts, manual verification) to prevent dependency on AI, preserving critical thinking and analytical skills, avoiding the pitfalls of "human automation bias" or declining agency noted in the summary.
- **Agency-centered design of interactions**: By framing AI use as a reflective exercise (not an automatic solution), researchers maintain control over their intellectual processes, for example, explicitly questioning AI outputs to ensure they align with personal analytical priorities and ethical frameworks.
- **Skill preservation through intentional practice**: Integrating self-reflective habits (e.g., auditing prompt engineering workflows) safeguards competencies like creativity and problem-solving while adapting to new demands (e.g., Responsible AI literacy), ensuring AI integration *enhances* rather than replaces human intellectual growth.
- **Ethical accountability in interpretation**: Researchers must cultivate rigorous meta-cognition, regularly assessing the credibility, biases, and contextual relevance of AI-generated insights, to prevent superficial engagement or obsolescence of traditional academic skills (e.g., manual literature review).

Critically, this disposition rejects *passive acceptance* of AI's influence. Instead, it requires researchers to treat every AI interaction as an opportunity for self-reflection: The integration of AI must be a facet of each scientific discipline, a practice in which the researcher actively questions their role, biases, and evolving agency within the system.

For instance, constructing prompts that demand explicit reflection on paradigmatic assumptions embodies this disposition. It transforms potential risks into ethical guardrails: by prioritizing self-awareness over convenience, researchers uphold intellectual integrity while responsibly leveraging AI's benefits.

In essence, responsible integrity in Self-Reflectiveness means treating AI as a mirror for personal development, not an extension of one's autonomy, ensuring that every interaction reinforces critical independence rather than diminishing it.

Considering Sustainability

Considering sustainability is supported by 19 documents (Butson & Spronken-Smith, 2024; Castillo-Martínez et al., 2024; Garg et al., 2024; Glerean & Silva, 2024; Helgesson, 2024; Hill et al., 2024; Ilegbusi, 2024; Khalifa & Albadawy, 2024; Kuiper, 2024; Magesh et al., 2024; Murphy & James, 2024; Picalho et al., 2025; Pretorius & Pretorius, 2025; Rietz & Maedche, 2021; Silva et al., 2024; Skyba et al., 2024; Yaroshenko & Iaroshenko, 2023; Yining et al., 2025; Yongtao et al., 2024). The responsible integrity disposition of Considering Sustainability refers to the ethical and proactive framework researchers must apply when selecting AI tools, platforms, and partnerships to ensure their choices actively advance environmental stewardship, social equity, and long-term human benefit, without compromising research rigor or personal agency. It mandates that sustainability is not an afterthought but a foundational criterion in all AI-related decisions. This disposition entails the following key dimensions:

- **Environmental accountability**: Prioritizing tools with transparent carbon footprints (e.g., avoiding high-energy datacenter dependencies) and actively seeking solutions like partnerships with companies investing in nuclear fusion for energy efficiency, ensuring AI adoption *reduces* ecological harm rather than exacerbating it.
- **Ethical due diligence**: Critically evaluating an AI vendor's practices to prevent exploitation, such as verifying that AI tools do not enable unauthorized data collection or commercialization of personal research assets, a safeguard against privacy violations and misuse.
- **Human-centric gap identification**: Using AI to uncover research opportunities that directly benefit humanity (e.g., studying long-term health impacts of insulin analogs across age groups), while ensuring these gaps align with sustainable development goals, not just technical convenience.
- **Integrative alignment**: Selecting tools and platforms that seamlessly blend into the researcher's workflow without compromising ethical standards. This includes prioritizing lightweight technologies that minimize data overhead and energy consumption.

Considering sustainability demands that researchers treat it as a dynamic practice: Every AI tool choice must be evaluated through the lens of its dual impact on the planet and people, requiring active mitigation of harm (e.g., data exploitation) while creating opportunities for equitable, scalable solutions.

For example, when researching the effects of nootropic drugs using AI, a responsible approach would prioritize tools that minimize computational waste *and* explicitly disclose how findings could inform sustainable health policies. This ensures sustainability becomes an operational ethic, not just a checklist item, ultimately strengthening research's capacity to serve humanity responsibly.

In essence, responsible integrity in Considering Sustainability means embedding environmental and ethical foresight into every decision about AI adoption, transforming potential trade-offs into opportunities for collaborative resilience.

Open Science/Open-Source

Open Science/Open-Source is supported by 13 documents (Alfirević et al., 2024; Burger et al., 2023; Charton et al., 2024; Chen et al., 2024; Chopra & Haaland, 2023; Fui-Hoon Nah et al., 2023; Glerean & Silva, 2024; Gokul et al., 2024; Ilegbusi, 2024; Khalifa & Albadawy, 2024; Limongi, 2024; Mohamed et al., 2025; Yongtao

et al., 2024). It treats scientific research not as a closed, proprietary process but as an open, collaborative endeavor in which data, code, methods, and model architecture are intentionally made accessible to ensure transparency, reproducibility, and collective progress.
This isn't merely about "sharing files". It's a foundational principle of trust: when researchers publish their full workflows on platforms like GitHub (for version-controlled repositories), OSF (for project management), or Chisquares.com/DataCite (for persistent identifiers), they enable others to verify results, reproduce findings independently, and build upon existing work without gatekeeping. This directly combats epistemic silos and strengthens scientific discourse. Key dimensions of this disposition include:

- **Technical Freedom Through Open Platforms:** Using open-source or open-weights AI models allows researchers to run tools locally, giving them absolute control over their data, eliminating risks of unauthorized access, surveillance, or commercial exploitation.
- **Ethical Accountability via Auditing:** Open science enables systematic model auditing, where fairness, accuracy, transparency, and bias can be evaluated by external parties using public codebases. Without openness, ethical issues like data leakage or hidden assumptions remain invisible to users and regulators alike.
- **Community-Centric Development:** For instance, tools such as PyGeoweaver and AEcroscopy embody open-source philosophies embedded with detailed metadata. This ensures reproducibility isn't optional; it's built into the design from day one.
- **Increased User Awareness & Policy Alignment:** As companies deploy generative AI, Open Science/Open-Source calls for stronger user education about privacy threats (e.g., data harvesting) and advocates for regulatory frameworks that protect personal information and prioritize public benefit over profit motives.

Ultimately, embracing Open Science/Open-Source helps distribute responsibility for scientific integrity across both individual researchers and the broader research community. It asserts that knowledge should not be hoarded but shared freely so innovation accelerates without sacrificing ethics or autonomy. In this light, releasing code isn't just best practice; it's an act of democratic integrity in science itself.

Skills

Interaction Engineering

Interaction Engineering is the ability to steer and optimize the collaborative dynamics between human users and machines (e.g., Large Language Models) by applying targeted techniques across multiple levels of software architecture. Unlike commonly known techniques confined to the user interface, Interaction Engineering operates vertically:

- **it encompasses interface-level methods such as prompt and context engineering** (Barrot, 2025; Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Cotton et al., 2024; Dönmez et al., 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Gupta, 2024; Hill et al., 2024; Karakose, 2023; Kuiper, 2024; Lieder & Schäffer, 2024; Magesh et al., 2024; Mudd et al., 2024; Nguyen et al., 2024; Nicholson Thomas et al., 2024; Parker et al., 2025; Picalho et al., 2025; Pretorius & Pretorius, 2025; Shukla et al., 2024; Silva et al., 2024; Tran et al., 2023; Yongtao et al., 2024);
- **model-level interventions including fine-tuning and model training** (Charton et al., 2024; Chopra & Haaland, 2023; Glerean & Silva, 2024; Gokul et al., 2024; Gupta, 2024; Mudd et al., 2024; Nicholson Thomas et al., 2024; Yining et al., 2025; Yongtao et al., 2024);
- and **infrastructure-level design based on the underlying AI architecture** (e.g., transformer design) that can be manipulated through programming (Charton et al., 2024; Chopra & Haaland,

2023; Glerean & Silva, 2024; N. Khan et al., 2024; Kuiper, 2024; Magesh et al., 2024; Mudd et al., 2024; Rietz & Maedche, 2021; Shukla et al., 2024; Yining et al., 2025; Yongtao et al., 2024).

By integrating these layers, Interaction Engineering holistically shapes the capabilities, alignment, and outcomes of human-machine interactions.

Designing Responsible/Ethical Workflow

The skill Designing Responsible/Ethical Workflow is supported by 26 documents (Barrot, 2025; Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Chopra & Haaland, 2023; Dönmez et al., 2023; Freiesleben, 2023; Garg et al., 2024; Glerean & Silva, 2024; Gokul et al., 2024; Helgesson, 2024; Karakose, 2023; Kuiper, 2024; Lieder & Schäffer, 2024; Limongi, 2024; Magesh et al., 2024; Mohamed et al., 2025; Mudd et al., 2024; Nicholson Thomas et al., 2024; Perkins & Roe, 2024a, 2024b; Picalho et al., 2025; Rietz & Maedche, 2021; Shukla et al., 2024; Silva et al., 2024; Yongtao et al., 2024). It refers to the deliberate and systematic integration of artificial intelligence into research processes while maintaining human agency, transparency, accountability, and ethical integrity. It involves constructing workflows considering the human-in-the-loop principle that treat AI not as a replacement for judgment but as an augmentation tool, guided by clear principles such as:

- **Transparency:** Ensuring all steps involving AI are documented so others can understand how decisions were made.
- **Bias Mitigation**: Proactively identifying and addressing potential algorithmic biases in both training data and model outputs. (e.g., Bryda and Costa, 2024)
- **Human Oversight**: Maintaining final ownership and epistemic stewardship of analysis, especially critical for qualitative insights, interpretation, or decisions affecting participants, and ensuring human validation before conclusions are drawn.
- **Privacy & Consent:** Explicitly obtaining informed participant consent when using AI to analyze sensitive information; respecting data sovereignty through secure local or research institutional execution where possible.
- **Ethical Review Standards**: Establishing protocols that require authors to disclose their use of AI tools in publications and subject manuscript reviews to uphold academic integrity.

Designing responsible/ethical workflows is iterative: researchers evaluate different options for integration of AI into the research design, test prompts, refine inclusion/exclusion criteria, and then manually validate findings. It also includes defining evaluation standards, such as setting benchmarks for acceptable performance, to ensure quality control throughout the research lifecycle.

Ultimately, this skill ensures that, while leveraging emerging AI capabilities such as AI agents (e.g., Hermes Agent or OpenClaw), academia advances knowledge sustainably without sacrificing trustworthiness or human values in favor of technological convenience alone.

Balancing and Controlling AI's Potential

The skill of Balancing and Controlling AI's Potential is supported by 18 documents (Alfirević et al., 2024; Barrot, 2025; Bhutta, 2024; Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Castillo-Martínez et al., 2024; Chopra & Haaland, 2023; Glerean & Silva, 2024; Helgesson, 2024; Khalifa & Albadawy, 2024; Limongi, 2024; Perkins & Roe, 2024a; Pretorius & Pretorius, 2025; Tran et al., 2023; Yining et al., 2025; Yongtao et al., 2024). It refers to the deliberate, context-sensitive decision-making process of determining when, and only when, AI should be used in research. It involves assessing whether

an AI tool offers a meaningful benefit over traditional methods while rigorously weighing its risks across multiple dimensions. Key aspects include:

- **Usefulness vs. Overreach:** Recognizing that AI is most valuable for automating repetitive or data-intensive tasks (e.g., literature screening, summarization), but must not replace core intellectual work, such as critical thinking, hypothesis generation, and interpretation, activities central to scientific creativity.
- **Ethical Boundaries:** Acknowledging risks like algorithmic bias in outputs, the potential for plagiarism through uncredited content generation, or erosion of cognitive skills due to overreliance. These concerns demand active mitigation strategies.
- **Transparency & Agency:** Ensuring AI is used as a support mechanism within human-led workflows rather than an autonomous decision-maker that diminishes researcher accountability and ownership.
- **Energy Efficiency vs. Research Trade-offs:** Considering environmental costs associated with large models' compute demands, particularly when running them remotely, and balancing these against research gains to ensure sustainability.

Ultimately, this skill requires researchers to maintain vigilant critical engagement throughout implementation: asking not just*, "Can I use AI?" but also, "Should I, and how will it affect my work's integrity and originality?*" It promotes a dynamic equilibrium between innovation and academic rigor, where technology serves humanity without supplanting it.

Report transparently

The skill Report transparently is supported by 18 documents (Bryda & Costa, 2024; Burger et al., 2023; Chopra & Haaland, 2023; Fui-Hoon Nah et al., 2023; Glerean & Silva, 2024; Helgesson, 2024; Hill et al., 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. A. Khan et al., 2023; Limongi, 2024; Mudd et al., 2024; Nicholson Thomas et al., 2024; Perkins & Roe, 2024a; Pretorius & Pretorius, 2025; Salvioni & Almici, 2024; Silva et al., 2024; Yongtao et al., 2024). It refers to the mandatory, ethically grounded practice of clearly documenting and disclosing how artificial intelligence was used throughout the research process, both during execution and in final reporting.
This includes providing detailed information such as:

- The specific AI tool(s) employed (e.g., GPT-4, NVivo),
- Its version number and configuration parameters,
- Exact prompts or queries issued to the model,
- Dates of usage and training data sources involved,
- How outputs were validated against human judgment,

The goal is not merely technical completeness but also trustworthiness, ensuring that other researchers can reproduce findings, evaluate their validity, and understand any limitations introduced by automated processes. Moreover, transparency extends beyond metadata: researchers must actively address discrepancies between human interpretations and AI outputs (Discrepancy Resolution) and validate results through critical engagement to maintain scientific rigor. This makes reporting more than a formality; it becomes an essential pillar of responsible scholarship in the age of intelligent agents.

It is important to note that AI cannot be granted authorship. Even if it generates text or structures arguments, the final responsibility for content accuracy, originality, and ethical compliance rests entirely with human authors; a principle upheld by major journals and research bodies worldwide.

Verifying AI Output

The skill of Verifying AI Output is supported by 28 documents (Alfirević et al., 2024; Barrot, 2025; Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Charton et al., 2024; Cotton et al., 2024; Dönmez et al., 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Helgesson, 2024; Hill et al., 2024; Karakose, 2023; Khalifa & Albadawy, 2024; Kuiper, 2024; Lieder & Schäffer, 2024; Magesh et al., 2024; Nicholson Thomas et al., 2024; Parker et al., 2025; Perkins & Roe, 2024b; Picalho et al., 2025; Pretorius & Pretorius, 2025; Silva et al., 2024; Tran et al., 2023; Wise et al., 2024; Yining et al., 2025; Yongtao et al., 2024). It refers to the critical, active process of verifying all content generated by artificial intelligence, whether during research execution or final reporting, to prevent the spread of misinformation and ensure methodological accuracy. It requires researchers to move beyond passive adoption. Instead, they must:

- **Actively cross-check outputs:** Compare AI-generated summaries, citations, references, or analysis with primary sources (e.g., peer-reviewed papers, databases) to confirm factual consistency.
- **Maintain human oversight:** Never accept AI content at face value. The researcher remains the ultimate decision-maker, interpreting results and validating the logic, ensuring alignment with research objectives and scholarly standards.
- **Iterate for reliability:** Refine prompts or re-run checks until outputs are precise, consistent, and contextually sound, recognizing that early versions often contain hallucinations or incomplete information.
- **Build analytical rigor:** Strengthen the researchers' ability to detect errors in machine-generated text through continuous practice, turning AI use into a training ground for deeper critical thinking about data provenance and source credibility.

Please note that if a deterministic algorithm exists to verify AI-generated output, it is recommended to use the deterministic alternative, as it is usually more energy-efficient and yields more consistent results.

Additionally, this skill extends beyond the individual researcher: editors and peer reviewers must manually inspect AI-assisted content before publication to uphold academic integrity. In short, verification is not optional; it's an essential component of trustworthiness in AI-augmented research. Without it, even advanced models risk propagating false knowledge into scientific discourse.

Integrate AI into Continuous Learning

The skill Integrate AI into Continuous Learning is supported by 13 documents (Barrot, 2025; Burger et al., 2023; Butson & Spronken-Smith, 2024; Castillo-Martínez et al., 2024; Karakose, 2023; N. A. Khan et al., 2023; Kuiper, 2024; Lieder & Schäffer, 2024; Nguyen et al., 2024; Parker et al., 2025; Perkins & Roe, 2024b; Pretorius & Pretorius, 2025; Ye et al., 2025). It refers to the practice of treating artificial intelligence not as a static tool or black box, but as an active, adaptive partner in one's intellectual development, transforming passive information consumption into dynamic, iterative learning. Rather than merely receiving outputs from platforms like SciSpace or Gemini, users actively engage with AI by:

- Asking targeted questions to explore concepts deeply (e.g., using "deep research" features),
- Receiving context-aware feedback tailored to their specific knowledge gaps and reasoning style, acting as a personalized tutor that adapts in real time.

This interaction could offer opportunities to build critical cognitive abilities: thinking about how to restructure their arguments, refining prompts for clarity, detecting biases in both human and machine logic, and assessing output quality through continuous reflection.

Ultimately, integrating AI into continuous learning enables researchers to be proactive architects of knowledge, using technology as a mirror that reflects their understanding back so they can build their professional capacities by incorporating AI.

Dialogical Co-Construction with AI

The skill Dialogical Co-Construction with AI is supported by 12 documents (Butson & Spronken-Smith, 2024; Hill et al., 2024; Khalifa & Albadawy, 2024; Lieder & Schäffer, 2024; Mudd et al., 2024; Nguyen et al., 2024; Nicholson Thomas et al., 2024; Parker et al., 2025; Perkins & Roe, 2024b; Pretorius & Pretorius, 2025; Silva et al., 2024; Tran et al., 2023). It refers to a dynamic, two-way collaborative process between humans and artificial intelligence, in which interaction transcends simple prompting or tool use and becomes an active co-creation of meaning, knowledge, and research outcomes.
Unlike static prompt engineering, which focuses on input syntax, the dialogical approach emphasizes real-time exchange: users don't just issue commands; they engage in iterative back-and-forth conversations with AI models. This involves:

- Offering high-level directives while adjusting for context,
- Receiving suggestions that evolve into new ideas or interpretations,
- Critically reflecting on outputs and iteratively refining inputs (e.g., "PromptFollowUp," "EditPrompt") until coherence and accuracy are achieved.

AI functions not as a passive executor but as an *interactive sounding board***,** helping researchers brainstorm, structure arguments, draft narratives, generate visuals like infographics, or translate complex concepts into accessible language. Crucially, it acts both as an assistant *and* a provocateur: offering alternative perspectives that challenge assumptions and stimulate deeper analysis.
The process is inherently recursive: each output triggers feedback loops in which users reconstruct their thinking based on AI-generated insights, then return to modify prompts with clearer examples or constraints. Moreover, this collaboration isn't limited to technical experts; even non-technical users can participate meaningfully through guided interaction patterns (e.g., method prompts that guide research logic).

From a hybrid intelligence perspective, this collaboration distributes cognitive load between humans and machines while maintaining humans' epistemic authority (Dellermann et al., 2019). However, further investigation is needed to see whether this approach increases or reduces biases.

Using Open-Source

The skill Using Open-Source is supported by 6 documents (Charton et al., 2024; Chopra & Haaland, 2023; Glerean & Silva, 2024; Gokul et al., 2024; Limongi, 2024; Mohamed et al., 2025). It refers to the practice of leveraging freely available, community-maintained artificial intelligence models and software tools, along with their associated platforms (such as GitHub), to conduct research with full transparency, control, and autonomy. This includes:

- **Accessing source code:** Navigating repositories on platforms like GitHub to understand how models are built, trained, and optimized.
- **Running locally:** Deploying open-source AI directly on personal devices rather than relying on cloud-based APIs, ensuring data privacy, avoiding corporate lock-in, and maintaining ownership of intellectual property (especially critical when handling sensitive or proprietary information).

- **Customization & adaptation:** Modifying pre-trained weights or architecture to fit specific research tasks such as named entity recognition, text classification, machine translation, or transcription, while preserving the ability to audit every step of development.

Tools such as PyGeoweaver, which combine open-source AI with intuitive graphical interfaces and multi-platform compatibility, exemplify how open science principles can be integrated into everyday workflows without requiring advanced technical expertise.

By embracing open-source ecosystems, researchers uphold the values of transparency, reproducibility, and collaborative innovation, allowing anyone to independently verify results, build upon existing work collectively, and avoid vendor dependencies that restrict freedom.

Ultimately, "Using Open-Source" is not just about technical navigation; it's a foundational commitment to democratizing access to research resources and ensuring agency over one's knowledge-generation process.

Disciplinary Skills

Disciplinary Skills are supported by 26 documents (Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Castillo-Martínez et al., 2024; Charton et al., 2024; Cotton et al., 2024; Dönmez et al., 2023; Freiesleben, 2023; Glerean & Silva, 2024; Gupta, 2024; Helgesson, 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. Khan et al., 2024; Lieder & Schäffer, 2024; Magesh et al., 2024; Nicholson Thomas et al., 2024; Parker et al., 2025; Perkins & Roe, 2024a, 2024b; Rietz & Maedche, 2021; Silva et al., 2024; Tran et al., 2023; Wise et al., 2024; Ye et al., 2025; Yining et al., 2025; Yongtao et al., 2024). They refer to the specialized, domain-specific abilities and professional judgment required of researchers to effectively apply artificial intelligence within their fields, particularly in tasks such as qualitative content analysis (e.g., coding), interpreting complex data, making ethical decisions, and critically evaluating AI outputs. Unlike generic technical proficiency, disciplinary skills encompass deep expertise rooted in subject matter, including:

- The ability to recognize subtle nuances in human language or behavioral patterns that AI models often miss, such as context-dependent meaning shifts or cultural bias.
- Expert-level understanding of research methodologies so researchers can verify whether AI-generated summaries, interpretations, or classifications reflect real-world validity and not hallucinated constructs.
- **Awareness of field-specific limitations**: e.g., knowing when an LLM might misrepresent psychiatric diagnostics despite strong performance metrics (as seen in human clinical raters being comparable to models).

These skills are essential for ensuring that AI remains a supportive tool, not a substitute for human insight. Researchers must use their disciplinary expertise to:

- Validate AI outputs against primary sources and theoretical frameworks,
- Provide iterative feedback loops during co-construction processes (e.g., adjusting prompts based on domain logic),
- Ensure all content aligns with academic integrity standards, avoiding plagiarism or data fabrication through careful oversight.

Ultimately, Disciplinary Skills represent the core intellectual authority that researchers bring to hybrid human-AI workflows: they enable context-sensitive reasoning, maintain scientific rigor across tools, and ensure that interpretations remain grounded in lived experience and scholarly tradition rather than just in

algorithmic prediction. Without them, even the most advanced AI systems risk producing misleading or ethically compromised work.

Selecting/Choosing adequate Tools/Models

The skill Selecting/Adequate Tools/Models is supported by 32 documents (Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Chen et al., 2024; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Gokul et al., 2024; Gupta, 2024; Helgesson, 2024; Ilegbusi, 2024; Khalifa & Albadawy, 2024; Kuiper, 2024; Limongi, 2024; Magesh et al., 2024; Murphy & James, 2024; Nicholson Thomas et al., 2024; Perkins & Roe, 2024a, 2024b; Picalho et al., 2025; Pretorius & Pretorius, 2025; Rietz & Maedche, 2021; Salvioni & Almici, 2024; Silva et al., 2024; Skyba et al., 2024; Tran et al., 2023; Yaroshenko & Iaroshenko, 2023; Ye et al., 2025; Yining et al., 2025; Yongtao et al., 2024). It refers to the deliberate, context-sensitive process of identifying and selecting specific artificial intelligence tools or models, such as large language models (LLMs), domain-specific platforms (e.g., Atlas.ti for coding), or specialized assistants (e.g., DALL-E 2 for visuals), that are best aligned with a given research task at hand. This involves more than generic selection; it requires critical evaluation based on functional fit, reliability, and ethical alignment:

- **Matching tools to precise needs:** For example, using Atlas.ti or NVivo for qualitative coding, where AI can assist in automatic categorization but still operates under human supervision.
- **Assessing performance trade-offs:** Recognizing that general LLMs like ChatGPT may lack up-to-date knowledge and show higher bias compared to alternatives such as Bard (Gemini), which has demonstrated stronger scholarly database accuracy in some benchmarks, though all require scrutiny.
- **Prioritizing transparency and control:** Favoring open-source models when handling sensitive data or requiring local execution, ensuring ownership of training data, and avoiding vendor lock-in.

Researchers apply structured frameworks, such as the ROBOT test (Reliability, Objective, Bias, Ownership, Type), as referenced by Murphy and James (2024), to evaluate tools across key dimensions, ensuring they meet criteria for objectivity, bias mitigation, intellectual property rights, and explainability.

Ultimately, Selecting Adequate Tools/Models is not a one-size-fits-all decision; it's an iterative strategy rooted in understanding each step of the research lifecycle, sensitive to each discipline's uniqueness, so that technology supports rather than undermines methodological integrity and intellectual responsibility.

Knowledge

Please note that each knowledge node is based on the document analysis and may contain outdated information, especially when referencing LLM versions such as GPT-4.

Human-Machine Relationship

The structured, declarative understanding that AI is not a passive instrument but an active collaborator, engineered through real-time interaction with humans, that participates in co-creation processes by providing iterative feedback, novel perspectives, and cognitive augmentation, while remaining subordinate to human judgment in matters of epistemic authority. This relationship is governed by principles of user agency (active prompt engineering), mutual refinement (prompt iteration based on AI output), transparency (checking sources and clarifying context), and critical verification (ensuring all AI-derived insights are contextualized before final adoption). Knowledge of the human-machine relationship is supported by 7 documents (Butson

& Spronken-Smith, 2024; Karakose, 2023; Kuiper, 2024; Lieder & Schäffer, 2024; Parker et al., 2025; Perkins & Roe, 2024b; Pretorius & Pretorius, 2025).

Data Handling/Structures

The declarative understanding that data must be systematically managed throughout its lifecycle, from raw collection to final use, using structured formats (e.g., flattening multi-dimensional inputs into 1D tensors), ensuring representativeness, completeness, and diversity in training sets through cleansing or synthetic generation, while actively identifying and mitigating bias originating from incomplete populations, monolingual sources, or factual inaccuracies. This process is supported by defined workflows that involve orchestration tools (e.g., FFmpeg for video, NumPy for analysis), prioritize data privacy through informed consent and re-identification risk awareness, and enable robust AI deployment through techniques such as Retrieval-Augmented Generation (RAG), which bind external context to model outputs. Knowledge of data structures and handling is supported by 14 documents (Charton et al., 2024; Chen et al., 2024; Chopra & Haaland, 2023; Fui-Hoon Nah et al., 2023; Glerean & Silva, 2024; Gokul et al., 2024; Gupta, 2024; Helgesson, 2024; Ilegbusi, 2024; N. A. Khan et al., 2023; Kuiper, 2024; Lieder & Schäffer, 2024; Mohamed et al., 2025; Picalho et al., 2025).

Prompt Structures

The structured, declarative understanding that high-quality human-AI interaction is governed by explicit prompt- and context-engineering principles, in which prompts must be clear, concise, logical, adaptive, and reflective to guide AI behavior effectively. This involves using modular architectures (e.g., breaking down complex tasks into sub-prompts), leveraging advanced techniques such as zero-shot prompting. The prompt structure can incorporate guidance on how the model should process the information it has just retrieved, using Retrieval-Augmented Generation (RAG) for contextual grounding, and incorporate iterative refinement through continuous negotiation of meaning and prompt revision based on prior outputs. Specific prompt types, such as research project, empirical, object theory, basic theory, or method prompts, are designed to align with domain-specific cognitive goals, and best practices include providing context upfront, enabling chain-of-thought reasoning, requesting AI-generated rationales, and ensuring that the final output is directly tied to user-defined constraints and epistemic expectations. Knowledge of prompt structures is supported by 21 documents (Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Chopra & Haaland, 2023; Cotton et al., 2024; Dönmez et al., 2023; Freiesleben, 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Hill et al., 2024; Karakose, 2023; Kuiper, 2024; Lieder & Schäffer, 2024; Magesh et al., 2024; Mudd et al., 2024; Nguyen et al., 2024; Nicholson Thomas et al., 2024; Parker et al., 2025; Pretorius & Pretorius, 2025; Shukla et al., 2024; Tran et al., 2023).

Software Architecture

The declarative understanding that AI systems are implemented as layered, modular software architectures, comprising front-end interfaces (e.g., user-interfaces from chatbots like ChatGPT or Gemini), back-end applications (e.g., PyGeoweaver with host/process/database/task modules), and specialized components such as RAG pipelines that allow models to access external data via databases for context-aware responses. These systems can be hosted centrally (e.g., OpenAI APIs) or locally (e.g., self-hosted LLMs powered by Ollama), each choice affecting control over data, security, and transparency. Core technical foundations include transformer-based neural networks (decoder-only architectures for modern LLMs), foundational models such as GPT-4 and Gemini, fine-tuned variants (e.g., CoAICoder, PaTAT), and open-source frameworks like PyTorch/TensorFlow or spaCy/NLP libraries that enable code generation, entity recognition, and

semantic similarity scoring. Integration is achieved through APIs in Python, while interpretability methods, both model-agnostic (XAI) and domain-specific, are applied to ensure explainable outputs from otherwise opaque black-box models. Architectural decisions are guided by functional requirements: whether tasks require generative output (ChatGPT), structured analysis (Cody's rule matching with Levenshtein distance), or local autonomy (self-hosted systems). Knowledge of software architecture is supported by 24 documents (Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Charton et al., 2024; Chopra & Haaland, 2023; Dönmez et al., 2023; Freiesleben, 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Gokul et al., 2024; Gupta, 2024; Hill et al., 2024; Kuiper, 2024; Magesh et al., 2024; Mudd et al., 2024; Nicholson Thomas et al., 2024; Perkins & Roe, 2024b; Rietz & Maedche, 2021; Shukla et al., 2024; Silva et al., 2024; Skyba et al., 2024; Ye et al., 2025; Yongtao et al., 2024).

Research Procedures/Methods

The declarative understanding that qualitative research remains grounded in traditional methodologies, such as the documentary method of interpretation involving sequential, step-by-step analysis for thematic structures and participant orientations, but is now augmented by AI-driven enhancements that automate preliminary tasks like literature gap detection, hypothesis generation, coding suggestions, and visualization. This integration operates through hybrid epistemologies, combining deductive frameworks (e.g., predefined codes) with inductive processes (e.g., emergent themes from data), with AI serving as a parallel co-analyst to accelerate iterative cycles of refinement rather than replace human judgment.
Research procedures are now structured around recursive loops involving (1) researcher input, (2) AI-generated output, (3) critical evaluation and prompt adjustment, as well as (4) refined interpretation, postulated to ensure both replicability via documented codes/books/guidelines and ethical rigor through ongoing expert oversight.
Tools such as CAQDAS platforms leverage visualization to transform complex qualitative data into interpretable patterns, while LLMs assist in summarizing guidelines or mapping patient-to-trial matches, functions that reduce cognitive load but do not eliminate the necessity of researcher agency for final validation. The knowledge of research procedures and methods is supported by 19 documents (Alfirević et al., 2024; Bryda & Costa, 2024; Burgui-Burgui, 2023; Glerean & Silva, 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. Khan et al., 2024; Lieder & Schäffer, 2024; Magesh et al., 2024; Mudd et al., 2024; Nguyen et al., 2024; Nicholson Thomas et al., 2024; Parker et al., 2025; Perkins & Roe, 2024a, 2024b; Pretorius & Pretorius, 2025; Rietz & Maedche, 2021; Yining et al., 2025; Yongtao et al., 2024).

Research Components/Resources

The declarative understanding that every decision regarding AI integration into research, such as the selection of a specific model (e.g., GPT-4 vs Med-PaLM 2), algorithmic approach (e.g., RAG, fine-tuning), prompt design strategy, database source, or versioned parameter configuration, is not arbitrary but must be explicitly documented and justified for reproducibility, transparency, and ethical accountability. This includes recording the exact date, time, input strings, model version numbers (e.g., GPT-4o, Gemini Ultra), system parameters ("temperature," "max tokens"), and workflow history in a structured format that traces how AI outputs evolved over iterations.
The rationale behind each choice, such as selecting Med-PaLM 2 for psychiatric assessments due to its domain expertise or using spaCy-based similarity scoring (Levenshtein distance) in Cody to align rule suggestions with text, is formally articulated to ensure trustworthiness and prevent misattribution of insights.
These components are interdependent: prompt engineering shapes model behavior, which depends on underlying algorithms such as transformer architectures, all of which are executed within a defined computational environment accessible via APIs or local runs (e.g., GitHub Copilot). Resources such as public databases, open-source tools (e.g., PyGeoweaver), and workflow management systems (WfMS) are selected for

their ability to support transparent data provenance while enabling human-in-the-loop oversight. Ultimately, this documentation serves both scientific integrity, enabling peer review, and compliance with journal standards requiring disclosure without crediting AI as a co-author. The knowledge of the research components and resources used in one's own research process is supported by 18 documents (Barrot, 2025; Bhutta, 2024; Bryda & Costa, 2024; Burger et al., 2023; Chopra & Haaland, 2023; Glerean & Silva, 2024; Gokul et al., 2024; Karakose, 2023; Kuiper, 2024; Lieder & Schäffer, 2024; Limongi, 2024; Nicholson Thomas et al., 2024; Perkins & Roe, 2024a; Rietz & Maedche, 2021; Shukla et al., 2024; Silva et al., 2024; Yining et al., 2025; Yongtao et al., 2024).

Research Standards and Quality Criteria

The declarative understanding that the integration of AI into qualitative research must operate under strict epistemic standards, specifically validity (accuracy in representing reality), reliability (consistency across iterations), and transparency, to ensure scientific integrity.
While AI tools can enhance consistency through standardized outputs, their outputs are only as trustworthy as the underlying algorithms and training data; thus, researchers remain ultimately responsible for validating all generated content against empirical evidence and human judgment. Quality criteria explicitly include identifying and mitigating algorithmic bias, particularly stemming from monolingual training sets or skewed population representations, and actively preventing ethical risks such as data fabrication, manipulation, or privacy violations through informed consent protocols and secure API use.
These standards demand that researchers treat AI not as a source of truth but as a cognitive extension whose outputs require rigorous scrutiny for coherence, logical consistency, and contextual plausibility. Furthermore, the absence of epistemic authority in AI (i.e., no inherent "truth claim") necessitates human verification at every stage to uphold academic responsibility: final interpretations must be grounded in researcher expertise, critical analysis, and adherence to ethical guidelines such as those requiring full disclosure without attributing authorship to models. The knowledge of research standards and quality criteria is supported by 27 documents (Bryda & Costa, 2024; Burgui-Burgui, 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Gokul et al., 2024; Helgesson, 2024; Hill et al., 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. A. Khan et al., 2023; Lieder & Schäffer, 2024; Limongi, 2024; Mohamed et al., 2025; Mudd et al., 2024; Murphy & James, 2024; Nguyen et al., 2024; Perkins & Roe, 2024a, 2024b; Pretorius & Pretorius, 2025; Salvioni & Almici, 2024; Shukla et al., 2024; Silva et al., 2024; Tran et al., 2023; Wise et al., 2024; Yining et al., 2025).

AI Limitations

The declarative understanding that current large language models (LLMs) operate with fundamental, well-documented constraints, most notably hallucination (fabricating non-existent citations, authors, or facts), limited reasoning capacity (inability to draw original inferences beyond pattern matching), and incomplete or biased knowledge bases derived from static training data. These limitations make AI outputs inherently non-reliable without human verification: even advanced models may produce plausible-sounding but factually incorrect statements, especially in high-stakes domains like law or medicine, and struggle with contextual nuance such as sarcasm, irony, or abstract humor.
Output quality is critically dependent on prompt design; poorly structured inputs result in nonsensical responses, while well-crafted prompts can only mitigate, not eliminate, the risk of error. Furthermore, AI detectors (e.g., GPTZero) are known to produce false positives and false negatives due to obfuscation techniques such as paraphrasing or stylistic mimicry, undermining their utility as standalone validation tools.
The absence of true epistemic authority means no model can be trusted to assert truth without external grounding or researcher correction. Over-reliance on these systems risks cognitive degradation: automation bias reduces human judgment, while diminished critical thinking undermines intellectual autonomy. The

knowledge of AI limitations is supported by 20 documents (Barrot, 2025; Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Castillo-Martínez et al., 2024; Freiesleben, 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Gupta, 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. A. Khan et al., 2023; Kuiper, 2024; Magesh et al., 2024; Perkins & Roe, 2024b; Picalho et al., 2025; Tran et al., 2023; Wise et al., 2024; Yongtao et al., 2024).

AI Tool Landscape

The declarative, structured understanding that the current ecosystem of AI tools spans diverse functional categories, ranging from conversational chatbots (ChatGPT, Gemini/Bard) and document intelligence platforms (Scite.ai, Dimensions) to specialized research assistants (Cody, Atlas.ti AI), coding aides (GitHub Copilot), literature review automation (Rayyan, Elicit), plagiarism detection (GPTZero), citation management (Zotero-integrated SciSpace), and multimodal generation (DALL-E 2), with distinct applications tied directly to domain needs such as qualitative analysis, academic writing, data extraction, or legal research.
This landscape includes both proprietary offerings (e.g., OpenAI, LexisNexis) and open-source alternatives such as Stable Diffusion, GPT-NeoX, and PyGeoweaver, though adoption varies with accessibility and integration depth. Key differentiators include model capabilities: Gemini's advanced search via web access versus OpenAI's closed architecture; prompt fidelity in AI responses (e.g., SciSpace's low hallucination rates); and user agency, some tools (like PaTAT or Scholastic) allow direct modification of outputs, enabling human-in-the-loop control.
Critical benchmarks such as the ROBOT test (Reliability, Objective, Bias, Ownership, Type) provide a standardized framework for evaluating tool quality beyond marketing claims. Tools are not monolithic; their utility depends on integration with workflows such as CAQDAS software or Otter.ai for audio transcription and AEcroscopy for microscope automation, demonstrating how AI fits into specific research processes rather than replacing them. The knowledge of the AI tool landscape is supported by 30 documents (Barrot, 2025; Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Chen et al., 2024; Chopra & Haaland, 2023; Dönmez et al., 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Gokul et al., 2024; Gupta, 2024; Helgesson, 2024; Hill et al., 2024; Ilegbusi, 2024; Khalifa & Albadawy, 2024; N. Khan et al., 2024; Magesh et al., 2024; Murphy & James, 2024; Perkins & Roe, 2024a; Picalho et al., 2025; Pretorius & Pretorius, 2025; Salvioni & Almici, 2024; Silva et al., 2024; Skyba et al., 2024; Yaroshenko & Iaroshenko, 2023; Ye et al., 2025; Yining et al., 2025; Yongtao et al., 2024).

It should be noted that these are only examples coded during the document analysis. However, the AI tool landscape is constantly and rapidly changing, so this knowledge and the tools mentioned in this paper should be seen as a snapshot and serve only as examples of 2020-2025.

AI's Requirements and Demands

The declarative understanding that deploying AI systems, especially large language models (LLMs) like ChatGPT-4 or Gemini, demands substantial physical, environmental, and infrastructural resources. This includes significant energy consumption: for example, generating a single response from ChatGPT 3.5 is estimated to consume approximately 0.5 liters of water per 5–50 prompts, reflecting the carbon footprint tied directly to compute-intensive inference. At scale, global AI data centers are projected to account for up to 100 terawatt-hours annually by 2027, an amount equaling Sweden's entire national electricity demand, driving urgent investment in clean energy solutions such as nuclear fusion and green hydrogen.
Beyond hardware demands (e.g., GPU clusters), operational constraints include access limitations: commercial models like ChatGPT require paid subscriptions for advanced features, including longer context windows, custom instructions, DALL-E 3 image generation, or custom GPTs; meanwhile, plugins are being discontinued due to platform shifts. For institutions prioritizing data sovereignty and privacy, local execution

using open-source alternatives (e.g., self-hosted LLMs) is essential despite the higher technical overhead. These requirements transform AI from a convenience into a strategic infrastructure decision in which cost, sustainability, and ethical responsibility intersect. The knowledge of AI's requirements and demands is supported by 3 documents (Bryda & Costa, 2024; Glerean & Silva, 2024; Hill et al., 2024).

Model's/ Tool's Capabilities and Features

The declarative understanding that AI models and tools offer a rich, functionally diverse set of capabilities, spanning contextual comprehension (e.g., ChatGPT's ability to understand user intent across interactions), multimodal generation (text, audio, image, video, 3D modeling), automated reasoning (hypothesis generation from data correlations or pattern recognition in medical records), and intelligent augmentation of human workflows.
Specific features include real-time deconstruction of scholarly arguments using NLP ("Find Concepts", "Extract Data"), dynamic learning via interaction history (SciSpace), interactive coding environments where users define units of analysis and add annotations with rule-based feedback (Cody), and natural language interfaces that allow asking questions without formal syntax (Scite.ai, Scopus AI).
Tools go beyond simple text generation: they provide structured outputs such as document outlines, emotional tone analysis for audience tailoring, smart citations with contextual metadata, and predictive modeling in scientific domains, including biomolecule design. Advanced tools like SenseMate or PaTAT offer on-demand theme suggestions with transparent rationales that users can inspect and override, enabling agency over AI output. Customized versions (e.g., Alumni AI Research GPT) adapt responses to user roles (author/editor/reviewer), while systems like Synergi use PDF highlights as seeds for generating research threads. Finally, integrated pipelines using GPT-4 or AE-GPT can autonomously generate scientific feedback comparable in accuracy and speed to human reviewers, providing actionable insights without requiring manual step-by-step intervention. The knowledge of the models' and tools' capabilities and features is supported by 35 documents (Barrot, 2025; Bryda & Costa, 2024; Burger et al., 2023; Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Castillo-Martínez et al., 2024; Chopra & Haaland, 2023; Cotton et al., 2024; Dönmez et al., 2023; Freiesleben, 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Gokul et al., 2024; Gupta, 2024; Helgesson, 2024; Ilegbusi, 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. A. Khan et al., 2023; Lieder & Schäffer, 2024; Limongi, 2024; Magesh et al., 2024; Mudd et al., 2024; Murphy & James, 2024; Nicholson Thomas et al., 2024; Picalho et al., 2025; Pretorius & Pretorius, 2025; Rietz & Maedche, 2021; Salvioni & Almici, 2024; Silva et al., 2024; Tran et al., 2023; Ye et al., 2025; Yining et al., 2025; Yongtao et al., 2024).

Disciplinary Knowledge

The declarative understanding that effective human-AI collaboration in research requires researchers to possess deep, domain-specific expertise, not as a passive background skill, but as an active cognitive framework for governance. This includes the knowledge of precise, contextually grounded research questions and hypotheses; interpret qualitative data with nuance based on patterns that have been seen before (e.g., identifying patterns beyond surface-level trends); critically evaluate AI-generated outputs against empirical knowledge, detecting hallucinations, biases, or logical inconsistencies; and validate model assumptions through rigorous analysis rooted in subject-matter logic.
Researchers must also master prompt engineering to align inputs with specific analytical goals, crafting queries that are both detailed enough for relevance and flexible enough to allow iteration, and possess the critical judgment required to distinguish when AI acts as an interlocutor (offering perspectives) versus an author (producing final text). Domain knowledge underpins every decision: selecting which literature to screen, adjusting prompts based on flaws in the output, or rejecting generated suggestions that misrepresent theory.

This disciplinary knowledge enables researchers not only to supervise but also to own the results, ensuring methodological integrity, ethical responsibility, and alignment with scholarly standards, even when AI processes vast volumes of data or automates complex workflows. Disciplinary knowledge is supported by 31 documents (Alfirević et al., 2024; Burgui-Burgui, 2023; Butson & Spronken-Smith, 2024; Castillo-Martínez et al., 2024; Charton et al., 2024; Cotton et al., 2024; Dönmez et al., 2023; Freiesleben, 2023; Fui-Hoon Nah et al., 2023; Garg et al., 2024; Glerean & Silva, 2024; Gupta, 2024; Helgesson, 2024; Hill et al., 2024; Karakose, 2023; Khalifa & Albadawy, 2024; N. Khan et al., 2024; N. A. Khan et al., 2023; Lieder & Schäffer, 2024; Magesh et al., 2024; Mudd et al., 2024; Nicholson Thomas et al., 2024; Parker et al., 2025; Perkins & Roe, 2024a, 2024b; Pretorius & Pretorius, 2025; Silva et al., 2024; Tran et al., 2023; Wise et al., 2024; Yining et al., 2025; Yongtao et al., 2024).

It should be noted that disciplinary knowledge varies across research disciplines, ranging from the standard literature used in every field of study to specialized, nuanced literature that addresses niche research topics. Therefore, this disciplinary knowledge is distinct for every researcher.

Open Science/Open-Source

The declarative understanding that open-source research is a foundational practice for ethical, transparent, and reproducible scientific advancement, encompassing not only the availability of freely accessible code (e.g., llamacpp, Makemore), algorithms, and libraries but also the communities, people, instructions, and infrastructure that sustain them.
This includes platforms like GitHub for version-controlled collaboration on model weights or research scripts; OSF, Chisquares.com, and DataCite as tools that enable persistent data sharing, project organization, and open-science workflows with full provenance tracking. The integration of open-source AI frameworks such as LangChain (for workflow orchestration), LangSmith (for monitoring), ChromaDB (for vector storage), and local execution models ensures that researchers retain ownership of their data and models without relying on proprietary cloud services or commercial lock-in.
This ecosystem fosters transparency by allowing third-party auditing of algorithms, promotes community-driven innovation through shared codebases, and aligns with core principles of scientific integrity, where reproducibility is not optional but engineered into the architecture from day one. Knowledge of open science and open-source is supported by 7 documents (Charton et al., 2024; Chopra & Haaland, 2023; Dönmez et al., 2023; Fui-Hoon Nah et al., 2023; Glerean & Silva, 2024; Ilegbusi, 2024; Limongi, 2024).

Final Remarks

This document should be seen as a work in progress. It should clarify the concepts introduced in Drechsler (2026) and invite other researchers to validate the RAIL model (version 0.9). If you have any questions, please contact me. Also, please register for a validation session (Registration: RAIL Model Validation Sessions). For a more detailed depiction of the competency elements identified in my doctoral project, including quotes from the codes and their sources, the reader of this paper must unfortunately wait until the publication of my thesis (scheduled for the beginning of 2027).

References

Alfirević, N., Rendulić, D., Fošner, M., & Fošner, A. (2024). Educational Roles and Scenarios for Large Language Models: An Ethnographic Research Study of Artificial Intelligence. *Informatics*, *11*(4), 78. https://doi.org/10.3390/informatics11040078

Barrot, J. S. (2025). Balancing Innovation and Integrity: An Emerging Technology Report on SciSpace in Academic Writing. *Technology, Knowledge and Learning*, *30*(1), 587–592. https://doi.org/10.1007/s10758-024-09802-w

Bhutta, A. H. (2024). The intersection of artificial intelligence and rehabilitation sciences: Promoting originality and integrity in research. *The Rehabilitation Journal*, *08*(04), 01–02. https://doi.org/10.52567/trehabj.v8i04.86

Brooks, L. (2023, August 15). *Enhancing Professional Skepticism: A Case Collection | Professional Accounting Centre*. https://www.utm.utoronto.ca/pac/enhancing-professional-skepticism-case-collection

Bryda, G., & Costa, A. P. (2024). TRANSFORMATIVE TECHNOLOGIES: ARTIFICIAL INTELLIGENCE AND LARGE LANGUAGE MODELS IN QUALITATIVE RESEARCH. *Revista Baiana de Enfermagem*, *38*. https://doi.org/10.18471/rbe.v38.61024

Burger, B., Kanbach, D. K., Kraus, S., Breier, M., & Corvello, V. (2023). On the use of AI-based tools like ChatGPT to support management research. *European Journal of Innovation Management*, *26*(7), 233–241. https://doi.org/10.1108/EJIM-02-2023-0156

Burgui-Burgui, M. (2023). *Artificial Intelligence in the automatic coding of interviews on Landscape Quality Objectives. Comparison and case study*. https://doi.org/10.48550/arXiv.2312.05597

Butson, R., & Spronken-Smith, R. (2024). AI and its implications for research in higher education: A critical dialogue. *Higher Education Research & Development*, *43*(3), 563–577. https://doi.org/10.1080/07294360.2023.2280200

Castillo-Martínez, I. M., Flores-Bueno, D., Gómez-Puente, S. M., & Vite-León, V. O. (2024). AI in higher education: A systematic literature review. *Frontiers in Education*, *9*, 1391485. https://doi.org/10.3389/feduc.2024.1391485

Charton, F., Ellenberg, J. S., Wagner, A. Z., & Williamson, G. (2024). *PatternBoost: Constructions in Mathematics with a Little Help from AI*. https://doi.org/10.48550/arXiv.2411.00566

Chen, Z., Chen, C., Yang, G., He, X., Chi, X., Zeng, Z., & Chen, X. (2024). Research integrity in the era of artificial intelligence: Challenges and responses. *Medicine*, *103*(27), e38811. https://doi.org/10.1097/MD.0000000000038811

Chopra, F., & Haaland, I. (2023). *Conducting Qualitative Interviews with AI*. https://doi.org/10.2139/ssrn.4572954

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Dönmez, I., Idil, S., & Gulen, S. (2023). *Conducting Academic Research with the AI Interface ChatGPT: Challenges and Opportunities*. *6*(2).

Drechsler, S. (2026). *The Responsible AI literacy (RAIL) competency model for researchers / Stephan Drechsler*. https://doi.org/10.17619/UNIPB/1-2504

Ennar, H., & Damak-Ayadi, S. (2024). Professional skepticism and auditors' judgments: Evidence from Tunisia. *Journal of Accounting and Management Information Systems*, *23*(3). https://doi.org/10.24818/jamis.2024.03007

Freiesleben, T. (2023). *What does explainable AI explain? …* https://dx.doi.org/10.5282/edoc.31933

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

Garg, S., Ahmad, A., & Madsen, D. Ø. (2024). Academic writing in the age of AI: Comparing the reliability of ChatGPT and Bard with Scopus and Web of Science. *Journal of Innovation & Knowledge*, *9*(4), 100563. https://doi.org/10.1016/j.jik.2024.100563

Glerean, E., & Silva, P. E. S. (2024). *Artificial Intelligence and Research Work*. Zenodo. https://doi.org/10.5281/ZENODO.10890289

Gokul, P., Sun, Z., & Achan, S. (2024). PyGeoweaver: Tangible workflow tool for enhancing scientific research productivity and FAIRness. *SoftwareX*, *27*, 101863. https://doi.org/10.1016/j.softx.2024.101863

Grant, A., Franklin, J., & Langford, P. (2002). The Self-Reflection and Insight Scale: A New Measure of Private Self-Consciousness. *Social Behavior and Personality: An International Journal*, *30*, 821–835. https://doi.org/10.2224/sbp.2002.30.8.821

Gupta, S. (2024). Can We Trust Them? A Critical Evaluation of AI-Generated Content Detection Tools. *International Journal of Scientific Research in Computer Science and Engineering*.

Helgesson, G. (2024). *Ethical aspects of the use of AI in research*. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-212651

Hill, J. E., Harris, C., & Clegg, A. (2024). Methods for using Bing's AI-powered search engine for data extraction for a systematic review. *Research Synthesis Methods*, *15*(2), 347–353. https://doi.org/10.1002/jrsm.1689

Ilegbusi, P. (2024). *Artificial Intelligence: The Game Changer in Scientific Research ...* https://dx.doi.org/10.60763/africarxiv/1516

Karakose, T. (2023). The Utility of ChatGPT in Educational Research—Potential Opportunities and Pitfalls. *Educational Process International Journal*, *12*(2). https://doi.org/10.22521/edupij.2023.122.1

Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update*, *5*, 100145. https://doi.org/10.1016/j.cmpbup.2024.100145

Khan, N. A., Osmonaliev, K., & Sarwar, M. Z. (2023). Pushing the Boundaries of Scientific Research with the use of Artificial Intelligence tools: Navigating Risks and Unleashing Possibilities. *Nepal Journal of Epidemiology*, *13*(1), 1258–1263. https://doi.org/10.3126/nje.v13i1.53721

Khan, N., Elizondo, D., Deka, L., & Molina-Cabello, M. A. (2024). Natural Language Processing Tools and Workflows for Improving Research Processes. *Applied Sciences*, *14*(24), 11731. https://doi.org/10.3390/app142411731

Kuiper, M. (2024). *WEBINAR: A practical guide to AI tools for life scientists*. https://doi.org/10.5281/ze-nodo.11206329

Lieder, F. R., & Schäffer, B. (2024). *Reconstructive Social Research Prompting (RSRP). Distributed Interpretation between AI and Researchers in Qualitative Research*. SocArXiv. https://doi.org/10.31235/osf.io/d6e9m

Limongi, R. (2024). The Use of Artificial Intelligence in Scientific Research with Integrity and Ethics. *Review of Artificial Intelligence in Education*, *5*(00), e0022. https://doi.org/10.37497/rev.artif.intell.educ.v5i00.22

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2024). *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools* (arXiv:2405.20362). arXiv. https://doi.org/10.48550/arXiv.2405.20362

Mohamed, B., Attou, Y., Seddik, M., & Abdelhamid, N. (2025). From Chatting to Cheating: How Can Ethical Considerations Be Ensured in this AI-Driven Research Era? *International Journal of Language and Literary Studies*, *7*(1), 287–297. https://doi.org/10.36892/ijlls.v7i1.1996

Mudd, A., Conroy, T., Voldbjerg, S., Goldschmied, A., & RPRT. (2024). *Chatting with Pythons: Using ChatGPT and Python computer code to screen abstracts for systematic literature reviews in complex disciplines*. https://vbn.aau.dk/da/publications/26b14f1d-26bb-480a-86b7-235fa0bd5d6f

Murphy, J. E., & James, K. (2024). *Generative AI Tools for Research*. https://hdl.handle.net/1880/119482

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, *49*(5), 847–864. https://doi.org/10.1080/03075079.2024.2323593

Nicholson Thomas, I., Roche, P., & Grêt-Regamey, A. (2024). Harnessing artificial intelligence for efficient systematic reviews: A case study in ecosystem condition indicators. *Ecological Informatics*, *83*, 102819. https://doi.org/10.1016/j.ecoinf.2024.102819

Parker, J. L., Richard, V. M., Acabá, A., Escoffier, S., Flaherty, S., Jablonka, S., & Becker, K. P. (2025). Negotiating Meaning with Machines: AI's Role in Doctoral Writing Pedagogy. *International Journal of Artificial Intelligence in Education*, *35*(3), 1218–1238. https://doi.org/10.1007/s40593-024-00425-x

Perkins, M., & Roe, J. (2024a). Academic publisher guidelines on AI usage: A ChatGPT supported thematic analysis. *F1000Research*, *12*, 1398. https://doi.org/10.12688/f1000research.142411.2

Perkins, M., & Roe, J. (2024b). The use of Generative AI in qualitative analysis: Inductive thematic analysis with ChatGPT. *Journal of Applied Learning & Teaching*, *7*(1). https://doi.org/10.37074/jalt.2024.7.1.22

Picalho, A. C., de Oliveira, G. R., & Cativelli, A. S. (2025). *Artificial intelligence in bibliographic searches in scientific databases: Comparing search expressions in ChatGPT, Copilot, and Gemini*.

Pretorius, L., & Pretorius, C. (2025). *Exploring ChatGPT's potential as a qualitative research partner: Researcher and participant perspectives on AI-generated insights*. https://doi.org/10.26180/28386053.v3

Rietz, T., & Maedche, A. (2021). Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445591

Rosales, C. (2024, February 1). The Decision Lab—Behavioral Science, Applied. *The Decision Lab*. https://thedecisionlab.com/insights/society/behavioral-science-is-weird-and-this-should-concern-us

Salvioni, D. M., & Almici, A. (2024). Ethics in Management Research and Artificial Intelligence. *Symphonya. Emerging Issues in Management*, (2), 50–65. https://doi.org/10.4468/2024.2.04salvioni.almici

Shukla, A. K., Terziyan, V., Tiihonen, T., & Autonomy. (2024). AI as a user of AI: Towards responsible autonomy. *Heliyon*, *10*(11). https://doi.org/10.1016/j.heliyon.2024.e31397

Silva, A. D. O., Janes, D. D. S., & Santos, R. (2024). GPT Alumni AI Pesquisa: A Practical Tutorial for the Adoption and Ethical Use of AI in Scientific Research. *Review of Artificial Intelligence in Education*, *5*, e033. https://doi.org/10.37497/rev.artif.intell.educ.v5i00.33

Skyba, V., Vozniuk, N., Likho, O., Vozniuk, S., & Buhaiev, O. (2024). Alternative Tools for Modern Agroecological Research. *International Conference of Young Professionals «GeoTerrace-2024»*, 1–5. https://doi.org/10.3997/2214-4609.2024510033

Tran, N., Chau Nguyen, D. N., Texas Tech University-HSC at Amarillo, Amarillo, TX, USA, H Nguyen, D., Cardiovascular Research Department, Methodist Hospital, Merrillville, IN USA, Phan, N., University of Medicine and Pharmacy at Ho Chi Minh City, Ho Chi Minh City, Vietnam, T Nguyen, D., Tan Tao University, School of Medicine, Long An Province, Vietnam, Nguyen, T., & Cardiovascular Research Department, Methodist Hospital, Merrillville, IN USA; Tan Tao University, School of Medicine, Long An Province, Vietnam. (2023). Utilizing ChatGPT in the Process of Crafting a Research Paper: A Comprehensive Guide. *TTU Journal of Biomedical Sciences*, *2*(1), 41–50. https://doi.org/10.53901/tjbs.2023.08.art05

Wise, B., Emerson, L., Van Luyn, A., Dyson, B., Bjork, C., & Thomas, S. E. (2024). A scholarly dialogue: Writing scholarship, authorship, academic integrity and the challenges of AI. *Higher Education Research & Development*, *43*(3), 578–590. https://doi.org/10.1080/07294360.2023.2280195

Yaroshenko, T. O., & Iaroshenko, O. I. (2023). Artificial Intelligence (AI) for Research Lifecycle: Challenges and Opportunities. *University Library at a New Stage of Social Communications Development. Conference Proceedings*, (8), 194–201. https://doi.org/10.15802/unilib/2023_294639

Ye, R., Varona, M., Huang, O., Lee, P. Y. K., Liut, M., & Nobre, C. (2025). *The Design Space of Recent AI-assisted Research Tools for Ideation, Sensemaking, and Scientific Creativity* (arXiv:2502.16291). arXiv. https://doi.org/10.48550/arXiv.2502.16291

Yining, H., Beam, A., Chibnik, L. B., & Torous, J. (2025). From statistics to deep learning: Using large language models in psychiatric research. *International Journal of Methods in Psychiatric Research*, *34*(1), e70007. https://doi.org/10.1002/mpr.70007

Yongtao, L., Checa, M., & Vasudevan, R. K. (2024). Synergizing human expertise and AI efficiency with language model for microscopy operation and automated experiment design*. *Machine Learning: Science and Technology*, *5*(2), 02LT01. https://doi.org/10.1088/2632-2153/ad52e9