American Journal of Multidisciplinary Research and Innovation (AJMRI)

ISSN: 2158-8155 (ONLINE), 2832-4854 (PRINT)

VOLUME 5 ISSUE 5 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Multidisciplinary Research and Innovation (AJMRI)

Volume 5 Issue 5, Year 2026
ISSN: 2158-8155 (Online), 2832-4854 (Print)
DOI: https://doi.org/10.54536/ajmri.v5i5.8515
https://journals.e-palli.com/home/index.php/ajmri

Between Methodological Expansion and Epistemic Boundary Dissolution: Generative AI, Distributed Epistemic Processes, and the Reorganization of Scientific Responsibility

Giovanni Vindigni[1*]

Article Information

Received: June 21, 2026

Accepted: September 05, 2026

Published: September 19, 2026

Keywords

Educational Research, Epistemic Agency, Generative AI, Large Language Models, Mixed Methods

ABSTRACT

Although individual studies have examined the capabilities of generative AI in annotation, coding, item construction, thematic analysis, and data simulation, there has so far been no cross-paradigm framework that systematically integrates technical performance, methodological validity, institutional embedding, and the attribution of scientific responsibility. Against this backdrop, the study examines under what conditions the delegation of knowledge-relevant research operations to large language models (LLMs) should be evaluated as a controlled methodological extension or as an epistemic dissolution of boundaries. Methodologically, it follows a critical integrative evidence synthesis. From 80 hits across the eight thematically differentiated search clusters and five additional articles identified through citation tracking, 58 publications remained in the core analytical corpus after removing duplicates, screening titles and abstracts, and reviewing full texts based on predefined criteria. The predefined criteria restricted the corpus to studies that empirically evaluated LLM-mediated research operations or examined their validity, bias, reproducibility, epistemic efficacy, or governance with direct or transferable relevance to social science and educational research. With regard to the synthesized findings, the evidence indicates that LLMs operationally scale quantitative methods but may in the process foster methodological pseudo-competence, non-classical measurement errors, and synthetic circularity. In qualitative designs, they expand the scope for contrastive interpretation but do not replace situated case understanding; in mixed-methods designs, they reduce translation costs without negating the independent logics of validity of the combined strands. Building on, but analytically distinct from, these synthesized findings, the study advances an exploratory conceptual framework in the form of an original five-level matrix of epistemic delegation, ranging from operational and infrastructural offloading to empirical surrogation. Each level is linked to specific epistemic risks and proportionate verification requirements. In addition, reflexive controllability is operationalized as a governance framework with eight dimensions. Both instruments can be directly applied to research design, quality assurance, institutional rule-making, and methodological training and are transferable across different empirical paradigms. Therefore, generative AI appears neither as a neutral tool nor as a responsible agent of knowledge but rather as a probabilistic epistemic actor within distributed knowledge arrangements, for which the ultimate scientific responsibility remains with identifiable individuals and institutions.

INTRODUCTION

Empirical social and educational research does not take place outside its technical media, institutional frameworks, and methodological formalizations. Questionnaire logic, transcription regimes, statistical software, databases, and computer-assisted qualitative analysis environments do not constitute a secondary sphere of instruments; rather, they contribute to the constitution of what can emerge as an observable object, a relevant difference, and communicable evidence. Nevertheless, the epistemic hierarchy of conventional methodological theory remained comparatively stable: technical systems processed data within human-defined operational frameworks; the determination of the subject matter, methodological justification, interpretation, and responsibility for the claim to validity were attributed to the researcher or the scientific collective. The diffusion of generative AI capable of dialogic interaction does not re-establish this technical nature of research but rather disrupts its previously assumed asymmetry.

In this context, large language models such as ChatGPT, Claude, or Gemini are not limited to the rule-based execution of pre-structured processing steps. Rather, they generate variant hypotheses, construct operationalizations, analyze codes, create category systems, annotations, summaries of material, simulated response patterns, and semantically coherent interpretations of heterogeneous findings. In doing so, the systems intervene in operations that, within quantitative research, are associated with measurement modeling and inferential validity; within qualitative research, with theoretical sensitivity and a situated understanding of others; and within mixed methods, with the reasoned relation of epistemologically heterogeneous data streams. It is precisely the empirically documented effectiveness of LLMs in annotation (Törnberg, 2025; Ziems *et al.*, 2024), deductive coding (Tai *et al.*, 2024), and inductive thematic structuring (De Paoli, 2024) that raises a question extending beyond the mere efficiency of methodological application. Rather, it destabilizes the established order of attribution within

[1] Professor, Dean of Studies, Diploma Hochschule – University of Applied Sciences, Germany, and Doctoral Advisor, Global Humanistic University, Anguilla and Middlesex University, United Kingdom
[*] Corresponding author's e-mail: giovanni@vindigni.de

which analytical operations and epistemic judgment were located in human research practice, even as scientific responsibility remains irreducibly assigned to researchers and institutions.

From a purely instrumentalist perspective, the definition of generative systems is insufficient because these systems produce orders of relevance and interpretive connections that cannot be fully subsumed under rules specified in advance by humans. However, it would be equally inadequate to elevate them to the status of autonomous subjects of knowledge, since probabilistic language models possess neither intentional references to objects, nor experience-based judgment, nor the capacity for responsibility. Between technical neutralization and anthropomorphic subjectification, a relational definition thus emerges: LLMs exert epistemic efficacy as actants within distributed epistemic arrangements, whose concrete operation arises from the interplay of training corpora, model architecture, alignment, prompt structure, research data, platform regimes, and human selection practices. Floridi's (2023) concept of "agency without intelligence" signifies this intermediate status, while Messeri and Crockett (2024) demonstrate that the rise in syntactically and semantically plausible research outputs can coincide with a decline in genuine subject matter comprehension. Epistemic boundary shifts thus manifest themselves not primarily in the occurrence of incorrect answers, but in the infrastructural generation of apparent validity without corresponding depth of justification.

Considering the current state of research, it can be noted first of all that this constellation has so far been addressed only in a fragmented manner. Systematic reviews focus primarily on application potential, academic integrity, data protection, bias, and higher education (Haltaufderheide & Ranisch, 2024; Qian, 2025; Yan *et al.,* 2024), while method-related studies evaluate limited functional areas such as deductive coding, thematic analysis, item development, synthetic samples, or statistical analysis (Bisbee *et al.,* 2024; Mathis *et al.,* 2024; Prandner *et al.,* 2025; Terry *et al.,* 2025). It remains, however, unclear how technical performance, methodological validity, institutional embedding, and accountability are mediated across established research paradigms. It is precisely on the quality of this mediation that it depends whether generative AI enables a controlled methodological expansion or produces an epistemic boundary dissolution within which operational feasibility and scientific justifiability diverge.

In this gap, the study offers a synthesis of evidence that is both paradigmatically contrasting and governance-oriented. The focus is not on the binary question of whether LLMs replace or support research, but rather on the conditions under which knowledge-relevant operations can be delegated without compromising the traceability of their prerequisites, the controllability of their variance, and the accountability for their consequences. Three research questions (RQ) are derived from this:

RQ1: How does generative AI reconfigure the distribution of epistemic functions among researchers, data, models, and institutional research infrastructures within quantitative, qualitative, and mixed-methods designs?

RQ2: Under what methodological, technical, and governance-related conditions can AI-mediated analysis and interpretation become valid, subject to reflexive control, and scientifically accountable?

RQ3: What shifts in competence and responsibility arise from the delegation of knowledge-generating operations for methodological training in higher education?

The guiding thesis is that generative AI does not eliminate the paradigmatic differences of empirical research within a unified methodological regime.

Rather, across these differences, it establishes a dimension of delegation in which the technical offloading of operations can gradually be transformed into semantic co-construction and, ultimately, into the surrogation of empirical references. As the system increasingly approaches the generation of categories, explanations, and meta-inferences, a quality assessment reduced to output performance loses its viability. What is required is a reflexive control architecture within which model variance, data provenance, cultural positioning, institutional access conditions, and human ultimate responsibility are brought together as constitutive components of scientific validity. Accordingly, the diagnosed epistemic dissolution should not be understood as the disappearance of methodological standards but rather as the risk of decoupling between operation, justification, and responsibility.

LITERATURE REVIEW

From Software-Supported Methodology to Generative Research Infrastructure

Even before the advent of generative AI, the digitization of empirical research had already brought about a lasting change in its epistemic processuality. Statistical systems such as SPSS, R, or Mplus translate theoretically specified models into formally executable operations; CAQDAS environments such as MAXQDA or ATLAS.ti structure data access, segmentation, coding, and retrieval. Their effectiveness is not limited to increased efficiency, especially since user interfaces, default values, and available analysis paths help determine which relationships become visible, comparable, and methodologically compatible. Nevertheless, the semantics of the operation remained fundamentally ex ante: regression specifications, codes, and comparison dimensions were defined by researchers before the software executed them within a predetermined or at least rule-bound procedure.

Generative models transform this operational logic of structuring because they make the boundary between execution and specification partially permeable. While their generativity enables the production of new categories, examples, formulations, and explanatory proposals, their generality connects otherwise separate research operations within the same model environment; furthermore, due to dialogic adaptivity, the result remains dependent on a sequential history of interaction.

Consequently, the specific research step is no longer fully prefigured by a stable software function; rather, it emerges from system instructions, prompt architecture, context windows, model version, sampling parameters, and previous outputs. What emerges is a probabilistic research infrastructure whose capabilities stem precisely from that semantic openness, which simultaneously limits its reproducibility and accountability (Chang *et al.,* 2024; van Dis *et al.,* 2023).

However, this infrastructure should be analyzed not as an isolated computational tool but as an institutionally embedded regime. Foundation models incorporate large-scale, selectively curated, and often incompletely documented training corpora, as well as corporate alignment and security decisions and economically controlled interfaces. Bender *et al.* (2021) demonstrate that statistical fluency constitutes neither communicative intentionality nor a reliable reference to the subject matter. Consequently, for scientific use, this situation implies that data provenance, operator interests, representational asymmetries, and technical access conditions should be treated not as external framework conditions but as operational components of knowledge production (Birhane *et al.,* 2023). Particularly under the current conditions of advancing platformization within a knowledge-economy-oriented environment, specifically within the present reality of life and working conditions, which is characterized by eclecticism due to the convergent implications of internationalization, informatization, and individualization, this sociotechnical and infrastructural interdependence is further intensified within tertiary education and research spaces, as model access, storage practices, version updates, and functional scopes are organized according to private-sector principles, thereby giving rise to a new form of infrastructural access conditionality (Vindigni, 2026a).

Against this backdrop, the concept of a 'tool' underestimates the system's pre-structuring effectiveness. LLM outputs distribute attention, normalize terminology, hierarchize bodies of knowledge, and prefigure subsequent argumentative decisions. Epistemically relevant risks therefore lie not exclusively in manifest hallucinations but equally in plausibilized selectivities whose coherence masks their contingency. Park *et al.* (2024) point to a reduction in cognitive diversity during repeated model-based idea generation; Peters and Chin-Yee (2025) demonstrate a systematic tendency to exceed the scope of scientific source texts. As a result, the generative infrastructure produces not only answers but also probability orders of what can be scientifically said.

Distributed Epistemicity Between Technical Agency and Human Ultimate Responsibility

A strong conception of epistemic agency presupposes the ability to weigh reasons, form beliefs, reflexively defend claims to validity, and take responsibility for the consequences of one's own judgments. Current LLMs do not meet these requirements. They lack both phenomenal experience and personal interest in knowledge, and they cannot affirm or revise the truth of a statement as a normative claim. As autonomous subjects of knowledge, they could therefore only be described as such if one confuses statistical connectivity with justifiable subjectivity.

In contrast, a relational conception shifts the analytical focus from internal intentionality to observable effectiveness within socio-technical epistemological arrangements. Actor-like relevance arises where system operations alter subsequent decisions: generated categories organize the perception of data, produced scripts realize specific model assumptions, and synthetic samples prefigure notions of population-based plausibility. The term epistemic actant refers to this technically mediated participation in the selection, ordering, and plausibilization of scientific statements, without attributing to the system a subject position capable of justification or accountability.

Thus, epistemic performance is defined as distributed, while scientific responsibility is defined as irreducibly asymmetrical. Model outputs arise neither autonomously nor entirely under human control. Rather, they result from a multi-stage conversion architecture involving training data, weightings, alignment, prompting, sampling, tool access, and context-based selection; at the same time, the internal transformation conditions of proprietary systems remain partially opaque. Researchers therefore cannot causally reconstruct every model operation but must take responsibility for the conditions of use, the verification of the output, and the scope of the claim to validity derived therefrom. Only individuals and institutions can ensure consent, weigh risks of harm, publicly justify their decisions, make corrections, and fulfill accountability obligations (Binz *et al.,* 2025; Leslie, 2025).

Distributed epistemicity therefore does not relativize human ultimate responsibility but rather clarifies its scope. Responsibility does not lie with a supposed internal decision of the model but rather with the institutionally authorized delegation, the selection of the model environment, the construction of the prompt, the level of validation, and the transformation of probabilistic outputs into scientific statements. Where this mediation becomes invisible and technical generativity appears as a self-legitimizing cognitive achievement, epistemic deregulation sets in.

Paradigmatic Difference as a Limit to Information Technology and Cyber-Physical Convergence

Quantitative, qualitative, and mixed-methods research differ not primarily in terms of data formats, but rather in their specific relationships between theory, the constitution of the object of study, observation, and the justification of validity. Quantitative designs formalize constructs, standardize observational relationships, and justify inferential generalizations. Qualitative designs reconstruct connections of meaning, practice, and context and make the researchers' positionality an integral

part of methodological reflexivity. Mixed methods relate these two strands of knowledge without dissolving their presuppositions into a single methodological language. Generative AI therefore does not intervene in a homogeneous methodological space but rather in heterogeneous epistemic structures of commitment.

Within quantitative designs, the model output materializes as code, an item, a classification, a parameter suggestion, or a synthetic observation; the central question of validation concerns measurement validity, estimation logic, and inferential robustness. Within qualitative designs, it appears as a category, thematic condensation, counter-interpretation, or theoretical connection; here, contextual sensitivity, openness, and reflexive traceability take center stage. Within mixed-methods designs, the issue shifts to the level of meta-inference: the fact that an LLM can process numerical and linguistic representations within the same vector space does not in itself justify the methodologically legitimate integration of their epistemological claims (Combrinck, 2024).

Consequently, this non-equivalence is corroborated by a heterogeneous yet convergent body of empirical finding. On the one hand, significant levels of agreement can be observed in narrowly specified annotations (Törnberg, 2025; Ziems *et al.,* 2024); on the other hand, these even coexist with socially systematic classification errors (Ashwin *et al.,* 2026; Lin & Zhang, 2026). Furthermore, it must be acknowledged that thematic analyses can generally give rise to central fields of significance, namely with regard to tendencies toward descriptive smoothing and culturally dominant interpretive readings, without taking into account the processes of inculturation, enculturation, acculturation, and assimilation in the respective intertextual context (De Paoli, 2024; Mathis *et al.,* 2024; Wachinger *et al.,* 2025). Accordingly, synthetic populations approximate aggregated response trends but fail to reliably capture minority positions, situational contingency, and variance structures (Argyle *et al.,* 2023; Bisbee *et al.,* 2024). From this heterogeneity, with regard to deductive rigor, i.e., including the determinants, indicators, and premises established here, there follows not a categorical rejection but rather the need for a paradigmatically situated delegation methodology that does not evaluate technical performance independently of the respective logic of validity but rather always requires a coherent evaluation, ex ante, here with regard to concept control; interim, here with regard to process control, and ex post with regard to outcome control.

MATERIALS AND METHODS

Research Design and Epistemological Scope

Designed as a critical-integrative evidence synthesis, this study considers a body of literature in which experimental performance comparisons, methodological validation studies, systematic reviews, analyses grounded in the philosophy of science, and governance-related positionings generate different types of evidence. In view of incommensurable tasks, models, outcome measures, and claims to validity, their meta-analytic homogenization would be justifiable neither statistically nor epistemologically. Instead of such standardization, the synthesis aims for a theory-guided comparability of heterogeneous findings by reconstructing their respective conditions of validity and relating them to the overarching structural problem of epistemic delegation. Thus, structured literature searching, deductive-inductive category formation, and paradigmatic contrast are not combined additively but are conducted as interrelated analytical steps within a critical-realist logic of mediation.

Literature Review, Selection Criteria, and Corpus Construction

By August 19, 2026, the literature review had been completed. The primary publication period was defined as January 2020 through August 2026; references from earlier periods were included only when they were of fundamental importance to the methodological or epistemological foundation of the study. Consensus served as the semantic scientific search space, while Crossref was used for DOI-based bibliographic verification; targeted queries of the respective publisher and journal platforms ensured that the "version of record" was prioritized. Operationally, eight cluster-specific searches were conducted in Consensus. From each of these search clusters, the ten records ranked as most relevant within the platform-mediated retrieval environment were transferred to the initial screening corpus, thereby yielding 80 records. Following the removal of six duplicates attributable to cross-cluster overlap, 74 unique publications remained. Forward and backward citation tracing identified five additional contributions whose methodological or epistemological relevance warranted their inclusion, increasing the screening corpus to 79 records. Title and abstract screening subsequently led to the exclusion of ten publications that did not satisfy the predefined eligibility criteria, leaving 69 contributions for full-text assessment. At this stage, eleven further publications were excluded because they constituted purely technical benchmarks without a reconstructible methodological basis for validity, clinical applications lacking transferable implications for social or educational research, or insufficiently documented conceptual contributions. The final analytical corpus therefore comprised 58 publications . A publication-level evidence matrix documenting the study design, data basis, AI system or model version, research operation, validation criteria, principal findings, and assignment to the five levels of epistemic delegation for each of the 58 publications is provided in Supplementary Table S1.

Criteria for inclusion were the empirical evaluation of an LLM-mediated research step, a methodologically explicit analysis related to the social, educational, or behavioral sciences, an investigation of validity, bias, reproducibility, or epistemic agency, or a systematic synthesis with direct applicability to research practice or methodological development. Excluded were duplicates of preprints and

versions of record, purely technical benchmarks without a reconstructible methodological basis for validity, clinical applications without transferable research implications, and insufficiently documented opinion pieces. Seven of the author's works on socio-technical governance, digital inclusion, AI-mediated inequality, cultural representation, and human-computer interaction were included not as self-validation but as theoretically contextualizing references (Vindigni, 2024a, 2024b, 2025a, 2025b, 2026a, 2026b, 2026c).

Although the research does not claim to be exhaustive in the sense of a traditional systematic review, it does aim for a methodologically controlled coverage of the evidence areas relevant to the research questions. Considering that the subject matter involves model versions, application practices, and publication statuses that change rapidly, such a limitation on scope appears necessary from an epistemological perspective. To prevent mere temporal currency from being conflated with evidentiary robustness, peer-reviewed versions of record were prioritized over corresponding preprints, while bibliographic metadata were independently verified through DOI-based cross-checking.

Theory-Guided Coding and Integrative Contrast

Within the operationalization of the synthesis, the 58 publications were subjected to a three-stage, recursively controlled coding procedure whose analytical validity was secured through methodological triangulation.

In the first stage, the publications were deductively assigned to the fields of quantitative research, qualitative research, mixed methods, research governance, and methods training. Inductively, the second stage identified recurring functional and problem complexes: code and item generation, annotation, category formation, theme reconstruction, synthetic data, heterogeneous evidence integration, model and prompt dependency, bias, data protection, reproducibility, and attribution of responsibility. In the third stage, these categories were not merely aggregated in tabular form but were contrasted along the dimensions of technical operation, epistemic validity claim, potential propagation of errors, and required intensity of control.

As an overarching analytical dimension, the level of epistemic delegation served as a framework. Here, this refers to the extent to which a system not only performs a human-specified operation but also selects relevant information, generates semantic mappings, structures interpretive spaces, or substitutes for empirical observation. Based on the cross-referenced relationships between the degree of delegation, paradigmatic validity logic, error propagation, and validation density, a five-level delegation matrix was developed. Systematically, the negative case test contrasted performance findings with evidence of bias, instability, and loss of validity. This serves to control for the fallacy central to the subject area, namely the tendency to directly equate technical output quality with epistemic quality.

RESULTS AND DISCUSSION

Quantitative Research: Operational Scaling and the Decoupling of Procedural Access and Methodological Competence

At this point, within quantitative research, generative AI initially functions as a translator between analysis intentions articulated in natural language and formally executable operations. LLMs transform requirements into R, Python, SPSS, or Stata syntax, diagnose programming errors, vary model specifications, and generate graphical representations. In light of the exploratory findings by Prandner *et al.* (2025), a new form of methodological accessibility can be identified here, within which operational access to advanced procedures is less severely limited by prior programming experience. By lowering technical barriers, analytical capabilities can be expanded, and feedback loops between research questions, code, and results can be accelerated in methodological training. However, it is precisely this operational accessibility that creates a structural risk of decoupling. Syntactic correctness indicates neither the appropriateness of the statistical model nor the validity of its assumptions; linguistically coherent interpretations of results can mask model uncertainty, effect size, multiple testing, or construct-irrelevant variance. Dialogic plausibility creates an epistemic proximity that can compensate for a deficiency in the ability to provide justification. In this regard, Lin (2023) and Chang *et al.* (2024) show that the ex post performance evaluation of generative systems depends on the benchmark, prompt design, contextualization, and evaluation metric. Therefore, the quality of quantitative results cannot be isolated from the output but must be tied back to reference code, controlled test data, assumption verification, sensitivity analysis, and interpretation grounded in subject-matter expertise. The democratization of access to the methodological process only becomes scientifically productive when it is distinct from a democratization of unfounded claims to validity.

With regard to the operational design of questionnaire items, scales, and test tasks, it can first be stated that the function of generative AI cannot be limited to the accelerated execution of development steps that have already been fully determined. Rather, there is an epistemically significant shift from the linguistic formulation of individual items to the prior structuring of what is even permitted within a measurement model as an observable expression of a theoretically defined construct. In particular, this is due to the ability of generative models to rapidly generate extensive item pools from prompt-based construct definitions, target group descriptions, difficulty parameters, and response formats; to systematically vary linguistic complexity; and to adapt formulations to anticipated characteristics of the respective target population. However, this scaling of the construct development process simultaneously shifts indicator-related preliminary decisions to an earlier stage in the model interaction: training-based probability orders influence which construct facets are linguistically represented, which semantic relations are prioritized, and

which response or interpretive horizons are excluded. Empirical comparative studies document, in this regard, a considerable degree of subject-matter and, in some cases, psychometric compatibility of AI-generated items with established measurement instruments, as well as significant efficiency potential compared to construction processes carried out exclusively by humans (Martin Kowal *et al.,* 2025; Terry *et al.,* 2025). Nevertheless, one cannot unconditionally conclude from this that the generated item pools possess independent psychometric validity. Semantic plausibility, subject-matter preference judgments, or an approximate factor structure do not substitute for the examination of content and construct validity, nor for the population-based investigation of reliability, measurement invariance, and differential item functioning. Consequently, in the development of scales for market and educational research, generative models do not attain the status of autonomous measurement constructors but rather that of hypothesis-generating construction tools whose productive role lies in the systematic expansion, variation, and contrast of possible operationalizations (Keane & McNaughton, 2026). Such pre-structuring does not itself confer psychometric validity upon a generatively produced item pool; rather, scientific robustness arises only through its incorporation into a multi-stage validation architecture comprising subject-matter content review, cognitive pretesting, population-based pilot testing, psychometric item analysis, and a documented human final decision. Within this methodologically delimited attribution of functions, Blanchard *et al.* (2025) systematize in this context the conditions under which generative AI may be incorporated into survey and experimental design. Precisely because AI-generated items remain linguistically prestructured test material rather than validated measurement instruments, their evaluation must extend beyond semantic plausibility to potential construct underrepresentation, differential item functioning, and cultural misfit.

As linguistically pre-structured test material, generatively produced item pools require expert judgments of content validity, theoretically grounded examination of construct representation, and empirical testing within the relevant taregt populations to establish their measurement quality. Models can reproduce dominant scale formulations without identifying construct underrepresentation, differential item functioning, or cultural misfit. If the same model type is used for generation, justification, and evaluation, a closed semantic confirmation loop emerges: regularities previously generated from the system's own probabilistic structure may then be misconstrued as independently validated evidence. Generative item construction therefore does not reduce the validation effort but shifts it from linguistic production to theoretical and population-based validity testing.

In the field of annotation and text classification, the epistemic shift described above becomes even more pronounced. Methodologically speaking, annotation by no means refers exclusively to the technical preprocessing of linguistic data but rather to the operational conversion process through which theoretically defined constructs that are inaccessible to direct observation are transformed into empirically analyzable categories. For example, when a generative language model assigns a text sequence a political orientation, an argumentative function, an emotional valence, or the presence of a social event, it consequently does not merely determine the formal classification of a document but also contributes to the empirical representation of the construct under investigation. All subsequent frequency distributions, group relations, regression models, and inferential conclusions are based on this conversion. Against this backdrop, the technical performance of language models acquires direct relevance to measurement methodology. Ziems *et al.* (2024) examine 13 language models using 25 tasks from computational social science and demonstrate that the zero-shot method can generate both taxonomic classifications and freely formulated coding justifications. While the models do not consistently outperform the most powerful, finely tuned benchmark models in classification tasks, they sometimes achieve remarkable agreement with human coding; with regard to the generation of explanatory justifications, their results in some cases even exceed the quality of the crowdworker texts used as references. Furthermore, for the classification of political social-media messages, Törnberg (2025) demonstrates that GPT-4 can achieve higher accuracy scores than human expert coders and supervised classifiers across different language areas and national contexts. However, such performance findings by no means justify the cross-domain superiority of generative annotation. Rather, the evidence reveals a pattern of performance that remains contingent on the characteristics of the task, the operational definition of the construct, and the underlying technical configuration, and whose scope of validity must therefore be established separately for each application context.

Codebook-based annotation is of particular methodological relevance in this context. Codebooks do not merely contain category labels; rather, they operationalize theoretical constructs through definitions, delimitation rules, inclusion and exclusion criteria, anchor examples, and guidelines for handling ambiguous cases. When translated into an instructional architecture, these components are converted into machine-processable decision conditions. In this process, the prompt partially assumes the function of an operational interface between theoretical conceptualization and categorical data production. Halterman and Keith (2026), however, demonstrate that even when codebooks are available, language models are sensitive to the order of categories, the linguistic formulation of definitions, and the distinction between category names and category meanings. Accordingly, a formally correct label output does not necessarily imply a construct-faithful application of the codebook. Rather, the model-mediated assignment introduces a second level of interpretation: First,

researchers translate a theoretical construct into coding rules; subsequently, the language model translates these rules into categorical decisions under the conditions of its training-based probability framework. As a result, annotation transforms from a supposedly neutral preparatory operation into a technically mediated measurement process, within which construct definition, instruction design, model architecture, context window, and decision format collectively determine the nature of the generated data.

Given this methodological status, an evaluation of generative annotation must not be limited to global accuracy values or aggregated agreement coefficients. Such metrics condense heterogeneous error patterns into a single overall value, thereby obscuring whether misclassifications are asymmetrically distributed across categories, occur more frequently with rare feature values, or disproportionately affect certain social groups. When class frequencies are unequal, even the preferential assignment of a dominant category can yield a high accuracy value, even though it is precisely the theoretically or socially relevant minority cases that are missed. Pangakis *et al.* (2023) demonstrate, using 27 annotation tasks from eleven social science datasets, that the performance of generative models varies significantly depending on the characteristics of the dataset, the conceptual difficulty of the respective category, and the design of the instructions. Lin and Zhang (2026) further point out that the results depend on prompt construction, category definition, and the provision of context. Finally, Ashwin *et al.* (2026) show that misclassifications are not necessarily randomly distributed but can covary with characteristics of the individuals under study. Where misclassification probabilities depend on group membership, language use, social position, or downstream covariates, the error can no longer be treated as non-differential measurement error. Rather, this misclassification results in a directed error structure through which distributions can be shifted, group differences can be exaggerated or leveled out, regression coefficients can be distorted, and standard errors as well as significance decisions can be misdetermined. Epistemically significant, therefore, is not merely how often a model is incorrect, but rather in which cases, under which configurations, and with what consequences for subsequent inference its erroneous decisions occur.

Methodologically, generative annotation must therefore be treated as a configuration-dependent measurement instrument whose validity cannot be derived from the plausibility of individual outputs but exclusively from a multi-stage examination of its measurement performance. First, a reference sample is required that is separate from the model development, manually coded, and stratified according to theoretically relevant categories and group characteristics. Such reference-based validation must, however, be extended to include a configuration-sensitive determination of sensitivity, specificity, positive and negative predictive values, class-specific calibration, and group-specific error profiles, accompanied by a systematic examination of variance across prompt formulations, category orders, context scopes, model families, and model versions. An ensuing error-propagation analysis must establish whether, and to what extent, the identified classification contingencies alter downstream statistical inference by recalculating the relevant estimates on the basis of both human reference codings and alternative model configurations.

Only through this process can one assess whether an empirical finding holds up against the contingencies of its technical production. Abdurahman *et al.* (2025) incorporate corresponding requirements for validity, replicability, robustness, and methodological transparency into a comprehensive testing framework for applications of language models in social science. In the absence of such a reconstruction-capable validation architecture, uncertainty is not eliminated but merely shifted from the visible negotiation of human coding decisions to a proprietary, version-dependent, and only partially transparent model configuration. Herein precisely lies the epistemic boundary dissolution of generative annotation: operational scalability and scientific validity appear technically intertwined, yet must remain strictly distinguished from one another methodologically.

At the extreme end of quantitative delegation lies the scenario in which generative language models are no longer involved solely in the processing, classification, or interpretation of previously collected data but rather technically generate the empirical units of observation, their response distributions, and their relational behavior themselves. This shift is epistemically significant because delegation now extends from individual analysis steps to the prior constitution of the supposedly empirical material: While in conventional research designs human respondents, documented actions, or observable interactions form the referential starting point for data production, within synthetic samples the statements, distributions, and behavioral sequences from which claims to knowledge are subsequently to be derived are generated from the linguistic, cultural, and social regularities of the model that have been embedded in the training corpus. Argyle *et al.* (2023) demonstrate, using the concept of "algorithmic fidelity," that demographically conditioned language models can approximate the response distributions of selected human subpopulations under certain research conditions. However, such a correspondence does not indicate an equivalence between synthetic response generation and socially situated experience, but rather, initially, merely the model's ability to reproduce statistically learned relationships between demographic attributions, linguistic expression patterns, and already represented attitudinal structures. Using the distinction made by Mou *et al.* (2026), this logic of surrogation can be extended to the modeling of interacting agent-based populations, within which dynamic trajectories and aggregated societal states emerge from technically defined role profiles, memory structures,

action rules, and interaction environments. Ashokkumar *et al.* (2026) further demonstrate that language models based on experimental descriptions can predict the results of numerous social science experiments with remarkable accuracy. However, even such predictive performance does not constitute an independent empirical observation in itself but rather points to the model's ability to transfer expectation structures that are already linguistically and culturally established to new experimental constellations. From this approach arises a simulation space conducive to new insights for the preliminary testing of instruments, theory-driven hypothesis generation, the variation of counterfactual assumptions, and the identification of potentially unexpected interaction patterns. However, its use remains methodologically permissible only on the condition that synthetically generated populations are explicitly identified as model-dependent heuristics, their results are externally validated against data collected from the real world, and their representational limitations regarding language, culture, social position, and historical contingency are disclosed. Otherwise, a synthetic circularity would arise, within which a language model first generates an artificial population from the social regularities contained in its training data and then presents that population's model-induced responses as supposedly new evidence for precisely those regularities. The highest degree of delegation thus denotes not merely a quantitative increase in technical support, but a categorical transition from model-mediated evaluation of empirical reality to its simulation-based surrogation, whose epistemological value remains tied to a continuing separation between a probabilistically generated order of expectations and socially produced observation.

From an ontological perspective, however, simulation must not be equated with the representation of real social experience. As Bisbee *et al.* (2024) argue, central tendencies can indeed be approximated, while distributions, minority positions, and context sensitivity remain unstable. What is simulated are not social subjects, but text-statistically reconstructed orders of expectations regarding subjects. Particularly where cultural stereotypes are deeply and consistently embedded in the training corpus, their plausibility can appear especially high. If such outputs are treated as newly collected evidence, social expectations that are already represented circulate back into research as supposed observations. Synthetic data should therefore be legitimized as simulation-based heuristics, not as an unconditional surrogate for empirical data collection.

Accordingly, the problem of quantitative boundary dissolution lies less in mere susceptibility to error than in the potential decoupling of the operational availability of procedures from methodological judgment. Methodological pseudo-expertise, in this context, does not refer to individual incompetence but rather to an infrastructure-generated constellation within which the performative accessibility of complex procedures renders their validity conditions invisible.

Qualitative Research: Algorithmic Pattern Plausibility and the Irreplaceability of Situated Understanding of the Other

In the qualitative paradigm, epistemic delegation achieves a particularly deep level of intervention because the researcher does not merely apply an external procedure but, through their theoretically informed preconceptions, their positionality, and their capacity for disruption, actively participates in the constitution of the object of knowledge. Category formation, case-based condensation, and theoretical abstraction are therefore not neutral processing steps but rather rule-guided acts of meaning attribution that remain open to the subject matter. As soon as a linguistic model establishes markers of relevance, applies codes, or relates themes, it therefore becomes a secondary interpretive operator within a distributed arrangement of knowledge.

In light of the available comparative studies, this operator can certainly be credited with substantial capabilities. De Paoli (2024) demonstrates that GPT-3.5 can reconstruct essential components of an inductive thematic analysis previously performed by humans; Tai *et al.* (2024) demonstrate, over 160 repeated runs, a comparatively systematic application of deductive coding rules. Studies using open models and health-related interview corpora confirm the scalability, operational consistency of explicitly formulated coding rules, and the ability to analytically organize extensive text corpora in a short time (Balt *et al.,* 2025; Mathis *et al.,* 2024; Qiao *et al.,* 2025; Wachinger *et al.,* 2025). Methodological added value thus manifests itself primarily in accelerated contrast analysis, in the search for divergent text passages, and in the iteration of possible category connections, as opposed to in a machine-authorized determination of meaning.

Within a narrowly defined and explicitly non-anthropomorphizing interpretation, the dialogic work with generative language models described by Hayes (2025) can be defined as a technically mediated arrangement for contrast and disruption, for which the metaphor of the critical friend can claim only heuristic status. The model is not attributed with friendship, critical capacity, or intentionally generated contradiction. Rather, what is described is its prompt-induced ability to generate, from the semantic space of possibilities within its training data, counter-categories, alternative relational frameworks, divergent case consolidations, negative hypotheses, and questions of consistency that have remained unconsidered in the human interpretive process carried out thus far. Following Hayes (2025), the methodological value of such a dialogical constellation therefore lies not in the transfer of interpretive authority to the technical system, but in the controlled generation of difference.

A discrepancy between human and model-generated coding only becomes productive of knowledge when it is treated neither as a machine correction nor as a mere shortcoming of human interpretation, but rather as an epistemic object of inquiry that is traced back to the as-yet-insufficiently-explicit presuppositions of category

formation. Where the model highlights different relations of meaning, shifts category boundaries, or proposes a competing case structure, this does not initially yield additional insight into the subject matter; rather, it creates a justifiable reason to reexamine decisions regarding inclusion and exclusion, theoretical presuppositions, implicit orders of normality, cross-case comparative frameworks, and possibly hastily established assumptions of coherence. In this context, methodological quality therefore results neither from the agreement between the two codings nor from the number of technically generated alternative interpretations, but rather from the documented, theoretically grounded decision, anchored in the original material, regarding why a machine-suggested interpretation is adopted, modified, or rejected. Thus, it is not the scope of the interpretation that is directly broadened, but rather, initially, the range of verifiable interpretive possibilities; the language model itself is unable to render a well-founded judgment regarding their appropriateness to the subject matter.

With the categorical distinction between probabilistic pattern plausibility and situated understanding of the other, the limits of this contrasting process are simultaneously defined. Situated understanding of others is not limited to the semantic processing of as extensive a textual context as possible but rather denotes a relational process of cognition in which theoretical preconceptions, social and institutional positioning, biographical experience, physical presence, practical familiarity with the field, the passage of time, and the formation of judgments based on comparative case studies all contribute to the reconstruction of others' systems of meaning. In contrast, linguistic models operate exclusively on the basis of representationally reduced sets of signs, whose probabilistic relations have not emerged either from direct participation in the sociality under investigation or from a biographically or field-practically acquired entanglement in its systems of meaning. Silence, irony, habitual taken-for-grantedness, performative contradictions, affective tensions, or context-specific power asymmetries, in particular, often derive their reconstructive meaning not from an isolatable linguistic marker, but from the discrepancy between what is said and what is left unsaid, between propositional content and the execution of action, and between the situational order of expectations and observed deviations. These connections alone cannot be derived from semantic coherence, because their meaning remains tied to social, historical, and interactive conditions that are represented in the text only in fragments or completely absent.

Under the production-logical conditions of generative language models, this difference is further exacerbated insofar as probabilistic text generation is geared toward producing coherent, consistent, and linguistically plausible continuations. Precisely where reconstructive research maintains ambiguity, does not hastily resolve discontinuities, and takes the unwieldy, the non-identical, or that which resists the case-internal order of expectations as the starting point for in-depth interpretation, the model's pursuit of coherence can lead to a semantic smoothing of the material. Plausibility then takes on the appearance of interpretive depth, even though only culturally dominant interpretive frameworks that are highly probable in the training dataset are being reproduced.

Hence, the model's epistemic competence remains limited to generating contrasting interpretive proposals and challenging human preconceptions. Subject-specific validity only comes from connecting these proposals back to original sequences, counterexamples, relevant knowledge from the field, theoretical frameworks, and the careful judgment of researchers. As a probabilistically operating epistemic actant, the language model can thus challenge the interpretive process and intensify its requirements for justification; however, situated understanding of others, case-specific validity checks, and scientific responsibility remain and must remain the responsibility of the human members of the epistemic arrangement.

For qualitative research, this results in a structural asymmetry of reflexivity. While human researchers can never make their social and institutional positioning completely transparent, they can make it explicit, allow it to challenge the interpretive collective, and justify it in relation to the data. In contrast, a proprietary model is incapable of providing such biographical or institutional accountability. Its unattainable model-based perspective must therefore be addressed through an external control architecture that interweaves prompt and context logs, contrastive model runs, negative case analyses, cross-references to original sequences, and the documented rejection of machine-generated interpretations. In this process, reflexivity is not transferred to the system but is reorganized as institutionalized observation of the distributed interpretive constellation.

Within this framework, data protection and research ethics are not merely secondary compliance issues. For example, interview transcripts often contain sensitive information that can be reidentified contextually despite formal anonymization. When such data is processed by external model providers, the group of epistemically and technically involved parties changes, which may exceed the scope of the original consent. Data minimization, purpose limitation, secure or local processing, and the control of model inferences that increase sensitivity thus also become conditions for the appropriateness of the research subject matter itself. Recurring problem areas in reviews, such as privacy, fairness, transparency, and diffusion of responsibility, therefore point simultaneously to methodological validity and not merely to legal admissibility (Haltaufderheide & Ranisch, 2024; Yan *et al.,* 2024).

Against this backdrop, qualitative quality, understood here in terms of the research ethical implications that must be observed, cannot be reduced to intercoder agreement with a human reference. Depending on the research

logic, contextual sensitivity, theoretical fruitfulness, case-specific consistency, the ability to handle negative cases, and the traceable transformation of confusion into a well-founded interpretation are decisive.

In this respect, generative AI becomes epistemically productive only insofar as it is incorporated into qualitative analysis as a contrastive interpretive resource within an intertextual constellation, namely one in which the relations among heterogeneous media and information-technological artefacts, as Hepp (2010) emphasizes, "are constituted through explicit references to one another and/or in the associations of the recipients" (p. 276, author's translation). Under these premises, the language model is to be conceptualized neither as an autonomous instance of interpretation nor as a source of self-validating meaning, but as an information-technologically mediated, culturally prestructured, and methodologically delimitable actant whose proposals differentiate the horizon of human judgment without themselves authorizing any of the interpretations generated therein.

Mixed Methods: Generative Conversion Architectures and the Persistence of Epistemological Non-Identity

Within the framework of mixed-methods designs, generative AI derives its specific methodological relevance not from a supposed dissolution of qualitative and quantitative epistemological logics, but from its role as a technically mediated conversion architecture between heterogeneous data, representation, and evidence systems. Here, conversion does not refer merely to the formal transfer of one data format into another. Rather, it refers to a sequence of knowledge-relevant transformation operations within which linguistically open utterances are segmented, semantically condensed, and converted into categorical variables; qualitative categories are related to numerical features; statistical distributions are retranslated into linguistically formulated explanatory hypotheses; and divergent structures of findings are prepared for their comparative arrangement in joint displays. Generative language models can facilitate such transitions with a level of operational density that has been virtually unattainable until now, thanks to their ability to process natural language, to relationally map heterogeneous units of information, and to represent data across formats.

With every conversion operation, however, decisions are simultaneously pre-structured regarding which semantic units are considered comparable, which categories appear quantifiable, which statistical relationships are given a linguistic interpretation, and which divergences remain visible within an overarching presentation of results. Technical transformability and epistemological integrability, therefore, do not denote synonymous properties: while data formats can be converted into one another and findings can be linguistically related to one another, qualitative reconstruction of meaning and quantitative correlation of features remain bound to different understandings of the subject matter, forms of evidence, and conditions of validity.

Based on the multi-stage workflow developed by Combrinck (2024), there is an application procedure demonstrated using real data, within which quantitative and qualitative sub-studies are first processed independently and their results are subsequently synthesized with the aid of generative AI. Methodologically significant is not so much the mere acceleration of individual work steps as the shift of translation tasks that were previously dependent on human input into a prompt-mediated infrastructure: category relations, convergences, complementary relationships, and contradictions can be proposed across a wider range of variations and organized for subsequent human review. What remains unchanged, however, is the obligation to justify why a technically generated relation may be considered a methodologically valid integration and not merely a linguistically plausible connection between heterogeneous findings. Precisely because generative models tend toward coherent condensation, there is a risk of transforming paradigmatically relevant differences into a semantically coherent overall representation, thereby smoothing over the very contradictions from which mixed methods derive a significant portion of their insights.

Evidence that must be distinguished from this is provided, in this context, by empirical mixed-methods studies whose research focus lies in the conditions and consequences of generative AI use, as Verma *et al.* (2026) deduce. In this regard, Verma *et al.* (2026) combine qualitative investigations of the conditions of use by researchers with a quantitative examination of theoretically derived determinants of acceptance and usage within a sequential design. Consequently, it is also understandable why Wang *et al.* (2026) prefer an exploratory sequential design to first identify the subjectively perceived mechanisms of AI-related competency development and, building on these findings, to quantitatively examine their relational structure. In their work, as part of their multi-stage integration approach, Zhang and Dong (2024) combine qualitative case studies with system dynamics and agent-based modeling to form a multi-level evidence architecture, within which institutional orders, social interaction structures, and technological frameworks are not merely juxtaposed in an additive manner but are analyzed in terms of their mutual conditions of constitution, feedback dynamics, and cross-level causal pathways.

From an epistemological perspective, the research frameworks cited do not point to a paradigm convergence brought about by generative AI, but rather to the design-specific complementarity of epistemologically independent strands of evidence. While qualitative analyses reconstruct semantic relationships, procedural mechanisms, and contextual frameworks, quantitative analyses reveal distributions, feature relationships, and model-based regularities; only theoretically grounded meta-inference legitimizes whether and, within what scope of validity, both sets of findings may be integrated into a common epistemological framework as convergent,

complementary, limiting, or divergent.

The above-mentioned conceptual framework defines infrastructural convergence as the IT-mediated reduction of the complexity of translation, relation-building, and representation that arises from the methodological intertwining of heterogeneous data streams. Persistent epistemological non-identity, by contrast, refers to the fundamental irreconcilability of divergent constitutions of objects, orders of evidence, and claims to validity, whose mere translation into a common machine-based form of representation neither establishes commensurability nor legitimizes an integrative epistemological context. In this heuristic derivation, generative AI is capable of generating integration options, identifying epistemically relevant transition zones, and pre-structuring joint displays; however, the meta-inferential decision regarding validity, namely which relational framework constitutes a context of knowledge appropriate to the subject matter, remains bound to theoretical justifiability, paradigmatically grounded judgmental competence, and institutionally attributable human ultimate responsibility.

However, the specific benefit lies not in the elimination of paradigmatic differences, but in the reduction of operational translation and relational costs. Models can highlight convergences, complementarities, and divergences between data strands, suggest qualitative cases for further exploration based on quantitative distributions, or derive preliminary items from exploratively developed categories. By doing so, they condense those transitional zones where mixed methods must demonstrate their integrative power. Technical transformability, however, does not in itself generate epistemological commensurability: a correlation and a biographical narrative remain bound to different understandings of the subject matter, forms of evidence, and conditions of validity.

It is precisely this generative capacity for coherent synthesis that can therefore pose a methodological risk. When divergent findings are linguistically smoothed into a plausible overarching narrative, the very contradiction from which mixed-methods research derives its specific insights may be lost. Under the influence of a preference for algorithmic coherence, integration then shifts from a theoretically grounded meta-inference to a representational artifact. Representational coherence must therefore be strictly distinguished from epistemological integration.

Accordingly, high-quality AI-mediated mixed-methods research requires a three-part validation framework. First, both data streams must be examined according to the criteria relevant to their respective paradigms. Next, the logic of integration must be disclosed in terms of timing, level, and theoretical function. Finally, the model-mediated conversion also requires its own scrutiny: What relationships did the system propose, which divergences were overlooked, which meta-inferences were rejected, and by which original data or counter-analyses was the adopted connection substantiated? Such a meta-inference audit is necessary to prevent technical convergence from devolving into epistemic deregulation. Indeed, while the LLM is capable of generating integration options, the legitimizing act of translating heterogeneous forms of evidence into a well-founded context of validity remains non-delegable.

Epistemic Delegation as a Gradual Conversion Architecture

As the findings to date suggest, a binary distinction between research with or without AI fails to holistically capture the fact that applications that appear similar from an information technology perspective can have vastly different effects on the constitution of scientific claims to validity. In the author's view, the decisive analytical factor is therefore not merely the presence of the system but rather the question of which epistemic function is outsourced and to what extent, which subsequent decisions it influences, and whether its results remain subject to correction by independent bodies. From the synthesis of the evidence, a delegation matrix has emerged that reconstructs this shift as a five-stage conversion architecture.

The graduated architecture presented in Table 1

Table 1: A Gradual Architecture of Epistemic Delegation in AI-Mediated Research

Level	Form of Delegation	Typical Applications	Epistemic Risk	Required Evaluation
1	Operational-Infrastructural Delegation	Transcription, formatting, syntax checking, translation	Local transmission errors, uncontrolled data sharing	Random-sample verification, validated processing, documented error profile
2	Heuristic-Generative Pre-Structuring	Search spaces, hypothesis variants, code and item drafts, visualization options	Apparent plausibility, premature narrowing of the search space	Independent source verification, variant comparison, technical justification for selection
3	Analytical Co-Processing	Annotation, deductive coding, statistical code, clustering, open-ended responses	Non-classical measurement errors, prompt and version sensitivity	External reference data, robustness testing, group-specific error analysis

4	Interpretive Co-Construction	Inductive theme formation, theoretical grounding, meta-inference, interpretation of results	Semantic smoothing, hegemonic patterns of interpretation, suppression of negative cases	Sequence backtracking, counter-interpretation, model contrast, reasoned final decision
5	Empirical Surrogation	Synthetic respondents, simulated interviews, simulated experiments	Circularity, substitution of social voices, apparent evidence	Designated as a simulation; no substitution of evidence without external empirical validation

Source: Author (2026).

is intended neither to establish a research-ethical hierarchy in the deontological sense nor to imply that progressively higher degrees of epistemic delegation constitute a linear advancement in methodological innovativeness. Methodologically explicit applications of interpretive co-construction may be more responsible than covert heuristic pre-structuring. Rather, a principle of proportionality of cumulative consequences is formulated: the more a delegated output contributes to the formation of subsequent decisions, the more rigorous independent validation, technical provenance, and human cross-checking must be in place. While a local transcription mistake impacts specific sections, an unverified thematic structure shapes the overall interpretive area; using empirical substitutes eventually changes what the research is actually about.

In this sense, the matrix above also clarifies the frequently used metaphor of the co-researcher, especially in the context of a virtualized consortium-based research approach. At the first two levels, it overestimates the system's epistemic function; at levels three and four, while it captures the system's operational involvement, it obscures the structural asymmetry between efficacy and accountability. At level five, it risks confusing simulated orders of expectation with the empirical presence of social subjects. From the perspective of the philosophy of science, the term probabilistic epistemic actant is therefore more appropriate: the system performs possibilities of knowledge without itself becoming capable of justification or accountability.

Reflexive Controllability as an Expanded Framework for AI-Mediated Research

Generative systems do not render the classical criteria of quantitative and qualitative research obsolete; however, they do alter the conditions under which these criteria can be applied. Objectivity, reliability, and validity, along with subject-appropriateness, reflexivity, and intersubjective comprehensibility, now also need to include a system that produces results based on probabilities, varies by version, and is influenced by private technology. Complete identity of results often remains unattainable under these conditions and would, even where technically approximable, remain insufficient as an isolated criterion for methodological quality. Within the framework developed here, reflexive controllability designates the overarching governance principle according to which AI-mediated epistemic delegations must remain amenable to human reconstruction, external validation, and justified intervention. Reflexive verifiability, by contrast, specifies the corresponding methodological requirement: third parties must be able to reconstruct the technical and epistemic conditions under which an output was generated, the degree of variance permitted or empirically examined, the manner in which that output entered subsequent decisions, and the points at which its validity was delimited through documented human judgment. The methodological operationalization of this overarching regime comprises eight interrelated control dimensions.

This control regime can be differentiated into eight interrelated dimensions. Model- and version-specificity requires documentation of the provider, model, access date, interface, and relevant system settings; the unspecific statement that ChatGPT was used does not constitute a replicable procedure. Prompt and context transparency encompasses system instructions, few-shot examples, the sequence of interactions, and the scope of the material provided, without exposing sensitive data to unprotected disclosure. Through data provenance and purpose limitation, the origin, processing, anonymization, permissible processing, and potential feedback to the provider become visible as components of the methodological decision-making framework.

In this scenario, externally anchored validation refers to the methodological comparison of model-generated outputs with intersubjectively verified reference codings, original datasets, and theoretically justified validity criteria, thereby excluding self-confirming loops inherent in the model as impermissible circular evidence production. Correspondingly, it must be determined through replicated runs, alternating prompt architectures, contrasting model configurations, and sensitivity tests whether a finding retains its validity stability in the face of the contingency of its technical genesis (Lin, 2026). Finally, the examination of group and context sensitivity requires a differentiation of error profiles according to language, cultural coding, and social position, since aggregated performance metrics can level out structural representational asymmetries and thereby reproduce epistemic injustices (Olaniyan *et al.,* 2026; Vindigni, 2025a).

Last but not least, reflexive decision provenance

documents not only inputs and outputs but also the grounds on which researchers have adopted, modified, or rejected machine-generated suggestions, specifically for the purpose of algorithmically operationalized plausibility checks. These must be coupled with a clear assignment of responsibility and liability: data protection, choice of methods, interpretation of results, and publication remain the responsibility of identifiable individuals and institutions, because a language model cannot assume responsibility for either authorship or peer review (de Cassai *et al.,* 2026). It is only through this relational connection that technical log data becomes a scientifically meaningful audit trail.

At this point, it becomes clear why transparency thus shifts from being an optional disclosure to a constitutive condition of scientific validity. When LLM use is undocumented, it creates more than a research ethics problem; it removes key analytical decisions from methodological scrutiny. Conversely, a complete appendix of prompts does not in itself constitute a chain of reasoning: thousands of logged interactions can be technically complete yet epistemically empty. Reflexive verifiability therefore refers to the institutionalized connection between technical provenance, subject-specific validation, and accountable judgment practices.

Table 2: Paradigmatic Functional Profiles and Priority Control Requirements

Research Logic	Primary AI Function	Key Added Value	Main Risk	Priority Quality Assessment
Quantitative	Coding, classification, item and data generation	Scaling and Access to Complex Procedures	Pseudo-expertise, measurement errors, synthetic circularity	Reference code, assumption verification, error and sensitivity analysis
Qualitative	Coding, theme formation, counter-interpretation	Variation in Perspective and Corpus Analysis	Decontextualization, interpretive smoothing, bias	Link to original data, negative cases, reflection log
Mixed Methods	Linking, transformation, joint display	Reduced Translation Costs and Systematic Comparison	Premature coherence, loss of paradigmatic difference	Separate strand validation, explicit integration logic, meta-inference audit
Methodology Training	Simulation, feedback, scaffold	Low-Threshold Exploration and Personalized Feedback	Illusion of competence, outsourcing of basic judgmental practice	Performance-based verification, obligation to provide justification, AI-free core competency

Source: Author (2026).

Research Governance: Infrastructure Dependency, Access Asymmetry, and Accountability

The incorporation of generative AI into research practice should not be construed as an isolated tool-related decision by individual researchers; rather, it constitutes an institutionally mediated reconfiguration of research instrumentation. Outside of scientific regulatory frameworks, proprietary models are trained, parameterized, and continuously modified; their training foundations, moderation rules, and version-specific interventions remain only partially transparent. At the same time, their use is tied to licensing models, organization-specific access rights, and geopolitically unevenly distributed computing resources. Therefore, the choice of model also determines which data may be processed, which operations are technically feasible, and to what extent a procedure remains reconstructible ex post. In this respect, research instrumentation takes on the character of a governance decision.

The linkage between technical mediation frameworks and access conditionality, as reconstructed by Vindigni (2026a) for platform-based higher education infrastructures, can be applied to AI-mediated research. This linkage is understood as the institutionally unequal distribution of the convertibility of formal system availability into actual research capability. High-performance enterprise environments, locally controlled models, and secured computing infrastructures afford different methodological possibilities than free, linguistically restricted, or data-opaque systems. Consequently, formal availability does not in itself signify epistemically effective accessibility: Where methodological participation remains tied to institutional financial power, technical sovereignty, and infrastructural resources, it is not merely usage options but the very prerequisites for scientific knowledge production that are allocated asymmetrically.

By superimposing infrastructural access conditionality with asymmetries in the corpus, language, and the order of knowledge, access conditionality coalesces into a multidimensional epistemic architecture of selection. The overrepresentation of resource-rich languages, hegemonic disciplines, and globally visible publication cultures results in unequal probability orders, within which dominant interpretive frameworks are granted greater model-based connectivity than marginalized forms of knowledge. Moreover, cultural and gender-related distortions thus do not represent contingent errors in an otherwise neutral model, but rather the

effects of upstream selection orders, visibility regimes, and weighting logics (Abdurahman *et al.,* 2024; Olaniyan *et al.,* 2026; Vindigni, 2025b). For educational and social research, therefore, the problem extends beyond the erroneous classification of underrepresented groups. Where their bodies of experience are already marginalized within the epistemic framework, their articulations are subject to a further conversion into model-compliant categories, through which existing invisibilities can be technically reproduced and scientifically reauthorized. In the author's view, methodological bias and epistemic injustice are thus not merely correlated problems but rather two manifestations of the same selection order, which is pre-structured in terms of infrastructure and representation.

At the organizational level, this framework must also be addressed through research governance that integrates risk-appropriate systems, data protection impact assessments, usage rights tiered according to data sensitivity and the degree of delegation, auditable repositories for audit trails, and responsibilities for model changes, incidents, and the suspension of procedures that can no longer be controlled. In addition, participatory processes must be established for those groups about which statements are generated using model-based categories. Within this analytical constellation, the human-centered orientation codified in DIN EN ISO 9241 provides a normatively compatible frame of reference, provided that usability is not reductively equated with mere operational convenience but is conceptualized, in accordance with DIN EN ISO 9241-11, as the context-dependent attainment of specified goals in terms of effectiveness, efficiency, and satisfaction, while simultaneously being related to the principles of task suitability, self-descriptiveness, and controllability specified in DIN EN ISO 9241-110 (Vindigni, 2024a, 2024b). According to this expanded interpretation, research quality is not a retrospective attribute of the research output; rather, it is a property of the entire socio-technical configuration.

Methodology Development: Between Cognitive Relief and the Erosion of Epistemic Autonomy

In the context of methodological training in higher education, the use of generative language models creates a didactic double bind between increasing operational accessibility and the potential externalization of epistemic agency. While lowered procedural thresholds, immediate feedback, and methodological scope for variation with a high density of iterations consolidate access to complex research operations, the same logic of reducing cognitive load can supplant the cognitive and metacognitive processes from which methodological judgment, the testing of assumptions, and the reasoned selection of procedures develop. Correspondingly, recent surveys document a rapidly increasing use of AI, accompanied by considerable heterogeneity regarding trust in the system, risk perception, and performatively demonstrable AI competence (Al-Abdullatif, 2024; Andersen *et al.,* 2025; Haroud & Saqri, 2025; Saúde *et al.,* 2024). Consolidated review evidence points equally to didactic potential as well as dependency risks, conflicts of integrity, and a possible decoupling of externally visible performance from internally available methodological understanding (Qian, 2025; Weng *et al.,* 2024; Yan *et al.,* 2024).

In the author's view, the normative standard for AI-mediated methodological training cannot, therefore, be the degree to which operational automation is maximized, but rather, exclusively, the safeguarding of human epistemic agency, understood as the non-delegable capacity for goal-setting, validity assessment, reasoned intervention, and the assumption of responsibility. In line with this, Wu *et al.* (2025) conceptualize the human-AI relationship as a symbiotic learning partnership whose epistemic quality remains dependent on human intentionality, continuous oversight, and corrective judgment. Yan *et al.* (2025) also point out that efficiency gains do not represent epistemically neutral increases in productivity, as they are accompanied by reconfigurations of cognitive, metacognitive, and epistemically autonomous processes. Within academic learning spaces, this gives rise to a new order of negotiation regarding epistemic authority, within which the validity of model-generated statements relative to scholarly literature, knowledge legitimized by instructors, and one's own judgmental practices must be justified (Jose *et al.,* 2025).

Consequently, disciplinary validity in higher education methodological training is achieved only through the relational interweaving of basic methodological competence, critical AI literacy, and reasoning-oriented application into an indivisible competence architecture, within which the operational use of the system remains subordinate to the independent reconstruction of its epistemic premises. The starting point is the ability to explicate measurement assumptions, interpretive premises, and integration logics even without technical assistance, because only a previously developed methodological judgment competence can distinguish between permissible procedural relief and the unnoticed delegation of decisions relevant to knowledge. Within this framework, methodological autonomy does not imply abstinence from generative models, but rather the ability to test, modify, or justifiably reject their outputs against original data, counterexamples, alternative theoretical explanations, and discipline-specific validity criteria. Along these lines, a sufficient understanding of the technical logic underlying the model's generation is required so that training dependency, probabilistic variance, bias, and platform interests are classified not as external contextual factors but as constitutive determinants of the genesis of results (Dilek *et al.,* 2025; Tzirides *et al.,* 2024).

From a didactic perspective, this necessitates a reorientation of performance assessment away from the isolated evaluation of manifest results and toward the reconstructible quality of the methodological reasoning through which those results are generated, critically examined, and ultimately justified. The primary focus of

evaluation should not be on whether the output is formally convincing, but rather on whether learners can provide a reasoned account of their choice of model, the limits of delegation, their verification strategy, the alternatives they have ruled out, and any remaining uncertainty. Contrastive arrangements are particularly well-suited for this purpose: human and machine codings are not reduced to mere agreement but are compared in terms of their implicit orders of relevance; generated analysis code must be examined for violations of assumptions; synthetic samples must be tested against real distributions; and divergent mixed-methods findings must be defended against the generative impulse toward premature harmonization. Dengel *et al.* (2023) demonstrate that LLMs themselves can become the subject of qualitative inquiry and thus a medium for simulated methodological reflection.

In this regard, it is advisable, particularly concerning higher education, not to design such training as product-specific, prompt-based instruction. Against this backdrop, interfaces and model families change faster than curricular frameworks; what remains permanently relevant is the ability to assess the delegability, scope, and controllability of a research step. In this regard, the central future skill is not the smoothest possible operation of generative systems, but rather the well-founded maintenance of epistemic autonomy under conditions of technology-mediated research. Here, methodological competence intersects with the question of who can substantively participate in future knowledge production and how digital assessment and learning infrastructures should be designed to be fair, accessible, and accountable (Vindigni, 2026b, 2026c).

Limitations of Scope and Subsequent Research Needs

The validity of this synthesis is initially limited by the extraordinary dynamics of change inherent in its subject matter. Performance, error profiles, and governance conditions are tied to specific model versions, interfaces, and dates of access; a validation that is robust today may lose its transferability due to silent system changes. Added to these limitations is the dominance of English-language and Western-institutionalized research contexts. From these factors arises the risk that the evidence base will inadvertently perpetuate the very linguistic and cultural asymmetries in representation that it critically seeks to examine. Finally, numerous comparative studies remain domain- and task-specific: high quality in political text annotation does not allow for conclusions regarding reconstructive interview interpretation, psychometric scale development, or integrative meta-inference.

In addition, the selected design imposes a limit on the scope. In this critical-integrative review, the focus is on theoretical synthesis and paradigmatic contrast, not on the exhaustiveness of a systematic review with meta-analytic aspirations. Its value lies in the reconstruction of recurring problem areas and the development of an overarching framework for evaluation; it does not involve statements about average effect sizes or a definitive ranking of specific models.

It is therefore advisable for future research to focus on the temporal, cultural, and method-specific contingencies of the subject matter itself. In this regard, pre-registered, multilingual, and model-comparative replications are needed that identify not only global performance metrics but also group-specific error patterns and their implications for inferential results. Qualitative evaluation designs must go beyond code agreement to examine context sensitivity, the ability to handle negative cases, and theoretical fruitfulness. For mixed methods, it must be experimentally investigated whether generative systems preserve divergence as a resource for knowledge or systematically transform it into artificial coherence. In the long term, longitudinal studies are needed to address how repeated use of AI alters methodological understanding, epistemic agency, and disciplinary diversity.

In this regard, scientific monoculturalization does not first arise at the level of instrumental standardization associated with the recurrent use of identical models; rather, it is already initiated at an epistemically antecedent stage, at which model-privileged modes of problem formulation, relevance selection, and plausibility assessment become normalized largely beneath the threshold of methodological reflexivity and thereby sediment into a collectively shared cognitive infrastructure of research (Messeri & Crockett, 2024).

The selection process that guides knowledge shifts even before an explicit methodological decision is made by semantically privileging certain research questions, making alternative approaches to the subject matter appear less compatible, and reshaping epistemic uncertainty through linguistic coherence. What is at stake, therefore, is not merely the diversity of the research instruments employed but the plurality of the theoretical presuppositions and epistemic differences from which scientific contradiction, disciplinary self-correction, and subject-appropriate knowledge emerge.

CONCLUSION

Overall, the analysis indicates that generative AI does not negate the distinct logics of validity in quantitative, qualitative, and mixed-methods research, but it does alter the distribution of operations relevant to knowledge generation. Quantitative applications enhance scalability and procedural access but can also generate methodological pseudo-competence, measurement errors, and synthetic circularity. In this regard, qualitative applications open up spaces for contrastive interpretation without replacing situated case understanding; mixed-methods applications facilitate the integration of heterogeneous data but do not legitimize hasty epistemic standardization. The five-level matrix of epistemic delegation illustrates that the intensity of scrutiny, documentation, and human control must increase cumulatively with the scope of subsequent decisions. In this context, reflexive controllability combines technical provenance, external validation, robustness and

sensitivity tests, documented decision-making processes, and a clear assignment of responsibility. Generative AI should therefore be defined as a probabilistic epistemic actant, not as a subject of knowledge capable of bearing responsibility. Scientific validity arises only when its infrastructure- and probability-based contributions are examined in accordance with the relevant paradigm, delimited, and accounted for by identifiable individuals and institutions.

Declarations

The author has no acknowledgements to declare.

Conflict of Interest: The author declares that no conflict of interest or competing financial interest exists.

Funding: No external or earmarked institutional funding was received for the conduct or preparation of this study.

Ethics Statement: As a critical-integrative evidence synthesis based exclusively on previously published material, the study involved neither the recruitment of human participants nor the collection of new participant-related or personal data. Ethical approval and informed consent were therefore not applicable.

Author Contributions: The author assumed sole responsibility for the conceptualization and methodological design of the study, the literature search and critical-integrative evidence synthesis, the development of the analytical framework, and the preparation and revision of the manuscript.

Research Integrity Statement: The study was conducted and reported in accordance with the applicable principles of research integrity, methodological transparency, and accountable documentation, particularly the Guidelines for Safeguarding Good Research Practice: Code of Conduct issued by the Deutsche Forschungsgemeinschaft (DFG, 2025).

REFERENCES

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus, 3*(7), pgae245. https://doi.org/10.1093/pnasnexus/pgae245

Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., & Dehghani, M. (2025). A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science, 8*(2). https://doi.org/10.1177/25152459251325174

Al-Abdullatif, A. M. (2024). Modeling teachers' acceptance of generative artificial intelligence use in higher education: The role of AI literacy, intelligent TPACK, and perceived trust. *Education Sciences, 14*(11), 1209. https://doi.org/10.3390/educsci14111209

Andersen, J. P., Degn, L., Fishberg, R., Graversen, E. K., Horbach, S. P. J. M., Schmidt, E. K., Schneider, J. W., & Sørensen, M. P. (2025). Generative artificial intelligence (GenAI) in the research process—A survey of researchers' practices and perceptions. *Technology in Society, 81*, 102813. https://doi.org/10.1016/j.techsoc.2025.102813

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Ashokkumar, A., Hewitt, L., Ghezae, I., & Willer, R. (2026). Large language models can predict the results of social science experiments. *Nature, 656*(8126), 115–122. https://doi.org/10.1038/s41586-026-10742-x

Ashwin, J., Chhabra, A., & Rao, V. (2026). Using large language models for qualitative analysis can introduce serious bias. *Sociological Methods & Research, 55*(3), 795–839. https://doi.org/10.1177/00491241251338246

Balt, E., Salmi, S., Bhulai, S., Vrinzen, S., Eikelenboom, M., Gilissen, R., Creemers, D., Popma, A., & Mérelle, S. (2025). Deductively coding psychosocial autopsy interview data using a few-shot learning large language model. *Frontiers in Public Health, 13*, 1512537. https://doi.org/10.3389/fpubh.2025.1512537

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., Shiffrin, R. M., Gershman, S. J., Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z., & Schulz, E. (2025). How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences, 122*(5), e2401227121. https://doi.org/10.1073/pnas.2401227121

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics, 5,* 277–280. https://doi.org/10.1038/s42254-023-00581-4

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis, 32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

Blanchard, S. J., Duani, N., Garvey, A. M., Netzer, O., & Oh, T. T. (2025). New tools, new rules: A practical guide to effective and responsible generative AI use for surveys and experiments in research. *Journal of Marketing, 89*(6), 119–139. https://doi.org/10.1177/00222429251349882

Chang, Y.-C., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology, 15*(3),

1–45. https://doi.org/10.1145/3641289

Combrinck, C. (2024). A tutorial for integrating generative AI in mixed methods data analysis. *Discover Education, 3*, 116. https://doi.org/10.1007/s44217-024-00214-7

de Cassai, A., Dost, B., Augoustides, J. G. T., Azamfirei, L., Alanoğlu, Z., Azi, L. M., Calvache, J. A., Cerny, V., De Hert, S., Eldawlatly, A., Farber, M. K., Sobreira-Fernandes, D., Fettiplace, M. R., Galante, D., Garg, R., Goldstein, H. V., Abad-Gurumeta, A., Gupta, L., Hemmings, H. C., … Zdanowski, S. (2026). Responsible use of large language models in manuscript authorship, peer review, and editorial processes: A Delphi consensus among editors-in-chief of anaesthesia and pain medicine journals (RULE-AP). *British Journal of Anaesthesia, 136*(5), 1625–1633. https://doi.org/10.1016/j.bja.2026.01.029

Dengel, A., Gehrlein, R., Fernes, D., Görlich, S., Maurer, J., Pham, H. H., Großmann, G., & Eisermann, N. D. G. (2023). Qualitative research methods for large language models: Conducting semi-structured interviews with ChatGPT and BARD on computer science education. *Informatics, 10*(4), 78. https://doi.org/10.3390/informatics10040078

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review, 42*(4), 997–1019. https://doi.org/10.1177/08944393231220483

Deutsche Forschungsgemeinschaft. (2025). *Guidelines for safeguarding good research practice: Code of conduct* (Version 3). https://doi.org/10.5281/zenodo.14281892

Dilek, M., Baran, E., & Aleman, E. (2025). AI literacy in teacher education: Empowering educators through critical co-discovery. *Journal of Teacher Education, 76*(3), 294–311. https://doi.org/10.1177/00224871251325083

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36,* 15. https://doi.org/10.1007/s13347-023-00621-y

Halterman, A., & Keith, K. A. (2026). Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *Political Analysis, 34*(2), 188–204. https://doi.org/10.1017/pan.2025.10017

Haltaufderheide, J., & Ranisch, R. (2024). The ethics of ChatGPT in medicine and healthcare: A systematic review on large language models (LLMs). *npj Digital Medicine, 7*, 183. https://doi.org/10.1038/s41746-024-01157-x

Haroud, S., & Saqri, N. (2025). Generative AI in higher education: Teachers' and students' perspectives on support, replacement, and digital literacy. *Education Sciences, 15*(4), 396. https://doi.org/10.3390/educsci15040396

Hayes, A. S. (2025). "Conversing" with qualitative data: Enhancing qualitative research through large language models (LLMs). *International Journal of Qualitative Methods, 24*, 1–19. https://doi.org/10.1177/16094069251322346

Hepp, A. (2010). *Cultural Studies und Medienanalyse: Eine Einführung [Cultural studies and media analysis: An introduction]* (3rd ed.). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92190-7

Jose, B., Cleetus, A., Joseph, B., Joseph, L., Jose, B., & John, A. K. (2025). Epistemic authority and generative AI in learning spaces: Rethinking knowledge in the algorithmic age. *Frontiers in Education, 10,* 1647687. https://doi.org/10.3389/feduc.2025.1647687

Keane, D., & McNaughton, R. B. (2026). Using generative AI to enhance psychometric scale development in market research. *International Journal of Market Research, 68*(2), 194–218. https://doi.org/10.1177/14707853251384769

Leslie, D. (2025). Does the sun rise for ChatGPT? Scientific discovery in the age of generative AI. *AI and Ethics, 5*(4), 3439–3444. https://doi.org/10.1007/s43681-023-00315-3

Lin, H., & Zhang, Y. (2026). Navigating the risks of using large language models for text annotation in social science research. *Social Science Computer Review, 44*(3), 403–427. https://doi.org/10.1177/08944393251366243

Lin, Z. (2023). Why and how to embrace AI such as ChatGPT in your academic life. *Royal Society Open Science, 10*(8), 230658. https://doi.org/10.1098/rsos.230658

Lin, Z. (2026). A validity-guided workflow for robust large language model research in psychology. *Behavior Research Methods, 58*(8), Article 216. https://doi.org/10.3758/s13428-026-03073-2

Martin Kowal, J., Hurley Bryant, K., Segall, D., & Kantrowitz, T. (2025). Harnessing generative AI for assessment item development: Comparing AI-generated and human-authored items. *International Journal of Selection and Assessment, 33*(3), e70021. https://doi.org/10.1111/ijsa.70021

Mathis, W. S., Zhao, S., Pratt, N., Weleff, J., & De Paoli, S. (2024). Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine, 255,* 108356. https://doi.org/10.1016/j.cmpb.2024.108356

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627,* 49–58. https://doi.org/10.1038/s41586-024-07146-0

Mou, X., Ding, X., He, Q., Wang, L., Liang, J., Zhang, X., Sun, L., Lin, J., Zhou, J., Huang, X., & Wei, Z. (2026). From individual to society: A survey on social simulation driven by large language model-based agents. *ACM Computing Surveys, 58*(11), 1–41. https://doi.org/10.1145/3800683

Olaniyan, Y. D., Martins, M. O., & Al Maqrashi, R.

H. (2026). Generative artificial intelligence and epistemic (in)justice: Perspectives from higher education students in the Global North and South. *Frontiers in Human Dynamics, 8,* 1790324. https://doi.org/10.3389/fhumd.2026.1790324

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated annotation with generative AI requires validation [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2306.00176

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods, 56*(6), 5754–5770. https://doi.org/10.3758/s13428-023-02307-x

Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science, 12*, 241776. https://doi.org/10.1098/rsos.241776

Prandner, D., Wetzelhütter, D., & Hese, S. (2025). ChatGPT as a data analyst: An exploratory study on AI-supported quantitative data analysis in empirical research. *Frontiers in Education, 9*, 1417900. https://doi.org/10.3389/feduc.2024.1417900

Qian, Y. (2025). Pedagogical applications of generative AI in higher education: A systematic review of the field. *TechTrends, 69*(5), 1105–1120. https://doi.org/10.1007/s11528-025-01100-1

Qiao, S., Fang, X., Wang, J., Zhang, R., Li, X., & Kang, Y. (2025). Generative AI for thematic analysis in a maternal health study: Coding semistructured interviews using large language models. Applied Psychology: *Health and Well-Being, 17*(3), e70038. https://doi.org/10.1111/aphw.70038

Saúde, S., Barros, J. P., & Almeida, I. (2024). Impacts of generative artificial intelligence in higher education: Research trends and students' perceptions. *Social Sciences, 13*(8), 410. https://doi.org/10.3390/socsci13080410

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods, 23*, 1–14. https://doi.org/10.1177/16094069241231168

Terry, J., Strait, G., Alsarraf, S., Weinmann, E., & Waychoff, A. (2025). Artificial intelligence in scale development: Evaluating AI-generated survey items against gold standard measures. *Current Psychology, 44*(20), 16339–16350. https://doi.org/10.1007/s12144-025-08240-w

Törnberg, P. (2025). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review, 43*(6), 1181–1195. https://doi.org/10.1177/08944393241286471

Tzirides, A. O., Zapata, G. C., Kastania, N. P., Saini, A. K., Castro, V., Ismael, S. A., You, Y.-L., Santos, T. A. D., Searsmith, D., O'Brien, C., Cope, B., & Kalantzis, M. (2024). Combining human and artificial intelligence for enhanced AI literacy in higher education. *Computers and Education Open, 6*, 100184. https://doi.org/10.1016/j.caeo.2024.100184

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature, 614*, 224–226. https://doi.org/10.1038/d41586-023-00288-7

Verma, S., Kashive, N., & Gupta, A. (2026). Examining predictors of generative-AI acceptance and usage in academic research: A sequential mixed-methods approach. *Benchmarking: An International Journal, 33*(6), 1821–1849. https://doi.org/10.1108/BIJ-07-2024-0564

Vindigni, G. (2024a). Enhancing human-computer interaction in socially inclusive contexts: Flow heuristics and AI systems in compliance with DIN EN ISO 9241 standards. *European Journal of Contemporary Education and E-Learning, 2*(4), 115–139. https://doi.org/10.59324/ejceel.2024.2(4).10

Vindigni, G. (2024b). Optimierung der HCI in sozial inklusiven Kontexten: Flow-Heuristiken und KI-Systeme gemäß DIN EN ISO 9241. *Zeitschrift für Sozialmanagement, 22*(2), 111–123.

Vindigni, G. (2025a). Data-driven disparities: How AI applications in education may perpetuate or mitigate inequality. *European Journal of Science and Modern Technologies, 1*(4), 4–54. https://doi.org/10.59324/ejsmt.2025.1(4).02

Vindigni, G. (2025b). Gender bias and cultural misrepresentation in AI: A critical inquiry into cross-cultural communication and algorithmic design. *European Journal of Applied Science, Engineering and Technology, 3*(3), 51–72. https://doi.org/10.59324/ejaset.2025.3(3).06

Vindigni, G. (2026a). Platformized private higher education institutions in the EHEA: Instructional infrastructure, access conditionality, and platform governance. *European Journal of Contemporary Education and E-Learning, 4*(4), 275–356. https://doi.org/10.59324/ejceel.2026.4(4).13

Vindigni, G. (2026b). Beyond compliance: Future Skills 2030, digital accessibility and democratic participation in higher education—A scoping review. *European Journal of Contemporary Education and E-Learning, 4*(4), 56–137. https://doi.org/10.59324/ejceel.2026.4(4).05

Vindigni, G. (2026c). Digital exams as socio-technical interaction systems: An ISO 9241-based framework for validity, fairness and governance. *American Journal of Education and Technology, 5*(3), 61–93. https://doi.org/10.54536/ajet.v5i3.8137

Wachinger, J., Bärnighausen, K., Schäfer, L. N., Scott, K., & McMahon, S. A. (2025). Prompts, pearls, imperfections: Comparing ChatGPT and a human researcher in qualitative data analysis. *Qualitative Health Research, 35*(9), 951–966. https://doi.org/10.1177/10497323241244669

Wang, J., Bai, B., & An, Q. (2026). Enhancing college

students' AI literacy through generative AI use: A mixed-methods investigation. *Frontiers in Psychology, 17*, 1728785. https://doi.org/10.3389/fpsyg.2026.1728785

Weng, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. F. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology, 40*(6), 37–55. https://doi.org/10.14742/ajet.9540

Wu, J.-Y., Lee, Y.-H., Chai, C. S., & Tsai, C.-C. (2025). Strengthening human epistemic agency in the symbiotic learning partnership with generative artificial intelligence. *Educational Researcher, 54*(6), 358–368. https://doi.org/10.3102/0013189X251333628

Yan, L., Pammer-Schindler, V., Mills, C., Nguyen, A., & Gašević, D. (2025). Beyond efficiency: Empirical insights on generative AI's impact on cognition, metacognition and epistemic agency in learning. *British Journal of Educational Technology, 56*(5), 1675–1685. https://doi.org/10.1111/bjet.70000

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Zhang, Y., & Dong, C. (2024). Exploring the digital transformation of generative AI-assisted foreign language education: A socio-technical systems perspective based on mixed methods. *Systems, 12*(11), 462. https://doi.org/10.3390/systems12110462

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics, 50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Supplementary Evidence Matrix: Generative Ai In Social And Educational Research

Supplementary Table S1. Publication-level evidence matrix of the 58 publications included in the analytical core corpus, comprising study design, data basis, AI system or model version, research operation, validation criteria, principal findings, and assignment to the five levels of epistemic delegation (XLSX).