

The Domain Expert as Analyst: Characterizing a Workflow Paradigm for AI- Mediated Applied Analysis, with a Worked Case in College Baseball

Lionel Mew
lmew@richmond.edu

Sean Scanlon
Sean.Scanlon@richmond.edu

University of Richmond
University of Richmond, VA 23173, USA

Abstract

A workflow paradigm has emerged in which a domain expert without formal statistical training can conduct defensible applied analytical work by using a large language model as an analytical partner. This paper characterizes the paradigm, articulates the specific competences it requires of the domain expert, and demonstrates its viability through a worked case study. Four competences are proposed as necessary to the paradigm's operation: specification, verification, moderation, and interpretation. The paradigm depends on genuine domain expertise; it is a force multiplier for expertise, not a substitute for it. It also depends on deliberate rather than credulous engagement with known LLM failure modes, including sycophancy, confabulation, and literal-interpretation-that-misses-statistical-method. The case study, an applied analysis of what distinguishes elite four-seam fastballs in NCAA Division I Power 4 conference play, illustrates the paradigm operating end-to-end over a five-week compressed capstone. Implications for pedagogy, hiring, and the boundary between practicing analytical work and methodological judgment are discussed.

Keywords: AI-mediated analysis, applied analytics, domain expertise, workflow paradigm, human-AI collaboration, analytical competences

The Domain Expert as Analyst: Characterizing a Workflow Paradigm for AI-Mediated Applied Analysis, with a Worked Case in College Baseball

Lionel Mew and Sean Scanlon

1. INTRODUCTION

Something has changed in how analytical work gets done, and the change has moved faster than the scholarly literature has yet acknowledged. The evidence is clearest where the change has been studied most directly. Developers routinely use large language models to write, debug, and refactor code that once required deeper implementation knowledge, with measurable effects on task completion (Peng et al., 2023; Vaithilingam et al., 2022). Zamfirescu-Pereira et al. (2023), observing non-AI-experts attempting to direct LLM behavior toward tasks they could not otherwise accomplish, found that they approached the work in ways that differed systematically from those of trained practitioners: implicitly rather than explicitly, opportunistically rather than systematically, and with more confidence than their results warranted. Data scientists appear increasingly to use LLMs to draft exploratory analyses, interpret model output, and produce visualizations without composing the underlying statistical scripts themselves. Working analysts across finance, healthcare, marketing, and academic research increasingly delegate the mechanical execution of statistical procedures to conversational interfaces, retaining for themselves the question framing, the interpretive judgment, and the verification of results. The population of people producing analytical deliverables is expanding rapidly to include domain experts who lack formal statistical training but who possess genuine expertise in the substantive question at hand.

This transition raises a question the analytical literature has not yet answered directly. What is the workflow paradigm under which such collaborations produce defensible output? The question is not whether domain experts should use LLMs for analytical work — they are already doing so, at scale, across every substantive domain that produces data. The question is what the practice looks like when it is done well, what competences it requires of the human partner, and what conditions distinguish rigorous AI-mediated applied analysis from fluent but shallow output vulnerable to characteristic LLM failure modes.

This paper offers a characterization of that paradigm. The central claim is that AI-mediated applied analysis, done well, is not a matter of the domain expert asking an AI to run an analysis while the AI does the work. Neither is it a matter of the domain expert deploying an AI as a search engine or calculator subordinate to their own analytical judgment. Rather, the paradigm constitutes a genuine collaboration in which each party contributes what they are structurally equipped to contribute. The domain expert brings the question, the data access, the operational definitions, the verification against domain knowledge, and the interpretation of findings. The LLM brings computational execution, methodological guidance, and the surfacing of analytical alternatives. Neither party could produce the work alone; the analysis requires competences distributed across the collaboration.

Four competences are proposed as necessary to the domain expert's participation in this paradigm: specification, verification, moderation, and interpretation. Each is characterized in Section 3. Together they constitute a competence profile distinct from that of the traditional analyst, whose training centered on the technical implementation of statistical procedures and the mathematical understanding required to select among them. The AI-mediated paradigm does not eliminate the value of that training — it remains essential for methodological judgment and for the doctoral-level evaluation of what counts as valid inference — but it opens applied analytical work to a broader population than traditional training could ever have credentialed.

To illustrate the paradigm in operation, Section 4 presents a worked case study. A fifth-year college baseball player without formal statistical training investigated what distinguishes elite four-seam fastballs from non-elite ones in 2026 NCAA Division I Power 4 conference play, using TruMedia/Trackman

data and an LLM as analytical partner over a five-week compressed capstone. The case is offered as an existence proof of the paradigm's viability under specified conditions, not as an empirical validation of the paradigm across cases.

The paper proceeds as follows. Section 2 reviews three bodies of literature relevant to the paradigm: trust in and use of LLMs for analytical work, the emerging literature on AI-assisted applied analysis, and the cognitive-psychology work on domain expertise and computational augmentation. Section 3 characterizes the workflow paradigm and its four competences. Section 4 presents the worked case study. Section 5 discusses conditions of viability and the failure modes that bound the technique. Section 6 addresses implications for pedagogy, hiring, and evaluation, and identifies future work.

2. LITERATURE REVIEW

The workflow paradigm characterized in this paper sits at the intersection of three bodies of scholarly work, each addressing part of the phenomenon but none characterizing it as such. The literature on trust in AI systems establishes that users adopt AI-mediated tools with varying degrees of calibrated reliance, and that miscalibration in either direction produces suboptimal outcomes. The emerging literature on AI-assisted analysis documents the growing use of LLMs for analytical tasks and characterizes some of the failure modes that attend that use. The cognitive-psychology literature on domain expertise and computational augmentation offers older but relevant frameworks for understanding the division of labor between human expertise and computational tools. The gap this paper addresses is the absence, across these three literatures, of a characterization of the specific workflow that emerges when a non-analyst domain expert uses an LLM as an analytical partner across a complete analytical arc.

2.1 Trust and Use of LLMs for Analytical Work

Trust in AI systems is a well-established construct in the information systems and human-computer interaction literatures. Lee and See (2004) established the foundational framework of trust calibration in automated systems, distinguishing appropriate trust (reliance matched to actual capability) from over-trust and under-trust, both of which produce suboptimal outcomes. Their framework anticipated many of the concerns that would later become central to LLM research, particularly the observation that neither uncritical acceptance nor blanket skepticism of automated systems constitutes appropriate engagement.

Empirical work on LLM trust specifically has extended these foundations. Money and Thanetsunthorn (2025) proposed the Acceptance-Trust Model, integrating self-efficacy, perceived control, perceived usefulness, perceived ease of use, and usage familiarity as predictors of trust in LLM outputs. Buğınca et al. (2021) demonstrated that users of AI decision-support tools frequently overrely on system output, accepting suggestions even when those suggestions are incorrect and the user would have decided better unaided. They further found that cognitive forcing functions — interventions that interrupt the interaction and require analytical engagement before output is accepted — measurably reduced overreliance, while participants rated those same interventions as less trustworthy and more effortful than simpler explanation interfaces. This finding bears directly on the paradigm characterized here: the practices that protect analytical work from overreliance impose friction that users are disposed to avoid.

Related work has characterized specific LLM failure modes relevant to analytical use. Sharma et al. (2023) documented sycophancy as a robust behavior in current frontier models, in which the system produces outputs oriented toward user agreement rather than accuracy. Ji et al. (2023) surveyed the phenomenon of hallucination in natural language generation, in which systems produce confident but fabricated content. Alkaiissi and McFarlane (2023) documented specific instances of hallucination in analytical contexts, showing that LLM outputs can mix accurate information with completely fabricated data in ways that are not detectable from the output alone. These failure modes constitute the risk landscape any AI-mediated analytical workflow must navigate.

What this literature has not yet characterized is the workflow through which a domain expert who lacks statistical training nonetheless produces defensible analytical output. The trust literature treats reliance as the outcome variable and does not address the operational sequence of specification, execution,

verification, and interpretation that constitutes an actual analytical project. The gap is not in the trust construct itself but in its application to the specific case of extended AI-mediated analytical work.

2.2 AI-Assisted Analysis and the LLM-as-Tool Literature

A smaller but growing body of work has begun to characterize the use of LLMs as analytical and technical tools directly. This work has concentrated on software development, where AI assistance arrived earliest and where controlled evaluation proved most tractable. It establishes two findings relevant to the present paradigm: that benefit from AI assistance is distributed unevenly across levels of user experience, and that such assistance shifts rather than removes the burden carried by the human partner.

Peng et al. (2023) conducted a controlled experiment with 95 professional developers on a standardized programming task and found that the group with access to GitHub Copilot completed the task 55.8 percent faster than the control group. The effect was heterogeneous: less experienced developers, and those who spent more hours per day writing code, benefited most. The authors note that the study examined a single standardized task and that gains may vary across task types. Vaithilingam et al. (2022) studied 24 programmers using the same tool and reached a more complicated conclusion. The tool did not reliably improve task completion time or success rate. Participants nonetheless preferred working with it, largely because it supplied a useful starting point. They also faced substantial difficulty understanding, editing, and debugging the code it generated, and that difficulty significantly hindered their effectiveness. Neither study was designed to characterize workflow, but together they establish that AI-mediated technical work redistributes effort toward comprehension and verification rather than simply reducing it, and that it therefore requires competences distinct from those of unassisted technical work.

Related work on LLM-as-judge applications has demonstrated that LLMs can perform evaluative tasks on model output at levels approaching human agreement under appropriate conditions. Zheng et al. (2023), introducing the approach systematically, report that GPT-4 acting as judge reached 85 percent agreement with expert human evaluators, higher than the 81 percent agreement observed among those evaluators themselves, while documenting characteristic biases that constrain the technique, including position bias, verbosity bias, and self-enhancement bias. That final constraint bears on the present paper: the same work establishing LLM evaluative capability also establishes that such systems favor their own output, which bounds how far verification can responsibly be delegated to the tool. This capability and its documented limits have implications for the verification competence discussed in Section 3, though the LLM-as-judge literature has focused primarily on automated evaluation of model outputs rather than on human-directed verification of analytical work.

The gap in this literature is characterization of the human contribution to AI-mediated analytical work in cases where the human is not a trained analyst. Existing work either studies trained professionals using AI to augment their existing capability, or studies laboratory settings in which the human role is minimized. The population of interest for this paper — domain experts with substantive knowledge but no statistical training — is largely absent from the literature, despite constituting an increasingly common configuration in practice.

2.3 Domain Expertise and Computational Augmentation

The cognitive-psychology literature on expertise offers older but foundational frameworks for understanding the division of labor between human judgment and computational tools. Ericsson and Charness (1994) established the empirical basis for expertise as domain-specific pattern recognition developed through extended practice, distinguishing genuine expertise from familiarity or surface knowledge. Klein (1998) documented naturalistic decision-making in expert practitioners, showing that expert judgment operates on cues and patterns that are often not consciously articulable but that produce reliably better outcomes than novice judgment. These frameworks are relevant to the current paradigm because they establish that domain expertise brings capabilities to analytical work that are not reducible to statistical training and are not substitutable by computational tools.

Parasuraman and Riley (1997) characterized four modes of human-automation interaction — use, misuse, disuse, and abuse — and argued that appropriate integration of human and automated

capability required attention to the distinct strengths of each. Their framework has been extended and refined across three decades of automation research, but its central insight remains relevant: computational tools and human judgment are complementary rather than substitutable, and effective integration requires deliberate attention to the division of labor between them.

The current paradigm extends this tradition into a new configuration. Where earlier automation research addressed the integration of expert practitioners with well-characterized computational tools, the AI-mediated paradigm involves non-analyst domain experts working with computational tools whose behavior is only partially characterizable. The classical frameworks of appropriate reliance do not translate directly, because the object of reliance has properties the earlier automation literature did not anticipate.

2.4 Synthesis and Gap

Across these three literatures, the constituent elements of the current paradigm have been separately characterized: trust and its calibration failures, LLM capabilities and failure modes, and the general principles of human-computational division of labor. What remains uncharacterized is the specific workflow that emerges when a non-analyst domain expert uses an LLM as an analytical partner across a complete analytical project, and the specific competences that workflow requires of the human partner. The gap is not in any single literature's treatment of its own subject but in the absence of a synthesizing framework that treats the paradigm as a distinct phenomenon warranting its own characterization. The remainder of this paper addresses that gap. The contribution is threefold. First, the paper names and defines four competences — specification, verification, moderation, and interpretation — that the paradigm requires of the human partner and that are not reducible to the technical competences of traditional analyst training. Second, it demonstrates through a worked case that these competences are observable and separable in practice rather than merely conceptual. Third, it identifies the conditions under which the paradigm produces defensible work and the failure modes that compromise it, which together constitute a set of testable propositions for subsequent empirical work.

3. THE WORKFLOW PARADIGM

The workflow paradigm characterized in this paper operates as a genuine collaboration between a domain expert and a large language model. Each party contributes what they are structurally equipped to contribute. The paradigm is misdescribed when it is framed either as autonomous AI analysis with a human observer or as human analysis with an AI subordinate. Both framings misrepresent where the intellectual work occurs.

The domain expert contributes the question, the data access, the operational definitions that translate the substantive question into an analytical form, the verification of outputs against domain knowledge and internal consistency, and the interpretation of findings in terms that matter for the substantive domain. The LLM contributes computational execution of statistical procedures, methodological guidance on what procedures are appropriate under what conditions, generation of visualizations and summaries, and the surfacing of analytical alternatives the domain expert may not have considered. Each contribution is necessary; neither is sufficient. Without the domain expert, the LLM has no anchoring question, no domain-informed verification, and no meaningful interpretation. Without the LLM, the domain expert lacks the computational execution that would let their domain knowledge engage the data at scale.

This distribution of contributions places specific demands on the domain expert. Four competences are proposed as necessary to the paradigm's successful operation: specification, verification, moderation, and interpretation. Each is characterized below with attention to what the competence requires, how it functions in the workflow, and what its failure looks like when the paradigm is deployed without it.

3.1 Specification

Specification is the competence of defining an analytical question with sufficient precision that a computational tool can execute it faithfully. The domain expert must translate a substantive question — which may be posed in the vocabulary of the domain rather than in analytical terms — into a specification that identifies the unit of analysis, the sample from which observations are drawn, the operational

definitions of key constructs, the variables of interest, and any constraints on inclusion or exclusion of observations. This translation is not a mechanical task. It requires judgment about how the substantive question maps onto measurable quantities, and it requires anticipating how imprecision in the specification will propagate into the analysis.

This difficulty is empirically documented. Zamfirescu-Pereira et al. (2023), studying how non-AI-experts design prompts for language models, found that non-experts treat task specification implicitly and opportunistically where trained practitioners treat it explicitly and systematically, and that non-experts are systematically overconfident about how well their specifications convey their intent. The specification competence therefore becomes more important in AI-mediated work, not less. A capable LLM will execute what it is asked to execute, including specifications that are ambiguous, incomplete, or subtly misaligned with the substantive question. The tool does not typically refuse imprecise specifications; it interprets them and produces output consistent with its interpretation. When the interpretation diverges from the domain expert's intent, the resulting analysis answers a subtly different question than the one that was asked, and the divergence may not be detectable from the output.

Failure of specification competence produces analytical output that is technically correct with respect to what was asked and substantively misaligned with what was meant. The failure is invisible from the output alone; only the domain expert, holding the substantive question, can detect the misalignment. This makes specification competence the first line of quality control in the workflow. Deliberate, explicit, complete specification of the analytical question is not an academic nicety in the AI-mediated paradigm; it is a load-bearing element of the work.

3.2 Verification

Verification is the competence of checking outputs against domain knowledge, internal consistency, and known properties of the data. The human factors literature supplies a name for what its absence produces. Sarter, Woods, and Billings (1997) describe automation surprises, in which an automated system behaves in ways its operator did not anticipate because the operator's mental model of the system has drifted from the system's actual behavior. Their central finding — that experienced operators remain susceptible, and that surprises cluster in non-routine situations — transfers directly to AI-mediated analysis, where fluent and well-formatted output invites the assumption that the tool has construed the task as the analyst intended. In the AI-mediated paradigm, the domain expert cannot rely on the tool's own claims of correctness; the tool may produce confident output that is fabricated (Ji et al., 2023), agreement-seeking rather than accurate (Sharma et al., 2023), or the product of a methodologically inappropriate procedure the tool selected without flagging the choice. Verification is the competence by which the domain expert protects the work from these failure modes.

Effective verification operates on multiple levels. At the level of summary statistics, headline numbers reported by the tool can be cross-checked against independent computation on the raw data using spreadsheet operations or manual calculation. At the level of effect directions and magnitudes, findings can be sanity-checked against what domain knowledge would predict; findings that contradict well-established domain knowledge warrant additional investigation before acceptance, and findings that too closely match domain expectations warrant investigation for the possibility that the tool has produced a plausible-sounding confabulation. At the level of methodology, the tool's choice of statistical procedure can be interrogated by asking the tool to explain its reasoning before executing the analysis, then evaluating the explanation against what the domain expert understands about the data.

The domain expert's substantive expertise is a legitimate verification tool in this paradigm, not a suspect basis for skepticism. When a domain expert with genuine expertise reads a finding as implausible, that reading is data. It may be wrong — domain knowledge is not infallible — but it is worth investigating before acceptance. The AI-mediated paradigm depends on the domain expert bringing their expertise to bear on verification, not on the domain expert deferring to the tool as the more authoritative source. Failure of verification competence produces work that looks defensible on the surface but that contains errors the tool's failure modes have introduced and that no one has checked.

3.3 Moderation

Moderation is the competence of deciding when to accept a tool output, when to push back, when to redirect the analysis, and when to reject a line of inquiry entirely. It is the judgment that distinguishes productive iteration from unmoored drift, and productive redirection from post-hoc rationalization of unwelcome findings. Moderation is closer than the other competences to what senior analysts in the traditional paradigm have always possessed: the tacit knowledge of when an analysis is done, when it is misdirected, and when it needs to be reframed.

In AI-mediated work, moderation becomes more visible than it is in traditional analytical work, because the tool will happily continue producing output regardless of whether it should. The tool does not typically indicate that the analysis has reached diminishing returns, that a finding warrants pause for interpretation before further extension, or that the direction of inquiry has drifted from the original question. These judgments must be made by the domain expert, and made explicitly, because the tool provides no natural stopping points.

The judgment has precedent in the automation literature. Parasuraman and Riley (1997) distinguish misuse, disuse, and abuse of automation as distinct failure patterns, each arising from a mismatch between operator judgment and actual system capability; moderation is the competence that determines which pattern obtains in AI-mediated analytical work. Effective moderation also requires the ability to push back on tool outputs without being deflected by the tool's tendency toward agreement, which Sharma et al. (2023) document as a systematic consequence of training on human preference data rather than an occasional lapse. A domain expert who reads a finding as questionable and presses the tool for defense should expect that a poorly moderated interaction will produce capitulation rather than defense. The tool will often revise its output in the direction of the user's objection, whether or not the objection is well-founded. Moderation competence includes the discipline to distinguish substantive engagement with a challenge from sycophantic accommodation of it, and to persist in the analysis when the tool's responses to challenge are unsatisfying.

3.4 Interpretation

Interpretation is the competence of translating statistical output into domain meaning. It is the least substitutable of the four competences because it requires embedded expertise that AI tools cannot generate from analytical output alone. A coefficient, an effect size, a p-value, or a confusion matrix carries no domain meaning on its own; meaning is imposed by a reader who understands what the statistical quantity implies for the substantive question that motivated the analysis. Without interpretation, analytical output is inert. With it, analytical output becomes actionable knowledge in the domain.

Interpretation is also the competence that most benefits the eventual audience of the analytical work. In applied contexts, the audience is usually not another analyst but a practitioner or decision-maker in the domain who cares about the substantive question and has limited patience for statistical formalism. The domain expert who can translate findings into the vocabulary of the domain — into implications for practice, for decision, or for further investigation — produces output that reaches its audience. The domain expert who cannot, or who defers to the tool's natural-language rendering of the findings, produces output that is technically complete but practically inert.

This competence, more than the others, is where domain expertise is irreplaceable in the paradigm. The tool can produce grammatically correct sentences that describe statistical findings, but it cannot substitute for the domain expert's judgment about what the findings mean for the practice of the field. An LLM describing swings and misses on a fastball can produce accurate prose. A former pitcher describing the same finding can produce prose that a coach will act on. The difference is not technical; it is the difference between analytical output and analytical knowledge.

3.5 What the Paradigm Is Not

Three misconceptions about the paradigm are worth explicitly rejecting.

First, the paradigm is not autonomous AI analysis with a human observer. In autonomous analysis, the tool takes a loosely specified question and produces an output that the human accepts or rejects. Under this framing, the human contribution is evaluation, not construction, of the work. This is not the paradigm characterized here. The paradigm requires the human to specify, verify, moderate, and interpret — each an active contribution to the analytical work, not a passive evaluation of tool output.

Second, the paradigm is not the domain expert using an AI as a search engine or calculator subordinate to their own analytical judgment. Under this framing, the tool is invoked to perform discrete tasks within an analytical framework the human has already constructed. This underestimates the tool's contribution. In the paradigm, the tool contributes methodological guidance the human may not possess, surfaces alternatives the human would not have considered, and produces methodological choices that the human can accept or redirect but rarely originates from scratch. The tool is a partner, not an instrument.

Third, and most importantly, the paradigm is not a substitute for domain expertise. The paradigm operates on the foundation of genuine substantive knowledge in the domain of inquiry. Attempts to deploy the paradigm without such expertise produce fluent but shallow output vulnerable to the tool's failure modes at every stage — misaligned specification, unverified findings, unmoored moderation, and inert interpretation. The paradigm is a force multiplier for domain expertise, not a replacement for it. The distinction matters because the paradigm's value in the field depends on it being understood correctly. Deployed on a foundation of expertise, the paradigm extends what domain experts can accomplish. Deployed without that foundation, it produces work whose surface fluency masks its substantive vacancy, and whose failures will accumulate faster than any evaluator without domain expertise can detect them.

4. CASE STUDY: ELITE FASTBALLS IN POWER 4 BASEBALL

To illustrate the workflow paradigm and its four competences operating end-to-end, this section presents a worked case. The case is offered as an existence proof of the paradigm's viability under specified conditions, not as an empirical validation across cases. A single well-executed instance cannot establish the general reliability of a paradigm; it can, however, demonstrate that the paradigm operates as characterized when its conditions are met, and it can illustrate what each of the four competences looks like when exercised deliberately by a domain expert without formal analytical training.

4.1 Context

The domain expert in this case is a fifth-year NCAA Division I baseball player at a medium national liberal arts university, completing a capstone independent study during a five-week compressed summer session. His undergraduate degree is in history. He has completed four introductory analytics courses at the survey level and has no coursework in statistics or research methods, no training in inferential technique, and no prior experience conducting an independent analysis from question to finding. His domain expertise, by contrast, is substantial: five years of collegiate baseball as both pitcher and hitter, two years of active work with the TruMedia/Trackman platform in his university's baseball program, and fluency in the modern pitching-analytics vocabulary through sustained engagement with Driveline, FanGraphs, Baseball Savant, and adjacent sources. The case is representative of the population of interest for this paper: a domain expert with genuine substantive knowledge, working access to relevant data, comfort with AI tools, and exposure to analytics well short of the training that would traditionally credential an analyst.

The research question was one he had held personally for years: what distinguishes an elite four-seam fastball from a non-elite one at the college level. The question mattered to him because a pitching coach had once told him, as a freshman, that his four-seam fastball was in a "dead zone" despite reasonable velocity. It mattered professionally because he intended to pursue a coaching career and wanted a defensible analytical basis on which to evaluate and develop pitchers. The question thus arrived at the capstone with both intellectual and vocational stakes attached.

4.2 Specification in Practice

The specification competence appears at several junctures in the case, but its most consequential exercise concerns the operational definition of "elite." The domain expert initially framed elite in outcome

terms as “a pitch that produces swings and misses in general, and more specifically, in the zone.” That intuition was substantively correct but required translation into a measurable form. Working with his faculty supervisor, elite was operationalized as the top quartile of the sample by in-zone whiff percentage, with the empirical cutoff to be determined from the data. The final threshold was 18.5 percent in-zone whiff, above which sat 34 pitchers whose average was 21.7 percent, compared with 12.5 percent for the 104 non-elite pitchers.

Two additional specification decisions carried the analytical weight of the study. First, he set a minimum threshold of 300 four-seam fastballs per pitcher. Initial exploration had shown in-zone whiff rates to be unstable at low pitch counts, with pitchers throwing fewer than 50 fastballs producing rates erratic enough to distort any group-level comparison. Rather than place the cutoff at the point where instability first became visible, he set it substantially higher. The threshold was chosen on domain rather than statistical grounds: 300 four-seam fastballs across a season of conference play characterizes a pitcher for whom the four-seam is an established primary offering, which is the population the research question concerns. The decision traded sample size for stability, and it was made deliberately in that knowledge. No sensitivity analysis across alternative threshold values was performed, and the stability of the findings under different cutoffs is therefore unknown. Second, when he pulled the raw data from TruMedia, he shifted from the pitch-level unit of analysis that had been provisionally specified in the learning contract to a pitcher-level aggregate, with one row per pitcher averaging that pitcher’s qualifying fastballs. He documented and defended the shift explicitly: because the research question concerned what makes an elite fastball pitcher, pitcher-level aggregation was the analytically appropriate unit; pitch-level analysis would have introduced within-pitcher variance the question did not require. This shift was substantive rather than clerical, and it exemplifies specification as active judgment rather than mechanical translation.

4.3 Verification in Practice

The verification competence appears most visibly in the ten-item verification log the domain expert maintained across the analytical phase and included as an appendix to his methodology paper. Each item consisted of a check, a method, and a result. Headline summary statistics were cross-checked against the raw CSV using Excel AVERAGEIF operations, confirming that the elite and non-elite group means reported by the AI tool matched independent computation. The elite group count was verified by independent sorting and row-counting. Effect directions were sanity-checked against baseball domain knowledge: higher velocity producing more swing-and-miss was confirmed as consistent with domain understanding; a flatter vertical approach angle produced by a lower release point was confirmed as mechanically consistent; and a near-zero effect for induced vertical break was flagged as surprising and noted as a finding rather than an error, explicitly resisting the temptation to treat an unwelcome result as a computational failure.

The most consequential verification move concerned horizontal break. The raw data showed horizontal break values ranging from -18.7 to $+18.4$ inches. To an untrained reader the aggregate group means might appear straightforwardly interpretable. Drawing directly on baseball domain knowledge, he recognized that the sign of horizontal break is opposite for left-handed and right-handed pitchers: a right-handed pitcher’s arm-side run appears as positive horizontal break, while a left-handed pitcher’s equivalent movement appears as negative. Aggregating across handedness produces a mean that is a mathematical artifact of the sample’s handedness composition rather than a meaningful measure of fastball movement. He inspected the range in the raw CSV, confirmed the handedness mixing, and excluded horizontal break from the inferential interpretation while retaining the descriptive number for transparency. This is verification competence operating on the substantive properties of the data, not merely on the arithmetic of the tool’s output. No general-purpose statistical check would have caught the issue; only a reader who understood what the numbers meant in the domain could recognize that the aggregation was inappropriate. The AI tool executed what it was asked to execute; the domain expert caught what the tool could not detect.

4.4 Moderation in Practice

The moderation competence is most visible in the domain expert’s treatment of the logistic regression’s modest predictive gain. The seven physical predictors, entered together as standardized coefficients, produced a model that correctly classified 77.5 percent of pitchers, compared with a null baseline

(predicting non-elite for every pitcher) of 75.4 percent. The gain over baseline was 2.1 percentage points — a marginal improvement in classification terms. The model was also underpowered for the number of predictors entered: 34 elite pitchers across seven predictors yields roughly five events per variable, against a conventional guideline of ten, so the individual coefficient estimates should be read as indicative rather than precise. This constraint reinforces rather than undercuts the reading developed below, since an underpowered model fitted to an incomplete set of predictors is precisely what would be expected to produce a small gain over baseline. Several explanations for the modest gain are consistent with the data and cannot be distinguished within this case. Relevant predictors may be absent from the model, as the interpretation below argues. But measurement error in the tracking data, the restriction of the sample to pitchers meeting the 300-fastball threshold, and the dichotomization of a continuous outcome at the quartile boundary would each independently depress apparent predictive performance. The reading developed below should be understood as the most substantively useful of several admissible explanations rather than as the only one available. A less moderated engagement with the finding would have led in one of two unproductive directions: either presenting the 77.5 percent accuracy as an impressive result while omitting the baseline, or dismissing the model as a failure and searching for reasons to disregard it.

He did neither. His methodology paper reports the model accuracy alongside the null baseline explicitly and characterizes the gain as “marginal improvement only,” language that reflects methodological honesty rather than salesmanship. More substantively, he moderated the finding by reframing what it meant: the marginal improvement did not represent a failure of the analysis but rather a substantive finding about the domain. Because the seven physical predictors together produced only a small improvement over baseline, most of what distinguishes an elite four-seam fastball from a non-elite one at the Power 4 level must lie outside the physical characteristics captured in the data. Sequencing, tunneling, command, deception, and contextual factors were absent from the model by construction, and the model’s marginal gain quantified the space those unmeasured factors occupy. This reframe transformed an apparent shortcoming of the analysis into a substantive insight about the limits of pitch-design metrics as standalone predictors, an insight that informs the coaching implications derived from the study and that constitutes the most defensible practical takeaway of the work. Moderation, in this instance, meant neither accepting nor rejecting the finding as it initially appeared, but deciding to read it against a framing that made it useful.

4.5 Interpretation in Practice

The interpretation competence appears throughout the deliverables, but is most concentrated in the coach-facing portfolio artifact produced as the terminal product of the capstone. Interpretation is where the analytical output leaves the language of statistics and enters the language of the domain. A coefficient of +0.97 for velocity, standardized, with an odds ratio of 2.64, is a statistical object. The interpretive move that renders it useful is the translation into coaching vocabulary: a pitcher throwing 93 miles per hour or above is meaningfully more likely to be in the elite tier; a 1.5 mile-per-hour gap in average velocity between the elite and non-elite groups yields a medium effect size and a substantial odds multiplier. That translation is not mechanical. It requires knowing which framings of the finding will land with a coaching audience, which units are used in dugout conversation rather than in statistical reports, and which comparisons are meaningful at the practitioner level.

The joint interpretation of vertical approach angle and release height further illustrates the competence at work. Read as independent predictors, they are statistical quantities that separately correlate with elite status. Read as a delivery profile — a pitcher releasing from a lower arm slot naturally produces a flatter downward trajectory — they become a coherent mechanical picture that a coach can recognize and, if desired, develop toward. The domain expert’s portfolio piece states the interpretation directly: “elite pitchers tend to release from a lower arm slot and arrive at the plate on a flatter plane, which, combined with higher velocity, makes the pitch harder to square up.” That is a sentence a pitching coach can read once and use. It carries the statistical finding without carrying the statistical language, and it does so because his domain expertise let him see the two variables as a single mechanical unit rather than as separate predictors. Interpretation, in this instance, produced something the tool could not have produced without domain-embedded guidance.

4.6 Deliverables and the Scope of the Case

The case produced three deliverables: a coach-facing portfolio artifact suitable for professional use, a formal methodology paper documenting the analytical process and including a verification appendix, and a recorded walk-through presentation aimed at a coaching-staff audience. Each deliverable served a distinct audience and purpose while drawing on the same analytical work. The portfolio artifact is what a coaching staff would read; the methodology paper is what a rigorous reviewer would examine to evaluate the artifact's credibility; the recorded presentation demonstrates that the domain expert can talk about the work in the register of the domain rather than the register of statistics.

The case is offered here as existence proof, not as generalization. A single successful instance cannot establish that the paradigm reliably produces defensible work across the population of possible domain experts, questions, and datasets. The limits are worth stating precisely. This is one domain expert, in one domain, working one question, under an unusually favorable configuration of conditions: deep substantive expertise, direct access to professional-grade instrumented data, and a research question held personally for years. Because the case was selected on a successful outcome, it cannot speak to base rates of success or failure, and it cannot separate which of the four competences were necessary from which merely co-occurred. The design supports an existence claim and the generation of propositions. It does not support inference about prevalence, about effectiveness relative to traditional analytical training, or about the relative contribution of the individual competences, each of which requires the comparative or experimental work outlined in Section 6.4. What it does establish is that the paradigm as characterized in Section 3, operated deliberately and by a domain expert with genuine substantive knowledge, produced work whose methodological quality is defensible on its own terms and whose interpretive quality is useful in the domain of application. The next section examines what specific features of the case supported that outcome and what failure modes would have compromised it under different conditions.

5. DISCUSSION: CONDITIONS OF VIABILITY AND FAILURE MODES

The paradigm's viability depends on conditions that were satisfied in the case study but that cannot be assumed to hold in every application. Two conditions in particular deserve explicit attention: the presence of genuine domain expertise, and deliberate rather than credulous engagement with the tool's failure modes.

5.1 The Paradigm Rests on Genuine Domain Expertise

Every competence the paradigm requires of the human partner draws on substantive knowledge of the domain in which the analysis operates. Specification requires knowing what constructs matter in the domain and how they can be operationally defined. Verification requires knowing what the numbers should look like if the analysis is well-executed and what would count as a red flag warranting investigation. Moderation requires knowing when a finding is worth developing further and when it has drifted from the substantive question. Interpretation requires knowing what the findings mean in the vocabulary of practitioners in the domain. Without genuine domain expertise, none of these competences can be exercised with the judgment the paradigm depends on.

This condition is the paradigm's single strongest boundary. Attempts to deploy the paradigm without genuine domain expertise produce fluent output that mimics analytical work without possessing its substantive foundation. The AI tool can produce grammatically well-formed prose describing statistical findings in any domain regardless of whether the operator understands the domain. What such output cannot do is exercise the specification, verification, moderation, and interpretation competences at levels that depend on knowing the subject matter. A user without domain expertise cannot recognize that horizontal break requires handedness-split interpretation; cannot recognize that the marginal improvement over baseline is a substantive finding about unmeasured variables rather than a shortcoming of the model; cannot translate statistical output into the vocabulary of practitioners because they do not speak the vocabulary. The paradigm operates as a force multiplier for expertise, not as a substitute for it.

5.2 Deliberate Engagement with LLM Failure Modes

The paradigm also requires deliberate rather than credulous engagement with characteristic LLM failure modes. Four such failure modes bore directly on the case and can be expected to bear on any application of the paradigm. Table 1 summarizes them.

Table 1. LLM failure modes bearing on AI-mediated analytical work

Failure mode	How it appears in analytical work	Mitigation in the case	Competence engaged
Hallucination	Summary statistics that look plausible but do not match the data; fabricated citations or factual claims	Systematic cross-checking of headline statistics against raw data by independent computation	Verification
Literal interpretation	The request as posed is executed rather than the request as intended; an inappropriate test is run without objection	Explicit specification of unit of analysis, sample, and operational definitions before execution	Specification
Sycophancy	Agreement with the user's framing irrespective of its merit; findings shaded toward the expected result	Requiring the tool to state its methodology and defend it before the result is accepted	Moderation
Capitulation under pushback	Revision of a correct output in the direction of an objection, whether or not the objection is well founded	Distinguishing substantive response to challenge from accommodation of it; persisting when responses are unsatisfying	Moderation

Hallucination is the tendency of LLMs to produce confident output that has no basis in the data (Ji et al., 2023). In an analytical context this appears as summary statistics that look plausible but do not match the underlying data, citations that do not exist, or factual claims that are fabricated. The mitigation in the case study was systematic cross-checking of headline statistics against the raw data using independent computation. This mitigation scales; every AI-mediated analysis can and should include a small set of independently verified anchor statistics.

Literal interpretation is the tendency of LLMs to execute the request as posed rather than the request as intended. When the user asks for a comparison the AI executes a comparison, whether or not the statistical test chosen is appropriate for the data at hand. The output is technically correct with respect to what was asked and substantively misaligned with what was meant, and the misalignment is invisible from the output alone. The mitigation the case employed, and the one that generalizes most usefully, is to ask the tool to explain its methodological choice before executing the analysis and to evaluate the explanation before accepting the output. This slows the workflow at exactly the moment it needs slowing, and it moves the methodological judgment from an implicit choice by the tool to an explicit choice by the operator.

Sycophancy is the tendency of LLMs to produce outputs oriented toward user agreement rather than accuracy (Sharma et al., 2023). In an analytical context this appears as unwarranted confirmation of user-suggested interpretations, as reluctance to disagree with proposed methodological framings, and as capitulation when the user pushes back on a correct finding. The mitigation is prompt discipline: framing questions in ways that invite criticism rather than validation, and reading agreement as a signal to probe further rather than as confirmation. The case employed this discipline explicitly and named it in the disclosure section of the methodology paper.

Capitulation under pushback is a specific instance of sycophancy that deserves separate mention. When the user challenges a tool output, the tool tends to revise in the direction of the challenge whether or not the challenge is well-founded. This produces analytical drift in which a correct finding is progressively modified toward an incorrect one because the user found the correct finding unwelcome. The mitigation is discipline about the epistemic status of pushback: challenge the tool when the finding conflicts with domain knowledge, but require the tool to defend its output substantively rather than accepting revised output that merely accommodates the challenge. The case employed this practice as part of its verification protocol.

5.3 What Would Have Compromised the Case

The case succeeded because both conditions held: genuine domain expertise on the part of the analyst, and deliberate engagement with the tool's failure modes. Absent either condition the outcome would have differed. Absent domain expertise, the horizontal break aggregation error would not have been caught; the marginal-improvement-over-baseline finding would not have been reframed into a substantive insight; and the coach-facing interpretation would have defaulted to statistical language that no coaching staff would use. Absent deliberate engagement with failure modes, hallucinated statistics might have entered the final deliverables; methodologically inappropriate analyses might have been executed and reported; and sycophantic accommodation might have distorted the interpretation of surprising findings. Each competence in isolation is necessary but not sufficient; the paradigm depends on their joint exercise, supported by the two viability conditions.

6. IMPLICATIONS AND CONCLUSION

6.1 Pedagogical Implications

The paradigm has direct implications for the training of non-analyst domain experts who will use AI-mediated analysis in their professional work. Traditional analyst training centered on the technical implementation of statistical procedures and the mathematical understanding required to select among them. The paradigm does not require the domain expert to acquire those competences at analyst-level depth, but it does require training in specification, verification, moderation, and interpretation as distinct and teachable skills. A pedagogy for the paradigm would treat these competences explicitly rather than assuming they emerge as byproducts of technical training. It would teach the discipline of explicit specification as a load-bearing skill, would introduce known LLM failure modes and the protocols that mitigate them, would exercise the judgment of when to accept and when to challenge tool outputs, and would develop the domain expert's ability to translate statistical output into practitioner vocabulary. None of these skills is currently central to either statistics education or AI-literacy programs, and their absence is likely to become increasingly visible as the paradigm spreads. The implication falls most directly on information systems and computing education. IS already occupies the ground between substantive domains and technical implementation, and is therefore better positioned than either statistics or computer science to house this training. IS curricula have long taught requirements elicitation, systems analysis, and the translation of business questions into technical specifications; specification competence as characterized here is a recognizable extension of that tradition to a new class of tool. Verification and moderation have analogues in the testing and quality-assurance content already present in most IS programs. What is missing is not a new subject area but the explicit treatment of these competences as transferable skills applied to AI-mediated analysis, taught to students who will use analytical tools without becoming analysts. A program could implement this within existing courses rather than by adding coursework: an analytics course that grades specification quality and verification protocol alongside analytical results, or a capstone that requires a verification appendix as a standard deliverable, would exercise the competences directly. This also reframes a question that AI-literacy curricula currently struggle to answer. Asking whether a student can produce analytical output without formal training in the underlying software is the wrong test; the useful question is whether the student can specify, verify, moderate, and interpret the work that results.

6.2 Implications for Hiring and Evaluation

Evaluators of AI-mediated analytical work face an inherent difficulty: they cannot cleanly determine where the domain expert's contribution ends and the tool's contribution begins. The origination question, in the traditional sense, cannot be answered from the deliverable alone. This paper suggests that the question is the wrong one. The evaluative question is not whether any specific insight originated with the human or the tool, but whether the human can carry the work: defend the specification, explain the verification, justify the moderation decisions, and extend the interpretation under substantive questioning. Ownership in the AI-mediated paradigm is demonstrated not by isolated authorship but by demonstrated capacity to engage the work at every level a critical audience would probe. Hiring and evaluation protocols for practitioners using AI-mediated analysis should be structured around demonstrated capacity rather than demonstrated origination.

6.3 The Boundary the Paradigm Does Not Cross

One boundary of the paradigm warrants explicit statement. The paradigm supports the practicing use of applied analytics by domain experts who lack traditional analyst training. It does not support the methodological evaluation of analytical work, the falsification of established findings, the development of new statistical methods, or the doctoral-level competences that concern the boundaries of what counts as valid inference in a discipline. Those competences remain the province of deep statistical and methodological training, which the paradigm does not replace and cannot substitute for. The paradigm expands the population of people who can produce defensible applied analytical work; it does not expand the population of people who can evaluate the framework within which that work is defensible. The distinction matters for the health of the field, because the paradigm's continued viability depends on the preservation of methodological judgment somewhere in the system. If everyone becomes a paradigm-enabled practitioner and no one is trained to evaluate practice at the methodological level, the paradigm loses its epistemic grounding.

6.4 Future Work

Four directions warrant investigation. First, empirical characterization of the paradigm across additional cases, domains, and demographic profiles of domain experts would establish which features of the case study generalize and which are specific to its particular configuration. Second, the failure modes discussed here — hallucination, literal interpretation, sycophancy, capitulation under pushback — have architectural sources in current LLM systems whose deeper mechanisms remain incompletely characterized in the applied literature; a companion program of work developing these mechanisms would deepen the foundation the paradigm rests on. Third, specific methodological extensions of the present case would strengthen its evidentiary value, the most direct being a sensitivity analysis of the findings across alternative inclusion thresholds, which would establish how far the substantive result depends on a specification choice made on domain rather than statistical grounds. A related question concerns what precedes the workflow entirely: whether a domain expert adopts an AI-mediated approach at all is an antecedent that the technology acceptance tradition is equipped to address and that the present paper does not treat. Fourth, the boundary between paradigm-enabled practice and methodological judgment identified in Section 6.3 will require ongoing attention as the paradigm spreads, particularly in disciplines that have historically drawn a firm line between practitioner and researcher roles.

6.5 Conclusion

A workflow paradigm has emerged in which domain experts without formal statistical training can conduct defensible applied analytical work in collaboration with large language models. This paper has characterized the paradigm as a genuine collaboration in which each party contributes what it is structurally equipped to contribute, has identified four competences the paradigm requires of the domain expert — specification, verification, moderation, and interpretation — and has demonstrated the paradigm's viability through a worked case in college baseball analytics. The paradigm depends on genuine domain expertise and on deliberate engagement with characteristic LLM failure modes; deployed without these conditions it produces fluent but shallow output whose failures accumulate faster than any evaluator without domain expertise can detect them. Deployed with these conditions it extends what domain experts can accomplish and opens applied analytical work to a broader population than traditional analyst training could have credentialed. The distinction between paradigm-enabled practice and methodological judgment remains, and the paradigm neither replaces nor substitutes for the deep statistical training on which methodological judgment depends. What the paradigm does is name and enable a mode of analytical work that was not previously available at scale, and that is now being practiced without adequate scholarly attention to what it requires, what it produces, and how it should be taught. This paper is offered as a step toward that attention.

7. REFERENCES

Alkaiissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), e35179. <https://doi.org/10.7759/cureus.35179>

- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188, 1–21. <https://doi.org/10.1145/3449287>
- Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, 49(8), 725–747.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
- Klein, G. (1998). *Sources of power: How people make decisions*. MIT Press.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
- Money, W. H., & Thanetsunthorn, N. (2025). A proposed study of factors moderating degree of trust in LLM and ChatGPT-like outputs. *Journal of Information Systems Applied Research and Analytics*, 18(4), 67–80.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
- Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590*.
- Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed., pp. 1926–1943). Wiley.
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., Bowman, S. R., ... Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.
- Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.
- Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 437, 1–21. <https://doi.org/10.1145/3544548.3581388>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.