Pedagogical Alignment, Adaptivity, and Analytics in Intelligent Tutoring Systems: A Systematic Review

Ilyass Houssam, Zouhair Chiba and Mounia Miyara
LIS Labs, Faculty of Sciences Aïn Chock, Hassan II University of Casablanca, Morocco

ilyasshoussam9@gmail.com (corresponding author)

https://doi.org/10.34190/ejel.24.4.4779

Abstract: Intelligent tutoring systems have changed quickly since large language models became widely available in 2023, raising a practical question for e-learning research: when a tutor is built on a general-purpose language model rather than on hand-encoded rules, what happens to its pedagogical grounding, learner adaptivity, and reproducibility infrastructure? Earlier reviews have examined these matters separately, but none has considered how they come together in the systems now being built. This systematic review addresses that gap through three analytical axes: pedagogical alignment with an instructional theory, adaptivity and learner modelling, and analytics standards supporting comparison and reproducibility. The review followed the SPAR-4-SLR protocol and used PRISMA 2020 as a reporting framework where the recoverable record allowed. Candidate records were screened, verified against their source documents, and coded conservatively against pre-specified criteria. Twelve primary studies were retained after source verification. Pedagogical alignment and adaptivity were each addressed by eleven of the twelve studies (92 percent), although implementation varied widely, from explicit frameworks such as Cognitive Apprenticeship, Productive Failure, and Socratic questioning to looser prompt-based behaviour. Analytics standards and reproducibility were the weakest axis, addressed by three studies (25 percent). No study combined all three axes with explicit uncertainty quantification or calibration of tutoring decisions. Controlled learning-outcome evidence was also scarce: only one study reported a controlled comparison, and its findings were preliminary. For e-learning practice, the review gives teachers, platform designers, and institutional decision-makers a realistic basis for expectation: current language-model tutoring systems are promising and often pedagogically plausible, but their learning benefits remain under-demonstrated and should be evaluated locally before large-scale adoption. For e-learning knowledge, the review contributes a source-verified synthesis of recent intelligent tutoring research and identifies a clear methodological gap: the field is progressing faster in system design than in shared evaluation, interoperable analytics, and calibrated decision-making. The paper concludes that future work should combine instructional theory, inspectable learner models, standardised logging, open benchmarks, and uncertainty-aware evaluation within the same tutoring systems.

Keywords: Intelligent tutoring systems, Large language models, Learner modelling, Learning analytics, Pedagogical alignment, Systematic review

1. Introduction

Intelligent tutoring systems have occupied educational technology research for close to four decades. The early motivation was Bloom's (1984) finding that one-to-one human tutoring produced gains of around two standard deviations over conventional classroom teaching, a result that set a target successive generations of computer-based instruction have tried to reach. Meta-analyses since then have found real but more modest effects. VanLehn (2011) reports a Cohen's d of about 0.76 for step-based tutoring against no tutoring but only 0.31 for answer-based systems, Steenbergen-Hu and Cooper (2014) report between 0.32 and 0.37 for college students, and Ma et al. (2014) and Kulik and Fletcher (2016) report comparable ranges across domains. These are useful effects, but they sit well short of two standard deviations.

The situation has shifted quickly since 2023. Large language models that hold a natural conversation have entered tutoring research at speed, and a system can now cover a wide range of subject matter and produce pedagogically plausible responses without the extensive rule encoding earlier systems required. The change is real, but it has not settled into a clear picture. It is not yet known whether a tutor built on a language model can match the pedagogical fidelity of one built on instructional theory, nor how a coherent account of what a learner knows can be maintained when the tutoring agent does not itself retain state. Comparison is harder still, because systems built in different research traditions rarely share outcome measures or data formats.

Several earlier reviews have examined parts of this territory. Zawacki-Richter et al. (2019) surveyed artificial intelligence in higher education and noted how rarely educators were involved, Roll and Wylie (2016) traced the field's longer trajectory, and Luckin et al. (2016) set out an early agenda for the area. None of these, however, has examined how pedagogical grounding, adaptive mechanisms, and reproducibility infrastructure come together within the language-model systems now being built. Part of the reason is disciplinary: computer

Cite this article: Houssam, I., Chiba, Z. and Miyara, M. 2026. "Pedagogical Alignment, Adaptivity, and Analytics in Intelligent Tutoring Systems: A Systematic Review", *Electronic Journal of e-Learning*, 24(4), pp 32-43, https://doi.org/10.34190/ejel.24.4.4779

scientists tend to focus on algorithms, learning scientists on instructional design, and educational data miners on the analysis of log data, and these communities seldom work in concert.

This review addresses that gap through three analytical axes. Pedagogical alignment concerns whether a system implements an instructional theory, such as scaffolding, mastery learning, Cognitive Apprenticeship, or Productive Failure, rather than merely naming one. Adaptivity and learner modelling concern how personalisation is achieved, whether through a classical probabilistic model such as Bayesian Knowledge Tracing, through retrieval-augmented generation, or through the emergent behaviour of a language model. Analytics standards and reproducibility concern a system's engagement with the infrastructure that allows replication and comparison, including interoperability standards such as the Experience API (xAPI) and a Learning Record Store, the release of source code, and the use of shared benchmarks. Three research questions follow. RQ1: to what extent, and with what fidelity, do these systems realise explicit pedagogical theories? RQ2: what forms of adaptivity and learner modelling are used, and how are they implemented? RQ3: which analytics and interoperability standards are adopted, and what do they allow in terms of comparison and replication?

A distinguishing feature of this review is that every included study was checked against its actual source document. As Section 2 explains, this step proved necessary: a first pass found that several frequently cited bibliographic records did not match the papers they named. The verified corpus is therefore smaller than the screening totals might suggest, but each study in it has been confirmed to exist and to report what is attributed to it. Section 2 describes the protocol, the verification procedure, and the coding framework. Section 3 presents the findings axis by axis and how the axes occur together. Section 4 interprets them against the meta-analytic record and for e-learning practice. Sections 5 and 6 set out the limitations and the priorities for future work.

2. Materials and Methods

The review followed the SPAR-4-SLR protocol (Paul et al., 2021), which divides the process into assembling, arranging, and assessing, and requires each decision to be documented. The protocol was chosen because its emphasis on logging decisions suited a corpus expected to mix classical tutoring systems with recent language-model systems that report their work in very different ways, and because its audit trail partly offsets the limitations of a review conducted by a single reviewer.

2.1 Reviewer Design, Verification, and Audit Trail

Screening, source verification, extraction, and coding were conducted by the lead reviewer and logged in an audit trail. To reduce classification error, the review used pre-specified inclusion and exclusion criteria, conservative binary coding, full-text source verification, and an evidence note for every coding decision. Ambiguous cases were coded as not addressing an axis unless direct textual evidence supported inclusion. The complete coding sheet is provided as Supplementary Table S1 (Appendix A). Although the design does not include an independent second-coder reliability estimate, and Cohen's kappa is therefore not reported, the audit trail allows readers to inspect the basis for each inclusion and coding decision, and independent double coding is identified in Section 5 as the principal extension for future work.

Source verification was central to the design. To strengthen source reliability, every candidate study was opened and checked against its bibliographic record before coding, so that the synthesis rests on verified full-text documents rather than on citation metadata. This check identified a set of records whose details did not match the documents they named: five of the files initially gathered were different papers from the ones cited, one described artificial-agent learning rather than human tutoring, and one had been posted after the search window. These records were not used as primary evidence, and their disposition is reported in Supplementary Table S2 (Appendix B). The synthesis therefore rests on the twelve source-verified studies.

2.2 Assembling: Scope, Search, and Screening

The scope was defined around the three axes above. A study was included if it was a peer-reviewed article, a full conference paper, or a substantial preprint; if it described or evaluated a working tutoring system, or analysed one in a way that bore directly on an axis; if it was in English and within the search window; and if it engaged with at least one axis in a substantive rather than a passing way. A study was excluded if it addressed an axis only superficially, dealt with purely technical matters unrelated to pedagogy or learning, offered only speculation, was a duplicate, fell outside the window, or, after verification, could not be confirmed to be the work that had been cited.

2.2.1 Search strategy

Six databases were searched, chosen to balance disciplinary coverage: Scopus and Web of Science Core Collection for indexed multidisciplinary and high-impact work, IEEE Xplore and the ACM Digital Library for computer science and human-computer interaction, SpringerLink for education and the learning sciences, and arXiv for the recent preprints in which much language-model tutoring first appears. The master query combined five groups of terms, joined with the Boolean operator AND: terms for intelligent tutoring and pedagogical agents; terms for pedagogical alignment, including named theories such as scaffolding, mastery learning, Cognitive Apprenticeship, and Productive Failure; terms for adaptivity and learner modelling, including knowledge tracing and retrieval-augmented generation; terms for learning analytics and interoperability standards, including xAPI and the Learning Record Store; and terms for the educational context. Each group was a set of synonyms joined with OR and wildcarded where appropriate. The query was then adapted to the field-tag conventions of each database. The full master query and its per-database adaptations are provided as supplementary material (Appendix C, Supplementary Table S3). After retrieval, records were deduplicated by matching on title and DOI, and backward and forward citation searches were run on the studies that passed full-text screening. Because the original platform-level search export was not recoverable at the revision stage, the upstream identification and deduplication counts are not reported; to preserve transparency, the review instead documents all full-text inclusion decisions and all source-verification exclusions in the supplementary material, and restricts the synthesis to the twelve studies whose source documents were opened and verified directly (Figure 1).

The main searches covered records available up to 30 June 2025, with backward and forward citation checks conducted during revision. Searches were restricted to English-language records within the review window. Each candidate source was opened and checked against its bibliographic record before coding, and a record was retained only where the source document matched the cited study and reported evidence relevant to at least one review axis. Because the original platform-level exports were unavailable at revision, the review does not reconstruct upstream identification or deduplication counts retrospectively. The supplementary materials instead document the recoverable audit trail (Appendices A to D): the search strategy, the source-verification decisions and exclusion reasons, the coding evidence, the quality assessment, and the final verified corpus.

2.2.2 Note on the uncertainty dimension

An earlier version of this review also examined whether systems quantified the uncertainty of their own decisions and calibrated their confidence against learner behaviour. Because no included study did so, that dimension is not reported as a separate results axis; it is taken up in the conclusion, where its absence is the clearest single gap the review identifies.

2.3 Arranging and Assessing

A structured form recorded each study's bibliographic details, design, domain, and participants, together with a coding of each axis as addressed (1) or not (0) and a short evidence note quoting or locating the supporting passage. Where it was unclear whether a study addressed an axis, it was coded as not addressing it, so the figures reported below are conservative. A study was coded as addressing pedagogical alignment only where it implemented or substantively operationalised an instructional theory, not where it merely named one; as addressing adaptivity where it personalised instruction on the basis of an inferred or retrieved learner state; and as addressing analytics where it engaged with interoperability standards, released usable code or data, or used a shared benchmark.

Each study was also scored against five quality criteria, each from 0 to 2 for a maximum of ten, set out in Table 1: clarity of the research questions, appropriateness of the methods, sufficiency of implementation detail, rigour of the evaluation, and clarity of reporting. Scores across the twelve studies ranged from 6 to 8, with a median of 7 (Appendix D, Supplementary Table S4). The two lowest-scored studies were the authoring platform and the tutor-assessment study, both of which evaluate something other than learning with students; they are retained in the synthesis but weighted lightly wherever a claim rests on several sources. Because the studies differed so much in outcomes, designs, and domains, they were synthesised narratively rather than pooled.

Table 1: Quality assessment rubric (five criteria, maximum score 10)

Criterion	0	1	2
Research questions clearly stated	No questions	Implicit or vague	Clear and specific

Criterion	0	1	2
Methods appropriate and described	Not described	Partly described	Fully described
Implementation detail sufficient	No detail	Some detail	Reproducible
Evaluation rigorous	No evaluation	Informal or small	Formal study
Findings clearly reported	Unclear	Partly reported	Fully reported

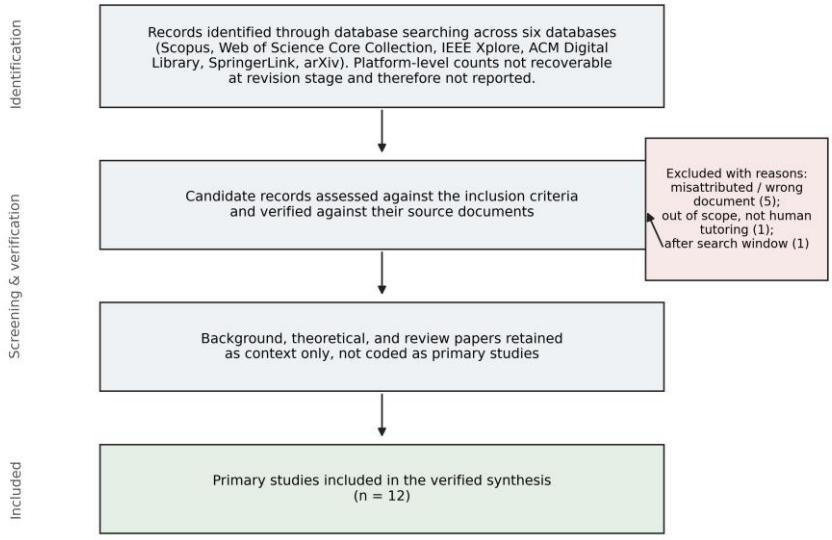

Figure 1: Flow of study selection and source verification. The diagram reports the recoverable source-verification pathway and the final verified corpus; upstream platform-level identification and deduplication counts were not reconstructed retrospectively, and the detailed search strategies, exclusion reasons, and coding decisions are provided in the supplementary materials (Appendices A to D)

3. Results

3.1 Characteristics of the Included Studies

The twelve verified studies span 2018 to 2025, and ten of them appeared in 2024 or 2025, which reflects the recent surge of language-model tutoring. Most are recent preprints rather than journal articles, a point returned to in the limitations. The domains were varied: language learning, programming, mathematics, electrical engineering, and multi-domain platforms. Most studies presented a system together with a technical or usability evaluation; only one reported a controlled comparison with a learning-outcome measure. Two of the twelve are included with the reservations noted in Section 2.3: one is an authoring platform rather than a tutor in use, and one assesses tutors' responses rather than tutoring learners directly. Both are retained for completeness and weighted lightly; a reader who excluded them would see ten studies and a slightly higher analytics proportion. They were retained because they directly evaluate tutor behaviour, authoring, or tutoring infrastructure relevant to language-model intelligent tutoring, but they are interpreted cautiously throughout, since they do not provide full learning-outcome evidence from a deployed tutoring system.

Table 2 summarises the included studies and their coding. The full study-by-study coding, with an evidence note for every judgment, is in Supplementary Table S1 (Appendix A).

By domain, three studies addressed language learning (Kwon et al., 2024; Liu et al., 2024; Liu et al., 2025), two mathematics (Puech et al., 2024; Kakarla et al., 2024), two electronics or electrical engineering (Graesser et al., 2018; Knievel, Bernhardt and Bernhardt, 2025), one programming (Li et al., 2025), and one the development of research skills (Degen and Asanov, 2025), with the remaining three operating across domains as platforms or testbeds (Smith, Gupta and MacLellan, 2024; Weitekamp, Siddiqui and MacLellan, 2025; Borchers and Shou, 2025). By design, the corpus is dominated by system descriptions paired with a technical evaluation, such as the accuracy of a component or an ablation of the architecture. One study reported a controlled comparison with a

learning measure (Degen and Asanov, 2025), one a usability study with instructors (Smith, Gupta and MacLellan, 2024), and one a systematic benchmarking of model behaviour against an existing tutoring system (Borchers and Shou, 2025). Only the established ElectronixTutor (Graesser et al., 2018) predates the recent wave; the other eleven are concentrated in 2024 and 2025. This distribution shapes what the review can conclude: it describes a young, design-led literature in which working systems are plentiful but controlled evidence of their effect on learning is still scarce.

Table 2: The twelve included studies and their coding on the three axes

Study	Year	Domain	Ped.	Adapt.	Analyt.
BIPED (Kwon et al.)	2024	Language (ESL)	Yes	Yes	No
CogGen (Li et al.)	2025	Programming	Yes	Yes	No
AITEE (Knievel, Bernhardt and Bernhardt)	2025	Electrical eng.	Yes	Yes	No
Socratic AI (Degen and Asanov)	2025	Research skills	Yes	Yes	No
Scaffolding tutor (Liu et al.)	2024	Language	Yes	Yes	No
SingaKids (Liu et al.)	2025	Language	Yes	Yes	No
Productive Failure tutor (Puech et al.)	2024	Mathematics	Yes	Yes	No
Apprentice Tutor Builder (Smith, Gupta and MacLellan)	2024	Multi-domain	No	Yes	No
Tutor assessment (Kakarla et al.)	2024	Mathematics	Yes	No	No
TutorGym (Weitekamp, Siddiqui and MacLellan)	2025	Multi-domain	Yes	Yes	Yes
ElectronixTutor (Graesser et al.)	2018	Electronics	Yes	Yes	Yes
Adaptivity benchmark (Borchers and Shou)	2025	Multi-domain	Yes	Yes	Yes

Note: Ped. = pedagogical alignment; Adapt. = adaptivity and learner modelling; Analyt. = analytics standards and reproducibility.

3.2 Distribution Across the Three Axes

Pedagogical alignment and adaptivity were each addressed by eleven of the twelve studies (92 percent). Analytics standards were addressed by three (25 percent). Figure 2 shows the distribution. The imbalance is the central descriptive finding: designing a system that is pedagogically informed and adaptive is now ordinary practice, but engaging with the shared infrastructure that would let one system be compared with another is not.

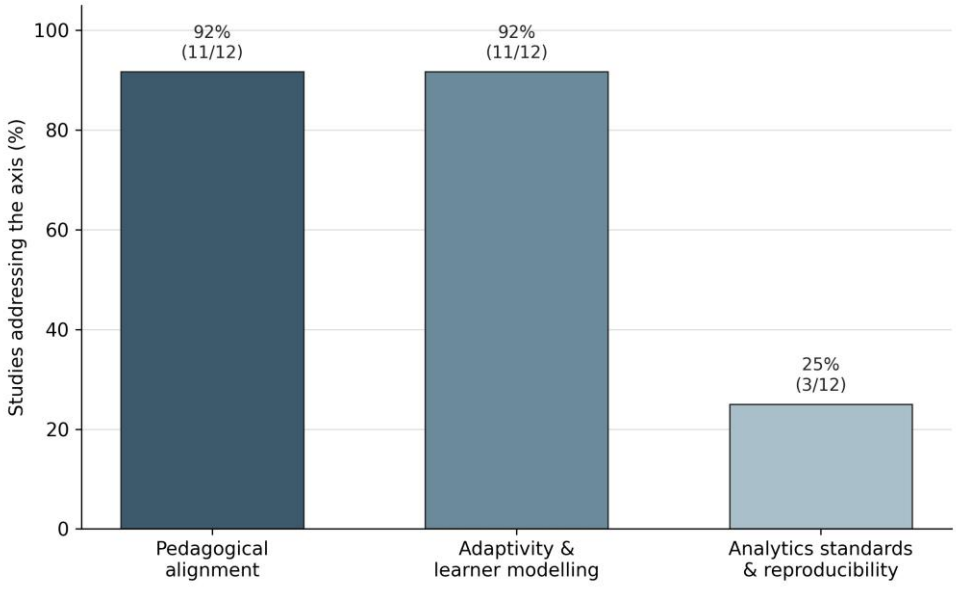

Figure 2: Proportion of the twelve included studies addressing each axis

3.3 Pedagogical Alignment (RQ1)

Most studies engaged with pedagogy, but the depth varied. The strongest cases implemented a named instructional theory in a way that shaped the system's behaviour. CogGen (Li et al., 2025) is built on the Cognitive Apprenticeship framework, mapping its six teaching moves, from modelling through coaching, scaffolding, articulation, and reflection to exploration, onto a domain-specific language that structures each generated turn around programming videos. The Productive Failure tutor of Puech et al. (2024) operationalises Kapur's account of productive failure as an explicit steering strategy, in which the tutor deliberately withholds direct instruction so that learners first generate and explore their own, often flawed, solutions. AITEE (Knievel, Bernhardt and Bernhardt, 2025) and the Socratic AI tutor of Degen and Asanov (2025) implement Socratic questioning to keep the initiative with the learner, the former in electrical engineering and the latter in the development of research questions, where it scaffolds the learner through an adapted PICOT structure. The two scaffolding-based systems (Liu et al., 2024; Liu et al., 2025) make graduated support explicit, fading prompts as competence grows; the more recent of the two, a multilingual dialogic tutor for young language learners, treats the calibration of scaffolding to proficiency as a design target in its own right.

BIPED (Kwon et al., 2024) takes a distinctive empirical route to pedagogical grounding. Rather than encoding a single theory, its authors annotated recordings of human tutoring to derive a lexicon of thirty-four tutor dialogue acts and nine student acts, then built a two-step system that first predicts an appropriate tutor act for the current turn and only then generates an utterance to realise it. This grounds the system's moves in observed tutoring practice and makes its pedagogical intent inspectable turn by turn, although the reported accuracy of act prediction is modest, which the authors acknowledge as a limit on how reliably the intended move is carried out.

Two patterns qualify the picture. First, pedagogical grounding was often asserted more firmly than it was tested: few studies checked whether a pedagogically structured design produced better learning than an unstructured one, so the link between the theory invoked and any measured benefit usually remains an assumption. Second, the systems built on language models tend to realise pedagogy through prompting rather than through encoded rules, so the intended strategy is encouraged rather than guaranteed, and several authors note that behaviour varies from one run to the next. The authoring platform of Smith, Gupta and MacLellan (2024) was coded as not implementing a pedagogical theory itself, since the pedagogy resides in the tutors a user builds with it rather than in the platform; its contribution is to lower the expertise needed to author a tutor at all, by letting a non-programmer demonstrate worked examples from which an apprentice-learning agent induces an expert model.

3.4 Adaptivity and Learner Modelling (RQ2)

Adaptivity was the most developed axis and took three broad forms. A first group keeps an explicit learner model. CogGen (Li et al., 2025) uses Bayesian Knowledge Tracing to estimate mastery and decide when to advance, so that the Cognitive Apprenticeship sequence is driven by an inferred learner state rather than by a fixed script. The testbed of Weitekamp, Siddiqui and MacLellan (2025) and the established ElectronixTutor (Graesser et al., 2018) both track mastery over knowledge components, the latter through a recommender that draws on a persistent student model combining cognitive and non-cognitive attributes. These systems inherit the interpretability of the classical tradition: the basis for each adaptive decision can be read off the model. A second group adapts through retrieval. AITEE (Knievel, Bernhardt and Bernhardt, 2025) uses a graph-based similarity measure to pull the most relevant lecture material into the prompt and pairs this with a circuit simulator, so that adaptation is a matter of supplying the model with the right context rather than of maintaining a model of the learner.

A third and now common group relies on the language model itself, with adaptivity emerging from the prompt context and the conversation history rather than from a maintained model of the learner (Degen and Asanov, 2025; Kwon et al., 2024; Liu et al., 2024; Liu et al., 2025; Puech et al., 2024). These systems adjust the depth of an explanation, the difficulty of a question, or the wording of a hint in response to what a learner has just said, and they do so fluently. What they generally do not do is represent the learner's state in a form that persists across a session or that can be inspected, which makes their adaptation harder to audit and, as several authors note, less stable from one interaction to the next.

Two limits run through the corpus. The systems that adapt through a language model rarely report a quantitative measure of how precisely they adapt, and almost none compares adaptation driven by a language model against a model-based control. The clearest evidence on this point comes from Borchers and Shou (2025), whose benchmarking study varied the contextual information given to several models across seventy-five real tutoring

scenarios and assessed the resulting moves for both adaptivity and pedagogical soundness. Even the best-performing model only marginally matched the adaptivity of an established tutoring system, and the model judged most pedagogically sound was also the least reliable at following instructions. That result suggests the gap between emergent and modelled adaptivity is real and measurable, and that fluency in dialogue should not be mistaken for precision in adaptation.

3.5 Analytics Standards and Reproducibility (RQ3)

Analytics and reproducibility infrastructure formed the weakest axis, addressed by three of the twelve studies. ElectronixTutor (Graesser et al., 2018) records activity to a Learning Record Store, the interoperability mechanism associated with the Experience API. TutorGym (Weitekamp, Siddiqui and MacLellan, 2025) is an open, shared testbed with a public code repository that lets different agents be evaluated against the same tasks. Borchers and Shou (2025) release open benchmarking code and a reproducible procedure. The remaining nine studies released, at most, a dataset or a dependency, and none adopted a community xAPI profile or a shared outcome measure that would let its results be placed alongside another system's.

The consequences run through the synthesis. No two of the language-model studies used the same outcome measure on a comparable population, which rules out any quantitative pooling, and the scarcity of open implementations makes external replication difficult. Where code or data were released at all, it was typically a model dependency or a one-off dataset rather than anything designed for reuse, and outcome reporting ranged from automatic accuracy metrics to small usability studies, with no shared definition of what a successful tutoring interaction is. The two pieces of genuine shared infrastructure in the corpus, TutorGym and the benchmark of Borchers and Shou, are recent and point to what a more cumulative version of the field might look like; notably, both come from research groups working on the evaluation problem itself rather than on a particular tutor, which may be why analytics is so much better served in their work than in the system-building studies that make up the rest of the corpus.

3.6 Integration Across Axes

Pedagogical alignment and adaptivity occurred together in ten of the twelve studies, the dominant pairing and an unsurprising one. Analytics co-occurred with each of the other two axes in only three studies. Figure 3 shows the full pattern. Three studies addressed all three axes at once: ElectronixTutor (Graesser et al., 2018), TutorGym (Weitekamp, Siddiqui and MacLellan, 2025), and the benchmarking study of Borchers and Shou (2025). It is worth noting that the most integrated work is either an established pre-language-model system or recent infrastructure built deliberately for comparison, rather than one of the new conversational tutors, which cluster on pedagogy and adaptivity and leave analytics aside.

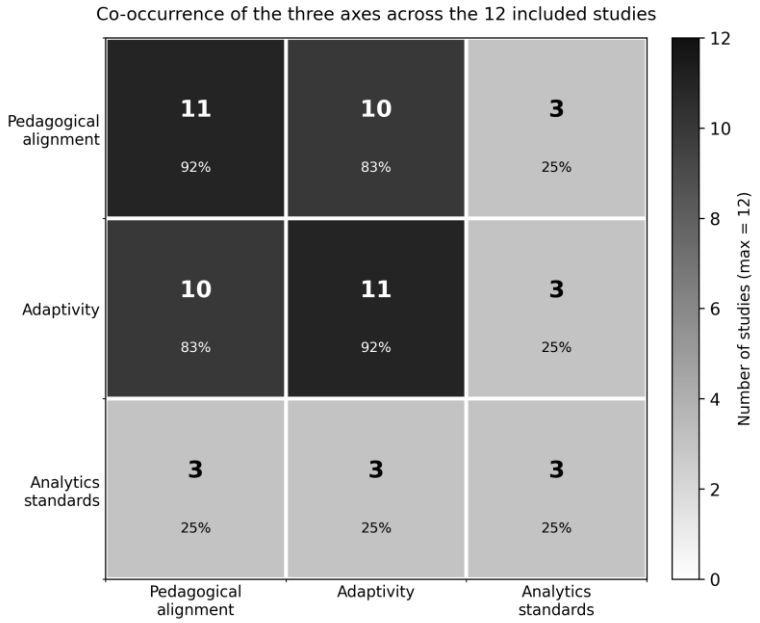

Figure 3: Co-occurrence of the three axes across the twelve included studies (diagonal cells give the total per axis)

This pattern is the review's clearest structural result. The newest conversational tutors are strong on pedagogy and adaptivity but treat analytics as an afterthought, while the studies that take analytics seriously are not themselves novel tutors. No single study brings the three strengths together, and, as the conclusion discusses, none adds a treatment of decision uncertainty on top. The field is therefore advancing along separate tracks: a system-building track that produces fluent, pedagogically framed tutors whose behaviour is hard to verify, and an evaluation track that produces the infrastructure to verify systems but few new tutors to run through it. The value of the most integrated work in the corpus lies in showing that the tracks can meet, but they do so rarely, and never yet in a single recent language-model tutor.

4. Discussion

4.1 Main Findings

The review describes a field that is technically sophisticated on the two axes its researchers attend to most, pedagogical alignment and adaptivity, and underdeveloped on the one that depends on cross-disciplinary infrastructure. Pedagogy is widely claimed but unevenly implemented, and in the language-model systems it is encouraged through prompting rather than guaranteed by design. Adaptivity is the most mature axis, yet the shift from modelled adaptivity to emergent, prompt-driven adaptation is exactly where the evidence is thinnest, and the one benchmarking study that bears on it suggests that language models still trail established systems in matching a learner's needs.

4.2 The Analytics gap

The clearest structural finding is that analytics standards are underdeveloped, and that this holds back the accumulation of knowledge. Most of the recent systems cannot be compared with one another, because they define outcomes differently and rarely share code or data. The obstacle is one of coordination rather than technology: the standards exist, and the two shared resources in the corpus show they can be used, but adopting them requires the kind of cooperation the field's fragmentation makes difficult. Even a small set of agreed benchmarks and a shared logging profile would change what comparison and synthesis are possible.

It is worth dwelling on why this matters more than it might first appear. A field that cannot compare its systems cannot tell which design choices are responsible for which outcomes, and so it cannot improve cumulatively; each new system starts an argument rather than settling one. The neighbouring field of educational data mining made progress in part by building shared datasets and repositories that allowed the same question to be asked of many systems, and the analytics tradition has long argued that interoperability is a precondition for that kind of advance (Baker and Siemens, 2014). The tutoring subfield has not yet followed. The two exceptions in this corpus are instructive: TutorGym is valuable precisely because it lets different agents be run against the same tasks, and the contribution of Borchers and Shou (2025) is as much the reusable benchmarking procedure as the particular result. Both point to a version of the field in which claims about adaptivity or pedagogy could be checked against a common yardstick rather than asserted system by system.

4.3 Comparison with the Meta-analytic Record

Where causal evidence exists, it is consistent with the wider record. VanLehn (2011), Steenbergen-Hu and Cooper (2014), Ma et al. (2014), and Kulik and Fletcher (2016) place intelligent tutoring in the moderate range, well below the two-sigma mark of expert human tutoring. The verified corpus adds little new causal evidence of its own. Only the Socratic AI study (Degen and Asanov, 2025) reports a controlled comparison, contrasting a Socratic tutor with an uninstructed chatbot on a pre-test and post-test design, and the authors describe their findings as preliminary and plan to extend the sample. The Productive Failure tutor of Puech et al. (2024) reports an encouraging field test, but the relevant comparison concerns the fidelity of the tutoring strategy rather than a controlled learning-outcome contrast, so it is not read here as an effect-size estimate. On present evidence, language-model tutors look likely to land in the same moderate range as their predecessors rather than to transform it, and the field still lacks the controlled comparisons that would settle the question.

4.4 Implications for e-Learning Practice

For researchers, the analytics axis is where the greatest leverage lies, since the alternative to shared evaluation is a continuing accumulation of case studies that cannot be compared.

For those who design systems, language-model tutors call for particular attention to pedagogical fidelity, whether through hybrid architectures that pair a language model with a persistent learner model or through carefully constrained prompting. The corpus hints at what such hybrids might look like: CogGen's use of Bayesian

Knowledge Tracing alongside a generative front end, and the way TutorGym and ElectronixTutor anchor adaptation in an explicit model of knowledge components, suggest that the interpretability of the classical tradition and the fluency of the generative one are not mutually exclusive. A tutor that generates dialogue freely but defers decisions about what to teach next to an inspectable learner model would address the main weakness of the purely prompt-based systems, namely that their adaptation cannot be audited or guaranteed.

For teachers and other practitioners introducing these tools, the review offers a realistic basis for expectation: the benefits are plausible but largely unproven, the systems work best as a supplement to classroom instruction rather than a replacement for it, and material aimed at practitioners should reflect those limits rather than present language-model tutoring as more transformative than the evidence supports. The honest message for a school or university weighing adoption is that these systems are promising and worth piloting, but that their effectiveness for a particular cohort is an open question that local evaluation, rather than vendor demonstration, should answer.

4.5 Beyond the Three Axes

Two issues fall outside the coding framework but bear on the field's overall effect. The first is privacy: conversational tutors process detailed records of interaction and material that can reveal sensitive cognitive or emotional states, and the privacy-protecting architectures and governance well developed elsewhere in artificial intelligence are little used in education. The second is equity of access: the promise of intelligent tutoring has been to widen access to personalised instruction, yet subscription platforms and the infrastructure they need risk directing the best tools towards the learners who need them least. The studies reviewed here do not address either question systematically, and a future review that treated them as analytical dimensions in their own right would add something the current literature lacks.

5. Limitations

Several limitations define the scope of the conclusions. First, the review was conducted by a single reviewer; although conservative coding, source verification, and an auditable extraction sheet reduced the risk of classification error, future work should extend the design through independent double coding and inter-rater reliability reporting. Second, the verified corpus is small and includes several recent preprints, reflecting the early stage of language-model tutoring research, so some findings may settle as that work reaches final publication. Third, controlled learning-outcome evidence remains scarce, so the review is best interpreted as a source-verified map of current design practice rather than as a pooled estimate of effectiveness. The three-axis framework is also one reasonable lens among several, and a framework built around affective computing, multimodal interaction, or accessibility would bring different features into focus.

Two reporting considerations follow from this. The certainty of the evidence is bounded by the corpus size, the predominance of preprints, and the scarcity of controlled studies, and some reporting bias is plausible, since novel or successful systems are more likely to be disseminated than unsuccessful classroom deployments. In addition, because the original platform-level exports were unavailable at revision, conventional identification and deduplication counts are not reported; the review therefore prioritises recoverable full-text verification, documented exclusions, and transparent study-level coding, which together support an auditable reading of the corpus.

6. Conclusion

On the evidence of this review, recent intelligent tutoring research is technically accomplished in grounding systems pedagogically and in making them adaptive, and underdeveloped in the shared infrastructure that lets results be compared and reproduced. The arrival of language-model tutors has widened what is possible in dialogue and subject coverage, but it has not yet been shown to lift learning outcomes beyond the moderate range established for earlier systems, and the most integrated work in the corpus is either an older system or new infrastructure built for comparison rather than one of the new conversational tutors.

One absence deserves to be named directly, because it cuts across all three axes. None of the twelve systems quantifies the uncertainty of its own decisions or calibrates its confidence against what learners actually do. Elsewhere in artificial intelligence, wherever decisions carry consequences, calibration metrics such as Expected Calibration Error and Brier scores are reported as a matter of course; in tutoring, where a confidently stated wrong answer can mislead a learner, they are absent. The methods are well established and the difficulty is one of attention rather than technique, which makes this the clearest single priority for the field.

The most productive direction, then, is to make the field more cumulative and more self-aware: to run controlled comparisons of language-model tutoring against tutoring grounded in instructional theory, to report calibration using the metrics already standard elsewhere, to develop shared benchmarks and logging profiles as a community, and to build systems that bring pedagogical theory, adaptive mechanisms, and standardised analytics together in one place. None of these alone will reach Bloom's two-sigma benchmark, but together they would let the field find out, for the first time, how close it can come. The central gap is therefore not simply that language-model tutors need better prompts, but that the next generation of intelligent tutoring systems must combine instructional theory, inspectable learner models, interoperable analytics, open evaluation protocols, and explicit uncertainty calibration within the same system.

AI Statement: The authors used an AI-assisted tool during two stages of this review. First, Claude was used to support the application of the SPAR-4-SLR protocol, specifically in organising and analysing candidate studies against the defined inclusion and exclusion criteria and in coding primary studies across the analytical axes. All screening decisions and coding judgements were verified and finalised by the authors. Second, Claude was used to assist with linguistic editing and grammatical enhancement of the manuscript. The intellectual content, research design, data interpretation, and conclusions are entirely the authors' own.

Ethics Statement: This systematic review analysed published or publicly available scholarly documents and did not involve human participants, intervention, or access to private learner records. Formal ethics approval was therefore not required.

Protocol and registration: No external review protocol was registered. The review followed the SPAR-4-SLR procedure, and the screening, verification, extraction, and coding decisions were logged in an internal audit trail. The audit trail, search strategy, coding sheet, quality assessment, and source-verification exclusion table are provided as supplementary materials (Appendices A to D).

Data availability: The data supporting this review consist of the search strategy, the study-selection audit trail, the source-verification exclusion table, the study extraction and coding sheet, and the quality assessment scores, all provided as supplementary files (Appendices A to D). No new learner-level dataset was created or analysed.

Competing interests: The authors declare no competing interests.

Author contributions: The authors jointly designed the review, contributed to manuscript revision, and approved the submitted version. The search, screening, source verification, extraction, coding, figures, and supplementary materials were prepared and checked through the documented audit trail.

References

Baker, R.S. and Siemens, G., 2014. Educational data mining and learning analytics. In: R.K. Sawyer, ed. *The Cambridge Handbook of the Learning Sciences*. 2nd ed. Cambridge: Cambridge University Press, pp. 253-272.

Bloom, B.S., 1984. The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6), pp. 4-16.

Borchers, C. and Shou, T., 2025. *Can large language models match tutoring system adaptivity? A benchmarking study*. arXiv:2504.05570.

Degen, P.-B. and Asanov, I., 2025. *Beyond automation: Socratic AI, epistemic agency, and the implications of the emergence of orchestrated multi-agent learning architectures*. Preprint, University of Kassel.

Dolata, M., Katsiuba, D., Wellnhammer, N. and Schwabe, G., 2023. Learning with digital agents: an analysis based on the activity theory. *Journal of Management Information Systems*, 40(1), pp. 56-95.

Graesser, A.C., Hu, X., Nye, B.D., VanLehn, K., et al., 2018. ElectronixTutor: an intelligent tutoring system with multiple learning resources for electronics. *International Journal of STEM Education*, 5, 15.

Holstein, K., Aleven, V. and Rummel, N., 2020. A conceptual framework for human-AI hybrid adaptivity in education. In: *Artificial Intelligence in Education*. Lecture Notes in Computer Science. Cham: Springer, pp. 240-254.

Kakarla, S., Thomas, D.R., Lin, J., Gupta, S. and Koedinger, K.R., 2024. *Using large language models to assess tutors' performance in reacting to students making math errors*. arXiv:2401.03238.

Knievel, C., Bernhardt, A. and Bernhardt, C., 2025. *AITEE - agentic tutor for electrical engineering*. arXiv:2505.21582.

Kulik, J.A. and Fletcher, J.D., 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research*, 86(1), pp. 42-78.

Kwon, S., Kim, S., Park, M., Lee, S. and Kim, K., 2024. *BIPED: pedagogically informed tutoring system for ESL education*. arXiv:2406.03486.

Li, W., Pea, R., Haber, N. and Subramonyam, H., 2025. *CogGen: a learner-centered generative AI architecture for intelligent tutoring with programming videos*. arXiv:2506.20600.

Liu, Z., Yin, S.X., Lee, C. and Chen, N.F., 2024. *Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions*. arXiv:2404.03429.

Liu, Z., Lin, G., Tan, H.L., Zhang, H., et al., 2025. *SingaKids: a multilingual multimodal dialogic tutor for language learning*. arXiv:2506.02412.
Luckin, R., Holmes, W., Griffiths, M. and Forcier, L.B., 2016. *Intelligence Unleashed: An Argument for AI in Education*. London: Pearson.
Ma, W., Adesope, O.O., Nesbit, J.C. and Liu, Q., 2014. Intelligent tutoring systems and learning outcomes: a meta-analysis. *Journal of Educational Psychology*, 106(4), pp. 901-918.
Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., et al., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
Paul, J., Lim, W.M., O'Cass, A., Hao, A.W. and Bresciani, S., 2021. Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *International Journal of Consumer Studies*, 45(4), pp. 1-16.
Puech, R., Macina, J., Chatain, J., Sachan, M. and Kapur, M., 2024. *Towards the pedagogical steering of large language models for tutoring: a case study with modeling productive failure*. arXiv:2410.03781.
Roll, I. and Wylie, R., 2016. Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26, pp. 582-599.
Shute, V.J. and Ventura, M., 2013. *Stealth Assessment: Measuring and Supporting Learning in Video Games*. Cambridge, MA: MIT Press.
Smith, G., Gupta, A. and MacLellan, C., 2024. *Apprentice tutor builder: a platform for users to create and personalize intelligent tutors*. arXiv:2404.07883.
Steenbergen-Hu, S. and Cooper, H., 2014. A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, 106(2), pp. 331-347.
VanLehn, K., 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), pp. 197-221.
Weitekamp, D., Siddiqui, M.N. and MacLellan, C.J., 2025. *TutorGym: a testbed for evaluating AI agents as tutors and students*. arXiv:2505.01563.
Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F., 2019. Systematic review of research on artificial intelligence applications in higher education - where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 39.

Appendix A. Supplementary Table S1: Study-by-study Coding Sheet

Study	Pedagogical alignment	Adaptivity	Analytics
BIPED (Kwon et al., 2024)	Yes	Yes	No
CogGen (Li et al., 2025)	Yes	Yes	No
AITEE (Knievel, Bernhardt and Bernhardt, 2025)	Yes	Yes	No
Socratic AI (Degen and Asanov, 2025)	Yes	Yes	No
Scaffolding tutor (Liu et al., 2024)	Yes	Yes	No
SingaKids (Liu et al., 2025)	Yes	Yes	No
Productive Failure tutor (Puech et al., 2024)	Yes	Yes	No
Apprentice Tutor Builder (Smith, Gupta and MacLellan, 2024)	No	Yes	No
Tutor assessment (Kakarla et al., 2024)	Yes	No	No
TutorGym (Weitekamp, Siddiqui and MacLellan, 2025)	Yes	Yes	Yes
ElectronixTutor (Graesser et al., 2018)	Yes	Yes	Yes
Adaptivity benchmark (Borchers and Shou, 2025)	Yes	Yes	Yes

Appendix B. Supplementary Table S2: Source-verification Exclusions

Exclusion category	n	Reason	Treatment
Mismatched / wrong document	5	Supplied source document did not match the cited bibliographic record	Excluded from primary synthesis
Not human tutoring	1	Concerned artificial-agent (multi-agent RL) learning, not human tutoring	Excluded from primary synthesis

Exclusion category	n	Reason	Treatment
Posted after search cut-off	1	Record fell outside the review search window	Excluded from primary synthesis
Total	7		Reported in the flow diagram

Appendix C. Supplementary Table S3: Search Strategy

Master query (five concept groups joined by AND; synonyms within each group joined by OR), adapted to each database's field syntax:

("intelligent tutoring system*" OR "AI tutor*" OR "pedagogical agent*" OR "educational chatbot*" OR "LLM tutor*" OR "dialogue-based tutor*")

AND (pedagog* OR "instructional design" OR "learning theory" OR scaffolding OR "mastery learning" OR "cognitive apprenticeship" OR "productive failure" OR Socratic)

AND (adaptiv* OR "learner model*" OR "student model*" OR "knowledge tracing" OR BKT OR IRT OR "retrieval-augmented" OR RAG OR personali*)

AND ("learning analytics" OR analytics OR xAPI OR SCORM OR "learning record store" OR LRS OR reproducib* OR benchmark* OR "open source")

AND (education OR learning OR student* OR "higher education" OR "K-12")

Scopus: TITLE-ABS-KEY(master query)

Web of Science Core Collection: TS=(master query)

IEEE Xplore: each group wrapped in "All Metadata": (...), joined with AND

ACM Digital Library: each group as [[All: term] OR ...], joined with AND

SpringerLink: master query in title/abstract/full-text advanced search

arXiv: all: terms per group, joined with AND, restricted to cat:cs.AI OR cs.CL OR cs.CY OR cs.LG OR cs.HC

Appendix D. Supplementary Table S4: Quality Assessment

Study	Total /10	Note
BIPED (Kwon et al., 2024)	7	Dialogue-act grounding; no learning-outcome trial
CogGen (Li et al., 2025)	7	Cognitive Apprenticeship + BKT; technical evaluation
AITEE (Knievel, Bernhardt & Bernhardt, 2025)	7	Socratic + RAG; knowledge-application benchmark
Socratic AI (Degen & Asanov, 2025)	7	Controlled comparison; preliminary
Scaffolding tutor (Liu et al., 2024)	7	Empirical case study
SingaKids (Liu et al., 2025)	7	Child-user evaluation
Productive Failure tutor (Puech et al., 2024)	7	Field test; not a controlled outcome effect
Apprentice Tutor Builder (Smith, Gupta & MacLellan, 2024)	6	Authoring platform; usability study (n=14)
Tutor assessment (Kakarla et al., 2024)	6	Assesses tutor responses, not tutoring learners
TutorGym (Weitekamp, Siddiqui & MacLellan, 2025)	8	Shared open testbed
ElectronixTutor (Graesser et al., 2018)	8	Established ITS; LRS logging
Adaptivity benchmark (Borchers & Shou, 2025)	8	Open, reproducible benchmarking method