International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Reliable Multimodal AI for Structured Knowledge Extraction and Study Material Generation in Real Classrooms: A Transparent Scoping Survey, Taxonomy, Benchmarks, and Research Roadmap

Soma Kiran Kumar Nellipudi[1], Nidhibehen Patel[2]

Submitted:03/02/2026 **Revised**: 05/03/2026 **Accepted:** 15/03/2026

Abstract: Educational knowledge in real classrooms is distributed across speech, slides, whiteboards, handwritten mathematics, code, and ad hoc diagrams. This makes accurate and persistent study support difficult even when recordings are available. Recent multimodal models and large language model (LLM) systems can summarize lectures and generate notes, but real deployment remains limited by alignment drift, OCR and ASR noise, incomplete extraction of formal STEM content, and hallucinations that can silently corrupt study artifacts. This paper presents a transparent scoping survey of a balanced 100-paper corpus organized into five clusters: multimodal lecture understanding, educational artifact generation, structured knowledge extraction, reliability and hallucination control, and benchmarks and evaluation. We explicitly treat the last two clusters as a transfer toolkit layer for classroom AI rather than as classroom-native systems. Beyond synthesis, the paper contributes: (1) a review protocol with an explicit audit trail and descriptive-count caveats; (2) a reliability-first classroom pipeline in which alignment is the operational core; (3) an operational intermediate representation (IR) with typed fields, evidence granularity, verification records, and abstention behavior; (4) a worked micro-example that carries a 30-second lecture snippet into evidence-linked flashcards; (5) a lecture-grounded versus resource-grounded verification matrix; and (6) a reviewer-ready multimodal faithfulness protocol for mixed evidence such as noisy board crops, OCR, and ASR. The result is a sharper, more operational roadmap for trustworthy classroom AI.

Index Terms—*Multimodal learning, lecture understanding, automatic note generation, educational knowledge graphs, retrieval-augmented generation, factuality, verification, benchmarks, trustworthy AI.*

I. Introduction

Real classroom knowledge is transient, multimodal, and unevenly captured. In a single lecture, the instructor may speak an informal definition, point to a slide bullet, derive an equation on the board, sketch a diagram, and type code in an editor. Students rarely miss only one modality; they miss the linkages between modalities. This is why raw lecture capture is not the same as reliable study support. Recordings preserve time, but students still need structure, retrieval, and verification to transform classroom events into reusable artifacts such as notes, concept maps, algorithm steps, flashcards, and question-answer pairs [4], [6], [13], [21], [22].

Current lecture summarization and note-generation systems partially address this problem, but they remain brittle in precisely the conditions that matter in real classrooms. Automatic speech recognition degrades under room noise and rapid domain-specific terminology; OCR on boards and projected slides is sensitive to blur, glare, handwriting, and occlusion; and multimodal alignment often drifts when the instructor elaborates away from the slide sequence or writes content not present in the deck [5], [10], [11], [14]-[19]. Many systems also stop at fluent summaries instead of producing structured, reusable, and checkable learning objects.

[1]*Senior Software Engineer, Interactive Communications International, Inc. | ORCID: 0009-0003-5400-5958*

[2]*Test Automation Engineer, Verizon | ORCID: 0009-0000-1958-266X*

Reliability is therefore the central deployment barrier. Educational artifacts are higher stakes than generic summaries because an unsupported definition, an omitted algorithm step, a wrong asymptotic complexity claim, or a silently repaired code snippet can directly distort learning. The factuality literature provides important tools—FactCC, QAGS, FRANK, SummaC, TRUE, TruthfulQA, HaluEval, SelfCheckGPT, RARR, and MiniCheck—but these resources were largely built for open-domain or text-only settings and do not directly solve multimodal classroom grounding [65]-[75]. The key research question is not whether these tools matter, but how they should be adapted when evidence comes from ASR spans, slide OCR blocks, board crops, diagrams, and course resources rather than a clean source paragraph.

This revised survey makes that adaptation explicit. We distinguish classroom-native literature from a transfer toolkit layer. Clusters C1–C3 address classroom capture, alignment, artifact generation, and structured extraction directly. Clusters C4–C5 supply reusable mechanisms—retrieval, factuality metrics, hallucination detectors, and evaluation suites—that must be operationalized for classroom evidence. Framing C4–C5 as a transfer layer resolves the conceptual mismatch between a multimodal classroom paper and a partially text-centric reliability toolkit: the text-centric literature is included because it contributes verification primitives, not because it is itself classroom-complete.

The paper makes six contributions. First, it clarifies the review as a transparent scoping survey rather than a prevalence-estimating systematic review and strengthens the audit trail around search, screening, coding, and corpus-level counts. Second, it centers alignment as the unifying technical core and categorizes alignment families by alignment unit, cues, drift-handling strategy, and metrics. Third, it operationalizes the IR with explicit fields, evidence granularity, and verification tasks for definitions, equations, code, and complexity claims. Fourth, it adds a professor-requested worked example that traces a 30-second lecture snippet into evidence spans, IR objects, and a cited flashcard. Fifth, it separates lecture-grounded from resource-grounded verification and specifies which artifact types require which evidence. Sixth, it tightens the evaluation framework by defining measurable dimensions and a mixed-evidence protocol for board-crop + OCR + ASR faithfulness. The remainder of the paper proceeds from protocol and positioning to pipeline, IR, corpus synthesis, benchmarks, evaluation, and deployment.

II. Review Protocol, Positioning, And Audit Trail

A. Review questions and corpus design

The manuscript is positioned as a structured scoping survey. The goal is not to estimate publication prevalence across the entire literature or to pool effect sizes. Instead, the goal is to compare method families under a common coding lens and to expose the design choices that matter for reliable classroom AI. The corpus is intentionally balanced across five clusters to avoid allowing publication-volume asymmetries to dominate the narrative.

The review is driven by four practical questions: RQ1 asks how classroom systems segment, align, and perceive long multimodal lectures; RQ2 asks how educational systems generate study artifacts and tutoring interactions from those signals; RQ3 asks what intermediate representations enable controllable extraction of definitions, procedures, claims, and prerequisites; and RQ4 asks which reliability mechanisms and benchmarks transfer effectively into classroom settings.

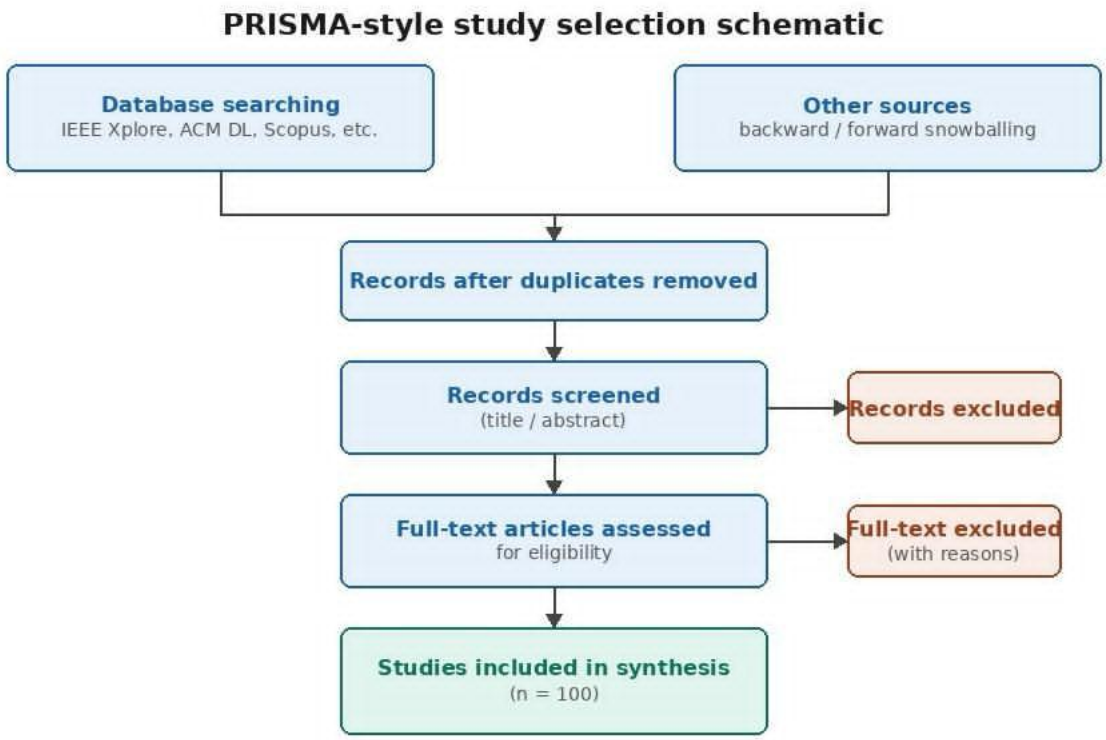

Figure 1. PRISMA-style selection schematic adapted for the balanced 100-paper scoping corpus; the accompanying audit trail consists of the query log, eligibility notes, coding sheet, and adjudication log.

Cluster	Representative query template	Primary databases	Inclusion focus
C1: Multimodal lecture understanding	(lecture OR classroom) AND (video OR audio) AND (segmentation OR alignment OR "whiteboard" OR OCR)	IEEE Xplore, ACM DL, CVF, ISCA	Segmentation, slide/board alignment, classroom capture, OCR/ASR, navigation
C2: Artifact generation	(lecture OR course) AND (notes OR flashcards OR "question generation" OR tutoring)	IEEE Xplore, ACM DL, SpringerLink	Note generation, tutoring, flashcards, question generation, study tools
C3: Structured extraction	(education OR document) AND (definition extraction OR prerequisite OR "knowledge graph" OR algorithm extraction)	ACL Anthology, ACM DL, IEEE Xplore	Definitions, prerequisite graphs, structured extraction, document understanding
C4: Reliability	(hallucination OR factuality OR retrieval OR	ACL Anthology,	Retrieval, factuality,

Cluster	Representative query template	Primary databases	Inclusion focus
transfer toolkit	grounded generation) AND (LLM OR summarization)	arXiv, IEEE Xplore	self-revision, grounded tutoring, contradiction detection
C5: Benchmarks and evaluators	(multimodal benchmark OR lecture dataset OR VQA OR evaluator) AND (education OR document OR video)	CVF, ACL Anthology, IEEE Xplore, arXiv	Lecture datasets, VQA, multimodal evaluation toolkits, metric suites

Table I. Transparent search strategy organized by cluster-specific query families.

B. Search, screening, and audit trail

Primary search sources were IEEE Xplore and ACM Digital Library, complemented by Scopus/Web of Science indexing, SpringerLink, ACL Anthology, arXiv, CVF Open Access, and ISCA proceedings where these sources were needed to capture NLP, vision, and speech papers adjacent to classroom AI. Query families were cluster-specific and combined classroom terms with task terms such as segmentation, alignment, note generation, definition extraction, hallucination detection, retrieval, and benchmarking. Backward and forward snowballing were used when a paper functioned as a bridge between clusters.

Eligibility followed a relevance-first logic. We included papers that either (i) operate directly on lecture, classroom, educational, or course materials, or (ii) contribute a mechanism that can be concretely transferred into educational artifact generation and verification. Studies were excluded when they were purely generic and offered no clear path to classroom evidence adaptation, when they focused only on administrative education analytics without content grounding, or when they did not expose enough methodological detail to support interpretive coding.

To improve auditability, the revised protocol separates title/abstract screening, full-text eligibility, and coding into distinct logged stages. Each retained paper receives one primary cluster label and multi-label tags for modalities, pipeline stages, alignment family, evidence type, evaluation style, and study quality. The intended protocol is dual screening with adjudication: a shared codebook is calibrated on a pilot subset, disagreements are logged, and final labels are reconciled by discussion with a third-author adjudication step when needed. Because the archival artifact currently available to this revision is a reconciled coding sheet rather than the raw double-coding worksheets, the paper reports descriptive corpus statistics only and does not claim meta-analytic precision. The protocol therefore makes the missing audit fields explicit—overlap size, coder agreement, disagreement counts, and adjudication log—even where the preserved material is limited.

Minimum audit package for future updates: query log; deduplication rules; title/abstract decisions; full-text exclusion reasons; reconciled coding sheet; and the adjudication record for disagreements. This revision makes those required fields explicit and limits corpus counts to descriptive summaries.

C. Coding schema, quality appraisal, and interpretation limits

The final corpus contains 100 studies, intentionally balanced as 20 papers per cluster. This design improves cross-cluster comparability but also changes the interpretation of counts: a statement such as “17/20 C2 papers contain human or user evidence” is descriptive of the curated corpus, not an estimate of the field-wide prevalence of that evaluation practice. Appendix A provides the paper-level mapping table used to derive all main-text corpus summaries.

We also record a lightweight quality rubric to inform narrative weighting rather than formal exclusion. Dataset realism, grounding explicitness, evaluation depth, and deployment relevance are

coded on a 0–2 scale and used to distinguish conceptually interesting but weakly validated papers from stronger deployment-oriented evidence. This helps the survey avoid treating all papers as methodologically equivalent when discussing roadmap priorities.

Field	Operational coding rule	Why it is recorded	Audit note
Primary cluster	Assign one dominant cluster per paper; use multi-label stage tags to avoid collapsing cross-cutting contributions.	Supports balanced corpus synthesis.	Disagreements should be adjudicated and logged.
Modalities	Mark all modalities actually used: audio, video, slides/screen, board/OCR, document/resource, code/logs.	Exposes text-centric vs multimodal coverage.	Needed for cross-cluster modality counts.
Pipeline stages	Tag any stage materially addressed: capture/OCR, segmentation, alignment, extraction, IR/structuring, generation, ground/verify, evaluation/benchmark.	Links papers to the reliability-first pipeline.	Multi-label coding reduces false exclusivity.
Alignment family	If alignment is central, tag slide-change, semantic, OCR-anchor, cross-modal embedding, or hybrid.	Operationalizes the professor-requested alignment taxonomy.	Record drift-handling strategy when available.
Evidence type	Mark lecture-grounded evidence, resource-grounded evidence, or both.	Separates 'what was said in class' from external course knowledge.	Critical for artifact-level verification rules.
Evaluation style	Record intrinsic metrics, human ratings, user study, learning outcome evidence, and benchmark release.	Avoids flattening all evaluations into one bucket.	Main-text counts are descriptive, not prevalence estimates.
Study quality	Score dataset realism, grounding explicitness, evaluation depth, and deployment relevance on a 0–2 rubric.	Guides interpretive weighting.	Rubric is reproduced in Appendix B.

Table II. Coding schema and audit fields used to derive corpus-level summaries.

III. Positioning The Survey And Defining The Transfer Toolkit Layer

Table III positions this survey relative to adjacent reviews in AI in education, affective tutoring, and personalized e-learning [28]-[30]. Those reviews provide useful context, but none organize the literature around verified study-artifact generation from noisy multimodal classroom evidence. The present survey therefore occupies a different niche: it connects capture, alignment, structured extraction, generation, and verification under a single reliability-first pipeline.

The most important rhetorical shift in this revision is to treat reliability and benchmark papers as a transfer toolkit layer. Text-centric factuality metrics, hallucination detectors, and multimodal evaluation suites are included because they contribute components that can be adapted to classroom evidence, not because they already solve classroom AI end to end. This framing prevents C4 and C5 from appearing appended and instead

makes them operational inputs to the classroom pipeline.

Review / survey	Main scope	What it covers well	What it does not center	Why the present survey differs
Affective tutoring review [28]	Emotion-aware tutoring systems	Learner affect, tutoring design	Classroom evidence linking, structured IR, verification	Our focus is evidence-grounded artifact generation from multimodal lectures.
AI in education review [29]	Broad AIEd overview	System landscape and educational context	Lecture capture, multimodal alignment, claim-level faithfulness	We narrow to content-grounded classroom AI and reliability.
Personalized e-learning review [30]	Adaptive e-learning systems	Personalization issues and deployment challenges	Board/slide evidence, multimodal extraction, verification	We connect capture, IR, and verification under one pipeline.
This survey	Transparent scoping survey of reliable classroom AI	Classroom-native clusters plus transfer toolkit layer	Prevalence estimation or meta-analysis	The goal is comparability, auditability, and operational guidance.

Table III. Positioning of the present survey relative to adjacent reviews.

Transfer toolkit family	Representative papers	Classroom evidence it plugs into	Adaptation required for real classrooms	Best-matched artifact types
Retrieval-augmented generation	[61]-[64], [76]-[80]	ASR spans, slide OCR blocks, board crops, LMS notes, textbook pages	Chunk evidence by time and region; separate lecture-grounded and resource-grounded indices; preserve citation pointers	Tutor answers, notes, FAQ, flashcards
Text factuality and consistency metrics	[65]-[70], [93]-[99]	Claim/evidence pairs derived from lecture segments	Convert multimodal evidence bundles into local support units; downweight noisy OCR/ASR	Claim checking for notes, summaries, explanations
Hallucination detection and self-revision	[71]-[75]	Generated study artifacts plus linked evidence	Use abstention thresholds tied to evidence quality; prevent unsupported auto-repair of STEM content	Self-checking notes, flashcards, code explanations
Knowledge-graph and KG-	[41]-[46], [76]-[80]	Concept graphs and course resources	Distinguish canonical course knowledge from	Tutoring, prerequisite

Transfer toolkit family	Representative papers	Classroom evidence it plugs into	Adaptation required for real classrooms	Best-matched artifact types
RAG tutoring			lecture evidence and inferred prerequisites	navigation, study planning
Multimodal benchmark suites and evaluators	[86]-[92]	Slides, board images, diagrams, charts, multi-page resources	Replace generic document or natural-image tasks with classroom pages and multimodal lecture context	Evaluation of board reasoning, slide QA, diagram explanation

Table IV. Transfer toolkit layer: how generic reliability and benchmark papers map onto classroom evidence types and study artifacts.

IV. Reliability-First Pipeline And Alignment As The Operational Core

Figure 2 summarizes the proposed reliability-first pipeline: capture and OCR, segmentation, alignment, structured extraction, IR structuring, artifact generation, grounding and verification, and benchmarking or deployment. Although these stages can be implemented jointly, alignment is the operational core because every downstream decision depends on whether speech, slides, board regions, and resource documents are connected at the right granularity [101].

In real classrooms, alignment is not a single task. Some systems need coarse slide-transition anchoring; others need semantic alignment between a spoken concept and a slide bullet; STEM settings often need OCR-anchor alignment between a spoken symbol and a handwritten board region; and multimodal retrieval systems use cross-modal embeddings to recover evidence when direct timeline alignment drifts. These families use different cues, fail in different ways, and should not be conflated under a single generic "alignment" label.

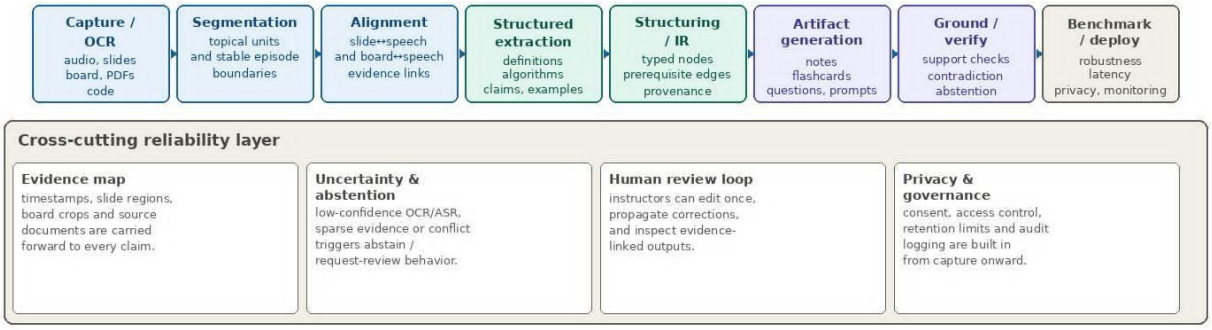

Figure 2. Reliability-first pipeline for multimodal classroom AI, emphasizing evidence linking, uncertainty handling, human review, and privacy as cross-cutting layers.

A. Alignment families and why they fail differently

Table V operationalizes the alignment taxonomy requested in review. Each family is defined by its alignment unit, primary cues, drift-handling strategy, and reported metrics. The design implication is straightforward: a deployable system will usually need more than one alignment family [102]. Transition anchors stabilize the timeline, semantic alignment recovers meaning, OCR anchors protect fine-grained evidence on the board, and cross-modal retrieval acts as a fallback when the lecture deviates from the linear slide order.

This perspective also changes evaluation. Alignment quality should be measured not only with alignment-specific scores such as transition F1 or temporal offset, but also by its downstream effect on artifact faithfulness [103]. A system can achieve acceptable segment overlap while still generating unsupported flashcards if the wrong evidence is linked to the right-looking sentence.

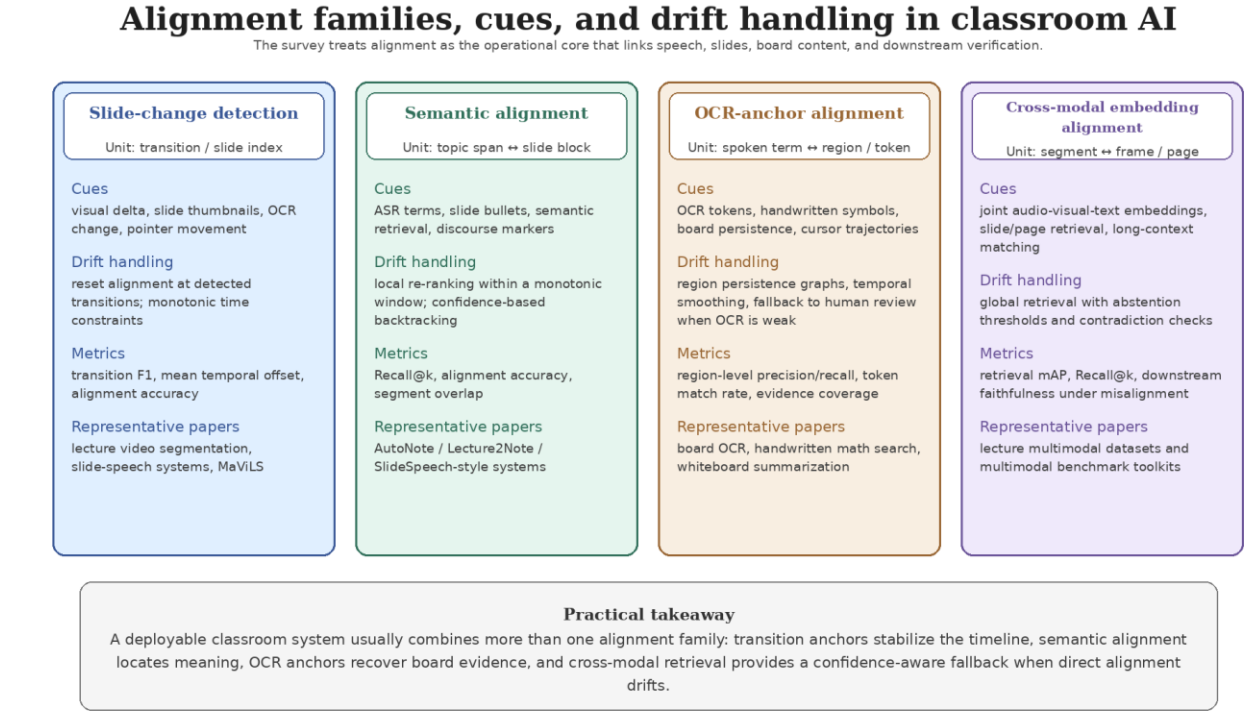

Figure 3. Alignment methods categorized by alignment unit, cues, and drift-handling strategy.

Alignment family	Alignment unit	Primary cues	Drift-handling strategy	Typical metrics	Representative classroom papers
Slide-change detection	Transition / slide index	Visual delta, thumbnails, OCR change, cursor movement	Reset anchors at transitions; enforce monotonic time	Transition F1, mean temporal offset	[2], [10], [21], [22], [84]
Semantic alignment	Topic span ↔ slide block	ASR terms, discourse markers, slide bullets, semantic retrieval	Local re-ranking inside a monotonic window; confidence backtracking	Alignment accuracy, Recall@k, segment overlap	[15], [21], [22], [83], [84]
OCR-anchor alignment	Spoken term ↔ board region / token	OCR tokens, handwritten symbols, region persistence	Temporal smoothing; require a second modality when OCR is weak	Token match rate, region precision/recall	[5], [11], [14], [16]-[19]
Cross-modal	Segment ↔	Joint audio-visual-	Global retrieval	mAP,	[81]-[84], [89]-

Alignment family	Alignment unit	Primary cues	Drift-handling strategy	Typical metrics	Representative classroom papers
embedding alignment	frame / page	text embeddings, frame/page retrieval	fallback with abstention thresholds	Recall@k, downstream faithfulness under drift	[92]

Table V. Alignment taxonomy requested in review: methods are organized by alignment family, cues, drift handling, and reported metrics.

V. OPERATIONAL INTERMEDIATE REPRESENTATION AND VERIFICATION TASKS

The survey argues that reliable classroom AI needs an explicit intermediate representation rather than direct transcript-to-summary generation. The IR should separate evidence, knowledge objects, pedagogy, and verification. Evidence units carry source pointers and confidence. Knowledge objects normalize the content into definitions, procedure steps, equations, code fragments, complexity claims, or prerequisite edges. Pedagogy fields describe why an object matters instructionally. Verification records preserve the logic by which an artifact sentence was supported, contradicted, or abstained [104].

The revised IR is intentionally field-level and operational. Table VI lists the minimum fields needed to make downstream artifacts auditable. Evidence granularity is explicit: ASR support is stored as sentence or 5–15 second spans; slide evidence is stored as region-level OCR blocks; board evidence is stored as persistent crops or step regions; and external resource evidence is stored at page, section, or block level. A claim cannot be considered grounded if it cites only a document-level source without a local pointer [105].

The JSON-like record below illustrates what a single verified segment looks like in the proposed IR. It is not presented as a fixed standard, but as a concrete schema that future benchmarks can instantiate. Figure 5 then shows the same logic visually in the professor-requested micro-example.

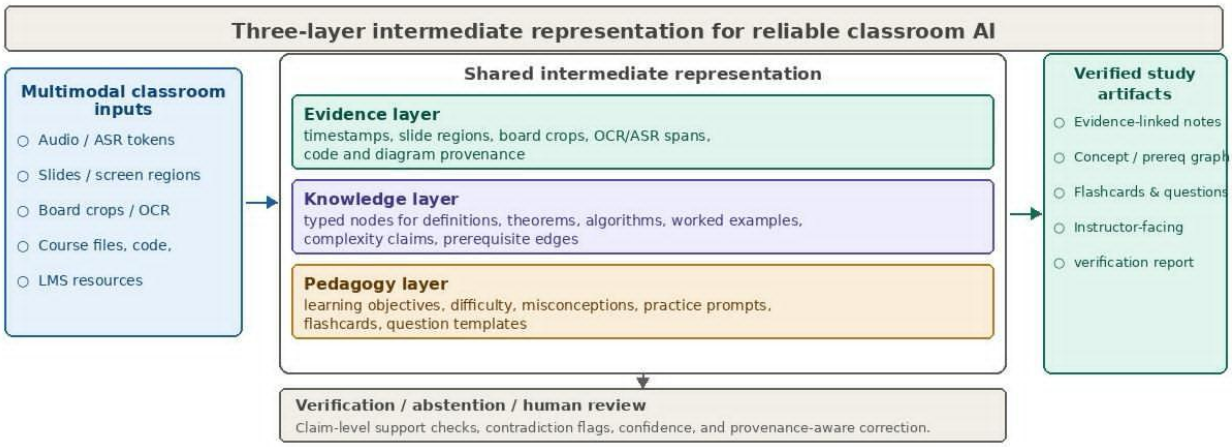

Figure 4. Layered intermediate representation linking multimodal evidence, structured knowledge, pedagogy, and verification.

IR layer	Required fields	Purpose	Example
Evidence span	evidence_id, modality, segment_id, time_start/end, frame_id/page_id, region_bbox, raw_text, normalized_text, confidence, source_uri	Makes every claim traceable to a local support unit.	ASR 12:22–12:27; slide 14:r3; board frame 18:402
Knowledge object	object_id, type, canonical_text, surface_form, granularity, supports[], prerequisites[], instructor_emphasis	Stores definitions, steps, equations, code, and concept relations independently of wording.	definition:BFS; algorithm_step:enqueue neighbor
Pedagogical annotation	learning_objective, difficulty, misconception_tag, practice_potential, artifact_priority	Connects extracted knowledge to study-material design.	high-priority exam concept; beginner misconception
Verification record	target_id, check_type, evidence_bundle, score, decision, contradictions[], abstain_reason, reviewer_action	Preserves how support was judged and why the system abstained or flagged conflict.	supported_with_caution because board OCR=0.62
Artifact unit	artifact_id, artifact_type, text, citations[], derived_from[], release_status	Carries citations into notes, flashcards, tutor replies, and quiz items.	flashcard citing E1+E2

Table VI. Field-level operational schema for the proposed classroom IR.

```
{
  "segment_id": "lec07_t1320_1350",
  "time_span": {"start_s": 1320, "end_s": 1350},
  "evidence": [
    {"id": "E1", "modality": "ASR", "time": [1322, 1327],
     "text": "Breadth-first search explores the graph level by level using
a queue.",
     "confidence": 0.91},
    {"id": "E2", "modality": "slide", "slide_id": 14, "region": "r3",
     "ocr_text": "BFS queue-based traversal; visit neighbors in FIFO
order.",
     "confidence": 0.97},
    {"id": "E3", "modality": "board_crop", "frame": "18:402",
     "ocr_text": "enqueue unvisited neighbors",
     "confidence": 0.62}
```

```
  ],
  "knowledge_objects": [
    {"id": "K1", "type": "definition",
     "canonical_text": "BFS visits nodes in nondecreasing distance from the
source.",
     "supports": ["E1", "E2"], "verification": "supported"},
    {"id": "K2", "type": "algorithm_step",
     "text": "Initialize a queue with the source node; then enqueue each
unvisited neighbor.",
     "supports": ["E2", "E3"], "verification": "supported_with_caution"}
  ],
  "artifact": {
    "type": "flashcard",
    "front": "What data structure does BFS use to explore nodes level by
level?",
    "back": "A queue.",
    "citations": ["E1", "E2"]
  }
}
```

Listing 1. JSON-like record illustrating a single verified lecture segment in the proposed IR.

A. Formal verification for STEM content

Operational IR design matters because the formal verification tasks differ by claim type. Definitions require semantic support and terminology normalization. Algorithm steps require order-sensitive evidence and omission checks. Equations require symbol-level transcription and algebraic consistency. Code requires syntax parsing and, when possible, executable tests. Complexity claims require support from either the lecture trace or an explicitly labeled canonical resource; otherwise the system must abstain rather than silently repairing the claim.

Equally important is the distinction between lecture-grounded and resource-grounded verification. Some artifacts are about what happened in the lecture and therefore must be justified by lecture evidence. Other artifacts are allowed to use external course resources, but the source must be labeled and the system must not blur canonical correction with lecture quotation. Table VIII makes these requirements explicit by artifact type.

Claim type	Minimum evidence granularity	Automatic verification task	Failure trigger	Human fallback
Definition / theorem statement	Sentence-level ASR span plus local slide or resource span	Entailment / equivalence check with terminology normalization	Claim introduces content absent from all cited evidence	Instructor or TA confirms canonical wording
Algorithm step / procedure	Ordered step spans from lecture, slide, or board region	Missing-step detection; order consistency; prerequisite consistency	Steps cited out of order or a key step unsupported	Reviewer inspects sequence and reorders or edits

Claim type	Minimum evidence granularity	Automatic verification task	Failure trigger	Human fallback
Equation / derivation	Symbol-level OCR/vision span plus spoken context	Symbol transcription, dimensional/algebraic consistency, variable reuse	Unknown symbol, sign error, unsupported transition	Manual equation review or canonical resource lookup
Code fragment / syntax claim	Code-screen region or typed resource block	Syntax parse, API existence, optional unit tests	Non-executable code or unsupported auto-correction	Human code review with explicit source labeling
Complexity claim	Lecture/resource support plus linked algorithm object	Check against extracted loop structure or labeled canonical source	Complexity bound appears without evidence or conflicts with cited source	Mark as resource-grounded correction or abstain

Table VII. Formal verification tasks for STEM-oriented classroom claims.

Artifact type	Lecture-grounded mandatory?	Resource-grounded allowed?	Minimum acceptable evidence	Reporting rule
Segment summary / "what the instructor said"	Yes	Only as supplemental context	ASR and/or slide/board evidence from the segment	Do not paraphrase external resources as if they were said in class.
Definition card	Preferred	Yes	Lecture support if presented in class; otherwise explicit course-resource citation	Label whether the wording is lecture-grounded or canonical resource-grounded.
Algorithm-step card / worked example	Yes if describing lecture flow	Yes for canonical normalization	Ordered lecture evidence plus optional external correction	Distinguish recorded lecture sequence from normalized textbook sequence.
Corrected code snippet	No	Yes	Lecture code region plus canonical course or API reference	Never present a repaired snippet as the original

Artifact type	Lecture-grounded mandatory?	Resource-grounded allowed?	Minimum acceptable evidence	Reporting rule
				classroom code.
Practice question / quiz item	No	Yes	Any clearly labeled lecture or course resource basis	Tag as derived material rather than quoted lecture content.
Tutor answer	Depends on user request	Yes	At least one cited local pointer; answer type must indicate lecture-grounded vs resource-grounded	Surface the grounding mode in the UI.

Table VIII. Lecture-grounded versus resource-grounded verification requirements by artifact type.

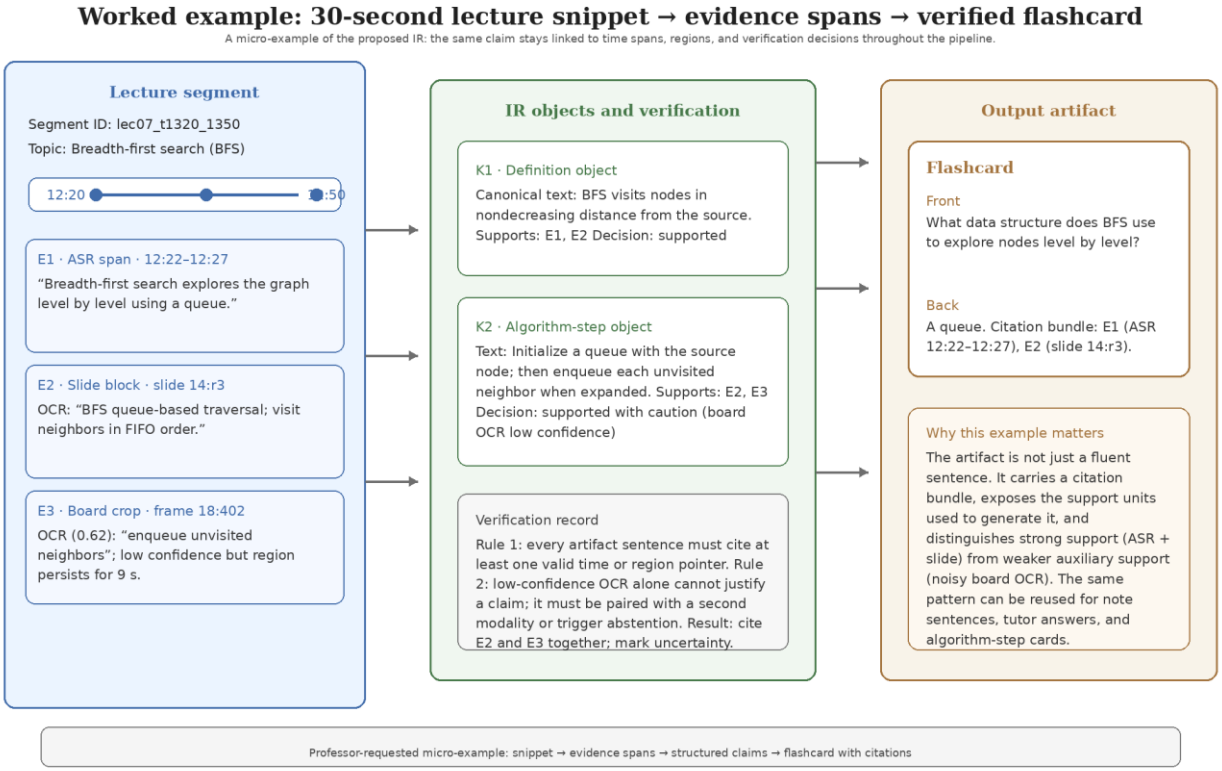

Figure 5. Worked example requested in review: a 30-second lecture snippet becomes evidence spans, structured claims, and a flashcard with citations.

VI. Cross-Cluster Synthesis Of The 100-Paper Corpus

The curated corpus reveals a clear division of labor. Cluster 1 papers are strongest on perception, segmentation, and alignment in long recordings. Cluster 2 focuses on user-facing artifacts such as notes, flashcards, tutoring interactions, and question generation. Cluster 3 contributes structured extraction and knowledge-graph building, but mostly from text and documents rather than live multimodal classroom evidence. Cluster 4 provides reliability mechanisms such as retrieval, factuality checks, self-revision, and grounded tutoring. Cluster 5 contributes datasets, benchmark suites, and evaluators that expose gaps in multimodal reasoning [61]-[92].

Table IX provides the corrected cross-cluster snapshot. Two points matter especially. First, Cluster 2 contains much richer human or user evidence than the earlier draft reported: 17/20 C2 papers include human ratings, user studies, or learning-oriented evidence, not 9/20. Second, Cluster 4 remains intentionally text-centric by corpus design, but that is now framed as a transfer toolkit layer rather than a parallel classroom literature. This rhetorical shift better matches what the cluster contributes.

The cross-cluster picture also clarifies where the field is still thin. Explicit reliability mechanisms are rare in classroom-native clusters, board-grounded evidence remains underrepresented outside perception papers, and most generation papers still evaluate fluency or utility more often than claim-level faithfulness. The implication is that the next generation of classroom AI systems should not merely improve models; they should integrate the pipeline stages more tightly.

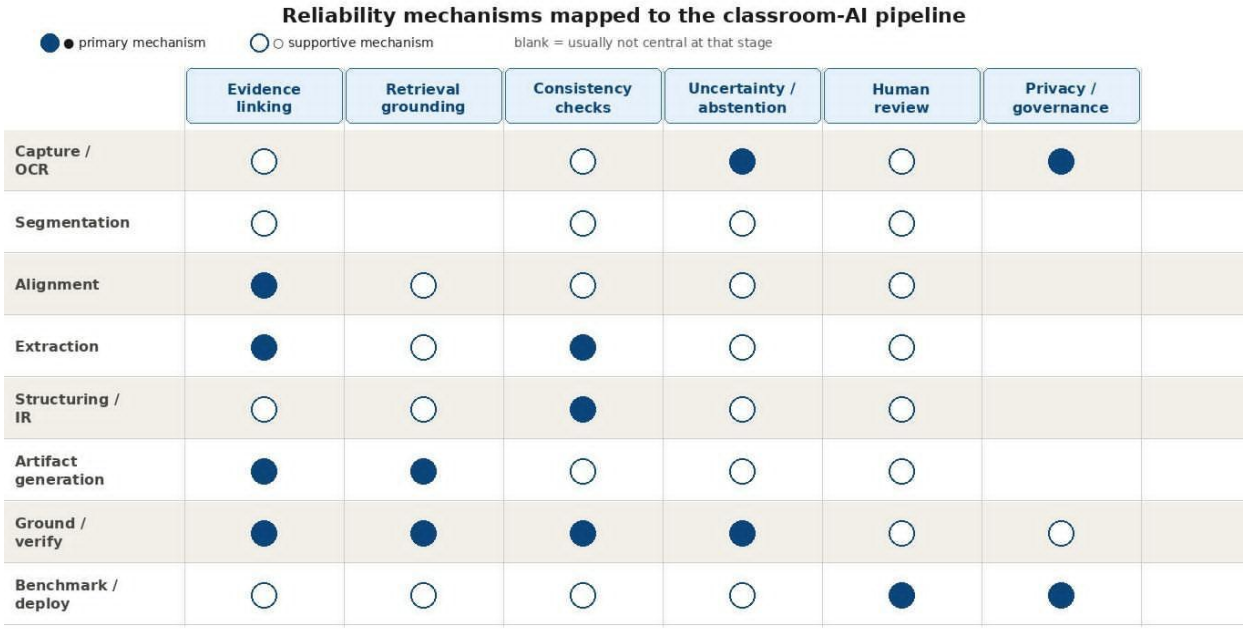

Figure 6. Reliability mechanisms mapped onto classroom-AI pipeline stages; C4 and C5 are interpreted as a transfer toolkit layer that feeds these mechanisms into classroom systems.

Cluster	Role in survey	Dominant modality profile	Explicit reliability mechanisms	Human / user / learning evidence	Primary contribution
C1	Classroom-native perception and alignment	17/20 include video; 5/20 include board/OCR	0/20	12/20	Segmentation, navigation, slide/board alignment
C2	Classroom-native artifact generation	16/20 text-centric; 3/20 slide-video linked	1/20	17/20	Notes, flashcards, tutoring, question generation

Cluster	Role in survey	Dominant modality profile	Explicit reliability mechanisms	Human / user / learning evidence	Primary contribution
C3	Structured extraction and IR building	18/20 text/document-centric	1/20	3/20	Definitions, prerequisite graphs, document understanding
C4	Transfer toolkit: reliability and grounding	20/20 text-centric by corpus design	20/20	13/20	RAG, factuality, self-revision, grounded tutoring
C5	Transfer toolkit: benchmarks and evaluators	15/20 text-centric, 7/20 video-based	2/20	4/20	Datasets, metric suites, multimodal evaluators

Table IX. Corrected cross-cluster corpus snapshot derived from Appendix A. The C2 human/user/learning evidence count is 17/20, consistent with the paper-level coding sheet.

VII. DATASETS, BENCHMARKS, AND MISSING EVALUATION INFRASTRUCTURE

Existing datasets cover pieces of the problem but rarely the full reliability loop. Slide-enriched lecture corpora and alignment benchmarks help with segmentation and slide-speech matching [81]-[84], while document VQA and multimodal benchmark suites stress visual understanding [86]-[92]. However, no widely used benchmark captures the end-to-end path from noisy lecture evidence to cited study artifacts with claim-level verification.

Table X therefore retains the original benchmark blueprints and adds an explicit mixed-evidence benchmark. ClassroomMM-StudyBench targets real classroom capture and end-to-end note generation. Board2Graph-STEM targets handwritten derivations, diagrams, and code structure. VerifiedNotes-QG targets instructor-verified notes, flashcards, and question generation. EvidenceMix-FaithBench targets the exact reviewer concern that generic metrics miss: claims supported by a combination of ASR, board crops, OCR, and course resources with differing quality.

Benchmark	Suggested scale	Modalities	Key annotations	Primary tasks	Core metrics
ClassroomMM-StudyBench	300–500 lecture hours; ≥20 courses	Video, audio, slides, board	Segments; slide↔speech; board regions + OCR; artifact citations	Segmentation, alignment, note generation, verification	F1 / Acc, WER/CER, evidence-support rate, citation precision
Board2Graph-STEM	50–100 hours STEM	Board video/images, audio	Step boundaries, symbol transcripts, diagram primitives, provenance	Handwriting/diagram parsing, derivation extraction, graph building	CER, step F1, graph accuracy, consistency checks

Benchmark	Suggested scale	Modalities	Key annotations	Primary tasks	Core metrics
VerifiedNotes-QG	5k–10k verified segments	Lecture evidence + course resources	Instructor-verified notes, flashcards, questions, contradiction labels	Artifact generation, tutoring, question generation	Faithfulness, pedagogical utility, human preference, abstention quality
EvidenceMix-FaithBench	2k–5k evidence bundles	ASR, board crop, OCR, slides, resources	Claim-level support/partial/contradiction/unverifiable labels plus evidence quality	Mixed-evidence faithfulness evaluation	Weighted evidence score, contradiction rate, calibration / abstention

Table X. Proposed benchmark suite, extended with an explicit mixed-evidence faithfulness benchmark.

VIII. EVALUATION FRAMEWORK FOR MULTIMODAL FAITHFULNESS AND PEDAGOGICAL VALUE

The evaluation framework is only useful if each dimension is measurable at a clear unit of analysis. Table XI therefore defines eight dimensions with explicit units and required evidence: claim faithfulness, citation precision, coverage, structural validity, robustness, uncertainty calibration and abstention, pedagogical utility, and reproducibility or reporting quality. This moves the framework from a broad wish list toward a reviewer-ready checklist.

The most difficult case is multimodal faithfulness when the evidence bundle contains a noisy board crop, imperfect OCR, and ASR context. The revision makes this operational instead of purely conceptual. We treat each claim as paired with an evidence bundle, quality-score each evidence item, and require either one high-quality sufficient support item or two mutually supporting medium-quality items. If the only support comes from low-confidence OCR or a weak ASR span, the system should not be rewarded for a fluent guess; it should be counted as an abstention or unverifiable claim.

Formally, for a claim c with evidence bundle $E(c) = \{e_i\}$, each evidence item receives a quality score q_i in [0,1] derived from OCR confidence, ASR confidence, image quality, and pointer validity. Annotators or automatic checkers assign $s_i \in \{1, 0.5, 0, -1\}$ for support, partial support, neutral, or contradiction. A weighted evidence score $WES(c) = \Sigma_i q_i s_i / \Sigma_i q_i$ is then computed. A claim is counted supported only when WES(c) exceeds a threshold and at least one citation resolves to a valid local pointer. A contradiction by a high-quality item dominates support. Claims justified only by low-quality evidence are scored as uncertainty-triggered abstentions rather than faithful generations. Table XII converts this logic into an annotation and scoring protocol.

Dimension	Unit of analysis	Operational definition	Primary metrics	Required evidence
Claim faithfulness	Sentence / answer /	Whether the claim is supported by a valid local evidence pointer	Evidence-support rate;	Lecture and/or

Dimension	Unit of analysis	Operational definition	Primary metrics	Required evidence
	flashcard back		contradiction rate	resource evidence bundle
Citation precision	Citation bundle	Whether cited pointers are the correct supporting locations	Citation precision / recall	Time spans, regions, page blocks
Coverage	Segment artifact	Whether key concepts and steps are included	Concept recall; missing-step analysis	Instructor outline or verified target
Structural validity	Outline / graph / procedure	Whether relations and order are coherent	Graph F1; order consistency; schema validity	Structured target representation
Robustness	Artifact under perturbation	Stability under ASR/OCR noise, slide drift, or missing modalities	Performance drop under perturbation	Controlled noisy inputs
Uncertainty calibration	Claim decision	Whether confidence or abstention tracks actual support	ECE / Brier; abstention precision	Support labels plus confidence values
Pedagogical utility	Learner interaction	Whether the artifact helps study or comprehension	Human ratings; task completion; learning gain	User study or classroom trial
Reproducibility and reporting	System / paper	Whether data, protocol, and auditing details are exposed	Checklist completeness	Query log, codebook, coding sheet, benchmark details

Table XI. Eight-dimension evaluation framework with measurable units and required evidence.

Step	Annotator or checker action	Decision labels	Scoring rule
1. Claim extraction	Split each artifact into atomic claims that can be individually cited.	claim_id	No multi-claim sentence is scored as

Step	Annotator or checker action	Decision labels	Scoring rule
			one unit.
2. Evidence bundle assembly	Collect cited ASR spans, slide blocks, board crops, and resource spans for each claim.	bundle complete / incomplete	Missing local pointer makes the claim automatically non-faithful.
3. Evidence quality scoring	Assign q in [0,1] from OCR confidence, ASR confidence, image clarity, and pointer validity.	high / medium / low quality	Low-quality evidence alone cannot justify support.
4. Support labeling	Judge each evidence item as support, partial support, neutral, or contradiction.	{1, 0.5, 0, -1}	High-quality contradiction overrides weak support.
5. Bundle sufficiency decision	Aggregate evidence items into WES(c)=Σq_i s_i / Σq_i.	supported / partial / contradicted / unverifiable / abstain	Supported requires threshold plus at least one valid local citation.
6. Artifact aggregation	Aggregate claim scores to notes, flashcards, tutor answers, or quiz items.	artifact faithful / mixed / unreliable	Report claim-level and artifact-level scores separately.

Table XII. Reviewer-ready protocol for multimodal faithfulness under mixed evidence (ASR + OCR + board crops + resources).

IX. DEPLOYMENT CHALLENGES AND RESEARCH ROADMAP

Several deployment blockers recur across clusters. Board capture remains hard because handwriting, erasure, and occlusion are exactly the conditions under which students most need structured support. Alignment remains brittle in open lecture flow where the instructor departs from the slide order. Verification is underdeveloped for STEM content because equations, diagrams, and code require different formal checks. And privacy remains a first-class systems concern because classroom capture intersects with consent, retention policy, and institutional governance.

The near-term roadmap is therefore systems-oriented. First, build evidence-rich benchmarks rather than only larger language models. Second, standardize IR schemas and verification reports so that outputs are portable across note generation, tutoring, and practice creation. Third, design interfaces that surface uncertainty rather than hiding it—e.g., note sentences that visibly show citation bundles and tutor answers that distinguish lecture-grounded from resource-grounded claims. Fourth, treat human correction as part of the model design, not as an afterthought.

In the longer term, the field should move toward course-aware agents that can answer questions at multiple grounding levels without conflating them: what the instructor said, what the canonical course material says, and what the model infers. Reliable classroom AI will depend on keeping those levels distinct while still allowing retrieval, normalization, and pedagogical adaptation.

X. CONCLUSION

Reliable classroom AI is not just a generation problem. It is an evidence-management problem spanning capture, alignment, structured extraction, verification, and human review. This survey revises the earlier draft by making the scope sharper, the methodology more transparent, the IR more

operational, and the evaluation framework more measurable.

The key message is simple: classroom AI becomes trustworthy only when every generated artifact can be traced back to the right evidence unit at the right granularity. That requirement turns alignment into the technical center of the pipeline, transforms C4–C5 into a transfer toolkit layer for classroom grounding, and motivates benchmarks that expose mixed-evidence failure modes rather than hiding them behind fluent outputs.

APPENDIX A. MASTER MAPPING OF THE 100-PAPER CORPUS

Table A1 is the audit-oriented master mapping sheet used to derive the main-text cluster summaries. It maps each paper to cluster, year, short title, modality profile, main pipeline stage(s), and reliability/evaluation posture.

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
1	C1	2018	A Complete System for Analysis of Video Lecture Based on Eye Tracking	Video	Alignment	Not reported; Acc/F1
2	C1	2021	A Framework for Lecture Video Segmentation from Extracted Speech Content	Audio, Video	Segmentation	Not reported; Acc/F1 or human study
3	C1	2023	A Smartphone-Based Multi-Functional Hearing Assistive System to Facilitate Speech	Audio	Capture/OCR	Not reported; WER/CER
4	C1	2020	A Video Analytic In-Class Student Concentration Monitoring System	Video	Alignment	Not reported; Acc/F1
5	C1	2021	Automated Detection of Handwritten Whiteboard Content in Lecture Videos for Summer	Board/OCR, Video	Generation and Capture/OCR	Not reported; Acc/F1, ROUGE/BLEU, or human rating
6	C1	2020	Automatic Detection of Mind Wandering from Video in the Lab and in the	Video	Alignment	Not reported; Acc/F1

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			Classroom			
7	C1	2019	Automatic Lecture Video Content Summarization with Attention-Based Recurrent Neu	Video	Generation	Not reported; ROUGE/BLEU or human rating
8	C1	2021	EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos	Video	Generation	Not reported; ROUGE/BLEU or human rating
9	C1	2024	Fast and Accurate Video Analysis and Visualization of Classroom Activities Using	Video	Alignment	Not reported; Acc/F1
10	C1	2024	Multimodal Speech Recognition Assisted by Slide Information in Classroom Scenes	Audio, Slides	Capture/OCR	Not reported; WER/CER
11	C1	2014	Ote-OCR-Based Text Recognition and Extraction from Video Frames	Text, Video	Extraction, Structuring, Capture/OCR	Not reported; Acc/F1 or human study
12	C1	2016	Usability Evaluation of a Video Conferencing System in a University's Classroom	Video	Capture/OCR	Not reported; Acc/F1 or human study
13	C1	2017	VideoMark: A Video-Based Learning Analytic Technique for MOOCs	Logs, Video	Alignment	Not reported; User/learning study

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
14	C1	2018	Visual Search Engine for Handwritten and Typeset Math in Lecture Videos and LaTeX	Board/OCR, Text, Video	Generation and Capture/OCR	Not reported; ROUGE/BLEU or human rating
15	C1	2020	Visual Summarization of Lecture Video Segments for Enhanced Navigation	Video	Alignment, Generation	Not reported; ROUGE/BLEU or human rating
16	C1	2017	Whiteboard Video Summarization via Spatio-Temporal Conflict Minimization	Board/OCR, Video	Generation and Capture/OCR	Not reported; ROUGE/BLEU or human rating
17	C1	2008	Whiteboard Content Extraction and Analysis for the Classroom Environment	Board/OCR	Extraction, Structuring, Capture/OCR	Not reported; Acc/F1
18	C1	2005	A Unified Text Extraction Method for Instructional Videos	Text, Video	Extraction and Structuring	Not reported; Acc/F1 or human study
19	C1	2014	Automatic Detection of Handwritten Texts from Video Frames of Lectures	Board/OCR, Text, Video	Capture/OCR	Not reported; Acc/F1
20	C1	2007	Summarization of Visual Content in Instructional Videos	Video	Generation	Not reported; ROUGE/BLEU or human rating
21	C2	2023	Semantic Navigation of PowerPoint-Based Lecture Video for AutoNote	Slides, Video	Alignment, Generation	Not reported; ROUGE/BLEU or human rating, WER/CER

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			Generation			
22	C2	2019	Lecture 2 Note: Automatic Generation of Lecture Notes from Slide-Based Educational	Slides, Text, Video	Generation	Not reported; ROUGE/BLEU or human rating
23	C2	2024	Voicense: AI-Powered Lecture Note Generation Tool	Audio	Generation	Not reported; ROUGE/BLEU or human rating, WER/CER
24	C2	2024	Automatic Running Notes Generation from Audio Lecture using NLP for Comprehensive	Audio, Text	Generation	Not reported; ROUGE/BLEU or human rating, User/learning study
25	C2	2019	A New Visual Interface for Searching and Navigating Slide-Based Lecture Videos	Slides, Video	Generation	Not reported; ROUGE/BLEU + user study
26	C2	2023	DIRECT: Toward Dialogue-Based Reading Comprehension Tutoring	Logs	Generation	Not reported; User/learning study
27	C2	2022	A Parametrized Comparative Analysis of Performance Between Proposed Adaptive and	Text	Generation	Verification: Acc/F1, User/learning study
28	C2	2020	The Transition From Intelligent to Affective Tutoring System: A Review and Open	Text	Generation	Not reported; User/learning study

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
29	C2	2020	Artificial Intelligence in Education: A Review	Text	Generation	Not reported; ROUGE/BLEU + user study
30	C2	2022	AI-Based Personalized E-Learning Systems: Issues, Challenges, and Solutions	Text	Generation	Not reported; User/learning study
31	C2	2023	EduQG: A Multi-Format Multiple-Choice Dataset for the Educational Domain	Text	Evaluation/Benchmark	Not reported; Dataset stats + baseline
32	C2	2022	Investigating Educational and Noneducational Answer Selection for Educational Qu	Text	Generation	Not reported; WER/CER
33	C2	2023	Mask and Cloze: Automatic Open Cloze Question Generation Using a Masked Language	Text	Generation	Not reported; ROUGE/BLEU + user study
34	C2	2019	Reverse SQL Question Generation Algorithm in the DB Learn Adaptive E-Learning Systems	Text	Generation	Not reported; User/learning study
35	C2	2024	Student-AI Question Co-Creation for Enhancing Reading Comprehension	Text	Generation	Not reported; ROUGE/BLEU + user study
36	C2	2023	Automatic Question	Text	Generation	Not reported; ROUGE/BLEU

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			Generation Using Natural Language Processing and Transformers			+ user study
37	C2	2025	AI-Powered Automatic Question Generation for Teachers	Text	Generation	Not reported; WER/CER
38	C2	2025	Utilizing Large Language Models for Developing Automatic Question Generation	Text	Generation	Not reported; ROUGE/BLEU + user study
39	C2	2025	Enhancing Assessments: A Comparative Study of T5 and BART Transformers for QG	Text	Generation	Not reported; ROUGE/BLEU + user study
40	C2	2023	FlashMe: Automatic Flashcard Generation	Text	Generation	Not reported; ROUGE/BLEU + user study
41	C3	2019	What Should I Learn First: Introducing LectureBank for NLP Education and Prerequisites	Text	Extraction and Structuring	Not reported; User/learning study
42	C3	2019	Inferring Concept Prerequisite Relations from Online Educational Resources	Text	Extraction and Structuring	Verification: Extraction F1/graph metrics
43	C3	2020	R-VGAE: Relational-Variational Graph Autoencoder for	Text	Extraction and Structuring	Not reported; User/learning study

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			Unsupervised Prerequisite C			
44	C3	2020	MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs	Logs	Extraction and Structuring	Not reported; Extraction F1/graph metrics
45	C3	2021	Constructing an Educational Knowledge Graph with Concepts Linked to Wikipedia	Text	Extraction and Structuring	Not reported; Extraction F1/graph metrics
46	C3	2024	ACE: AI-Assisted Construction of Educational Knowledge Graphs with Prerequisite	Text	Extraction and Structuring	Not reported; Extraction F1/graph metrics
47	C3	2020	SemEval-2020 Task 6: Definition Extraction from Free Text with the DEFT Corpus	Text	Extraction and Structuring	Not reported; Extraction F1/graph metrics
48	C3	2020	Deep Learning-based Extraction of Algorithmic Metadata in Full-Text Scholarly Do	Text	Extraction and Structuring	Not reported; User/learning study
49	C3	2020	SciREX: A Challenge Dataset for Document-Level Information Extraction	Text	Evaluation/Benchmark; Extraction; Structuring	Not reported; Dataset stats + baseline
50	C3	2020	S2ORC: The Semantic Scholar Open Research Corpus	Text	Extraction and Structuring	Not reported; Extraction F1/graph metrics
51	C3	2020	LayoutLM: Pre-training of Text	Text, Video	Extraction and Structuring	Not reported; Extraction

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			and Layout for Document Image Understanding			F1/graph metrics
52	C3	2021	LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding	Text, Video	Extraction and Structuring	Not reported; Extraction F1/graph metrics
53	C3	2022	LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking	Text, Video	Extraction and Structuring	Not reported; Extraction F1/graph metrics
54	C3	2022	DiT: Self-supervised Pre-training for Document Image Transformer	Text, Video	Extraction and Structuring	Not reported; Extraction F1/graph metrics
55	C3	2021	StrucTexT: Structured Text Understanding with Multi-Modal Transformers	Text	Extraction and Structuring	Not reported; Extraction F1/graph metrics
56	C3	2023	Donut: Document Understanding Transformer without OCR	Text	Capture/OCR	Not reported; Extraction F1/graph metrics
57	C3	2021	TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models	Text	Capture/OCR	Not reported; Extraction F1/graph metrics
58	C3	2021	DocFormer: End-to-End Transformer for Document Understanding	Text	Extraction and Structuring	Not reported; Extraction F1/graph metrics
59	C3	2023	UDOP: Unifying	Text	Extraction and Structuring	Not reported;

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			Vision, Text, and Layout for Universal Document Processing			Extraction F1/graph metrics
60	C3	2023	Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding	Video	Extraction and Structuring	Not reported; Extraction F1/graph metrics
61	C4	2020	Retrieval-Augmented Generation for Knowledge-IntensiNLP Tasks	Text	Extraction, Structuring, Generation, Ground/Verify	RAG/Grounding; Factuality metrics + human eval
62	C4	2020	REALM: Retrieval-Augmented Language Model Pre-Training	Text	Ground/Verify	RAG/Grounding; Factuality metrics + human eval
63	C4	2020	Dense Passage Retrieval for Open-Domain Question Answering (DPR)	Text	Ground/Verify	RAG/Grounding; WER/CER
64	C4	2021	Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering	Text	Ground/Verify	RAG/Grounding; WER/CER
65	C4	2020	On Faithfulness and Factuality in Abstractive Summarization	Text	Generation/Ground/Verify	Factuality: ROUGE/BLEU or human rating
66	C4	2019	Evaluating the Factual Consistency of Abstractive Text Summarization	Text	Generation/Ground/Verify	Factuality: ROUGE/BLEU or human rating

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			(FactCC)			
67	C4	2020	Asking and Answering Questions to Evaluate the Factual Consistency of Summaries	Text	Ground/Verify	Factuality; WER/CER
68	C4	2021	Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for	Text	Evaluation/Benchmark; Generation; Ground/Verify	Factuality: Dataset stats + baseline, ROUGE/BLEU, or human rating
69	C4	2022	SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization	Text	Generation	Factuality, verification, Acc/F1, ROUGE/BLEU, or human rating
70	C4	2022	TRUE: Re-evaluating Factual Consistency Evaluation	Text	Evaluate/Benchmark; Ground/Verify	Factuality: Factuality metrics + human evaluation
71	C4	2022	TruthfulQA: Measuring How Models Mimic Human Falsehoods	Text	Ground/Verify	Reliability evaluation: Factuality metrics + human evaluation
72	C4	2023	HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models	Text	Evaluate/Benchmark; Ground/Verify	Factuality: Dataset stats + baseline
73	C4	2023	SelfCheckGPT: Zero-Resource Black-Box Hallucination	Text	Ground/Verify	Factuality: Acc/F1

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			Detection for Generative Lar			
74	C4	2023	RARR: Researching and Revising What Language Models Say, Using Language Models	Text	Ground/Verify	Reliability evaluation: Factuality metrics + human evaluation
75	C4	2024	MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents	Text	Ground/Verify	Factuality: Factuality metrics + human evaluation
76	C4	2023	How to Build an Adaptive AI Tutor for Any Course Using Knowledge Graph-Enhanced	Text	Extraction, Structuring, Ground/Verify	RAG/Grounding; Factuality metrics + human eval
77	C4	2023	AI-TA: Towards an Intelligent Question-Answer Teaching Assistant Using Open Source	Text	Ground/Verify	Reliability eval; WER/CER
78	C4	2024	YA-TA: Yet Another Teaching Assistant: A Case Study on Using Large Language Mode	Text	Ground/Verify	Reliability evaluation; User/learning study
79	C4	2024	Improving Assessment of Tutoring Practices using Retrieval-Augmented Generation	Text	Generation/Ground/Verify	RAG/Grounding: User/learning study
80	C4	2023	Retrieval-Augmented Generation to Improve Math	Text	Generation/Ground/Verify	RAG/Grounding; WER/CER

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			Question-Answering: Trade-offs Between			
81	C5	2023	Lecture Presentations Multimodal Dataset: Towards Understanding Multimodality in	Video	Evaluation/Benchmark	Not reported; Dataset stats + baseline
82	C5	2024	M3AV: A Multimodal, Multigenre, and Multipurpose Audio-Visual Academic Lectured	Audio, Video	Evaluation/Benchmark	Not reported; Dataset stats + baseline
83	C5	2024	SlideSpeech: A Large-Scale Slide-Enriched Audio-Visual Corpus	Audio, Slides, Video	Evaluation/Benchmark	Not reported; Baseline + metrics
84	C5	2024	MaViLS: a Benchmark Dataset for Video-to-Slide Alignment, Assessing Baseline Acc	Audio, Slides, Video	Alignment; Evaluation/Benchmark; Capture/OCR	RAG/Grounding: Dataset stats + baseline
85	C5	2022	Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question A	Text	Evaluation/Benchmark	Not reported; WER/CER
86	C5	2021	DocVQA: A Dataset for VQA on Document Images	Text, Video	Evaluation/Benchmark	Not reported; Dataset stats + baseline
87	C5	2023	SlideVQA: A Dataset for Document Visual	Slides, Text, Video	Evaluation/Benchmark	Not reported: Dataset stats + baseline,

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			Question Answering on Multiple Images			WER/CER
88	C5	2022	ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical	Logs, Video	Evaluation/Benchmark	Not reported: Dataset stats + baseline, WER/CER
89	C5	2024	MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark	Text	Evaluation/Benchmark	Not reported; Dataset stats + baseline
90	C5	2023	MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models	Text	Evaluation/Benchmark	Not reported; Dataset stats + baseline
91	C5	2024	MMBench: Is Your Multi-modal Model an All-Around Player?	Text	Evaluation/Benchmark	Not reported; Baseline + metrics
92	C5	2024	VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models	Text	Evaluation/Benchmark	Not reported; Baseline + metrics
93	C5	2021	SummEval: Re-evaluating Summarization Evaluation	Text	Evaluation/Benchmark Generation	Not reported; ROUGE/BLEU or human rating
94	C5	2004	ROUGE: A Package for Automatic Evaluation of	Text	Evaluation/Benchmark	Not reported; Baseline + metrics

Ref#	Cluster	Year	Short title	Modalities	Stage(s)	Reliability & evaluation
			Summaries			
95	C5	2020	BERTScore: Evaluating Text Generation with BERT	Text	Generation	Not reported; Baseline + metrics
96	C5	2021	QuestEval: Summarization Asks for Fact-based Evaluation	Text	Evaluation/Benchmark; Generation	Factuality: ROUGE/BLEU or human rating
97	C5	2022	Towards a Unified Multi-Dimensional Evaluator for Text Generation	Text	Generation	Not reported; Baseline + metrics
98	C5	2020	BLEURT: Learning Robust Metrics for Text Generation	Text	Evaluation/Benchmark; Generation	Not reported; User/learning study
99	C5	2021	BARTScore: Evaluating Generated Text as Text Generation	Text	Generation	Not reported; Baseline + metrics
100	C5	2025	MaRginalia: Enabling In-person Lecture Capturing and Note-taking Through Mixed R	Text	Generation	Not reported; ROUGE/BLEU or human rating

Table A1. Master mapping of the 100-paper corpus to cluster, modalities, pipeline stages, and reliability/evaluation evidence.

APPENDIX B. LIGHTWEIGHT STUDY-QUALITY APPRAISAL RUBRIC

This rubric is used for interpretive weighting only. It helps distinguish conceptually useful papers from stronger deployment-oriented evidence without converting the review into a formal risk-of-bias meta-analysis.

Dimension	0	1	2
Dataset realism	Synthetic or weakly described data	Curated or partially realistic data	Real classroom or clearly deployment-relevant data
Evaluation depth	Single intrinsic metric only	Intrinsic metrics plus limited human analysis	Multi-level evaluation including human or learning

Dimension	0	1	2
			evidence
Reproducibility detail	Minimal implementation or data detail	Partial task/data/method detail	Clear tasks, data, baselines, and replicable protocol
Reliability disclosure	No confidence / grounding discussion	Acknowledges errors or uncertainty qualitatively	Explicit grounding, verification, abstention, or factuality analysis

Table B1. Lightweight appraisal rubric used for narrative weighting.

APPENDIX C. AUDITABILITY AND REPORTING CHECKLIST

Auditability item	Why it matters	Status in this revision / action expected
Query log and source list	Allows readers to reconstruct search boundaries	Included at the strategy level in Table I; full working log should accompany future updates.
Eligibility and exclusion rules	Prevents ad hoc corpus construction	Included in Section II with relevance-driven inclusion logic.
Primary cluster + multi-label coding sheet	Makes corpus statistics auditable	Released conceptually via Appendix A structure.
Dual-screening and adjudication protocol	Strengthens reproducibility of coding decisions	Protocol is specified; authors should append observed overlap size and agreement statistics from working sheets if available.
Quality rubric	Separates strong evidence from weak deployment claims	Included in Appendix B.
Descriptive-count caveat	Prevents overclaiming prevalence from a balanced corpus	Explicitly stated in Section II and Table IX discussion.

Table C1. Auditability checklist for future revisions or public corpus updates.

References

[1] Xuebai Zhang, Shyan-Ming Yuan, Ming-Dao Chen, and Xiaolong Liu, "A Complete System for Analysis of Video Lecture Based on Eye Tracking," IEEE Access, 2018. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8438455

[2] Dipesh Chand and Hasan Ogul, "A Framework for Lecture Video Segmentation from Extracted Speech Content," 2021 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), 2021. https://www.researchgate.net/profile/Dipesh-Chand/publication/350294257

[3] Alan Chern et al., "A Smartphone-Based Multi-Functional Hearing Assistive System to Facilitate Speech Recognition in the Classroom," IEEE Access, 2023. https://ieeexplore.ieee.org/document/7938619

[4] Mu-Chun Su et al., "A Video Analytic In-Class Student Concentration Monitoring System," IEEE Transactions on Consumer Electronics,

2020. https://ieeexplore.ieee.org/abstract/document/9610134

[5] Bhargava Urala Kota et al., "Automated Detection of Handwritten Whiteboard Content in Lecture Videos for Summarization," IEEE Access, 2021. doi: https://par.nsf.gov/servlets/purl/10113238

[6] Nigel Bosch and Sidney K. D'Mello, "Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom," IEEE Transactions on Affective Computing, 2020. doi: https://ieeexplore.ieee.org/document/8680698

[7] Muhammad Bagus Andra and Tsuyoshi Usagawa, "Automatic Lecture Video Content Summarization with Attention-Based Recurrent Neural Network," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019. https://ieeexplore.ieee.org/abstract/document/8834514

[8] H. Zeng, X. Shu, Y. Wang, Y. Wang, L. Zhang, T.-C. Pong, and H. Qu, "EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos," IEEE Trans. Vis. Comput. Graph., vol. 27, no. 7, pp. 3168–3181, 2021. https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=6366&context=sis_research

[9] Venkatesh Jatla, Sravani Teeparthi, Ugesh Egala, Sylvia Celedon-Pattichis, and Marios S. Pattichis, "Fast and Accurate Video Analysis and Visualization of Classroom Activities Using Multiobjective Optimization of Extremely Low-Parameter Models," IEEE Access, 2025. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10988841

[10] Jingen Li, Jiatian Mei, Di Wu, Mingtao Zhou, and Lin Jiang, "Multimodal Speech Recognition Assisted by Slide Information in Classroom Scenes," 2024 7th International Conference on Video and Image Processing (ICVISP), 2025. https://ieeexplore.ieee.org/abstract/document/10959642

[11] Shashank Shetty, Arun S. Devadiga, S. Sibi Chakkaravarthy, and K. A. Varun Kumar, "Ote-OCR Based Text Recognition and Extraction from Video Frames," 2014 IEEE 8th International Conference on Intelligent Systems and Control (ISCO), 2014. doi: https://www.researchgate.net/profile/Shashank-Shetty-3/publication/301405380

[12] Md. Saifuddin Khalid and Md. Iqbal Hossan, "Usability Evaluation of a Video Conferencing System in a University's Classroom," in Proc. 19th Int. Conf. Comput. Inf. Technol. (ICCIT), Dhaka, Bangladesh, Dec. 2016, pp. 184–189. https://www.researchgate.net/publication/305904926

[13] Nen-Fu Huang, Hao-Hsuan Hsu, So-Chen Chen, Chia-An Lee, Yi-Wei Huang, Po-Wen Ou, and Jian-Wei Tzeng, "VideoMark: A Video-Based Learning Analytic Technique for MOOCs," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), 2017. https://ieeexplore.ieee.org/abstract/document/8078738

[14] Kenny Davila and Richard Zanibbi, "Visual Search Engine for Handwritten and Typeset Math in Lecture Videos and LaTeX Notes," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018. https://pdfs.semanticscholar.org/3a9e/29504ce39568ca64c6e27335aae6ce6eb751.pdf

[15] M. R. Rahman, S. Shah, and J. Subhlok, "Visual Summarization of Lecture Video Segments for Enhanced Navigation," in Proc. 2020 IEEE Int. Symp. Multimedia (ISM), Dec. 2020, pp. 154–157, https://arxiv.org/pdf/2006.02434

[16] Kenny Davila and Richard Zanibbi, "Whiteboard Video Summarization via Spatio-Temporal Conflict Minimization," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1727-1734, 2017. https://cs.rit.edu/~rlaz/files/Kenny_ICDAR_2017.pdf

[17] D. Dickson, C. V. Sharma, and K. Kwok, "Whiteboard Content Extraction and Analysis for the Classroom Environment," 2008 IEEE International Symposium on Multimedia, pp. 131-138, 2008. https://www.researchgate.net/profile/Allen-Hanson-2/publication/221558684

[18] Z. Tang and J. R. Kender, "A Unified Text Extraction Method for Instructional Videos," 2005 IEEE International Conference on Image Processing (ICIP), vol. 2, pp. II-1088-II-1091, 2005. https://www.researchgate.net/profile/Lijun-Tang/publication/224622476

[19] S. Banerjee, S. Kundu, and B. B. Chaudhuri, "Automatic Detection of Handwritten Texts from Video Frames of Lectures," 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 479-484, 2014. https://ieeexplore.ieee.org/abstract/document/6981089

[20] M. A. Choudary and S.-F. Liu, "Summarization of Visual Content in Instructional Videos," IEEE Transactions on Multimedia, vol. 9, no. 7, pp. 1443-1455, 2007. https://www.researchgate.net/profile/Chekuri-Choudary/publication/3424658

[21] Chengpei Xu, Wenjing Jia, Ruomei Wang, Xiangjian He, Baoquan Zhao, Yuanfang Zhang, "Semantic Navigation of PowerPoint-Based Lecture Video for AutoNote Generation," IEEE Transactions on Learning Technologies, 2023. https://ieeexplore.ieee.org/abstract/document/9927330

[22] Chengpei Xu, Ruomei Wang, Shujin Lin, Xiaonan Luo, Baoquan Zhao, Lijie Shao, Mengqiu Hu, "Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos," IEEE ICME 2019, 2019. https://www.researchgate.net/profile/Baoquan-Zhao/publication/334997213

[23] A.W.R.P. Karunarathna, T.U.M.N. Premarathna, R.G.S. Dilshan, W.A.K.H.R. Wanniarachchi, Y.M.C.N. Bimsara, I.T.S. Piyatilake, "Voicense: AI-Powered Lecture Note Generation Tool," IEEE ICITR 2024, 2024. https://ieeexplore.ieee.org/abstract/document/10857774

[24] A. Madhavi, A. Chilakamarri, C. Jupudi, S. Madanaboina, and S. Sriram, "Automatic Running Notes Generation from Audio Lecture using NLP for Comprehensive Learning," in Proc. 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT), 2024https://ieeexplore.ieee.org/abstract/document/10723991

[25] Baoquan Zhao, Songhua Xu, Shujin Lin, Ruomei Wang, Xiaonan Luo, "A New Visual Interface for Searching and Navigating Slide-Based Lecture Videos," IEEE ICME 2019, 2019. https://www.researchgate.net/profile/Baoquan-Zhao/publication/334997587

[26] Jin-Xia Huang, Yohan Lee, Oh-Woog Kwon, "DIRECT: Toward Dialogue-Based Reading Comprehension Tutoring," IEEE Access, 2023. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10003215

[27] N. Singh, V.K. Gunjan, M.M. Nasralla, "A Parametrized Comparative Analysis of Performance Between Proposed Adaptive and Personalized Tutoring System 'Seis Tutor' With Existing Online Tutoring System," IEEE Access, 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9755124

[28] M.A. Hasan, N.F.M. Noor, S.S.B. Ab Rahman, M.M. Rahman, "The Transition From Intelligent to Affective Tutoring System: A Review and Open Issues," IEEE Access, 2020. https://ieeexplore.ieee.org/document/9252896

[29] Lijia Chen, Pingping Chen, Zhijian Lin, "Artificial Intelligence in Education: A Review," IEEE Access, 2020. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9069875

[30] M. Murtaza, Y. Ahmed, J. A. Shamsi, F. Sherwani, and M. Usman, "AI-Based Personalized E-Learning Systems: Issues, Challenges, and Solutions," IEEE Access, vol. 10, pp. 81323-81342, 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9840390

[31] Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, Thomas Demeester, "EduQG: A Multi-Format Multiple-Choice Dataset for the Educational Domain," IEEE Access, 2023. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10051840

[32] Tim Steuer, Anna Filighera, Thomas Tregel, "Investigating Educational and Noneducational

Answer Selection for Educational Question Generation," IEEE Access, 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9791321

[33] Shoya Matsumori, Kohei Okuoka, Ryoichi Shibata, Minami Inoue, Yosuke Fukuchi, Michita Imai, "Mask and Cloze: Automatic Open Cloze Question Generation Using a Masked Language Model," IEEE Access, 2023. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10024779

[34] Kanokwan Atchariyachanvanich, Srinual Nalintippayawong, and Thanakrit Julavanich, "Reverse SQL Question Generation Algorithm in the DBLearn Adaptive E-Learning System," IEEE Access, 2019. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8703745

[35] Ming Liu, Jingxu Zhang, Lucy Michael Nyagoga, Li Liu, "Student-AI Question Co-Creation for Enhancing Reading Comprehension," IEEE Transactions on Learning Technologies, 2024. https://ieeexplore.ieee.org/abstract/document/10321718

[36] R. M. Elshiny and A. Hamdy, "Automatic Question Generation Using Natural Language Processing and Transformers," in Proc. 2023 International Conference on Computer and Applications (ICCA), 2023, pp. 1-6. https://ieeexplore.ieee.org/abstract/document/10401848

[37] Sugiyanto Yoannatan Widjaja, Alfa Yohannis, "AI-Powered Automatic Question Generation for Teachers," IEEE SIML 2025, 2025. https://www.researchgate.net/profile/Sugiyanto-Yoannatan-W/publication/393937138

[38] A. J. Winata, D. J. Surjawan, and V. C. Mawardi, "Utilizing Large Language Models for Developing Automatic Question Generation in Education," in Proc. 2025 International Conference on Advancement in Data Science, E-Learning and Information System (ICADEIS), 2025 https://ieeexplore.ieee.org/abstract/document/10933227

[39] P. Preetha, G. Sivakamasundari, and K. Srimathi, "Enhancing Assessments: A Comparative Study of T5 and BART Transformer for QG," in Proc. 2025 International Conference on Computing, Communication, and Multimedia (ICCMC), 2025. https://ieeexplore.ieee.org/abstract/document/11140610

[40] N. Nair, S. Pikle, S. Save, R. Varghese, and K. Sonawane, "FlashMe: Automatic Flashcard Generation," in Proc. 14th Int. Conf. Computing Communication and Networking Technologies (ICCCNT), 2023, https://ieeexplore.ieee.org/abstract/document/10308164

[41] Irene Li et al, "What Should I Learn First: Introducing LectureBank for NLP Education and Prerequisite Chain Learning," AAAI, 2019. https://arxiv.org/abs/1811.12181

[42] Sudeshna Roy et al, "Inferring Concept Prerequisite Relations from Online Educational Resources," AAAI, 2019. https://arxiv.org/abs/1811.12640

[43] Irene Li et al, "R-VGAE: Relational-Variational Graph Autoencoder for Unsupervised Prerequisite Chain Learning," COLING, 2020. https://aclanthology.org/2020.coling-main.99.pdf

[44] Jifan Yu et al, "MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs," ACL, 2020. https://aclanthology.org/2020.acl-main.285.pdf

[45] Fu-Rong Dang et al, "Constructing an Educational Knowledge Graph with Concepts Linked to Wikipedia," Journal of Computer Science and Technology, 2021. https://jcst.ict.ac.cn/fileup/1000-9000/PDF/2021-5-18-0328.pdf

[46] Dr. Mehmet Cem Aytekin, Yücel Saygın, "ACE: AI-Assisted Construction of Educational Knowledge Graphs with Prerequisite Relations," Journal of Educational Data Mining, 2024. doi: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/737

[47] Sasha Spala et al, "SemEval-2020 Task 6: Definition Extraction from Free Text with the DEFT Corpus," SemEval, 2020. https://aclanthology.org/2020.semeval-1.41.pdf

[48] Safder et al, “Deep Learning-based Extraction of Algorithmic Metadata in Full-Text Scholarly Documents,” https://e-space.mmu.ac.uk/625933/8/Deep%20Learning-based%20Extraction%20of%20Algorithmic%20Metadata%20in%20Full-Text%20Scholarly%20Documents%20e.pdf

[49] Sarthak Jain et al., “SciREX: A Challenge Dataset for Document-Level Information Extraction,” in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020. https://aclanthology.org/2020.acl-main.670.pdf

[50] Iz Beltagy et al, “S2ORC: The Semantic Scholar Open Research Corpus,” ACL, 2020. https://aclanthology.org/2020.acl-main.447.pdf

[51] Yang Xu et al, “LayoutLM: Pre-training of Text and Layout for Document Image Understanding,” KDD, 2020. https://dl.acm.org/doi/pdf/10.1145/3394486.3403172

[52] Yang Xu et al, “LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding,” / 2020 preprint, 2021. https://aclanthology.org/2021.acl-long.201.pdf

[53] Yupan Huang et al, “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking,” 2022. https://dl.acm.org/doi/pdf/10.1145/3503161.3548112

[54] Junlong Li et al, “DiT: Self-supervised Pre-training for Document Image Transformer,” 2022. https://dl.acm.org/doi/pdf/10.1145/3503161.3547911

[55] Yulin Li et al, “StrucTexT: Structured Text Understanding with Multi-Modal Transformers,” arXiv, 2021. https://dl.acm.org/doi/pdf/10.1145/3474085.3475345

[56] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park, “OCR-Free Document Understanding Transformer,” in Proc. European Conference on Computer Vision (ECCV), 2022. doi: 10.1007/978-3-031-19815-1_29.

[57] Minghao Li et al, “TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models,” arXiv, 2021. https://arxiv.org/pdf/2109.10282

[58] Srikar Appalaraju et al, “DocFormer: End-to-End Transformer for Document Understanding,” https://openaccess.thecvf.com/content/ICCV2021/papers/Appalaraju_DocFormer_End-to-End_Transformer_for_Document_Understanding_ICCV_2021_paper.pdf

[59] Zineng Tang et al., "Unifying Vision, Text, and Layout for Universal Document Processing,” CVPR, 2023. https://openaccess.thecvf.com/content/CVPR2023/papers/Tang_Unifying_Vision_Text_and_Layout_for_Universal_Document_Processing_CVPR_2023_paper.pdf

[60] Kenton Lee et al, “Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding,” ICML, 2023. https://proceedings.mlr.press/v202/lee23g/lee23g.pdf

[61] Aleksandra Piktus et al, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[62] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, “REALM: Retrieval-Augmented Language Model Pre-Training,” https://proceedings.mlr.press/v119/guu20a/guu20a.pdf

[63] Vladimir Karpukhin et al, “Dense Passage Retrieval for Open-Domain Question Answering (DPR),” EMNLP, 2020. https://aclanthology.org/2020.emnlp-main.550.pdf

[64] Gautier Izacard, Edouard Grave, “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering (Fusion-in-Decoder / FiD),” https://aclanthology.org/2021.eacl-main.74.pdf

[65] Joshua Maynez et al., “On Faithfulness and Factuality in Abstractive Summarization,” ACL, 2020. doi: 10.18653/v1/2020.acl-main.173. https://aclanthology.org/2020.acl-main.173/?utm_source=chatgpt.com

[66] Wojciech Kryściński et al, “Evaluating the Factual Consistency of Abstractive Text Summarization (FactCC),” arXiv, 2019. doi: 10.48550/arXiv.1910.12840. https://arxiv.org/abs/1910.12840

[67] A. Wang, K. Cho, and M. Lewis, “Asking and Answering Questions to Evaluate the Factual Consistency of Summaries (QAGS),” in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 5008–5020, doi: 10.18653/v1/2020.acl-main.450. https://aclanthology.org/2020.acl-main.450/

[68] Artidoro Pagnoni et al, “Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics,” NAACL, 2021. doi: 10.18653/v1/2021.naacl-main.383. https://aclanthology.org/2021.naacl-main.383/

[69] Philippe Laban et al, “SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization,” TACL, 2022. doi: 10.1162/tacl_a_00453. https://aclanthology.org/2022.tacl-1.10/

[70] Or Honovich et al, “TRUE: Re-evaluating Factual Consistency Evaluation,” NAACL, 2022. doi: 10.18653/v1/2022.naacl-main.287. https://aclanthology.org/2022.naacl-main.287/

[71] Stephanie Lin et al, “TruthfulQA: Measuring How Models Mimic Human Falsehoods,” paper circulated as arXiv, 2022. doi: 10.48550/arXiv.2109.07958. https://arxiv.org/abs/2109.07958

[72] Junyi Li et al, “HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models,” arXiv, 2023. doi: 10.48550/arXiv.2305.11747. https://arxiv.org/abs/2305.11747

[73] Potsawee Manakul et al, “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,” EMNLP, 2023. doi: 10.18653/v1/2023.emnlp-main.557. https://aclanthology.org/2023.emnlp-main.557/

[74] Luyu Gao et al, “RARR: Researching and Revising What Language Models Say, Using Language Models,” ACL, 2023. doi: 10.18653/v1/2023.acl-long.910.

[75] Liyan Tang et al, “MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents,” arXiv, 2024. doi: 10.48550/arXiv.2404.10774. https://arxiv.org/abs/2404.10774

[76] C. Dong, Y. Yuan, K. Chen, S. Cheng, and C. Wen, “How to Build an Adaptive AI Tutor for Any Course Using Knowledge Graph-Enhanced Retrieval-Augmented Generation (KG-RAG),” arXiv:2311.17696, 2023. doi: 10.48550/arXiv.2311.17696. https://arxiv.org/abs/2311.17696

[77] Y. Hicke, A. Agarwal, Q. Ma, and P. Denny, “AI-TA: Towards an Intelligent Question-Answer Teaching Assistant Using Open-Source Large Language Models,” arXiv:2311.02775, 2023. doi: 10.48550/arXiv.2311.02775. https://arxiv.org/abs/2311.02775

[78] D. Yang, S. Lee, M. Kim, J. Won, N. Kim, D. Lee, and J. Yeo, “YA-TA: Yet Another Teaching Assistant: A Case Study on Using Large Language Models for Learning Python,” arXiv:2409.00355, 2024. doi: 10.48550/arXiv.2409.00355. https://arxiv.org/abs/2409.00355

[79] Zifei FeiFei Han et al, “Improving Assessment of Tutoring Practices using Retrieval-Augmented Generation,” arXiv, 2024. doi: 10.48550/arXiv.2402.14594. https://arxiv.org/abs/2402.14594

[80] Zachary Levonian et al, "Retrieval-Augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference,” arXiv, 2023. doi: 10.48550/arXiv.2310.03184. https://arxiv.org/abs/2310.03184

[81] Dong Won Lee et al, “Lecture Presentations Multimodal Dataset: Towards Understanding Multimodality in Educational Videos,” ICCV (IEEE/CVF), 2023. doi: 10.1109/ICCV51070.2023.01838. https://openaccess.thecvf.com/content/ICCV2023/papers/Lee_Lecture_Presentations_Multimodal_Dataset_Towards_Understanding_Multimodality_in_Educational_Videos_ICCV_2023_paper.pdf

[82] Zhe Chen et al, “M3AV: A Multimodal, Multigenre, and Multipurpose Audio-Visual Academic Lecture Dataset,” ACL (Long),

2024. doi: 10.18653/v1/2024.acl-long.489. https://aclanthology.org/2024.acl-long.489/

[83] Haoxu Wang et al, "SlideSpeech: A Large-Scale Slide-Enriched Audio-Visual Corpus," ICASSP (IEEE) + 2023 arXiv preprint, 2024. doi: 10.1109/ICASSP48485.2024.10448079; 10.48550/arXiv.2309.05396. https://arxiv.org/abs/2309.05396

[84] Katharina Anderer et al, "MaViLS: a Benchmark Dataset for Video-to-Slide Alignment, Assessing Baseline Accuracy with a Multimodal Alignment Algorithm Leveraging Speech, OCR, and Visual Features," Interspeech, 2024. doi: 10.21437/Interspeech.2024-978. https://www.isca-archive.org/interspeech_2024/anderer24_interspeech.pdf

[85] Pan Lu et al, "Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering," NeurIPS, 2022. doi: 10.48550/arXiv.2209.09513. https://proceedings.neurips.cc/paper_files/paper/2022/file/11332b6b6cf4485b84afadb1352d3a9a-Paper-Conference.pdf

[86] Minesh Mathew et al, "DocVQA: A Dataset for VQA on Document Images," WACV (IEEE/CVF); dataset introduced in 2020, 2021. doi: 10.1109/WACV48630.2021.00225. https://ieeexplore.ieee.org/document/9423358

[87] R. Tanaka, K. Nishida, K. Nishida, T. Hasegawa, I. Saito, and K. Saito, "SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images," https://arxiv.org/pdf/2301.04883

[88] Ahmed Masry et al, "ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning," 2022 (arXiv preprint; widely used benchmark), 2022. doi: 10.48550/arXiv.2203.10244. https://arxiv.org/abs/2203.10244

[89] Xiang Yue et al, "MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI," CVPR (IEEE/CVF) (original arXiv 2023), 2024. doi: 10.48550/arXiv.2311.16502. https://arxiv.org/abs/2311.16502

[90] Chaoyou Fu et al., "MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models," 2023 (arXiv benchmark paper), 2023. doi: 10.48550/arXiv.2306.13394. https://arxiv.org/abs/2306.13394

[91] Yuan Liu et al., "MMBench: Is Your Multi-modal Model an All-Around Player?," ECCV (LNCS), 2024. doi: 10.1007/978-3-031-72658-3_13. https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/00959.pdf

[92] Haodong Duan et al., "VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models," 2024 (arXiv + ACM MM tooling), 2024. doi: 10.48550/arXiv.2407.11691. https://arxiv.org/abs/2407.11691

[93] Alexander R. Fabbri et al., "SummEval: Re-evaluating Summarization Evaluation," TACL, 2021. doi: 10.1162/tacl_a_00373. https://arxiv.org/abs/2007.12626

[94] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Text Summarization Branches Out (Workshop of ACL), 2004. https://aclanthology.org/W04-1013/

[95] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in Proc. International Conference on Learning Representations (ICLR), 2020. doi: 10.48550/arXiv.1904.09675. https://iclr.cc/virtual_2020/poster_SkeHuCVFDr.html

[96] Thomas Scialom et al., "QuestEval: Summarization Asks for Fact-Based Evaluation," EMNLP, 2021. doi: 10.18653/v1/2021.emnlp-main.529. https://arxiv.org/abs/2103.12693

[97] Ming Zhong et al., "Towards a Unified Multi-Dimensional Evaluator for Text Generation," EMNLP, 2022. doi: 10.18653/v1/2022.emnlp-main.131. https://arxiv.org/abs/2210.07197

[98] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning Robust Metrics for Text Generation," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 7881-7892. doi: 10.18653/v1/2020.acl-main.704. https://aclanthology.org/2020.acl-main.704/

[99] W. Yuan, G. Neubig, and P. Liu, "BARTScore: Evaluating Generated Text as Text Generation," arXiv:2106.11520, 2021. doi: 10.48550/arXiv.2106.11520. https://arxiv.org/abs/2106.11520

[100] Leping Qiu et al., "MaRginalia: Enabling In-person Lecture Capturing and Note-taking Through Mixed Reality," CHI, 2025. doi: 10.1145/3706598.3714065. https://dl.acm.org/doi/10.1145/3706598.3714065

[101] P. A. Diaz Munoz, "Interdisciplinary design practices in contemporary architectural development: Integrating creativity and functionality," Evolutionary Studies in Imaginative Culture, vol. 5, no. 2, pp. 1–9, 2021.

[102] D. Puthiya, "Strategic AI transformation initiatives for scalable business expansion," Journal of Information Systems Engineering and Management, vol. 6, no. 2, pp. 1–12, 2021.

[103] A. Kejriwal, "High-stakes negotiation frameworks in cross-functional project environments," International Journal of Environmental Sciences, vol. 7, no. 1S, pp. 20–27, 2021.

[104] R. Chhibber, "Strategic leadership in partner sales networks for enterprise market expansion," Journal of International Crisis and Risk Communication Research, vol. 4, no. 3, pp. 467–475, 2021.

[105] G. A. Ascanio, "Wellness-driven design development in luxury residential architecture: Spatial, social, and environmental dimensions," Journal of Information Systems Engineering and Management, vol. 6, no. 1, pp. 1–10, 2021.