

The Verification Gap in AI Accessibility: A Meta-Analysis of Architectural Paradigms and Agency Deficits

Madjid Sadallah

Universite Claude Bernard Lyon 1, CNRS, INSA Lyon,
LIRIS, UMR5205
69622 Villeurbanne, France
madjid.sadallah@liris.cnrs.fr

Benoît Encelle

Universite Claude Bernard Lyon 1, CNRS, INSA Lyon,
LIRIS, UMR5205
69622 Villeurbanne, France
benoit.encelle@liris.cnrs.fr

Abstract

AI-powered accessibility systems have shifted from transcription to interpretation, creating a verification paradox: blind and low-vision (BLV) users, who most rely on AI-generated narratives, lack perceptual redundancy to detect errors such as hallucinations or semantic drift, resulting in an agency deficit. Through a systematic meta-analysis of 17 empirically evaluated AI accessibility systems (2019–2025), we examine how architectural choices shape verification affordances. We introduce the Verification-Cognition Balance Index (VCBI), a composite metric quantifying trade-offs between validation rigor, user agency, complexity, and risk. Our analysis uncovers weak overall balance across the field, with transformer-based systems showing higher risk density than classical computer vision with no gain in validation rigor. Only hybrid specialized architectures—constraining AI to narrow semantic subtasks—achieve good balance in our corpus, though this finding is preliminary given the small number of such systems evaluated. We argue for a necessary paradigm shift in evaluation from system accuracy to user verifiability. Our findings suggest that for high-stakes accessibility, task-specific constrained AI outperforms general-purpose models in epistemic safety and user empowerment.

CCS Concepts

• **Human-centered computing** → **Accessibility**; *Empirical studies in HCI*; • **Computing methodologies** → *Artificial intelligence*.

Keywords

Accessibility, Blind and Low Vision, Artificial Intelligence, Verification, Meta-Analysis

ACM Reference Format:

Madjid Sadallah and Benoît Encelle. 2026. The Verification Gap in AI Accessibility: A Meta-Analysis of Architectural Paradigms and Agency Deficits. In *Web4All 2026 23rd International Web for All Conference (W4A '26)*, April 13–14, 2026, Dubai, United Arab Emirates. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3800424.3800451>

1 Introduction

Assistive technologies (AT) for blind and low-vision (BLV) users have historically operated under a *transcription paradigm*: optical

character recognition (OCR) transduces pixel arrays into character streams [11], screen readers linearize interface hierarchies into serialized auditory outputs [19], and tactile graphics encode spatial relations through embossed representations [29]. Across these modalities, systems functioned primarily as sensory transducers, rendering inaccessible visual information into alternative perceptual forms while preserving the user as the ultimate locus of semantic interpretation. The epistemic division of labor was unambiguous: systems rendered information accessible; users interpreted it.

The integration of deep learning and generative AI has introduced a fundamental shift from transcription to *interpretation*. Contemporary systems now generate natural-language narratives of statistical visualizations (VizAbility [12]), synthesize semantic summaries of complex tabular data (TableNarrator [24]), produce contextually enriched descriptions of visual content (Edukoi [13]), or render scientific diagrams as accessible learning materials (LectureAssistant [2]). This transformation repositions AT from neutral conduits to *epistemic intermediaries*, fundamentally altering the verification relationship between user and system.

For sighted users, AI-generated interpretations coexist with direct perceptual access to the source material, enabling rapid cross-validation. For BLV users, by contrast, the AI often constitutes the primary—and sometimes sole—information source. This asymmetry becomes particularly consequential given well-documented failure modes in vision-language models—including hallucination, semantic drift, misclassification, and contextual errors—which may remain undetectable without ground-truth visual access [15, 16]. When an AI misreports a data trend, fabricates details, or introduces factual inaccuracies, BLV users lack the perceptual redundancy available to sighted counterparts.

This creates a *verification paradox*: those who most need AI interpretation are least equipped to verify it. The practical consequences of this paradox become clear in routine analytical tasks. Consider a blind data analyst accessing a bar chart comparing quarterly sales across five regions. The system narrates: “*Sales peaked in Q3 across all regions, with the Northeast showing the strongest growth at 47%*”. The interpretation is fluent and authoritative—yet if the AI has hallucinated the 47% figure, misread axis labels, or conflated data series, the analyst has no independent means of detecting the error. Recent empirical work underscores this vulnerability: BLV users detect only approximately 50% of recognition errors in controlled studies [14]. These findings reveal a pronounced verification gap and fundamental epistemic vulnerability in AI-mediated access.

Traditional approaches to appropriate reliance in human-AI interaction—such as confidence scores, explanations, or alternative hypotheses—presume users retain independent epistemic access or



This work is licensed under a Creative Commons Attribution 4.0 International License. W4A '26, Dubai, United Arab Emirates

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2372-8/26/04

<https://doi.org/10.1145/3800424.3800451>

auxiliary verification channels [4, 35]. In accessibility contexts, this assumption does not hold: verification mechanisms must operate under the same sensory and cognitive constraints that initially necessitated AI mediation. Despite these heightened epistemic stakes, systematic analyses of verification affordances in AI-powered assistive technologies remain largely absent. Prior work has documented isolated failure risks (e.g., hallucinations in AlternAtive [27], misclassification in ChartVi [23]), but has typically treated verification as a binary property rather than a continuous, multidimensional capability. Evaluation practices remain centered on *system accuracy* rather than *user verifiability*—that is, whether users can recognize when the AI fails.

This study addresses three research questions through a systematic meta-analysis of empirically evaluated AI accessibility systems:

- **RQ1 (Profiles):** What verification-cognition profiles emerge from validation rigor, user agency, complexity, and risks in AI accessibility systems?
- **RQ2 (Architectural):** How does verification-cognition balance vary across AI paradigms (classical CV, transformer-based, hybrid), and which characteristics best support verification affordances?
- **RQ3 (Design):** What verification-aware design patterns and implementation choices distinguish systems with stronger verification affordances?

Our investigation yields four main contributions: (1) an empirical characterization of state-of-the-art AI accessibility systems, offering the first comprehensive analysis of verification affordances across architectural paradigms; (2) the introduction of the *Verification-Cognition Balance Index (VCBI)*, a composite metric to formalize and quantify the equilibrium between system capabilities and user verification challenges; and (3) insights into architectural trade-offs and capability-risk dynamics, and (4) a set of empirically grounded design principles to guide future development of more verifiable and agency-enhancing AI accessibility tools.

2 Related Work

2.1 AI Architectures and Verification

AI accessibility systems for BLV users span three architectural paradigms with distinct verification implications. *Rule-based and classical computer vision* systems rely on deterministic algorithms or early CNNs to produce constrained outputs where failures typically manifest as detectable errors [20]. *Transformer-based generative systems* generate rich semantic interpretations but produce hallucinations and plausible-yet-false outputs indistinguishable from valid ones without ground-truth access [16]; alignment techniques for multimodal models remain an active area of research [8]. *Hybrid specialized architectures* constrain AI to well-defined subtasks within structured workflows, offering a potential balance between enhanced capability and improved verifiability through human oversight and modular control [32].

Research on human-AI collaboration indicates that users struggle to calibrate appropriate trust without independent verification channels [4, 18]. However, these frameworks typically assume users have access to *alternative information sources*—an assumption that breaks down in accessibility contexts where the AI constitutes the primary or sole perceptual channel.

2.2 Explainability and User Agency

Explainability techniques such as saliency maps remain largely inaccessible to BLV users as they rely on visual representations [26]. Uncertainty quantification could aid calibration [10], yet such metrics are rarely exposed in commercial models.

Shneiderman’s human-centered AI framework [32] calls for moving beyond passive consumption toward greater supervisory control and user agency. Yet the practical implementation of such agency in AI systems designed for accessibility remains largely underexplored. While cognitive forcing functions and well-designed explanations have been shown to foster appropriate reliance on AI outputs [4], these techniques generally assume that users can meaningfully assess the system’s limitations—an assumption that relies on independent verification channels, which are frequently absent in BLV contexts. This reinforces the verification paradox.

2.3 Gap Analysis

While extensive research examines appropriate reliance on AI [4, 18] and explainability techniques [26], these frameworks assume users retain independent epistemic access—an assumption invalid in BLV contexts where AI constitutes the sole perceptual channel. Current evaluations predominantly emphasize *system accuracy* metrics (e.g., precision, recall, BLEU scores) rather than *user error detection capacity*. As systems evolve from classical computer vision (producing detectable classification errors) to transformer-based generative models (prone to undetectable semantic fabrications and hallucinations), failure modes become increasingly subtle and challenging to contest without ground-truth access—yet evaluation methodologies have not evolved accordingly.

To address this gap, this work introduces a composite metric quantifying verification affordances across AI architectural paradigms, integrating validation rigor, user agency, complexity, and risk into a unified verification-cognition framework.

3 Methodology

3.1 Corpus Construction

To systematically identify relevant AI-powered accessibility systems and the associated empirical evidence, we conducted a scoping review following the *PRISMA-ScR* guidelines [33]. We searched four major academic databases—ACM Digital Library, IEEE Xplore, SpringerLink, and PubMed—using structured Boolean queries that combined terms related to visual impairment, assistive technologies, visual content access, and empirical evaluation.

Studies were included if they: (1) reported empirical evaluation with BLV participants (no proxies or simulations), (2) focused on technologies enabling access to visual content (images, charts, diagrams, interfaces, documents), (3) employed AI/ML as a core functional component, and (4) were published as peer-reviewed full papers between 2015 and 2025.

This initial search and screening process yielded a parent corpus of 66 papers. From these, we further selected studies in which artificial intelligence or machine learning constituted a core functional component and where evaluation involved direct participation of BLV users (no proxy or simulation-based assessments). This curation resulted in a critical corpus of 17 papers, each representing an

empirically validated AI-powered accessibility system, spanning the period 2019–2025.

We categorize the 17 systems into three architectural paradigms. Four systems required reclassification from their raw coding label to match the paper’s taxonomy: the Android math application, the multimodal DNN framework, and the educational dialogue system were reclassified as *Classical CV*, while KASI (coded ‘cnn’) was reclassified as *Hybrid Specialized*; the ground-truth mapping is provided in Table 4.

- **Classical CV & Rule-Based (n=7):** Systems employing deterministic algorithms or convolutional neural networks (pre-transformer era) with constrained, structured outputs. This includes EyeMath [9], ChartVi [23], an educational dialogue system [3], an Android math application [1], a chart extraction pipeline [7], a multimodal DNN framework [17], and Edukoi [13]. Errors in these systems—such as misclassifications or extraction failures—are often detectable through domain knowledge or consistency checks.
- **Transformer-Based (n=8):** Systems using LLMs or vision-language models for generative interpretation and narrative production. Examples include AltCanvas [22], AltGen [31], AlternAtIve [27], MAIDR AI [30], LectureAssistant [2], MIUI [36], TableNarrator [24], and VizAbility [12]. While enabling rich semantic descriptions, these models introduce risks of hallucinations and semantic drift that can be difficult to detect without ground-truth access.
- **Hybrid Specialized (n=2):** Systems where AI contributes to semantic interpretation but is constrained to specific sub-tasks within a larger, structured workflow. The two systems in this category—KASI [34] and TableView+ [21]—show that controlled AI integration, rather than general-purpose generative modeling, can yield stronger verification affordances.

3.2 Data Extraction and Coding

3.2.1 Coding Framework. A multidimensional coding scheme was applied to each paper, encompassing AI/ML characteristics, documented risks and mitigations, system topology, validation methodology, user agency and interaction design, methodological quality, results, participant profiles, reproducibility, output modalities, ethical considerations, cognitive load, usability, and contextual factors. This scheme provided a structured basis for analyzing technical specifications, evaluation rigor, and verification affordances. A second trained coder independently coded 20% of the corpus. Inter-rater agreement reached Cohen’s $\kappa = 0.83$ for the qualitative coding items. Disagreements were resolved through discussion and consensus.

3.2.2 Derived Metrics and Scoring Rubrics. To enable quantitative comparison across the 17 systems, we developed a multidimensional scoring framework. Each system was coded across five primary metrics using a combination of ordinal scales (for structural and interaction properties) and continuous indices (for methodological and risk assessments), all normalized to a 0–10 range where appropriate.

Topology Complexity (C, scale 1–4). This metric assesses the structural sophistication of information processing. For BLV users,

higher topology often correlates with a higher cognitive-epistemic burden during verification. When a system’s affordances spanned two levels, intermediate values (e.g., 2.4, 3.6) were assigned to reflect partial fulfillment of the higher stage.

- **Level 1 (Discrete):** Single-function mapping with no branching logic (e.g., direct OCR).
- **Level 2 (Hierarchical):** Linear pipelines where information is transformed sequentially across multiple stages.
- **Level 3 (Network):** Convergent processing or multi-modal fusion where multiple information paths intersect.
- **Level 4 (Ecosystem):** Complex interacting subsystems with feedback loops where outputs influence prior processing stages.

User Agency (A, scale 1–4). Adapted from Shneiderman’s human-centered AI framework [32], this scale measures the degree of user control. In accessibility, agency is not merely about navigation but about the *epistemic capacity* to challenge AI interpretations.

- **Level 1 (Passive Consumption):** Fixed system outputs with no interaction; the user is a passive recipient.
- **Level 2 (Guided Exploration):** Pre-defined navigation pathways through constrained interfaces (e.g., simple keyboard drilling).
- **Level 3 (Steered Collaboration):** Bidirectional interaction, including natural language querying, clarification, and output contestation.
- **Level 4 (Full Epistemic Control):** Direct access to underlying source data, uncertainty estimates (e.g., confidence scores), and the ability to override AI reasoning.

Validation Rigor (V, 0–10 points). To assess methodological rigor, we developed a 10-point rubric spanning five sub-dimensions. This score captures the strength of the evidence underpinning a system’s accuracy, as well as the transparency with which it reports errors.

- **Study Design** (max 3.0): Mixed-methods or within-subjects (3.0); between-subjects with controls (2.0); qualitative-only (1.0); pilot or case study (0.5).
- **Statistical Completeness** (max 2.0): Full inferential testing with effect sizes and 95% CIs (2.0); partial inferential reporting (1.0); descriptive statistics only (0.5).
- **BLV Sample Adequacy** (max 2.0): $n \geq 15$ confirmed BLV participants (2.0); $10 \leq n < 15$ (1.5); $n < 10$ (1.0); token inclusion or $n \leq 3$ (0.5).
- **Protocol Reproducibility** (max 2.0): Comprehensive specification with materials archived (2.0); partially described (1.0); minimal detail (0.5).
- **Quality Framework** (max 1.0): Use of validated instruments (MMAT, SUS, NASA-TLX) (1.0); informal or ad-hoc rubrics (0.5).

Risk Density (R, 0–10 points). We identified and weighted documented failure modes to assess the risk of user misreliance. Unlike accuracy, this metric captures the *potential for undetectable error*. Scores are a weighted sum (capped at 10) of:

- **Critical Risks** (2.0 pts): Direct safety hazards, systemic bias, or privacy violations.

- **High Risks** (1.5 pts): Semantic hallucinations or fabrications indistinguishable from ground truth without visual access.
- **Moderate Risks** (1.0 pt): Detectable misclassifications or extraction errors (e.g., OCR "garbage" text).
- **Low Risks** (0.5 pt): Technical adoption barriers or platform dependencies.

AI Integration (0–10 points). This continuous metric assesses the centrality of AI to the system's function. It aggregates scores for architectural specificity (transformer vs. rule-based), model transparency, technique diversity (e.g., RAG + LLM), purpose clarity, and multimodal sophistication.

Justifications for individual system scores across these metrics are provided in Table 4 (Appendix).

3.2.3 The Verification-Cognition Balance Index (VCBI). We introduce the *Verification-Cognition Balance Index* (VCBI), a composite metric that quantifies the balance between verification affordances and cognitive-epistemic burden in AI accessibility systems. Verification affordances—comprising validation rigor and user agency—represent a system's capacity to support users in detecting and contesting AI errors. Cognitive-epistemic burden—comprising architectural complexity and documented risks—represents the mental effort required to verify AI outputs and the probability of encountering failures that demand such verification.

Mathematical Formulation. The VCBI combines verification affordances in the numerator and cognitive-epistemic burden in the denominator as follows:

$$\text{VCBI}(V, A, C, R) = \frac{V^{0.85} \times A^{0.85}}{1 + (C + 1.5R)^{0.65}} + 0.08 \cdot \mathbb{I}_{[V \geq 0.7 \wedge R \geq 0.6]}$$

where $V, A, C, R \in [0, 1]$ represent validation rigor (VR), user agency, topology complexity, and risk density (RD), normalized as: $V = \text{VR}/10$, $A = (\text{agency} - 1)/3$, $C = (\text{topology} - 1)/3$, $R = \text{RD}/10$.

The indicator function implements a *resilience bonus*: systems maintaining high validation ($V \geq 0.7$) despite elevated risks ($R \geq 0.6$) receive a +0.08 bonus. No system met both thresholds.

A critical mathematical property follows directly from the normalization: when Agency = 1, $A = 0$ and the numerator collapses to exactly zero, yielding VCBI = 0.000 regardless of validation quality. This is the case for the multimodal DNN framework (Agency = 1), whose VCBI $\approx$ 0.000 despite adequate validation rigor (VR = 6.4)—an extreme illustration of the agency deficit.

Design Rationale. The VCBI structure embodies three design principles:

- (1) **Verification requires both validation and agency.** The multiplicative numerator ensures deficiencies in either dimension cannot be fully compensated by the other. A system scoring 0.9 on validation but only 0.3 on agency yields $0.9^{0.85} \times 0.3^{0.85} = 0.33$, below balanced moderate scores $0.6^{0.85} \times 0.6^{0.85} = 0.42$.
- (2) **Cognitive burden compressed relative to affordances.** The denominator exponent of 0.65 causes $(C + 1.5R)^{0.65}$ to grow sub-linearly. Verification capacity emerges primarily from what systems provide (V, A) rather than what they avoid (C, R).

- (3) **Risk weighted above complexity.** The coefficient 1.5 on risk density prioritizes reliability over simplicity: each failure mode demands cognitive resources for detection, imposing greater verification burden than architectural sophistication alone.

Coefficient Calibration and Sensitivity. Coefficients were selected through iterative testing across the shared numerator exponent ($\{0.7, 0.8, 0.85, 0.9, 1.0\}$), the denominator exponent ($\{0.5, 0.65, 0.8\}$), and the risk weight ($\{1.0, 1.5, 2.0\}$). A linear numerator exponent (1.00) was rejected because it awards equal marginal weight to any gain in V or A regardless of starting level, causing the numerator to dominate over the burden denominator among high-performing systems; the chosen value (0.85) encodes diminishing returns so that the denominator remains the primary differentiator at the top of the scale. Values below 0.80 compress mid-range differentiation excessively. The denominator exponent 0.65 achieves the intended sub-linear burden growth; the risk weight 1.5 ensures high-risk systems are penalised proportionally more than equivalent-complexity low-risk systems. Across all 12 parameter configurations tested, rank ordering proved highly stable: no system shifted by more than two positions, and cluster assignments (Section 4.1.4) were unchanged in all configurations, supporting the robustness of the substantive findings. Per-system VCBI scores are derived directly from the Appendix inputs and reported in Table 4.

3.2.4 Statistical Analysis. Descriptive statistics (means, SD, ranges, 95% CI, t -distribution, $df = 16$) were computed for each metric. Quartiles use the Type 6 method (Minitab/SAS/SPSS convention: position = $(n+1) \cdot p$, linear interpolation), yielding Q1 = 0.113, median = 0.173, Q3 = 0.214 for the VCBI distribution. Cluster analysis (Ward method, Euclidean distance) and non-parametric Spearman's rank correlations (ρ) between independent components were examined, as variables like Agency and Topology are ordinal. Correlations with VCBI were avoided due to mathematical circularity. To control for the inflation of Type I error due to multiple comparisons, we applied the Benjamini-Hochberg (FDR) correction ($m = 6$) to all pairwise Spearman correlations among the four independent components: Agency $\times$ Topology, Agency $\times$ Risk, Agency $\times$ Validation, Risk $\times$ Validation, Topology $\times$ Risk, and Topology $\times$ Validation. Given small sample size ($N = 17$), analyses emphasize descriptive patterns, effect sizes (Cohen's d), and qualitative design pattern extraction rather than statistical significance testing. Where p -values are reported, they serve as supplementary indicators of pattern strength rather than confirmatory tests, acknowledging limited statistical power for detecting effects smaller than $d \approx 0.9$.

We further caution that non-significant correlations in a sample of $N=17$ should not be interpreted as evidence of independence: with this sample size, correlations below $|r| \approx 0.48$ are undetectable at conventional power levels ($1-\beta=0.80$, $\alpha=0.05$). Non-significant results are therefore reported as *no evidence of association in this sample*, not as evidence of independence.

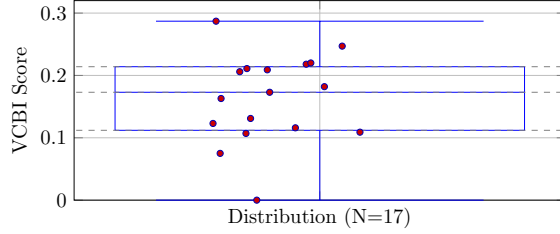
4 Results

4.1 RQ1: Verification-Cognition Profiles

4.1.1 Descriptive Statistics. Table 1 summarizes the central tendency, dispersion, and range of each primary metric across the 17

Table 1: Descriptive Statistics for Primary Metrics (N=17)

Metric	Mean	SD	95% CI	Range
VCBI (0–1)	0.163	0.071	[0.127, 0.199]	0.000–0.287
User Agency (1–4)	2.59	0.62	[2.28, 2.90]	1.0–3.0
Validation Rigor (0–10)	5.54	1.51	[4.77, 6.31]	2.0–8.0
Risk Density (0–10)	5.71	2.30	[4.53, 6.89]	1.8–10.0
AI Integration (0–10)	5.50	2.14	[4.40, 6.60]	2.55–9.0
Topology Complexity(1–4)	1.89	0.70	[1.55, 2.24]	1.0–3.0

**Figure 1: VCBI distribution across 17 AI accessibility systems. Median=0.173, IQR=[0.113, 0.214], range [0.000, 0.287].**

systems. Topology Complexity concentrates at lower levels: Level 1 accounts for 7 systems (41.2%), Level 2 for 8 systems (47.1%), Level 3 for 2 systems (11.8%), and no system reaches Level 4. User Agency skews toward Level 3: 11 systems (64.7%) achieve Steered Collaboration, 5 (29.4%) Guided Exploration, 1 (5.9%) Passive Consumption, and none reach Level 4 (Full Epistemic Control). Three salient patterns emerge: The mean VCBI of 0.163 (95% CI [0.127, 0.199]) indicates only marginal verification-cognition balance. User agency averages 2.59/4.0 with no system reaching Level 4 (Full Epistemic Control). Risk density averages 5.71/10 (SD=2.30), reflecting widespread documentation of potential failures.

4.1.2 VCBI Distribution Analysis. Figure 1 presents the VCBI distribution across all systems. The median VCBI of 0.173 (IQR=[0.113, 0.214]) indicates that the typical AI accessibility system in our corpus achieves moderate verification-cognition balance. Substantial variation exists, with scores spanning from effectively zero (multimodal DNN framework, VCBI= 0.000, a mathematical consequence of Agency= 1, c.f. Section 3.2.3) to 0.287 (TableView+ [21]).

The distribution exhibits several notable characteristics:

- **Central concentration:** The majority of systems (n=13, 76.5%) cluster within one standard deviation of the mean (0.094–0.232), suggesting convergence toward moderate balance in current design practices.
- **Lower tail cases:** Two systems (11.8%) fall substantially below the first quartile with VCBI below 0.08, representing fundamental verification-cognition failures. These systems—the multimodal DNN framework (0.000) and TableNarrator (0.075)—exhibit orthogonal failure modes detailed in Section 4.3.1.
- **Limited upper achievement:** Seven systems (41.2%) exceed 0.20: TableView+ (0.287), educational dialogue (0.247), AltCanvas (0.220), VizAbility (0.218), KASI (0.211), LectureAssistant (0.209), and Edukoi (0.206). The educational dialogue system, despite being rule-based, achieves the second-highest

score through exceptional validation rigor (VR= 7.7, highest in corpus), showing that strong evaluation can compensate for architectural simplicity. Three transformer systems also surpass 0.20, indicating that high balance is possible across paradigms.

- **Absence of excellence:** No system approaches the theoretical upper range (VCBI> 0.29). We emphasise that the theoretical maximum of the formula is 1.0 (a system with VR= 10, Agency= 4, Topology= 1, RD= 0). The value 0.287 is the empirical upper bound of our corpus, indicating that even the strongest evaluated system attains only ~29% of the theoretically attainable score—evidence of systematic design barriers rather than a scale artefact.

This polarization—concentration in the moderate range with severe low-scoring cases and rare upper achievers—suggests divergent design philosophies within the field. The rarity of high-VCBI systems indicates that achieving strong verification-cognition balance requires intentional design choices that remain uncommon in current practice.

4.1.3 The AI Capability–Risk Coupling. A striking pattern emerges: AI Integration correlates strongly with risk density ($\rho = 0.722$, $p_{adj} < 0.01$), indicating more sophisticated AI architectures systematically introduce greater documented risks. This reflects *fundamental capability-risk coupling*: generative AI systems produce outputs spanning vast possibility spaces, enabling rich interpretations but also fabrications indistinguishable from valid outputs [16]. This suggests a *verification ceiling*: beyond certain AI sophistication, user-facing verification becomes increasingly difficult without perceptual access.

4.1.4 Empirical Design Profiles via Hierarchical Clustering. To identify empirical design patterns beyond distributional statistics, we performed hierarchical cluster analysis using Ward’s method with Euclidean distance on normalized VCBI components (V, A, C, R). The dendrogram cutting analysis suggested an optimal four-cluster solution (pseudo-F=4.2), explaining 68% of the total variance. Given the sample size relative to the four-dimensional feature space, these clusters are treated as exploratory archetypes rather than definitive sub-populations. The emergence of singleton or doublet clusters (Clusters 3 and 4) reflects the high granularity inherent in clustering a small corpus, highlighting the sensitivity of these segments to exceptional individual cases.

Cluster Interpretation. Cluster 1 (*Moderate Mainstream*, 58.8%) represents the dominant paradigm: functional but unremarkable balance through moderate validation paired with steered collaboration (Agency Level 3). Spanning diverse architectures (classical CV, transformers, hybrid), these systems neither fail critically nor excel. The upper bound of this cluster (0.247) is achieved by the educational dialogue system, a Classical CV/rule-based system that attains this score through exceptional validation rigor (VR= 7.7, the highest in the corpus) rather than architectural specialization—illustrating that evaluation quality can compensate for architectural limitations. Notably, one transformer system (AlternAtive) and one Classical CV system (Android math app) also fall into this cluster, indicating that moderate balance is attainable across paradigms when validation and agency are sufficiently aligned.

Table 2: Empirical Design Profiles from Hierarchical Clustering (Ward’s Method, N=17)

Cluster (%)	n	Systems and Characteristics
Cluster 1: Moderate Mainstream (58.8%)	10	<i>KASI</i> [34], <i>AltCanvas</i> [22], <i>Edukoi</i> [13], <i>MAIDR AI</i> [30], <i>LectureAssistant</i> [2], <i>MIUI</i> [36], <i>educational dialogue</i> [3], <i>VizAbility</i> [12], <i>AlternAtive</i> [27], <i>Android math app</i> [1] Mean VCBI=0.194 [0.109–0.247]; balanced validation (5.59) and high agency (2.90); represents typical satisficing design approaches.
Cluster 2: Constrained Strugglers (23.5%)	4	<i>AltGen</i> [31], <i>ChartVi</i> [23], <i>EyeMath</i> [9], <i>chart pipeline</i> [7] Mean VCBI=0.119 [0.107–0.130]; constrained by low agency (2.00) despite moderate validation (5.47); risk density (5.40) slightly above the Classical CV subsample average (4.24), without compensatory verification support. Three of four use Classical CV architectures (AltGen is Transformer), suggesting legacy patterns prioritizing automation over empowerment.
Cluster 3: Critical Imbalance (11.8%)	2	<i>multimodal DNN framework</i> [17], <i>TableNarrator</i> [24] Mean VCBI=0.037 [0.000–0.075]; orthogonal failure modes—agency deficiency (VCBI= 0.000, Agency= 1) vs. validation failure (VCBI= 0.075, VR= 2.0)—both insufficient for verification.
Cluster 4: High Performers (5.9%)	1	<i>TableView+</i> [21] VCBI=0.287; validation excellence (8.0) compensates for moderate agency (3.0) and controlled complexity (2.4); demonstrates achievable upper bound.

Cluster 2 (*Constrained Strugglers*, 23.5%) show low agency (mean= 2.00/4.0) despite adequate validation. Three of four use Classical CV architectures (AltGen is Transformer), suggesting legacy patterns prioritizing automation over empowerment. Risk density (5.40) is above the Classical CV subsample average (4.24), without compensatory mechanisms, creating verification vulnerability.

Cluster 3 (*Critical Imbalance*, 11.8%) exhibit orthogonal failures: the multimodal DNN (VCBI= 0.000) has validation (6.4) without agency (1.0)—the numerator collapses to exactly zero; TableNarrator (VCBI= 0.075) has agency (3.0) without validation (2.0). Both demonstrate that verification requires *both* dimensions.

Cluster 4 (High Performers, 5.9%) contains only TableView+ (VCBI= 0.287). Its singularity indicates that exceptional verification-cognition balance remains rare in current practice. While this isolation marks the system as a benchmark, it also reflects the sensitivity of small-sample clustering to high-performing cases, which Ward’s method tends to isolate individually rather than absorb into broader groups.

Architectural Distribution Across Clusters. Classical CV systems distribute across clusters: 3 in Cluster 1, 3 in Cluster 2, 1 in Cluster 3. Transformer-based systems show higher variance: 6 in Cluster 1, 1 in Cluster 2, 1 in Cluster 3. Both hybrid systems appear in high-performing clusters (*KASI* in Cluster 1, *TableView+* in Cluster 4). While consistent with the hypothesis that specialized architectures facilitate stronger verification affordances, this pattern is based on only two systems and should be interpreted as preliminary.

Architectural paradigm shows a stronger descriptive association with VCBI variance ($R^2 \approx 0.18$) than publication year ($R^2 \approx 0.07$), both

Table 3: Verification-Cognition Balance by Architectural Paradigm (N=17)

Metric	Classical CV (n=7)	Transformer (n=8)	Hybrid (n=2)
<i>Verification Affordances</i>			
VCBI (mean)	0.141	0.162 ^a	0.249
VCBI [min–max]	[0.000, 0.247]	[0.075, 0.220]	[0.211, 0.287]
Validation Rigor (mean)	5.81	5.05	6.55
User Agency (mean)	2.14	2.88	3.00
Agency Level 4 achieved	0/7	0/8	0/2
<i>Risk Profile</i>			
Risk Density (mean)	4.24	7.64 ^b	3.15
Risk Density [min–max]	[1.80, 5.50]	[4.60, 10.00]	[2.30, 4.00]
Hallucination documented	1/7 (14.3%)	7/8 (87.5%)	0/2 (0%)
<i>Methodological Profile</i>			
Topology Complexity (mean)	1.57	2.04	2.45
Design pattern adoption	3/7 (42.9%)	8/8 (100.0%)	1/2 (50.0%)

^a VCBI gap vs Classical CV: Cohen’s $d = 0.31$ (small). ^b Risk gap vs Classical CV: Cohen’s $d = 2.43$ (very large). Effect sizes for Hybrid omitted ($n = 2$, unstable). Group comparisons via Mann-Whitney U test (two-tailed); see Section 4.1.4. Hybrid scores: VR mean = $(5.1+8.0)/2$; Topology mean = $(2.5+2.4)/2$. Design pattern adoption = at least one verification-aware pattern implemented.

estimated via OLS regression of VCBI on dummy-coded paradigm and continuous publication year, respectively. To compare Classical CV ($n = 7$) and Transformer-based ($n = 8$) groups, Mann-Whitney U tests revealed significant differences in Risk Density ($U = 3.0$, $p = 0.005$). Agency showed a borderline difference ($U = 11.0$, $p \approx 0.05$, exact test), which should be interpreted with caution given complete separation of values between groups. No significant differences were found for Topology Complexity ($U = 17.0$, $p = 0.203$), VCBI ($U = 24.0$, $p = 0.643$), or Validation Rigor ($U = 24.5$, $p = 0.685$). Thus, risk profile—not verification-cognition balance—is the main differentiator between paradigms.

4.1.5 Relationships Between Independent Components. To assess relationships between independent components (Agency, Topology, Risk, and Validation) while avoiding circularity with the VCBI, we examined pairwise Spearman correlations. Agency and Topology show a strong positive association (Spearman’s $\rho = 0.669$, raw $p = 0.0003$), which survives the Benjamini-Hochberg correction ($p_{adj} = 0.0018$, **). Agency and Risk also exhibit a moderate positive correlation ($\rho = 0.503$, raw $p = 0.0048$, $p_{adj} = 0.0144$, *). This indicates that systems with greater user control tend to present more structurally complex information pipelines and also face higher documented risks. The remaining four pairs show no significant associations after BH correction: Agency×Validation ($\rho = 0.151$, $p_{adj} = 0.0546$), Risk×Validation ($\rho = -0.134$, $p_{adj} = 0.0546$), Topology×Risk ($\rho = 0.181$, $p_{adj} = 0.0546$), and Topology×Validation ($\rho = 0.208$, $p_{adj} = 0.0546$). These results indicate that, after statistical correction, two of the four dimensions (Agency and Topology, and Agency and Risk) are significantly associated, while Validation operates largely independently of the others.

4.2 RQ2: Architectural Patterns in Verification Affordances

Table 3 reveals systematic differences across paradigms, spanning verification affordances, risk profiles, and methodological practices. Three critical patterns emerge:

Pattern 1: Capability-Risk Coupling Without Validation Improvement. Transformer-based systems document 80% higher risk density than Classical CV (7.64 vs. 4.24, $d = 2.43$, very large effect), with hallucinations reported in 7 of 8 systems (87.5%) versus one among Classical CV (the multimodal DNN framework, which documents hallucination as a risk in its LLM comparator evaluation). Yet this capability leap brings no validation improvement: Classical CV systems maintain marginally higher average validation rigor (5.81 vs. 5.05, $d = 0.52$), while the VCBI gap remains small (0.162 vs. 0.141, $d = 0.30$). Crucially, Classical CV systems demonstrate that deterministic architectures with controlled output spaces achieve weak synergy (mean VCBI= 0.141) *primarily due to limited agency* (mean= 2.14/4.0) despite relatively lower risks—not because of the risks themselves.

Pattern 2: Agency Progression But Persistent Ceiling. Agency increases from Classical CV to more sophisticated paradigms (2.14 → 2.88), representing a 34% improvement. Yet as shown in Table 3, no system across *any* paradigm achieves Level 4 (Full Epistemic Control): the *Agency Level 4 achieved* row reads 0/7, 0/8, and 0/2 respectively. Transformer systems like MAIDR AI [30] and VizAbility [12] implement query-based interaction (Level 3) but do not expose confidence estimates, reasoning traces, or raw data access—capabilities their architectures could readily support. The field prioritizes task efficiency over epistemic scrutiny.

Pattern 3: Hybrid Specialized as Implementation Proof Cases. The two hybrid systems achieve the highest observed scores across most affordance dimensions: VCBI mean of 0.249, the lowest risk density (3.15), zero documented hallucinations, and strong design pattern adoption (1/2, 50% vs. 100% for transformers and 43% for Classical CV). However, this result is strictly descriptive and must not be interpreted as a statistically validated paradigm victory. With only two instances, these systems act as high-performance proof cases of implementation excellence rather than a robust population:

- **AI constrained to semantic subtasks:** Both KASI [34] (tactile piece recognition) and TableView+ [21] (semantic table extraction) leverage AI for narrow, well-defined tasks within structured workflows. This minimizes the error surface and ensures failures manifest as detectable misclassifications rather than undetectable hallucinations.
- **Exceptional methodological rigour:** The high performance likely reflects the rigorous evaluation standards of their authors (MMAT scores ≥ 85 , BLV samples $n \geq 25$, extensive mixed-methods protocols) rather than an inherent property of hybridity alone.

Both systems were published in 2022–2023, before widespread generative AI deployment pressure, and target constrained domains. These characteristics—domain specificity, publication timing, and architectural choice—are confounded, precluding attribution of the observed advantage to architecture alone. Replication across a larger and more diverse sample of hybrid systems is necessary to validate these exploratory findings.

4.2.1 Architectural Analysis Summary. Three key findings emerge:

- (1) **Architectural determinism:** AI paradigm explains approximately 18% of VCBI variance ($R^2 = 0.178$, OLS), while publication year explains approximately 7% ($R^2 = 0.07$, OLS),

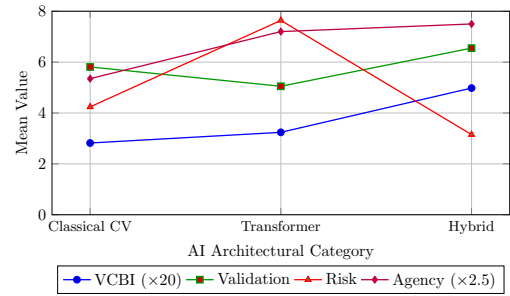


Figure 2: Verification-cognition balance across AI architectural paradigms (N=17). VCBI×20 and Agency×2.5 scaled for visibility.

suggesting that neither architectural choices nor temporal trends alone strongly predict verification affordances. The VCBI difference between Classical CV and Transformer paradigms is negligible and non-significant ($U = 24.0$, $p = 0.643$), while the difference in Risk Density remains statistically supported ($U = 3.0$, $p = 0.004$); the Agency difference is borderline significant ($U = 11.0$, $p = 0.049$).

- (2) **The verification ceiling:** Beyond transformer-scale generative models, user-facing verification becomes increasingly difficult without perceptual access. Each capability increment brings fundamentally new failure modes.
- (3) **Specialization as mitigation:** The two hybrid systems demonstrate that controlled AI scope within task-specific architectures can achieve stronger verification affordances, suggesting a promising direction for future investigation rather than a confirmed design principle.

4.3 RQ3: Design Patterns for Verification-Aware Systems

4.3.1 Comparative Analysis: High vs. Low VCBI Systems. We contrast TableView+ (VCBI= 0.287, Cluster 4) with the two critically imbalanced cases from Cluster 3.

TableView+ achieves the strongest balance through compensatory equilibrium: high validation rigor (8.0/10) offsets moderate agency (3.0/4) and limited architectural complexity (2.4/4). Its performance is driven by rigorous mixed-methods evaluation with BLV participants, customizable yet constrained interaction affordances, explicit documentation of failure modes, and a hierarchical pipeline architecture that localizes AI errors.

By contrast, the two low-VCBI systems exhibit orthogonal failure modes. The multimodal DNN framework (VCBI= 0.000) attains reasonable validation rigor (6.4/10) but offers minimal user agency (1.0/4), collapsing the VCBI numerator to exactly zero. Conversely, TableNarrator (VCBI= 0.075) provides substantial agency (3.0/4) through interactive exploration but lacks adequate validation (2.0/10), rendering its outputs epistemically unreliable. Together, these cases illustrate a core design constraint: verification requires both validation rigor and user agency, and a deficit in either dimension critically undermines epistemic safety.

4.3.2 Design Pattern Extraction.

Pattern 1: Provenance Transparency (Sensemaking Support). Provenance transparency—grounded in Russell et al.’s sensemaking loop [28]—enables users to trace AI-generated interpretations back to their evidential basis. Observed implementations include: *hierarchical drill-down* (TableView+ [21]); *sonified data access* (chart pipeline [7]), providing a direct auditory mapping; and *query-based retrieval* (MAIDR AI [30], VizAbility [12]). Adoption was strikingly low: one system (5.9%) implemented full provenance; three (17.6%) offered partial implementation; 76.5% provided none. An optimal design follows *progressive disclosure* [25]: trustworthy summary by default, with on-demand drill-down pathways.

Pattern 2: Uncertainty Signaling (Calibration Support). Uncertainty signaling extends confidence communication [35] to non-visual modalities. Promising implementations include: *prosodic modulation* [5]; *semantic hedging* (“likely,” “possibly”); *earcon tiers* [6]; and *explicit querying*. Despite its critical role in mitigating over-reliance, no system in our corpus exposed confidence estimates to users. Four systems (23.5%) computed internal confidence metrics but failed to surface them. Uncertainty must be *passively perceivable*: requiring users to actively query confidence for each AI statement creates prohibitive interaction burden.

Pattern 3: Challenge Interactions (Agency Support). This pattern operationalizes Shneiderman’s “Collaborate” level [32] for episodic scrutiny: *open-ended querying* (MAIDR AI [30], VizAbility [12]); *structured alternatives* (ChartVi [23]); and *undo/redo with explanations*. Only two systems (11.8%) fully implemented open-ended querying; four (23.5%) offered structured alternatives; 64.7% provided no contestation mechanisms. The educational dialogue system’s authors note that without such controls, LLMs proved inadequate for consistently describing FSA [3].

4.3.3 Architecture-Specific Design Pattern Adoption. Classical CV systems show moderate adoption (3/7, 42.9%). Transformer-based systems demonstrate universal adoption (8/8, 100.0%). Hybrid systems show partial adoption (1/2, 50.0%). Transformer systems document 80% higher risk density (7.64 vs 4.24, $d = 2.43$) with a negligible VCBI difference (0.162 vs 0.141, $d = 0.30$) and show no improvement in validation rigor (5.05 vs 5.81 for Classical CV, $d = 0.52$). Despite 100% pattern adoption, transformer-based designs still create significant verification gaps—isolated design patterns are insufficient; verification must be an architectural priority from the outset.

4.3.4 The Agency-Verification Link: Quality Over Quantity. Agency implementation quality matters more than quantity. We distinguish *verification-relevant agency*—questioning outputs, drilling to source data, contesting interpretations—from *task-relevant agency*—filtering, sorting, customizing presentation. Both MAIDR AI and TableView+ score 3.0/4.0 on agency, but MAIDR AI’s question-answering capability directly supports verification while TableView+’s table customization supports efficiency. A refined metric distinguishing *epistemic control* from *interface control* would better predict verification capacity.

5 Discussion

5.1 Architectural Determinism and the Verification Ceiling

Classical CV & Rule-Based: Constrained but Underutilized. These systems ($n=7$, mean VCBI= 0.141) achieve weak synergy despite controlled risks (mean= 4.24/10) due to limited agency (mean= 2.14/4.0). Poor agency affordances prevent users from exercising verification potential—as illustrated by the multimodal DNN framework [17], which achieves adequate validation (6.4/10) but VCBI exactly zero due to its passive consumption design (Agency= 1.0). This case demonstrates that even potentially verifiable AI fails without interaction affordances.

Transformer-Based: Capability Without Verifiability. These systems ($n = 8$, mean VCBI= 0.162) show negligibly higher balance than Classical CV (0.141), and this difference is non-significant ($U = 24.0$, $p = 0.643$). They achieve marginally improved agency (2.88/4, +34% vs Classical CV; $U = 11.0$, $p \approx 0.05$) but face significantly elevated risks (7.64/10, +80%; $U = 3.0$, $p = 0.005$) without validation improvements—mean validation rigor (5.05/10) is lower than Classical CV (5.81/10, $U = 24.5$, $p = 0.685$). Among transformer systems, 87.5% (7/8) document hallucination concerns versus 14.3% (1/7) of Classical CV systems.

The educational dialogue system (VCBI= 0.247, Classical CV) represents a notable exception: it achieves the second-highest score in the corpus through exceptional evaluation rigor (VR= 7.7), not architectural specialization. This case reinforces the core VCBI insight—validation quality and agency jointly determine verification capacity—and confirms that even rule-based systems can achieve strong balance when evaluation standards are sufficiently rigorous, albeit rarely in current practice.

Hybrid Specialized: Emerging Proof Cases. The hybrid systems in our corpus ($n = 2$) achieve the strongest balance (mean VCBI= 0.249). However, with only two systems, these findings represent proof cases of implementation excellence rather than a statistically generalizable architectural victory. The high performance of KASI [34] and TableView+ [21] likely stems from a synergy between constrained semantic subtasks and exceptionally rigorous evaluation protocols. While these systems demonstrate that controlled AI integration can yield superior verification affordances, the sample size is too small to calculate reliable variance or to isolate the architectural effect from individual researcher rigor.

Design Implication. Although the hybrid specialized approach achieved the highest average VCBI in our corpus, this outcome reflects exemplary implementation cases rather than a statistically generalizable superiority of the paradigm itself. With $n = 2$, the success of KASI and TableView+ likely reflects synergy between narrow-domain semantic constraints and the methodological rigor of their authors (both reporting MMAT scores ≥ 85). These systems serve as verifiable design benchmarks, suggesting that constraining AI to well-defined subtasks provides a promising roadmap for future AT. However, further research with a larger sample is required to decouple the benefits of specialized architectures from the evaluation standards observed in these cases.

5.2 The Agency-Verification Gap

Mathematical sensitivity analysis shows agency has equal marginal impact on VCBI as validation rigor ($\approx 8\%$ improvement per 10% increase). Yet empirically, agency remains underdeveloped (corpus mean = 2.59/4.0, Classical CV = 2.14, Transformer = 2.88, Hybrid = 3.00), with no system achieving Level 4. Improving agency may be more feasible than further increasing validation rigor for many research teams.

5.3 Risk Documentation Without Mitigation

Despite comprehensive risk documentation (mean = 5.71/10), no system provides user-facing mitigation mechanisms. Among transformer systems documenting hallucination (87.5%), none implement detection capabilities. Risks are treated as algorithmic accuracy problems rather than verification challenges addressable through interaction design. No system earned the VCBI resilience bonus ($V \geq 0.7 \wedge R \geq 0.6$).

5.4 Independence of Design Dimensions

Our analysis reveals that two of the four VCBI dimensions are significantly associated: Agency and Topology ($\rho = 0.669$, $p_{adj} = 0.0018$) and Agency and Risk ($\rho = 0.503$, $p_{adj} = 0.0144$). Validation Rigor, however, operates largely independently of the other three dimensions, showing no significant correlations after correction. This suggests that systems with higher user agency tend to have more complex information pipelines and also face greater documented risks, but validation strength does not systematically co-vary with these factors. The partial independence confirms that the VCBI captures distinct facets of verification-cognition balance, with Agency acting as a central hub linked to both complexity and risk, while Validation remains an orthogonal pillar that must be intentionally addressed.

5.5 From Accuracy to Verifiability

Among our 17 systems, 82.4% report accuracy metrics yet none measure error detection rates. This reveals a *methodological blind spot*: we measure what AI gets right, not what users can verify. We propose three evaluation criteria for verification-aware AI. **Error Detection** measures users' ability to identify injected errors without external validation (target: $>70\%$ detection rate). **Scrutiny Friction** counts the interaction steps required to access the evidential basis for an interpretation (target: <5 steps). **Contestation Success** measures the percentage of user challenges eliciting a meaningful AI response (target: $>50\%$). Future papers should report at least one of these verification metrics alongside traditional accuracy measures.

5.6 Limitations and Future Directions

Our sample size ($N = 17$) limits statistical power and increases the risk of instability in multivariate analyses. Specifically, performing hierarchical clustering in a four-dimensional space on a micro-sample may lead to over-segmentation, as the algorithm becomes highly sensitive to unique system profiles. Consequently, our cluster assignments should be viewed as a preliminary taxonomy intended to provide exploratory insights into design trends rather than a finalized population model. VCBI component circularity precludes

correlating individual components with the composite index. Our agency metric may underrepresent verification-specific affordances by conflating task-oriented and epistemic agency. Architectural classification required interpretive judgment for hybrid systems.

Our corpus focuses on visual information accessibility, excluding navigation and mobility domains. Geographic and linguistic diversity is limited: 76.5% of systems originate from US or European institutions. Adopting verification-aware designs faces practical barriers: commercial LLM APIs restrict access to uncertainty estimates; non-visual uncertainty signaling lacks shared conventions.

Future directions include: verification-aware protocols measuring error detection rates; refined agency metrics distinguishing epistemic from task-level control; corpus expansion (target $N \geq 50$); causal studies disentangling design choices from resource effects; adaptive interfaces scaling scrutiny to task stakes; longitudinal studies of trust calibration; and cross-domain extensions.

6 Conclusion

This work establishes verification affordances—the capacity for users to detect and contest AI errors—as a critical yet underexamined dimension of AI accessibility systems. Through the VCBI, we provide the first quantitative framework for assessing how architectural choices, validation practices, user agency, and risk exposure jointly shape epistemic safety for users who cannot independently validate AI outputs. Our analysis reveals a field at an architectural crossroads. Widespread adoption of transformer-based generative models has expanded AI capabilities while introducing failure modes fundamentally harder to verify without perceptual access. The rarity of systems achieving strong verification-cognition balance (41.2% exceeding VCBI=0.20 in our corpus, with two hybrid systems and two Classical CV systems—the educational dialogue and Edukoi—among the seven high-achievers) indicates that designing for verifiability requires intentional architectural and interaction choices that remain uncommon in current practice.

We argue for a paradigm shift: verification affordances must be design requirements, assessed by metrics measuring users' ability to detect AI failures, not just successes. For high-stakes accessibility applications, the question is not whether AI can perform a task, but whether users can maintain epistemic agency over AI-mediated information. The VCBI framework provides a methodological starting point for this shift. Although bounded by corpus size and domain scope, our findings establish that verification-cognition balance is quantifiable and achievable through deliberate design. As AI systems increasingly mediate access to information for those who cannot independently verify their outputs, ensuring users can appropriately trust—and appropriately distrust—these systems becomes an ethical imperative alongside a technical one.

Acknowledgments

This work is part of the COOBRA project (Coopération entre voyants et déficients visuels brailleistes pour apprendre et travailler ensemble), funded by the Agence Nationale de la Recherche (grant ANR-24-SARP-0006) under the Science with and for Society programme (SAPS-RA-RP2).

References

- [1] Rina Agustina, Nurul Farida, and Muhammad Irfan. 2024. Learning Mathematics Outcomes Using Android for Blind Students Based on Newman's Theory. *Journal of Education and Learning (EduLearn)* 18, 3 (2024), 1056–1065.
- [2] Katharina Anderer, Karin Müller, Lukas Strobel, Matthias Wölfel, Jan Niehues, and Kathrin Gerling. 2025. Making Lecture Videos Accessible for Students who are Blind or have Low Vision through AI-Assisted Navigation and Visual Question Answering. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)*. Association for Computing Machinery, New York, NY, USA, Article 56, 15 pages. doi:10.1145/3663547.3746349
- [3] Pier Felice Balestrucci, Elisa Di Nuovo, Manuela Sanguinetti, Luca Anselma, Cristian Bernareggi, and Alessandro Mazzei. 2024. An educational dialogue system for visually impaired people. *IEEE Access* 12 (2024), 150502–150519. doi:10.1109/ACCESS.2024.3479883
- [4] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. AAAI Press, Honolulu, HI, USA, 2–11.
- [5] Susan E. Brennan and Marilyn Williams. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language* 34, 3 (1995), 383–398. doi:10.1006/jmla.1995.1018
- [6] Stephen A. Brewster, Peter C. Wright, and Alistair D. N. Edwards. 1993. An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) (CHI '93). Association for Computing Machinery, New York, NY, USA, 222–227. doi:10.1145/169059.169179
- [7] Jinho Choi, Sanghun Jung, Deok Gun Park, Jaegul Choo, and Niklas Elmquist. 2019. Visualizing for the Non-Visual: Enabling the Visually Impaired to Use Visualization. *Computer Graphics Forum* 38, 3 (2019), 249–260. doi:10.1111/cgfc.13686
- [8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., Red Hook, NY, USA, 49250–49267.
- [9] Kanlayanee Dumkasem, Padchaya Srisingchai, and Prapaporn Rattanamrong. 2019. Eyemath: Increasing accessibility of mathematics to visually impaired readers. In *2019 23rd International Computer Science and Engineering Conference (ICSEC)*. IEEE, Piscataway, NJ, USA, 197–202. doi:10.1109/ICSEC47112.2019.8974682
- [10] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*. PMLR, New York, NY, USA, 1050–1059.
- [11] Gregory L. Goodrich, Richard R. Bennett, William De l'aune, Harvey Lauer, and Leonard Mowinski. 1979. Kurzweil Reading Machine: A Partial Evaluation of Its Optical Character Recognition Error Rate. *Journal of Visual Impairment & Blindness* 73, 10 (1979), 389–399.
- [12] Joshua Gorniak, Yoon Kim, Donglai Wei, and Nam Wook Kim. 2024. VizAbility: Enhancing Chart Accessibility with LLM-based Conversational Interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) (UIST '24). Association for Computing Machinery, New York, NY, USA, Article 89, 19 pages. doi:10.1145/3664777.3676414
- [13] Lucrezia Guiotto Nai Fovino, Anita Zanella, Luca Di Mascolo, Michele Ginolfi, Nicolò Carpi, Francesco Trovato Manunola, and Massimo Grassi. 2024. Evaluating the effectiveness of sonification in science education using Edukoi. *Personal and Ubiquitous Computing* 28, 5 (2024), 693–711.
- [14] Jonggi Hong and Hernisa Kacorri. 2024. Understanding How Blind Users Handle Object Recognition Errors: Strategies and Challenges. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) (ASSETS '24). Association for Computing Machinery, New York, NY, USA, Article 63, 15 pages. doi:10.1145/3663548.3675635
- [15] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155
- [16] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (2023), 1–38. doi:10.1145/3571730
- [17] Edward Kim, Connor Onweller, and Kathleen F. McCoy. 2021. Information Graphic Summarization using a Collection of Multimodal Deep Neural Networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, Piscataway, NJ, USA, 10188–10195. doi:10.1109/ICPR48806.2021.9412146
- [18] Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. arXiv:2112.11471 [cs.HC] <https://arxiv.org/abs/2112.11471>
- [19] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What frustrates screen reader users on the web: A study of 100 blind users. *International Journal of Human-Computer Interaction* 22, 3 (2007), 247–269.
- [20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* 521, 7553 (2015), 436–444. doi:10.1038/nature14539
- [21] Hae-Na Lee and Vikas Ashok. 2022. Customizable tabular access to web data records for convenient low-vision screen magnifier interaction. *ACM Transactions on Accessible Computing (TACCESS)* 15, 2 (2022), 1–22.
- [22] Seonghee Lee, Maho Kohga, Steve Landau, Sile O'Modhrain, and Hari Subramonyam. 2024. AltCanvas: A Tile-Based Editor for Visual Content Creation with Generative AI for Blind or Visually Impaired People. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) (ASSETS '24). Association for Computing Machinery, New York, NY, USA, Article 70, 22 pages. doi:10.1145/3663548.3675600
- [23] Prerna Mishra, Santosh Kumar, Mithilesh Kumar Chaube, and Urmila Shrawankar. 2022. ChartVi: Charts summarizer for visually impaired. *Journal of Computer Languages* 69 (2022), 101107.
- [24] Ye Mo, Gang Huang, Liangcheng Li, Dazhen Deng, Zhi Yu, Yilun Xu, Kai Ye, Sheng Zhou, and Jiajun Bu. 2025. TableNarrator: Making Image Tables Accessible to Blind and Low Vision People. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–17.
- [25] Don Norman. 2013. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, New York, NY, USA.
- [26] Chukwunonso Henry Nwokoye, Maria JP Peixoto, Akriti Pandey, Lauren Pardy, Mahadeo Sukhai, and Peter R Lewis. 2024. A survey of accessible explainable artificial intelligence research. arXiv:2407.17484 [cs.HC]
- [27] Giacomo Pedemonte, Maurizio Leotta, Marina Ribaudo, et al. 2025. Improving web accessibility with an llm-based tool: A preliminary evaluation for stem images. *IEEE ACCESS* 13 (2025), 107566–107582.
- [28] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*. ACM, Amsterdam, The Netherlands, 269–276.
- [29] William Schiff and Emerson Foulke (Eds.). 1982. *Tactual Perception: A Sourcebook*. Cambridge University Press, Cambridge.
- [30] JooYoung Seo, Sanchita S Kamath, Aziz Zeidieh, Saairam Venkatesh, and Sean McCurry. 2024. Maird meets ai: Exploring multimodal llm-based data visualization interpretation by and with blind and low-vision users. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 1–31.
- [31] Yixian Shen, Hang Zhang, Yanxin Shen, Lun Wang, Chuanqi Shi, Shaoshuai Du, and Yiyi Tao. 2025. Altgen: Ai-driven alt text generation for enhancing epub accessibility. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence*. ACM, New York, NY, USA, 78–83.
- [32] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction* 36, 6 (2020), 495–504.
- [33] Andrea C. Tricco et al. 2018. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine* 169, 7 (2018), 467–473. doi:10.7326/M18-0850
- [34] Sarah E Wegwerth, Gianna J Manchester, and Julia E Winter. 2023. A feasibility study of the kasi learning system to support independent use of stem diagrams by students with visual impairments. *Journal of Visual Impairment & Blindness* 117, 2 (2023), 162–174.
- [35] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona, Spain, 295–305.
- [36] Yue Zhao. 2025. MIUI: Make Interaction Usable on Information Visualization for People with Visual Impairments. In *2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, Piscataway, NJ, USA, 104–109.

A Per-System Score Justifications

Table 4 presents all coded scores with concise justifications and computed VCBI values. Scores are derived from the rubrics defined in Section 3; VCBI values are computed via the formula in Section 3.2.3.

Table 4: Per-System Scores, VCBI, and Justifications (N=17).

System	Arch.	VR (0–10)	Ag. (1–4)	Topo. (1–4)	RD (0–10)	VCBI (0–1)	Key justification
KASI [34]	Hyb.	5.1	3	2.5	2.3	0.211	VR: Mixed methods, $n=10$ BVI, MMAT=85; partial inferential reporting (no effect sizes) caps statistical completeness. Ag.: Multimodal querying + contextual elaboration (L3); no confidence exposure (L4 absent). Topo.: Predominantly hierarchical pipeline with one conditional branch for piece-recognition routing (2.5). RD: STEM chemistry domain (molecular models) and task-specific CV confine error space; minor occlusion failures and hardware dependency documented.
AltCanvas [22]	Trans.	6.5	3	2.0	7.0	0.220	VR: Within-subjects study with BLV; inferential stats partially reported; no formal quality checklist. Ag.: Structured alternative views + clarification queries (L3); no model-confidence access. Topo.: Tile-based editor with canvas→tile parent-child hierarchy (N2); spatial coordinates not preserved in non-visual output (2.0). RD: Hallucination (1.5×2), visual element misidentification (1.0×2), bias on non-standard charts (1.5).
AltGen [31]	Trans.	5.8	2	1.0	8.0	0.116	VR: Between-subjects; descriptive stats only; partial BLV inclusion. Ag.: Output navigation among fixed alternatives; no querying (L2). Topo.: Single-function pipeline processing individual images without spatial relationships; N1 discrete (1.0). RD: Multi-stage CLIP/ViT/GPT transformer pipeline introduces hallucination risk (1.5×2); misclassification of complex images (1.0×2); coverage limits (0.5×2).
Educ. Dialogue [3]	CLCV	7.7	3	3.0	5.5	0.247	VR: Mixed methods with formal checklist (MMAT); inferential stats reported; adequate BLV inclusion; highest VR in corpus. Ag.: Bidirectional FSA dialogue (L3); rule-based AIML engine (103 rules) with LLMs evaluated as external comparators only. Topo.: FSA represented as connected states with transitions (N3 topological graph connectivity); 3.0. RD: LLM comparators show hallucination in FSA descriptions (1.5×2), cross-turn inconsistency (1.0); deployed AIML system uses controlled rule-based responses—risk score reflects evaluation of LLM comparators.
ChartVi [23]	CLCV	6.1	2	1.7	4.9	0.131	VR: Task-based evaluation; partial inferential reporting; misclassification rates empirically documented. Ag.: Keyboard navigation of chart elements; no contestation (L2). RD: Overlapping bars/legend misclassification (1.0×3).
TableView+ [21]	Hyb.	8.0	3	2.4	4.0	0.287	VR: Within-subjects, $n=25$ BLV, MMAT=90; full inferential stats + effect sizes (2.0/2.0); formal checklist (1.0/1.0); minor protocol deduction (stimuli not publicly archived) yields 8.0. Ag.: Hierarchical drill-down + semantic queries (L3); embedding scores unexposed (L4 absent). Topo.: Sequential extraction pipeline with one semantic-routing branch for cell-grouping decisions (2.4). RD: Domain-specific term misembedding (1.5), inconsistent cell grouping (1.0×2), non-standard layout failure (1.0).
Edukoi [13]	CLCV	4.7	3	1.0	4.8	0.206	VR: Controlled study, adequate BLV inclusion; partial stats; no formal checklist. Ag.: Gesture-based bidirectional navigation (L3); semantic layer is deterministic/rule-based. Topo.: Pixel-to-sound conversion without spatial relationships; N1 discrete (1.0). RD: Hand-tracking failure under variable lighting (1.0×2); limited math-content vocabulary (0.5×2).
EyeMath [9]	CLCV	5.1	2	2.3	4.5	0.107	VR: Task-based evaluation; partial stats; small BLV n . Ag.: Step-by-step keyboard navigation; no querying (L2). RD: Handwritten/cursive notation failure (1.0×2); limited operator coverage (0.5×2).
AlternAtive [27]	Trans.	3.1	3	1.6	10.0	0.109	VR: Preliminary pilot; limited BLV n ; descriptive-only statistics; no checklist—lowest VR among transformer systems. Ag.: LLM-powered alternative generation supports follow-up queries and contestation of STEM descriptions (L3). RD: Maximum score—hallucination central (1.5×2); STEM element misidentification (1.0×2); coverage limits (0.5×2); no mitigation documented.
Multimodal DNN [17]	CLCV	6.4	1	1.0	4.0	0.000	VR: Controlled study with inferential tests; adequate BLV inclusion; no quality framework. Ag.: Single fixed output per input; no navigation or querying (L1). Agency= 1 sets A=0, collapsing the VCBI numerator to exactly zero despite VR= 6.4. Topo.: Template-based extraction of discrete elements with sequential ordering; N1 discrete (1.0). RD: Misclassification across visual domains (1.0×3); limited generalisability (1.0); hardware barriers (0.5×2); constrained CNN output space limits undetectable failure modes.
Android Math [1]	CLCV	5.8	2	1.0	1.8	0.173	VR: Pilot evaluation; partial stats; small BLV n ; Newman’s theory framework provides structured assessment rubric. Ag.: Structured expression navigation; no querying (L2). Topo.: Sequential mathematical problem steps (Newman’s theory); no spatial relationships; N1 discrete (1.0). No custom ML/CV component: standard Android TalkBack STT with rule-based AST traversal for LaTeX parsing. Classified CLCV by convention (non-neural pipeline). RD: Minimal—narrow scope; handwritten notation failure (1.0); platform dependency (0.5×2).
MAIDR AI [30]	Trans.	6.0	3	3.0	8.3	0.182	VR: BLV user study; partial inferential stats; no checklist. Ag.: Open-ended LLM queries, contestation supported (L3); no confidence/data-array exposure. Topo.: Box plot statistical relationships (medians, quartiles, outliers) across categories without metric coordinates; N3 topological (3.0). RD: Hallucination in trend descriptions (1.5×2); non-standard chart misinterpretation (1.0×2); API dependency (0.5×2).
LectureAsst. [2]	Trans.	6.4	3	2.0	8.0	0.209	VR: BLV user study; partial stats; no checklist; RAG pipeline provides partial reproducibility evidence. Ag.: Section navigation + follow-up dialogue (L3). Topo.: Hierarchical structure with chapters and summaries (N2); parent-child content organisation (2.0). RD: Scientific diagram misinterpretation (1.5×2); equation hallucination (1.5); out-of-STEM brittleness (1.0×2).
MIUI [36]	Trans.	4.2	3	2.3	4.6	0.163	VR: BLV user study; limited BLV n ; absent inferential statistics; no checklist. Ag.: NL navigation + mode-switching + clarification (L3). RD: UI-label hallucination (1.5×2); layout misinterpretation (1.0×2); OS dependency (0.5×2).
TableNarrator [24]	Trans.	2.0	3	2.4	8.3	0.075	VR: BLV study included; study design poorly specified (1.0/3.0); descriptive-only statistics (0.5/2.0); sample adequacy partial (0.5/2.0); insufficient protocol detail (0.0/2.0); no checklist (0.0/1.0). Sum: 1.0+0.5+0.5+0.0+0.0 = 2.0. Ag.: Multi-turn interactive exploration (L3). RD: Hallucinated summaries (1.5×2); semantic drift across turns (1.5); structural misinterpretation (1.0).
Chart Pipeline [7]	CLCV	4.9	2	1.0	4.2	0.123	VR: BLV study; quantitative metrics + partial inferential reporting; limited procedural detail. Ag.: Sonified playback of pitch-mapped data; passive delivery without on-demand querying (L2). Topo.: Discrete bar/line/pie data point extraction with no inter-element spatial relations preserved; N1 discrete (1.0). RD: Segmentation failure on low-contrast charts (1.0×3); 2D-only coverage (0.5×2).
VizAbility [12]	Trans.	6.4	3	2.0	6.9	0.218	VR: BLV user study; partial stats; no checklist. Ag.: LLM follow-up queries on visualisation narratives (L3); no underlying data/confidence access. Topo.: Olli tree view displaying hierarchical chart navigation structure (N2); hierarchical drill-down (2.0). RD: Statistical narrative hallucination (1.5×2); axis label misinterpretation (1.0×2); API latency (0.5×2).
Mean (SD)		5.54 (1.51)	2.59 (0.62)	1.89 (0.70)	5.71 (2.30)	0.163 (0.071)	

Arch.: CLCV=Classical CV/Rule-Based; Trans.=Transformer-Based; Hyb.=Hybrid Specialized. **Columns:** VR=Validation Rigor (0–10); Ag.=User Agency (1–4); Topo.=Topology Complexity (1–4); RD=Risk Density (0–10); VCBI computed via Section 3.2.3 formula, rounded to 3 decimal places. AltGen is classified as Transformer (CLIP/ViT/GPT pipeline). Educ. Dialogue is classified as CLCV (rule-based AIML; LLMs as external comparators only).