The Verification Gap in AI Accessibility

A Meta-Analysis of Architectural Paradigms & Agency Deficits

Madjid Sadallah & Benoît Encelle

LIRIS — Université Claude Bernard Lyon 1 / CNRS / INSA Lyon

W4A '26 · Dubai, UAE · April 13–14, 2026

ANR COOBRA – ANR-24-SARP-0006

doi:10.1145/3800424.3800451

A Fundamental Shift: From Transcription to Interpretation

Then — Sensory Transduction

OCR → character streams **Screen readers** → audio

Tactile graphics → embossed representations

Systems render. ***Users*** *interpret.*
Epistemic division of labour was unambiguous.

Now — Epistemic Intermediary

AI generates natural-language **interpretations**: chart narratives (VizAbility), table summaries (TableNarrator), enriched image descriptions (Edukoi)...

The AI ***is*** *now the interpretation.*

The Verification Paradox

Sighted users: AI interpretation + direct perceptual access → easy cross-validation.

BLV users: AI = **sole information source**.

Hallucination, semantic drift, factual errors go **undetected**.

"Those who most need AI interpretation are least equipped to verify it."

Empirical evidence: BLV users detect only ∼**50%** of recognition errors (Hong & Kacorri, ASSETS '24).

Three Research Questions

1. **RQ1 — Profiles.** What verification-cognition profiles emerge from validation rigor, user agency, complexity, and risk?
2. **RQ2 — Architecture.** How does balance vary across Classical CV, Transformer, and Hybrid paradigms?
3. **RQ3 — Design.** What patterns distinguish systems with stronger verification affordances?

Four Contributions

1. First comprehensive analysis of **verification affordances** across AI accessibility paradigms ($N = 17$ systems, 2019–2025)
2. The **Verification-Cognition Balance Index (VCBI)** — a composite metric formalising the equilibrium between affordances and cognitive-epistemic burden
3. Architectural trade-off analysis & capability–risk dynamics
4. Empirically grounded **design principles** for verifiable, agency-enhancing AT

4 databases
ACM DL · IEEE · Springer · PubMed

↓

66 papers screened

↓

17 papers retained
empirical BLV · AI core component

Three Architectural Paradigms:

Classical CV & Rule-Based $n = 7$

Deterministic / pre-transformer CNNs. Constrained outputs; errors typically **detectable**.
EyeMath · ChartVi · Edukoi · chart pipeline · . . .

Transformer-Based $n = 8$

LLMs / vision-language models. Rich semantic output; hallucinations **indistinguishable** from valid outputs.
VizAbility · TableNarrator · AltCanvas · MAIDR AI · . . .

Hybrid Specialized $n = 2$

AI constrained to narrow subtasks within structured workflows. Failures manifest as **detectable misclassifications**.
KASI · TableView+

The Verification-Cognition Balance Index (VCBI)

Mathematical Formulation

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

All inputs $\in [0, 1]$: V=VR/10, A=(Agency $-$ 1)/3, C=(Topo $-$ 1)/3, R=RD/10

Why these coefficients? Exponent 0.85 encodes *diminishing returns* on V and A; the denominator stays the main differentiator among top systems. Tested over 12 configurations; rank-order shifts $\leq$2 positions.

Key property: Agency$= 1 \Rightarrow A = 0 \Rightarrow$ **VCBI**$= 0$ regardless of validation quality.

Even well-evaluated AI fails without interaction affordances.

Numerator — Affordances

V: **Validation Rigor** (0–10)
Study design · stats completeness · BLV sample · reproducibility · quality framework

A: **User Agency** (1–4 scale)
1 Passive 2 Guided 3 Steered 4 Full Epistemic

Denominator — Burden

C: **Topology Complexity** (1–4)
Discrete → Hierarchical → Network → Ecosystem

R: **Risk Density** (0–10, weighted by detectability)
Critical (2pt) · Hallucination (1.5pt) · Moderate (1pt) · Low (0.5pt)

Descriptive Statistics (N=17):

Metric	Mean	Range
VCBI (0–1)	0.163	0.000–0.287
User Agency (1–4)	2.59	1–3
Validation Rigor (0–10)	5.54	2.0–8.0
Risk Density (0–10)	5.71	1.8–10.0
Topology (1–4)	1.89	1.0–3.0

0%
reached Agency Level 4
(full epistemic control)

29%
max VCBI vs theoretical max

Three Salient Patterns:

Central Concentration

76.5% of systems cluster within 1 SD of the mean (0.094–0.232) → field converges toward *moderate, unremarkable* balance.

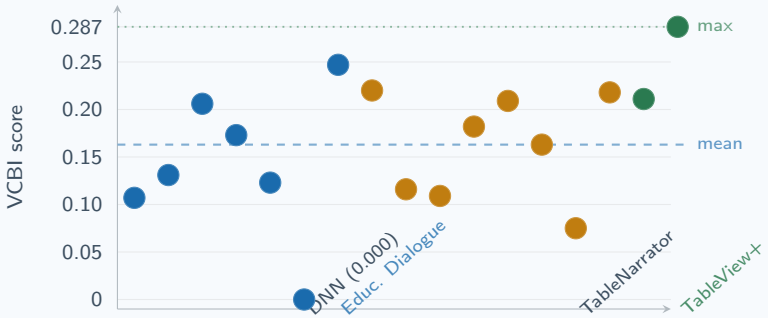
Lower-Tail Failures

2 systems (11.8%) with VCBI < 0.08:
Multimodal DNN (0.000): Agency= 1 collapses numerator.
TableNarrator (0.075): Agency= 3 but VR= 2.0 — rich interaction over unreliable output.

AI Capability–Risk Coupling

AI Integration ↔ Risk Density: $\rho = 0.722$, $p_{\text{adj}} < 0.01$.
More capable AI = harder-to-detect failures.

RQ1 — VCBI Scores Across All 17 Systems



- **Classical CV** $n = 7$
- **Transformer** $n = 8$
- **Hybrid** $n = 2$

Lower tail (VCBI<0.08):
Multimodal DNN (0.000)
TableNarrator (0.075)

Both represent design *failures*, but for **opposite** reasons.

Upper bound: TableView+ (0.287) — only ≈**29%** of the theoretical maximum of 1.0.

RQ1 — Four Empirical Design Profiles

Cluster 1 · 58.8%

Moderate Mainstream

VCBI: 0.109–0.247
Mean: **0.194**

Balanced validation + Steered Collaboration (A=3). Mixed architectures.

KASI · VizAbility · AltCanvas · MAIDR AI · ...

Cluster 2 · 23.5%

Constrained Strugglers

VCBI: 0.107–0.130
Mean: **0.119**

Low agency (A=2) despite adequate validation. Legacy pattern: automation over empowerment.

AltGen · ChartVi · EyeMath · chart pipeline

Cluster 3 · 11.8%

Critical Imbalance

VCBI: 0.000–0.075
Mean: **0.037**

Orthogonal failures: DNN has V but no A; TableNarrator has A but no V.

Multimodal DNN · TableNarrator

Cluster 4 · 5.9%

High Performer

VCBI: **0.287**
(corpus upper bound)

Exceptional validation (VR= 8.0) compensates moderate agency (A=3) and controlled complexity.

TableView+

Verification requires **both** V *and* A — a deficit in either dimension cannot be recovered by the other (Cluster 3; mathematical proof).

Metric	Classical CV ($n = 7$)	Transformer ($n = 8$)	Hybrid ($n = 2$)
VCBI (mean)	0.141	0.162[a]	**0.249**
Validation Rigor	5.81	5.05	**6.55**
User Agency	2.14	2.88	3.00
Risk Density	4.24	**7.64**[b]	3.15
Hallucination documented	1/7 (14%)	**7/8** (88%)	0/2

[a] VCBI gap vs Classical CV: $d = 0.31$ (small), $U = 24.0$, $p = 0.643$ (n.s.) [b] Risk gap: $d = 2.43$ (very large), $U = 3.0$, $p = 0.005$

Pattern 1 — Capability–Risk Coupling

Transformers: **80% higher risk** (d=2.43, very large) with **no** validation improvement (d=0.52). VCBI gap negligible and non-significant.

Pattern 2 — Agency Ceiling

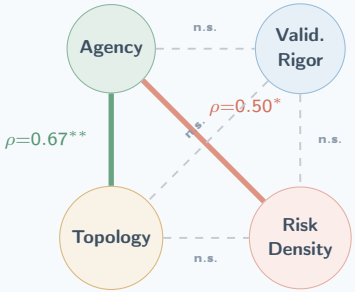
Agency improves (+34%) Classical CV→Transformer, but **no system** across any paradigm reaches Level 4. Confidence estimates withheld despite architecture supporting them.

Pattern 3 — Hybrid Proof Cases

Highest VCBI, zero hallucinations, lowest risk. But n=2: *proof cases*, not a validated paradigm victory. Architecture explains only R^2=0.178 of VCBI variance.

Independence of the Four VCBI Dimensions

Pairwise Spearman correlations (BH-corrected, N=17):



** p_{adj}=0.0018 * p_{adj}=0.0144 BH correction, m=6

What the graph reveals

Agency is a central hub: higher-agency systems tend toward more complex pipelines (**A–C**) and higher documented risks (**A–R**).

Validation Rigor is orthogonal to the other three dimensions — it does *not* co-vary with Agency, Risk, or Topology.

Design implication: Raising Validation Rigor is a methodological commitment, not an architectural constraint — any team, on any paradigm, can do it. Conversely, **agency raises risk exposure**: richer interaction needs stronger validation to compensate.

Caution: with N=17, $|\rho| < 0.48$ is undetectable at standard power. Non-significance is *absence of evidence*, not evidence of independence.

Pattern	Description	Full adoption
Provenance Transparency	Drill-down, sonified data access, query-based retrieval. Trace AI output back to its evidential basis.	5.9%
Uncertainty Signaling	Prosodic modulation, semantic hedging (*"likely"*), earcon tiers, explicit querying.	**0%**
Challenge Interactions	Open-ended querying, structured alternatives, undo/redo with explanations.	11.8%

No system exposed confidence estimates to users. 4 systems (23.5%) computed internal metrics but **withheld** them.

Uncertainty must be *passively perceivable* — requiring active querying per statement creates prohibitive interaction burden.

Agency quality > Agency quantity.

MAIDR AI and TableView+ both score A=3, but MAIDR AI's open-ended querying supports *epistemic* scrutiny, while TableView+'s customisation supports *task efficiency*. Same score — fundamentally different verification value.

From Accuracy to Verifiability

The Methodological Blind Spot

82.4% of the 17 systems report accuracy metrics.

0% measure user error detection rates.

We measure what AI gets right,
not what users can **verify**.

Agency has equal marginal impact on VCBI as validation rigor (≈8% per 10% increase).

Yet agency (mean 2.59/4.0) is the most underdeveloped dimension — and the most **actionable**: it is an *interaction design* problem, not a data collection problem.

Three Proposed Evaluation Criteria

1. **Error Detection Rate**
   Can users identify *injected* errors without external validation?
   Target: $> 70\%$ detection rate

2. **Scrutiny Friction**
   How many steps to access the evidential basis for an AI interpretation?
   Target: < 5 interaction steps

3. **Contestation Success**
   % of user challenges eliciting a meaningful AI response.
   Target: $> 50\%$

Future papers should report at least one of these alongside traditional accuracy metrics.

Limitations & Future Directions

Current Limitations

- **Sample size (N=17):** limits statistical power; clustering in 4D space is sensitive to outliers — treat profiles as *exploratory archetypes*.
- **Agency metric conflation:** task-level agency (sorting, filtering) vs. epistemic agency (contestation, drill-down) receive the same score.
- **Geographic bias:** 76.5% of systems from US or European institutions.
- **Domain scope:** visual information accessibility only — navigation and mobility excluded.
- **Practical barriers:** commercial LLM APIs restrict access to uncertainty estimates; non-visual signaling lacks shared conventions.

Future Directions

- **Corpus expansion** to $N \geq 50$ for reliable multivariate analysis and cluster validation.
- **Refined agency metric** separating epistemic control from interface control — better predicts verification capacity.
- **Verification-aware evaluation protocols:** deploy Error Detection Rate, Scrutiny Friction, Contestation Success in future user studies.
- **Causal studies** disentangling architectural effects from researcher methodological rigour (confounded in hybrid systems).
- **Adaptive interfaces** scaling verification scrutiny to task stakes + longitudinal trust calibration studies.
- **Cross-domain extension** to navigation, mobility, STEM.

Take-home Message

Gap: field achieves only marginal balance (mean VCBI=0.163; max=0.287 $\approx$ 29% of theoretical max).

Cause: transformers bring 80% higher risk with no VCBI improvement. Risk profile differentiates paradigms.

Path forward: hybrid specialisation is promising; improving *agency* may be the most feasible lever.

The Central Argument

The question for high-stakes accessibility is not:

"Can AI perform the task?"

but:

"Can users maintain epistemic agency over AI-mediated information?"

Ensuring appropriate trust *and* appropriate distrust is both an ethical and a technical imperative.

Verification affordances must be **design requirements**, assessed by metrics measuring users' ability to detect AI failures, not just successes.

ANR COOBRA (ANR-24-SARP-0006) · doi:10.1145/3800424.3800451