Evidence Before Eloquence

Designing auditable retrieval-augmented assistants for public services

Nadia Rahman Victor Chen

Civic Systems Laboratory
Illustrative Research Presentation

LetX Research Forum ▪ August 2026

From a plausible answer to a defensible decision

1. The trust gap
2. Assurance architecture
3. Evaluation
4. Deployment

Public-service answers carry consequences beyond the chat window

- Policies change across jurisdictions, programs, and effective dates.

Failure mode

Treating retrieval as background context allows unsupported claims to survive into the final answer.

Core question

Can every consequential sentence be traced to current, authorized evidence?

Public-service answers carry consequences beyond the chat window

- Policies change across jurisdictions, programs, and effective dates.
- A fluent answer can conceal missing evidence or an outdated source.

Failure mode

Treating retrieval as background context allows unsupported claims to survive into the final answer.

Core question

Can every consequential sentence be traced to current, authorized evidence?

Public-service answers carry consequences beyond the chat window

- Policies change across jurisdictions, programs, and effective dates.
- A fluent answer can conceal missing evidence or an outdated source.
- Staff need a short response; reviewers need the complete decision trail.

Failure mode

Treating retrieval as background context allows unsupported claims to survive into the final answer.

Core question

Can every consequential sentence be traced to current, authorized evidence?

Public-service answers carry consequences beyond the chat window

- Policies change across jurisdictions, programs, and effective dates.
- A fluent answer can conceal missing evidence or an outdated source.
- Staff need a short response; reviewers need the complete decision trail.
- The design target is therefore useful under time pressure and auditable afterward.

Failure mode

Treating retrieval as background context allows unsupported claims to survive into the final answer.

Core question

Can every consequential sentence be traced to current, authorized evidence?

Three requirements turn retrieval into an assurance layer

1. Bound the corpus

Index only approved sources, preserve effective dates, and attach ownership metadata.

2. Verify each claim

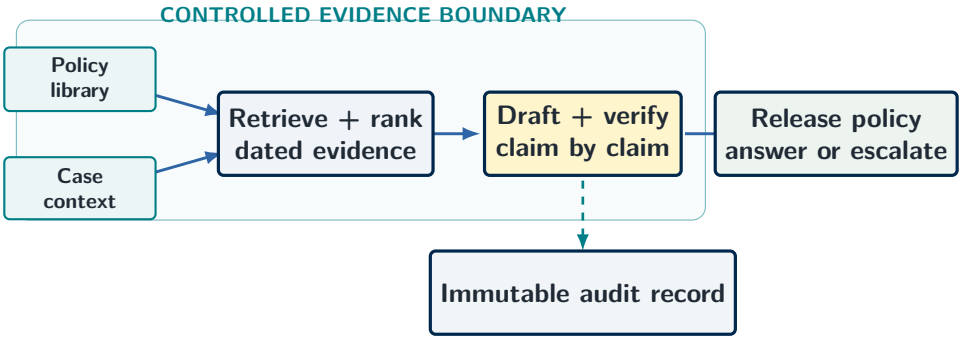
Separate candidate generation from evidence checking before any answer is released.

3. Preserve the trace

Store query, source version, passages, verdicts, and the final response together.

Retrieval finds evidence; verification decides whether that evidence is enough.

The release gate sits between generation and delivery



Unsupported claims never reach the delivery channel; uncertain cases follow a defined escalation path.

A bounded score makes the release policy explicit

For answer a with claims $c_1, \ldots, c_m$, define

$$Q(a) = \sum_{i=1}^{m} w_i \, s(c_i, e_i) - \lambda U(a) - \gamma A(a), \qquad \sum_{i=1}^{m} w_i = 1,$$

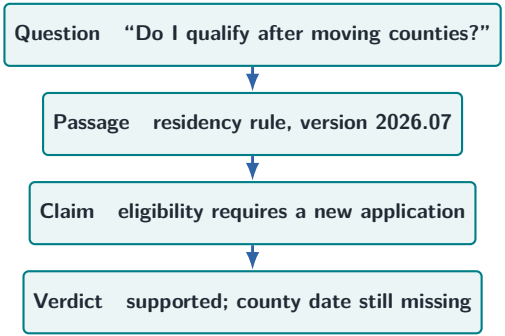
where

- $s(c_i, e_i) \in [0, 1]$: support from cited evidence,
- $U(a)$: unresolved ambiguity or missing context,
- $A(a)$: source-age penalty after a policy update,
- λ, γ: risk-calibrated penalties.

Release rule

Deliver only when $Q(a) \geq \tau$, every high-impact claim has direct support, and no source is outside its validity window; otherwise ask for context or escalate.

The audit trace explains both the answer and the refusal



What the user sees

A concise answer, a next action, and citations to the governing policy.

What blocks release

The move date determines which version applies, so the assistant asks one targeted question.

Evaluation separates answer quality from operational safety

- **1,200 scenarios**: routine, ambiguous, outdated, and adversarial queries.
- **Three systems**: generation only, retrieval augmented, and gated verification.
- **Double review**: policy specialists label support and acceptable escalation.
- **Update drill**: 10% of source documents change midway through the test.

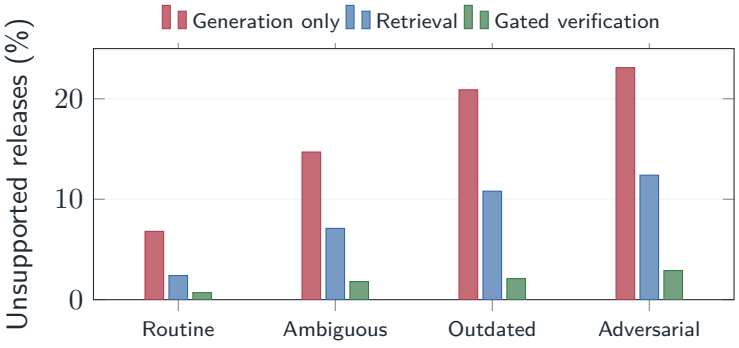
Primary measures

- claim support rate,
- unsafe-release rate,
- useful-answer rate,
- median latency.

Evaluation principle

A safe escalation counts as correct behavior, not as an unanswered query.

Verification sharply reduces unsupported releases



Result

The release gate cuts unsupported answers by **85%** relative to retrieval alone.

Trade-off

Median latency rises by 310 ms because claims are checked before release.

Safety improves without collapsing service usefulness

System	Supported claims	Unsafe release	Useful outcome	Latency p50
Generation only	84.2%	16.4%	91.8%	0.62 s
Retrieval augmented	93.7%	8.1%	**94.1%**	0.88 s
Gated verification	**98.6%**	**1.2%**	92.9%	1.19 s

Why usefulness holds

Targeted clarification recovers many ambiguous cases instead of refusing them outright.

Operational reading

The small usefulness change buys a sevenfold reduction in unsafe releases.

Values are illustrative and included to demonstrate the presentation template.

A verified answer can still fail the person receiving it

Coverage

Missing policy documents cannot be repaired by a stronger verifier; corpus owners need explicit coverage targets and update service levels.

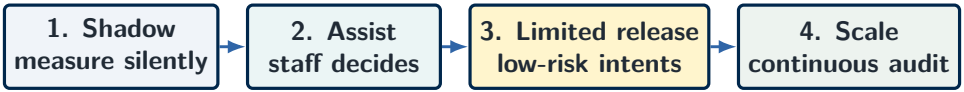
Interpretation

Evidence may support a sentence while the overall guidance remains misleading because of omitted exceptions or inaccessible language.

Accountability

Human review remains mandatory for appeals, conflicting authorities, and decisions that materially affect rights or benefits.

Deployment advances only when evidence supports the next step



Entry gate

Approved corpus, trace retention, red-team set, and named policy owners.

Promotion gate

Stable safety metrics across user groups, topics, and policy updates.

Stop condition

Roll back when unsafe release or stale-source rates breach their limits.

Evidence-first design changes what “good” looks like

1. **Constrain the evidence.** Authorized, versioned sources define the assistant’s knowledge boundary.
2. **Verify the claims.** A separate release gate checks support, ambiguity, and freshness.
3. **Keep the trace.** Every consequential output remains reproducible and reviewable.
4. **Escalate deliberately.** Uncertainty triggers clarification or human judgment, not improvisation.

Takeaway

The safest assistant is not the one that always answers; it is the one that knows what evidence permits it to say.

Questions?

Evidence before eloquence.

Contact research@civicsystems.example
Project example.org/civic-assurance