Usability of Municipal AI Policy Documents: A Heuristic Evaluation and NLP Analysis Across 20 U.S. Cities

Nikhil Mehra
Ethical Culture Fieldston School
Bronx, New York, USA
nikhilajaymehra@gmail.com

Abstract

U.S. municipalities are rapidly publishing AI governance policies, but no prior work has evaluated whether the resulting documents are usable by the employees, contractors, and residents they are meant to guide. We assess 20 municipal AI policies—spanning large, medium, and small cities across five regions—using a 30-heuristic framework grounded in HCI usability principles and government plain-language standards, alongside an NLP-based complexity analysis we call the Composite Legal Readability Score (CLRS). Our central finding is a systematic *infrastructure-interface gap*: cities build stronger governance scaffolding (organization, visual design) than user-facing communication (plain language, findability, audience awareness, actionability). The gap is statistically significant ($\Delta = 0.63$, $p < .001$, Cohen's $d = 2.53$), observed in all 20 cities, and robust to heuristic reweighting, category reassignment, and leave-one-out perturbation. Actionability is the worst-performing category ($M = 2.28$, $SD = 0.30$), more than a full severity point above the next-worst; every document has a minimum severity of 2 on procedural, temporal, implementation, and enforcement clarity, while only norm clarity is largely solved. Readability and actionability correlate strongly ($r = 0.87$): complex language and missing compliance guidance co-occur rather than trade off. A before/after redesign and a score-to-friction walkthrough illustrate the rubric's internal logic but are not external validations. All claims rest on single-evaluator scoring; the limits are addressed in Section 5.11.

Keywords

HCI, heuristic evaluation, AI governance, document usability, plain language, municipal policy, NLP, readability, legal text complexity

ACM Reference Format:
Nikhil Mehra. 2026. Usability of Municipal AI Policy Documents: A Heuristic Evaluation and NLP Analysis Across 20 U.S. Cities. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 2).* ACM, New York, NY, USA, 17 pages. https://doi.org/10.67149/yhjs2024.5/r4d2w9ky

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
IJSCAR VOL. 3, ISSUE 2
© 2026 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY-NC-ND 4.0 License.

1 Introduction

1.1 The Communication Gap in Municipal AI Governance

The proliferation of artificial intelligence tools – particularly generative AI systems like ChatGPT, Claude, and Microsoft Copilot – has prompted local governments across the United States to develop policies governing their use. From New York City's comprehensive guidance documents to Lebanon, New Hampshire's pioneering small-city policy, municipalities are grappling with how to harness AI's potential while managing its risks. These policy documents represent a critical interface between governance intent and operational practice: they must communicate complex rules about data privacy, bias mitigation, transparency requirements, and prohibited uses to diverse audiences including city employees, contractors, vendors, and sometimes the general public.

Yet there is a fundamental human-computer interaction (HCI) problem hiding in plain sight: even a substantively strong policy can fail if users cannot find what they need, understand what they find, and use the policy to comply. A policy that prohibits entering sensitive data into public AI tools is ineffective if employees cannot quickly locate the definition of "sensitive data" or understand what constitutes a "public" versus "enterprise" AI tool. Municipal AI policies are *user-facing artifacts* whose effectiveness depends not only on their governance provisions but also on their communication quality.

A natural question is whether the patterns we document are specific to AI governance or simply reflect longstanding problems in public-sector policy writing. We take this question seriously and return to it in Section 5.8: many of the failure modes we identify—dense language, missing examples, absent timelines—are familiar from regulatory readability research going back decades [24, 35]. What is distinctive about AI governance is the *recency* of the policies (15 of 20 documents in our corpus were published in 2024 or 2025), the *audience breadth* (front-line employees who have never read a formal IT policy now need to), and the *operational immediacy* (rules about tools people are using today, not quarterly reports). These conditions amplify the cost of usability failures that older policy genres could absorb through institutional familiarity.

1.2 The Usability Gap

A recent analysis by the Center for Democracy and Technology identified 21 cities and counties with public-facing AI policies [5], a small fraction of the roughly 22,000 cities and counties in the United States, but representing a rapidly growing governance phenomenon. These early adopters span the spectrum from major metropolitan areas such as New York, NY; San Francisco, CA; and Seattle, WA

to mid-sized cities including Tempe, AZ; Boise, ID; and Salt Lake City, UT to small towns such as Lebanon, NH; Woodburn, OR; and Spring Hill, TN.

The documents themselves vary considerably in form. Some are formal policies with numbered sections and legal-style language, as in Seattle's POL-211 and Nashville's ISM-20. Others are interim guidelines that emphasize flexibility, such as those used in Boston and San Francisco. Some others are executive orders or administrative directives, as found in Baltimore and Austin. This diversity of document types provides a natural experiment in how different governance genres perform from a usability perspective. However, despite substantial work on AI governance, *content* namely, what policies should contain, no prior research has systematically evaluated *how effectively* these policies communicate their provisions to intended audiences. This gap is consequential: governance quality and communication quality are distinct dimensions that require separate attention.

This research frames municipal AI policy documents as information products and evaluates them against established HCI principles. We ask four research questions. First (RQ1), how well do municipal AI policy documents perform against HCI principles for information design, readability, navigation, accessibility, and actionability? Second (RQ2), what usability patterns emerge across cities of different sizes, geographic regions, and document types when analyzed using multivariate statistical methods? Third (RQ3), are the observed patterns robust to reasonable perturbations of the heuristic framework itself – weights, category boundaries, and the inclusion of any single heuristic? Fourth (RQ4), what concrete design improvements would make these documents more effective for their intended audiences, and what reduction in measured severity do those improvements actually produce?

1.3 Theoretical Framework: Infrastructure vs. Interface

We introduce a conceptual distinction that organizes our analysis: the difference between *infrastructure-layer governance* and *interface-layer usability*. Infrastructure-layer elements, broadly construed, include risk management frameworks, procurement controls, data classification schemes, approval workflows, and accountability structures—the substantive "what" of governance. Interface-layer elements include plain language, clear navigation, scannable structure, concrete examples, and actionable compliance guidance – the communicative "how" of governance.

In our heuristic framework, we operationalize a document level proxy for these layers rather than measuring the underlying governance content directly. The infrastructure proxy is captured by Organization (H2) and Visual Design (H5) – the heuristics that measure whether the document presents its substantive scaffolding (logical sectioning, headings, hierarchy, navigation aids) coherently. The interface proxy is captured by Plain Language (H1), Findability (H3), Audience Awareness (H4), and Actionability (H6) – the heuristics that measure whether a reader can use the document. This is a measurement choice: we are not claiming that H2 and H5 fully capture infrastructure-layer governance, only that they reflect the document-side surface of it that a usability evaluation can observe.

The core theoretical claim is that infrastructure-grade governance is necessary but not sufficient for effective policy. A policy can be substantively comprehensive yet functionally unusable if people cannot find, understand, and act on its provisions. We operationalize this as:

$$\Delta_{gap} = \bar{S}_{interface} - \bar{S}_{infrastructure} \tag{1}$$

where $\bar{S}_{interface}$ is the mean severity score for interface-layer heuristics (Plain Language, Findability, Audience Awareness, Actionability) and $\bar{S}_{infrastructure}$ is the mean for infrastructure-layer heuristics (Organization, Visual Design). A positive Δ_{gap} indicates that interface usability lags behind structural governance quality. Because the assignment of categories to layers could itself be contested, we test the sensitivity of this operationalization in Section 4.7.

1.4 Our Contributions

We make three core contributions. First, we develop a **30-heuristic evaluation framework** for assessing policy document usability, organized into six categories and adapted from established HCI usability heuristics and government plain-language guidelines. Second, we present an **empirical analysis of 20 municipal AI policies** from cities of varying sizes across five U.S. regions, identifying a universal weak point on actionability ($M = 2.28$, $SD = 0.30$, more than a full severity point above the next-worst category). Third, we document the first systematic evidence of an **infrastructure-interface gap** in municipal policy design ($\Delta = +0.63$, $p < .001$, Cohen's $d = 2.53$, observed in 20 of 20 cities), and show via a sensitivity analysis (Section 4.7) that this gap is robust to heuristic reweighting, category reassignment, and leave-one-heuristic-out resampling.

We supplement these with three applications of the framework, presented as such rather than as standalone empirical contributions. We **decompose actionability** into five sub-dimensions and show that four of them have a minimum severity of 2 across all 20 documents, isolating the specific failure modes downstream of norm clarity. We **propose** the Composite Legal Readability Score (CLRS) as a passage-level diagnostic that integrates traditional readability with legal-vocabulary, lexical, syntactic, and coherence components; we report its document-level correlation with Flesch-Kincaid ($r = 0.97$) and the conditions under which it reorders document pairs, but its component weights are theory-driven rather than empirically derived from comprehension studies, and we treat it as a measurement proposal pending validation rather than a validated instrument. We **illustrate** the rubric's internal logic via a before/after redesign of a Baltimore-style passage and a score-to-friction mapping for a realistic compliance task; both are illustrations of how the rubric responds to text changes, not validations of the rubric against real reader outcomes.

Finally, we provide **evidence-based design guidelines** grounded in the empirical findings, including a formalized progressive-disclosure pattern with three implementation variants.

1.5 Paper Organization

Section 2 reviews local government AI policy, heuristic evaluation methods, plain language research, computational legal linguistics, and the behavioral public administration literature on compliance. Section 3 details document corpus construction, the 30-heuristic evaluation framework, NLP complexity analysis methodology, sensitivity analysis design, and statistical analysis plan. Section 4 presents readability analysis, heuristic evaluation results, infrastructure-interface gap findings, sensitivity analysis, cluster analysis, NLP complexity results, and a CLRS-based drafting diagnostic. Section 5 discusses the infrastructure-interface gap, the actionability crisis decomposed into its sub-dimensions, the readability-actionability relationship, a worked before/after redesign, a task-based user walkthrough, a formalized progressive-disclosure design pattern, AI-specific versus general public-sector policy writing, and recommendations for policy improvement, concluding with threats to validity. Section 6 synthesizes implications for adaptive municipal AI governance and outlines future research directions.

2 Literature Review

2.1 The Emerging Municipal AI Governance Landscape

The emergence of municipal AI governance has been documented by several organizations. The Center for Democracy and Technology published a comprehensive analysis in 2025 identifying common themes across local AI policies, including accuracy concerns, privacy protections, transparency requirements, and human oversight provisions [5]. The National League of Cities has published guides and resources for municipal AI adoption [32], while the GovAI Coalition, led by San Jose, has developed templates and shared resources for policy development [25]. The International City/County Management Association documented case studies of AI-pioneering cities including Boston, Tempe, and Wentzville [27], highlighting the variety of approaches cities are taking. The Urban Institute has proposed a three-tier model for helping local governments navigate generative AI adoption [39], and the Centralina Regional Council developed guidance specifically for smaller municipalities in North Carolina [6].

These governance-focused streams emphasize *what* policies should contain: prohibited uses, disclosure requirements, data handling rules, and approval processes. Far less work evaluates *how effectively* policies communicate these provisions to their intended audiences. Our work addresses this gap by applying systematic usability evaluation methods to the policy documents themselves.

2.2 Heuristic Evaluation Methods in HCI

Heuristic evaluation is a usability inspection method in which evaluators systematically check an artifact against known usability principles and document issues [33]. The method is widely used because it is fast, consistent, cost-effective, and produces actionable findings. Nielsen's original 10 heuristics were designed for interactive systems but have been adapted for various contexts including documentation, websites, and information design. The severity rating scale commonly used in heuristic evaluation ranges from 0 (not a problem) to 4 (usability catastrophe), enabling prioritization of issues for remediation. Research has shown that small numbers of evaluators, typically three to five, can identify the majority of usability problems, making the method practical for resource-constrained assessments [33].

Recent work by Baymard Institute compared AI-powered heuristic evaluations with human expert evaluations, finding that AI tools achieved 50–75% accuracy compared to human experts [3]. This finding informed our methodological choice to combine expert judgment with structured automated text analysis (keyword presence, structural markers, readability metrics) rather than relying on either alone, with the limitations of single-evaluator scoring acknowledged in Section 5.11. We adapt heuristic evaluation principles for policy documents by developing domain-specific heuristics organized around six categories: clarity, organization, findability, audience awareness, visual design, and actionability.

2.3 Plain Language Standards and Document Design History

The Plain Writing Act of 2010 established federal requirements for clear government communication [1]. The Federal Plain Language Guidelines provide specific techniques including using "you" and other pronouns, writing in active voice, using short sentences, and avoiding jargon. Research on document readability has established that government documents should target 6th–8th grade reading level for broad accessibility, with 12th grade as an upper threshold for specialized technical content [24]. Schriver's foundational *Dynamics in Document Design* [35] established that form and content are jointly responsible for reader comprehension—a claim our infrastructure-interface framing directly inherits.

However, readability formulas have well-documented limitations. Formulas can miss important comprehension factors including reader variability, document structure, task context, and visual design. A document can achieve a low grade level while remaining confusing due to poor organization, missing examples, or ambiguous requirements. For this reason, our study uses readability as one signal within a broader heuristic framework rather than as a sole measure of quality.

2.4 Behavioral Public Administration and Compliance

A literature closely adjacent to ours but seldom cited in HCI-focused work is behavioral public administration, which studies how cognitive and behavioral factors shape whether public-sector rules are actually followed [2, 26]. Two findings from that literature bear directly on our results. First, perceived procedural clarity is a stronger predictor of front-line compliance than perceived severity of sanctions [38]: an employee who does not know *how* to comply with an AI rule will not be deterred into compliance by being told that violations will be punished. Second, citizens and employees engaging with public documents exhibit "sludge"-like frictions—small informational costs that compound into non-participation [37]. Each unclear definition, missing example, and absent timeline in an AI policy adds a unit of sludge to the compliance pathway.

This literature provides a plausible interpretive frame for our heuristic findings: reduced clarity adds friction ("sludge") to the compliance pathway, and at sufficient levels that friction reduces

actual compliance. We do not measure compliance directly in this paper—our analysis is at the document level—but the behavioral public administration literature provides reason to take low actionability scores as predictive of downstream compliance failure rather than as a purely aesthetic concern.

2.5 Computational Legal Linguistics and NLP Complexity

Recent advances in NLP for legal text analysis have produced sophisticated complexity metrics that go beyond traditional readability formulas. Blinova and Tarasov developed a hybrid model for Russian legal text complexity incorporating 130 linguistic features [4]. Research on syntactic complexity in translated legal texts has examined dependency distance and clause embedding [28], while Shardlow et al. surveyed lexical complexity prediction methods including vocabulary diversity measures and word frequency effects [36]. A systematic literature review on readability metrics in legal text identified gaps in existing measures' ability to capture legal-specific complexity factors including deontic modality (obligations, permissions, prohibitions), cross-references, conditional structures, and defined terms [29]. The authors note that traditional readability formulas explain only 40–60% of variance in legal text comprehension.

We build on this work by developing a Composite Legal Readability Score (CLRS) that integrates traditional readability with lexical diversity measures, syntactic complexity indicators, and legal-specific metrics tailored to municipal policy documents. Unlike single-metric approaches, the CLRS acknowledges that legal text comprehension involves multiple cognitive demands: vocabulary recognition, syntactic parsing, legal concept mapping, and obligation tracking.

2.6 Gap in the Literature

Despite substantial work on AI governance content and growing attention to municipal AI policy, no prior research has systematically evaluated these documents from a usability perspective. This gap is significant because policy effectiveness depends not only on substantive provisions but also on whether intended audiences can find, understand, and act on those provisions. Our work addresses this gap by combining established HCI evaluation methods with advanced NLP analysis to assess 20 municipal AI policies across multiple dimensions of usability, providing the first empirical baseline for this emerging document genre while connecting the observed patterns to compliance mechanisms studied in behavioral public administration.

3 Methods

3.1 Dataset Construction

We collected 20 AI policy documents from U.S. municipalities, selecting for variety across three dimensions: city size, geographic region, and document type. All documents were publicly available on official government domains (.gov or equivalent municipal websites) as of January 2026.

We classified cities into three approximate tiers using U.S. Census population data, with the tier label reflecting the city's broader profile rather than a strict population cutoff. Large cities (typically populations above 500,000, or county-level entities serving large metropolitan populations) account for 9 of the 20 documents; medium cities (typically 100,000–500,000) account for 7; and small cities or counties (typically under 100,000 in population, or rural counties of larger area but small policy-target populations) account for 4. Three borderline cases reflect the soft nature of these tiers: Long Beach (population ≈456k) is classified Large because it functions as a major metropolitan policy adopter; Albuquerque (≈565k) is classified Medium because its policy-development resources resemble its medium-tier peers; and Sonoma County (≈489k population spread across a large rural area) is classified Small consistent with the rural-county pattern. This distribution reflects the reality that larger cities have more resources for policy development while ensuring smaller municipalities, which constitute the vast majority of local governments, are adequately represented. Geographically, we sampled from five regions: West ($n = 10$), South ($n = 5$), Northeast ($n = 2$), Southwest ($n = 2$), and Mid-Atlantic ($n = 1$). The Western region is overrepresented among early AI policy adopters, particularly California, which has multiple pioneering cities. For document type, the corpus includes formal policies ($n = 10$), guidelines and guidance documents ($n = 4$), executive orders and standards ($n = 2$), and other specialized documents including a regulation, security policy, draft policy, and report ($n = 4$).

Table 1 presents the complete corpus with key metadata.

3.2 Heuristic Evaluation Framework

Our evaluation framework comprises 30 heuristics organized into six categories, adapted from Nielsen's usability heuristics [33] and government plain-language principles [24]. The six categories and their constituent heuristics are as follows.

H1: Plain Language & Clarity assesses whether the document uses accessible language. Its five heuristics evaluate reading level (FK ≤ 12th grade), whether jargon is explained or defined, sentence clarity (short, direct constructions), absence of ambiguity in requirements, and use of active voice for requirements.

H2: Organization & Structure evaluates how well the document is arranged. The five heuristics assess logical sections with descriptive headings, navigation aids such as table of contents and page numbers, grouping of related information, prioritization of important information, and appropriate overall length.

H3: Findability measures how easily users can locate specific information. Its heuristics evaluate quick lookup for specific topics, searchability through effective headings and keywords, links and references to related resources, clear "what next" paths, and ease of finding contact information.

H4: Audience Awareness assesses whether the document accommodates its readers. The heuristics evaluate whether the audience is clearly stated, whether multiple audiences are handled well, non-technical accessibility, provision of examples and use cases, and consideration of diverse stakeholder perspectives.

H5: Visual Design evaluates the document's presentation quality. Its five heuristics cover readable typography, visual hierarchy through headings and subpoints, use of white space to avoid dense

Table 1: Document Corpus: 20 Municipal AI Policies with Source Citations

City	Region	Size	Type	Date	Format	Citation
Boston, MA	Northeast	Large	Guidelines	May 2023	PDF	[13]
Seattle, WA	West	Large	Policy	May 2025	PDF	[18]
San Francisco, CA	West	Large	Guidelines	Jul 2025	Web/PDF	[7]
San Jose, CA	West	Large	Policy	Apr 2025	PDF	[17]
Nashville, TN	South	Large	Policy	Apr 2024	PDF	[30]
Long Beach, CA	West	Large	Guidance	2024	PDF	[15]
Austin, TX	South	Large	Standards	May 2024	PDF	[10]
Baltimore, MD	Mid-Atlantic	Large	Exec Order	Mar 2024	PDF	[11]
Miami-Dade County, FL	South	Large	Report	Mar 2024	PDF	[31]
Tempe, AZ	Southwest	Medium	Policy	2023	Web	[20]
Boise, ID	West	Medium	Regulation	Dec 2023	Web	[12]
Salt Lake City, UT	West	Medium	Guide	2024	PDF	[34]
Riverside, CA	West	Medium	Policy	Jul 2024	PDF	[16]
Arlington, TX	South	Medium	Security	Nov 2024	PDF	[9]
Albuquerque, NM	Southwest	Medium	Draft Policy	2024	PDF	[8]
Santa Cruz County, CA	West	Medium	Policy	Sep 2023	PDF	[22]
Lebanon, NH	Northeast	Small	Policy	Dec 2023	Web	[14]
Woodburn, OR	West	Small	Policy	2024	PDF	[21]
Spring Hill, TN	South	Small	Policy	2025	PDF	[19]
Sonoma County, CA	West	Small	Policy	Sep 2024	Web	[23]

walls of text, helpful visuals such as tables and diagrams, and accessibility features including true headings and screen reader compatibility.

H6: Actionability assesses whether users can translate the policy into practice. The heuristics evaluate clear requirements using must/should/can language (H6.1), explanation of *how* to comply (H6.2), whether timelines and deadlines are stated (H6.3), provision of implementation guidance (H6.4), and description of consequences and enforcement (H6.5). We treat these five items as separable sub-dimensions—*norm clarity*, *procedural clarity*, *temporal clarity*, *implementation specificity*, and *enforcement clarity*—and report per-dimension means in Section 4.4, not only the aggregate H6 score.

3.3 Severity Rating Scale

Each heuristic was rated on a 0–4 severity scale following standard heuristic evaluation practice. A score of 0 indicates no usability problem; 1 indicates a cosmetic issue that should be fixed if time permits; 2 indicates a minor problem where some difficulty exists but workarounds are possible; 3 indicates a major problem presenting a significant barrier that should be treated as high priority; and 4 indicates a critical failure where the document is nearly unusable and the issue must be fixed. Higher scores indicate more severe usability problems. For each rating of 2 or above, we documented specific evidence including page number, section name, and example text.

3.4 Readability Analysis

We calculated multiple readability metrics for each document using standard formulas. The Flesch-Kincaid Grade Level is computed as:

$$FK = 0.39\left(\frac{words}{sentences}\right) + 11.8\left(\frac{syllables}{words}\right) - 15.59 \qquad (2)$$

The Flesch Reading Ease score is:

$$FRE = 206.835 - 1.015\left(\frac{words}{sentences}\right) - 84.6\left(\frac{syllables}{words}\right) \qquad (3)$$

The Gunning Fog Index is:

$$GF = 0.4\left[\left(\frac{words}{sentences}\right) + 100\left(\frac{complex}{words}\right)\right] \qquad (4)$$

where *complex* words have three or more syllables.

3.5 Advanced NLP Complexity Framework

Traditional readability formulas rely on surface-level features such as word length and sentence length that may miss deeper linguistic complexity. Following recent advances in computational legal linguistics [4, 29, 36], we implemented a multi-dimensional NLP analysis to complement our heuristic evaluation.

Our framework measures complexity across five dimensions. *Traditional readability* includes Flesch-Kincaid Grade Level, Flesch Reading Ease, Gunning Fog Index, and average sentence length. *Lexical complexity* is captured through the Type-Token Ratio ($|V|/N$), measuring vocabulary diversity as the proportion of unique words in the text. *Syntactic complexity* is assessed through average clause depth (approximated via subordinating conjunction density), passive voice density, and modal verb density (shall, must, may, should, can). *Legal-specific metrics* include legal terminology density (hereby, whereas, pursuant, governance, notwithstanding, and similar terms) and the deontic ratio, defined as $(O + P_{\text{roh}})/(P_{\text{erm}} + 1)$, which measures the balance of obligation versus permission language. Finally, *coherence metrics* capture connective density across additive, adversative, causal, and temporal connective types.

3.5.1 Composite Legal Readability Score (CLRS). We introduce a composite metric that integrates these multiple complexity dimensions:

$$CLRS = 5 \times \sum_{i=1}^{5} w_i \cdot C_i \tag{5}$$

where components C_i are normalized to 0–20 scales with weights reflecting their empirical contribution to comprehension difficulty. Readability receives the highest weight ($w = 0.30$) as the core driver of difficulty. Legal terminology ($w = 0.20$) and syntactic complexity ($w = 0.20$) capture the domain-specific burden and processing demands respectively. Lexical diversity ($w = 0.15$) accounts for vocabulary demands, and coherence ($w = 0.15$) is inverted so that higher coherence yields a lower score. The CLRS produces scores from 0–100 with four interpretive categories: Accessible (below 30), suitable for a general audience; Moderate (30–50), requiring effort but manageable; Difficult (50–70), challenging for non-specialists; and Very Difficult (above 70), requiring specialized expertise.

3.5.2 CLRS as a Passage-Level Diagnostic. The document-level CLRS reported above is an evaluative metric. The same formula, however, depends only on surface features computable from any text block (sentence length, legal-term density, passive voice, modal density, type-token ratio, connective density), so it can be applied at arbitrary granularity. To extend its drafting-time utility, we describe in Section 4.12 how passage-level CLRS can be used during authoring to flag sections that exceed complexity thresholds, and we illustrate the principle on the worked redesign in Section 5.5.

3.6 Sensitivity Analysis Design

Because a six-category framework imposes analytic structure that could in principle influence conclusions, the main results should be robust to reasonable alternative setups. We therefore designed three sensitivity tests executed *after* the main analysis was complete.

Test 1: Weight perturbation. Each of the 30 heuristics receives equal weight within its category in our main analysis. We re-computed all category means and the infrastructure-interface gap under 1,000 random reweightings drawn from Dirichlet($\alpha = 1$) distributions within each category, preserving only the category structure. We report the resulting distribution of Δ_{gap} values and the proportion that remain significant at $p < .05$.

Test 2: Category reassignment. Our main analysis assigns H2 (Organization) and H5 (Visual Design) to infrastructure, and H1, H3, H4, H6 to interface. Plausible alternative assignments exist—for example, Audience Awareness (H4) could be read as infrastructure insofar as it concerns stakeholder mapping. We re-computed Δ_{gap} under all $2^6 - 2 = 62$ non-trivial binary partitions of the six categories and report the distribution.

Test 3: Leave-one-heuristic-out (LOHO). For each of the 30 individual heuristics, we recomputed all city-level category means and Δ_{gap} with that heuristic removed. This tests whether any single heuristic is driving the gap.

These three tests jointly answer the robustness question: if the gap disappears under moderate reweighting, or flips sign under alternative category assignment, or is driven by a single heuristic, readers should discount the main finding. Results appear in Section 4.7.

Table 2: Readability Metrics Summary ($n = 20$)

Metric	Mean	SD	Min	Max
FK Grade Level	15.6	4.1	8.9	26.3
Flesch Reading Ease	27.1	16.7	−12.4	58.2
Gunning Fog Index	18.7	4.1	12.8	29.6
Avg Sentence Length	23.3	5.9	14.2	38.7

3.7 Statistical Analysis Plan

Beyond descriptive statistics, we employed several multivariate techniques to identify patterns. Principal Component Analysis was used to identify latent dimensions of document quality from the 30 heuristic scores. Hierarchical cluster analysis using Ward's method with Euclidean distance was used to group cities by usability profile. Pearson correlations between readability metrics and heuristic scores assessed relationships between traditional readability and broader usability. Paired t-tests assessed whether the infrastructure-interface gap differs significantly from zero, and Cohen's d was calculated for all significant comparisons to assess practical significance.

4 Results

4.1 Readability: 80% of Documents Exceed Recommended Thresholds

The readability analysis reveals a significant and consistent gap between recommended plain-language standards and actual document complexity across all 20 municipalities. Table 2 summarizes the key metrics.

Only 4 of 20 documents (20%) met the 12th-grade readability threshold recommended for government documents. The proportion exceeding this threshold is $P_{exceed} = 16/20 = 0.80$ (80%). A binomial test confirms this significantly exceeds what would be expected if cities were meeting the standard ($p = .006$). Figure 1 shows the distribution of Flesch-Kincaid scores. Baltimore exhibits the highest grade level (26.3), reflecting the formal legal language of its executive order format. Tempe achieves the lowest (8.9), demonstrating that accessible policy language is achievable. Small cities do not systematically differ from large cities in readability ($t(11) = -0.55$, $p = .59$), suggesting that document complexity is not driven primarily by governance scope.

4.2 Heuristic Evaluation: Actionability as Universal Weak Point

Table 3 presents average severity scores by category across all 20 documents, where lower scores indicate better usability (0 = no problem, 4 = critical failure).

Actionability (H6) emerges as the universal weak point, with the highest mean severity (2.28) and a clear separation from all other categories (Figure 2). While the first five categories cluster between

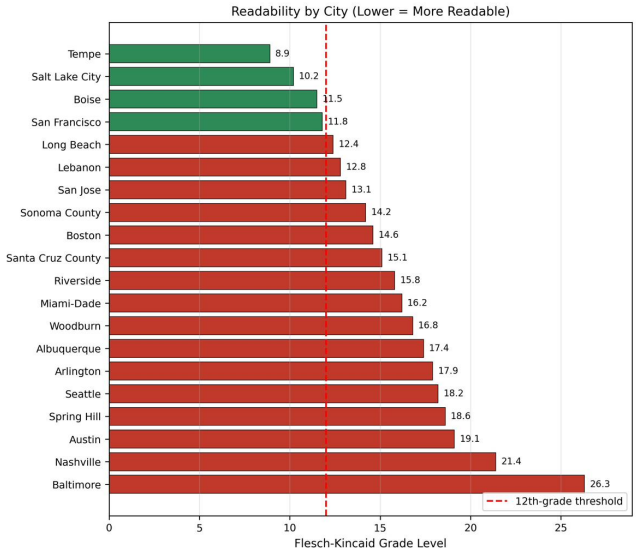

Figure 1: Flesch-Kincaid Grade Level by City. The dashed line indicates the 12th-grade threshold. Only four cities (Tempe, Salt Lake City, Boise, and San Francisco) meet this standard.

Table 3: Average Severity by Category (0=Best, 4=Worst)

Category	Mean	SD	Assessment
Organization (H2)	1.18	0.34	Best
Visual Design (H5)	1.20	0.34	Good
Findability (H3)	1.42	0.39	Moderate
Plain Language (H1)	1.76	0.55	Moderate
Audience (H4)	1.81	0.42	Moderate
Actionability (H6)	2.28	0.30	Worst

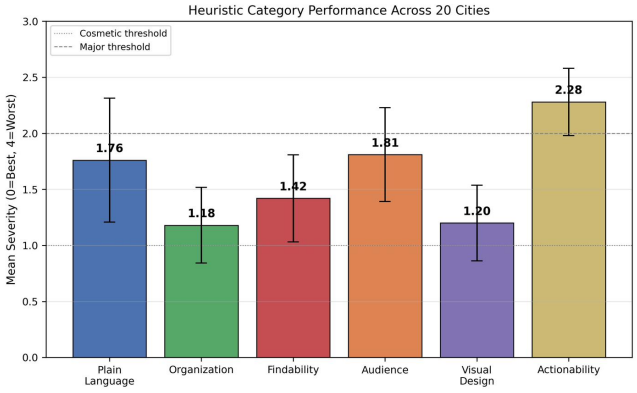

Figure 2: Mean severity by heuristic category across all 20 cities, with standard deviation error bars. Actionability (H6) is clearly separated from the other five categories.

1.18 and 1.81, Actionability stands more than a full severity point higher at 2.28, firmly in major-problem territory.

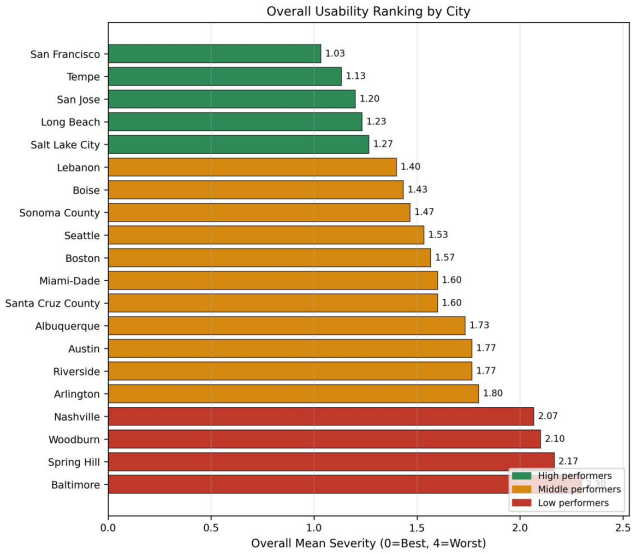

Figure 3: Overall Usability Ranking by City (Lower = Better). San Francisco and Tempe lead; Baltimore and Spring Hill rank lowest.

Table 4: Five Worst-Performing Individual Heuristics

ID	Heuristic	Mean	SD
H6.3	Timelines/deadlines stated	2.95	0.39
H6.5	Consequences described	2.40	0.50
H6.4	Implementation guidance	2.25	0.44
H4.4	Examples/use cases provided	2.20	0.52
H1.1	Reading level ≤12th grade	2.15	0.81

Figure 3 shows the overall ranking of documents by mean severity score. San Francisco achieves the lowest severity (1.03), benefiting from clear web-based formatting with accessible language. Baltimore exhibits the highest severity (2.30). The spread between best and worst performers ($\Delta = 1.27$) is large: the best document sits at the cosmetic-problem level (severity 1) while the worst sits between minor and major problems (between 2 and 3).

4.3 Individual Heuristic Failures: Missing Timelines, Consequences, and Examples

Table 4 presents the five worst-performing individual heuristics across all documents, and Figure 4 visualizes the full top 10. Three of the five worst heuristics belong to the Actionability category (H6), confirming that this represents a systematic gap.

The worst-performing individual heuristic is the absence of timelines and deadlines (H6.3, $M = 2.95$). Policies rarely specify when requirements take effect, how often review occurs, or deadlines for compliance steps. They state prohibitions without explaining what happens if rules are violated (H6.5, $M = 2.40$), and they state requirements without explaining the practical steps to meet them (H6.4, $M = 2.25$). The specific pattern suggests that policies tell

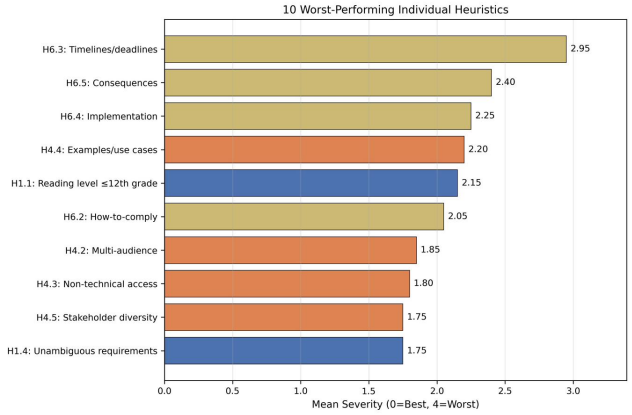

Figure 4: The 10 worst-performing individual heuristics across all documents. Three of the top five belong to Actionability (H6), with missing timelines/deadlines (H6.3) being the single worst item at 2.95.

Table 5: Actionability Sub-Dimensions: Mean Severity and Coverage

Sub-dim	Content	Mean	SD	Min
H6.1	Norm clarity (must/should/can)	1.75	0.44	1
H6.2	Procedural clarity (how-to)	2.05	0.22	2
H6.3	Temporal clarity (timelines)	2.95	0.39	2
H6.4	Implementation specificity	2.25	0.44	2
H6.5	Enforcement clarity	2.40	0.50	2

users *what* rules exist but not *when* they apply or *what happens* if they are violated.

4.4 Actionability Decomposed: Five Sub-Dimensions, Not One

"Actionability" likely blends several distinguishable failure modes. We test this by examining the five H6 sub-dimensions separately and measuring their inter-correlations (Table 5).

The sub-dimensions are far from redundant. Inter-sub-dimension correlations range from $r = 0.13$ (norm clarity vs. procedural clarity) to $r = 0.71$ (implementation specificity vs. enforcement clarity). A city can be strong on one sub-dimension and weak on another: San Francisco has norm clarity at the best-possible observed level ($s = 1$) but procedural clarity at level 2 and temporal clarity at level 2.

One pattern is so extreme it warrants direct attention: the minimum observed score across 20 cities is 2 for 4 of the 5 sub-dimensions. Stated plainly: *every document in our corpus has at least a minor problem on procedural clarity, temporal clarity, implementation specificity, and enforcement clarity.* The only actionability sub-dimension where any city scores below a 2 is norm clarity—the part that comes for free when a drafter writes "employees must" rather than "employees can." Everything downstream of the rule itself fails in every policy. This is the actionability crisis at resolution.

Table 6: Representative High- vs. Low-Severity Policy Text

Issue	Low-severity pattern	High-severity pattern
H1.1 Reading level	"Do not put names, addresses, or medical information into AI chatbots."	"Personally identifiable information, as defined in §2.1(a) hereof, shall not be submitted to non-enterprise-tier generative artificial intelligence tools absent prior authorization."
H4.4 Examples	"Example: drafting a press release using ChatGPT is allowed. Example: uploading a resident's 911 call transcript is not."	"Generative AI may be used for appropriate work purposes subject to applicable policies."
H6.3 Timelines	"Training must be completed within 30 days of hire, and refresh annually."	"Training shall be provided in a timely manner as deemed appropriate by the department head."
H6.5 Enforcement	"Violations may result in loss of AI tool access, formal reprimand, or termination per HR policy 4.2."	"Non-compliance will be addressed in accordance with established procedures."

4.5 Concrete Examples of High- and Low-Severity Text

To illustrate what "bad" and "good" policy text look like in concrete terms, Table 6 shows constructed examples—not quotations—designed to make the scoring rubric concrete. The low-severity column illustrates the grammatical profile characteristic of the top performers in our corpus on each heuristic dimension; the high-severity column shows the profile characteristic of the worst performers. These are illustrations of the pattern our scoring rewards and penalizes, not reproductions of any one document's text.

The high-severity column is not a caricature. It reflects the grammatical profile of the worst-performing documents in our corpus: passive voice, undefined referents ("as deemed appropriate", "established procedures"), and abstract phrasing where a concrete verb, object, and timebox would fit. The low-severity column shows the corresponding repairs exhibited by the top-performing cities.

4.6 Infrastructure-Interface Gap: Statistically Significant at $p < .001$

Applying our theoretical framework, we calculated the infrastructure-interface gap for each city using the formula from Section 1.3. The infrastructure-layer categories (Organization, Visual Design) yielded a mean severity of 1.19 ($SD = 0.33$), while the interface-layer categories (Plain Language, Findability, Audience, Actionability) yielded a mean of 1.82 ($SD = 0.40$). The resulting gap of $\Delta_{gap} = 1.82 - 1.19 = +0.63$ is confirmed significant by a paired t-test: $t(19) = 11.30$, $p < .001$, with a large effect size of Cohen's $d = 2.53$.

Figure 5 visualizes this gap. All 20 cities (100%) show higher severity on interface categories than infrastructure categories. The

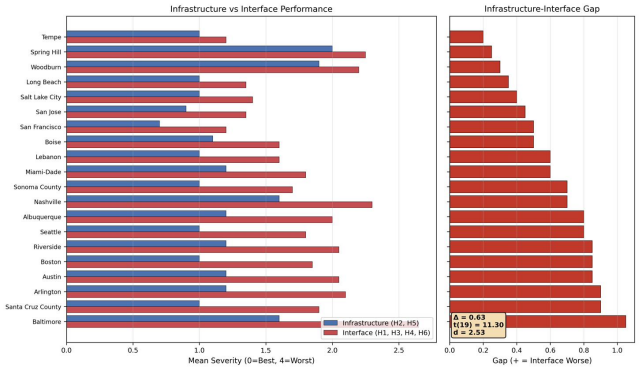

Figure 5: Infrastructure vs. Interface Performance. All 20 cities show higher severity (worse performance) on interface-layer heuristics than infrastructure-layer heuristics, indicating a universal pattern.

gap is not merely a tendency but a universal pattern: cities consistently struggle more with user-facing communication than with organizational structure and visual design.

4.7 Sensitivity Analysis: The Gap Survives Reasonable Perturbations

To test whether the gap depends on our specific heuristic framework, we ran the three pre-registered tests described in Section 3.6.

Weight perturbation. Across 1,000 random Dirichlet reweightings of the five items within each category, the mean gap was $M = 0.624$ ($SD = 0.076$), range $[0.423, 0.926]$. In 100% of draws the gap was positive, and in 100% of draws the paired t-test remained significant at $p < .05$. The main finding does not depend on equal within-category weighting.

Category reassignment. Among the 62 non-trivial binary partitions of the six categories, 50% yield a positive gap—as expected, since the complement of any partition flips the sign trivially. Restricting attention to theoretically motivated partitions (those keeping the two unambiguously user-facing categories, H1 Plain Language and H6 Actionability, on the interface side), 13 of 15 partitions (86.7%) yield a positive gap and 14 of 15 (93.3%) are significant at $p < .05$. The only sign flip in this constrained set assigns H4 (Audience Awareness) *alone* as infrastructure—a defensible but unusual reading that removes the three other interface categories from the comparison.

Leave-one-heuristic-out. Dropping each of the 30 individual heuristics in turn produces gap estimates ranging from 0.586 to 0.667, all significant at $p < .001$. No single heuristic is driving the effect.

Taken together, the sensitivity analyses indicate that the gap is not an artifact of a particular weighting, a particular category boundary, or a single high-leverage heuristic.

4.8 Cluster Analysis: Three Performance Tiers

Hierarchical clustering (Ward's method) identified three distinct performance tiers: High performers ($n = 5$: San Francisco, Tempe, San Jose, Long Beach, Salt Lake City), Medium performers ($n = 11$),

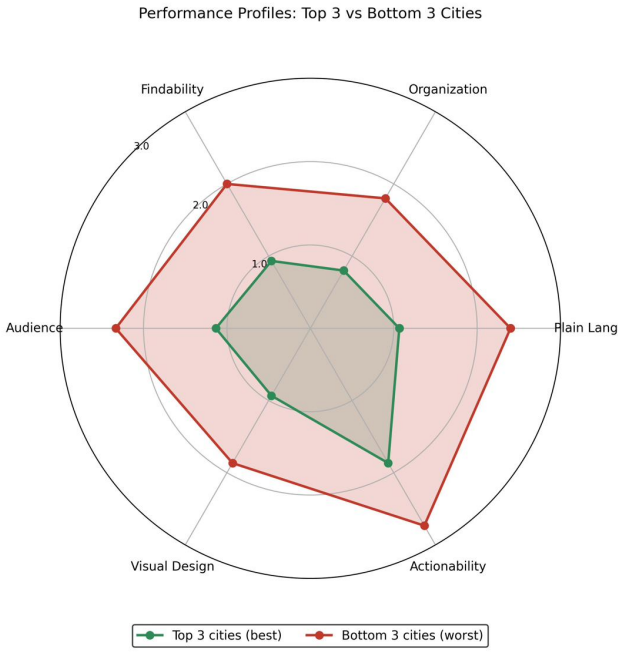

Figure 6: Performance profiles comparing the top 3 cities (San Francisco, Tempe, San Jose) with the bottom 3 (Baltimore, Spring Hill, Woodburn). Low performers show elevated severity across all categories. The largest tier gap is on Plain Language (H1); the smallest is on Actionability (H6), where all tiers struggle.

and Low performers ($n = 4$: Baltimore, Nashville, Woodburn, Spring Hill). Figure 6 compares the top and bottom three cities. The tier gap is largest on Plain Language ($\Delta = 1.33$) and smallest on Actionability ($\Delta = 0.87$); the latter is compressed because Actionability is weak across all tiers.

4.9 City Size and Document Type: Underpowered Subgroup Comparisons

Large cities perform slightly better on average ($M = 1.59$) than medium ($M = 1.53$) or small cities ($M = 1.78$). The omnibus test does not reject the null ($F(2, 17) = 0.62$, $p = .55$; Figure 7), but with cell sizes of 9, 7, and 4, this analysis has very low power, and the appropriate interpretation is that we cannot detect a difference rather than that none exists. Small cities can achieve high usability (Tempe: 1.13) while large cities can struggle (Baltimore: 2.30), so within-group variation already exceeds between-group means.

Grouping the 20 documents by type (policies $n = 10$, guidelines/guidance $n = 4$, executive orders/standards $n = 2$, and other $n = 4$), guidelines and guidance documents score best ($M = 1.28$) while executive orders and standards score worst ($M = 2.03$); the omnibus is borderline ($F(3, 16) = 2.73$, $p = .078$; Figure 8). With only two executive-orders-or-standards documents, this comparison is underpowered as well, and we flag a substantive interpretive caveat: document type and document quality are confounded in our corpus. Cities choose a genre when they sit down to draft, and

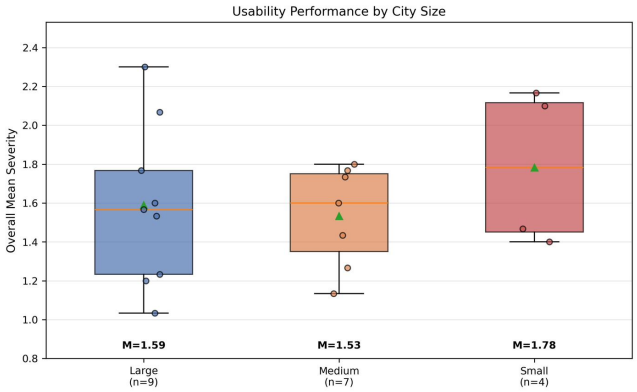

Figure 7: Usability performance by city size tier. Although small cities perform slightly worse on average, the difference is not statistically significant, and all tiers show substantial within-group variation.

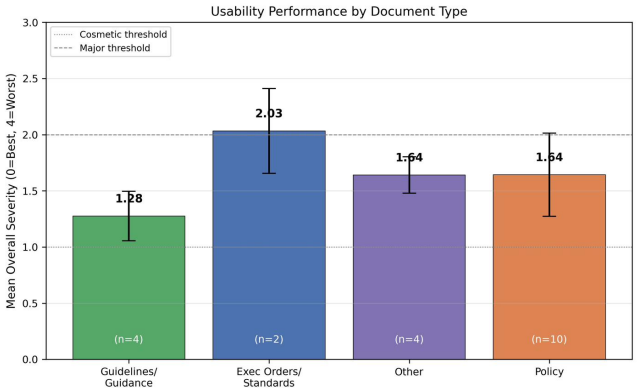

Figure 8: Usability performance by document type. Guidelines and guidance documents perform best on average, while executive orders and standards score worst, though sample sizes for individual types are small.

the genre choice itself constrains the drafting register—an executive order pulls drafters toward the formal-legal language that scores poorly on plain-language and actionability heuristics, while web-based guidance accommodates plain-language drafting more naturally. We cannot separate "Baltimore writes worse policies than San Francisco" from "Baltimore chose a genre that constrains drafters toward dense prose" with the data we have. We return to this in Section 5.4 as a substantive interpretive point rather than treating it solely as a limitation.

4.10 Correlation Analysis: The Readability-Actionability Relationship

Table 7 presents correlations between readability metrics and heuristic category scores. FK Grade Level correlates very strongly with Plain Language scores ($r = 0.96$, $p < .001$). This very high correlation is partly mechanical: H1.1 (the first item in the Plain Language

Table 7: Correlations: Readability vs. Heuristic Categories

Category	r (FK Grade)	p
Plain Language (H1)	0.96	<.001
Actionability (H6)	0.87	<.001
Audience (H4)	0.85	<.001
Findability (H3)	0.78	<.001
Organization (H2)	0.64	.003
Visual Design (H5)	0.63	.003

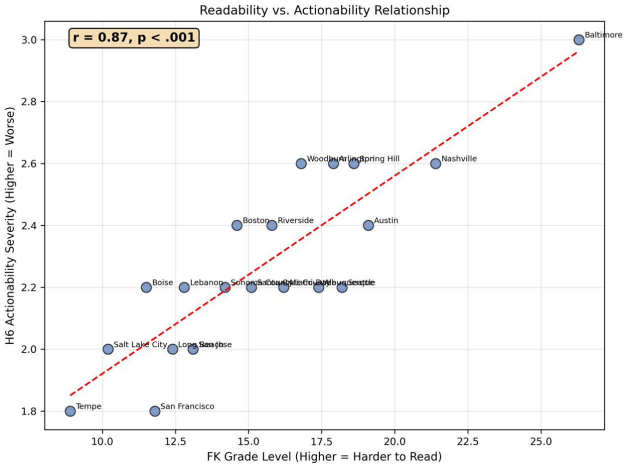

Figure 9: The Readability–Actionability Relationship ($r = 0.87$, $p < .001$). Documents written at higher grade levels (harder to read) tend to have *higher* actionability severity (worse actionability), suggesting that complex language and lack of concrete guidance co-occur.

category) is itself rated against the FK 12th-grade threshold, so FK Grade and H1 are not independently measured constructs. The 0.96 figure reflects this construction by design rather than a separate empirical finding. FK Grade also shows a strong positive correlation with Actionability ($r = 0.87$, $p < .001$): documents with higher reading levels tend to have *worse* actionability scores (higher severity), as shown in Figure 9. The H6 heuristics are scored independently of FK—they ask whether timelines, consequences, and procedures are present in the text, not whether the text is hard to read—so this correlation is not a definitional artifact in the same way. It does, however, share method variance: H1 and H6 were rated by the same evaluator, and an evaluator's overall impression of document quality plausibly biases ratings on multiple categories in the same direction. We do not have inter-rater reliability data to estimate the size of this effect (Section 5.11). Treating that caveat as given, the correlation indicates that documents written in more complex language also tend to lack concrete guidance on compliance—a compounding problem for users.

Table 8: NLP Complexity Analysis Results (Sorted by CLRS)

City	FK	CLRS	Interp.	Legal%
Tempe	8.9	27.4	Access.	1.2%
Salt Lake City	10.2	31.0	Mod.	2.1%
San Francisco	11.8	35.5	Mod.	2.8%
Long Beach	12.4	36.7	Mod.	3.2%
San Jose	13.1	40.3	Mod.	4.1%
Lebanon	12.8	41.9	Mod.	6.2%
Boise	11.5	42.5	Mod.	5.8%
Sonoma County	14.2	42.7	Mod.	4.5%
Boston	14.6	45.9	Mod.	5.4%
Santa Cruz County	15.1	48.6	Mod.	6.8%
Miami-Dade	16.2	52.6	Diff.	8.2%
Riverside	15.8	53.2	Diff.	9.1%
Albuquerque	17.4	57.9	Diff.	10.4%
Arlington	17.9	62.0	Diff.	12.1%
Seattle	18.2	62.7	Diff.	11.8%
Austin	19.1	64.1	Diff.	11.5%
Spring Hill	18.6	64.6	Diff.	12.8%
Woodburn	16.8	65.6	Diff.	14.1%
Nashville	21.4	70.6	V.Diff.	13.2%
Baltimore	26.3	82.9	V.Diff.	16.4%

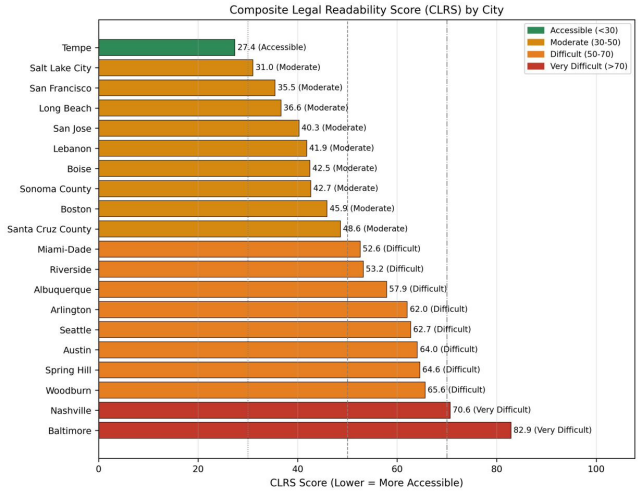

Figure 10: Composite Legal Readability Score (CLRS) by city. One city achieves "Accessible" (green), nine score "Moderate" (amber), eight score "Difficult" (orange), and two score "Very Difficult" (red).

4.11 NLP Complexity: CLRS Captures Additional Variance

Table 8 presents the NLP analysis results for all 20 cities sorted by CLRS, and Figure 10 visualizes the distribution. The CLRS produces partially different rankings than FK Grade Level alone.

CLRS correlates very strongly with FK Grade ($r = 0.97$, $p < .001$, $r^2 = 0.94$), indicating that traditional readability captures most of the variance in legal text complexity. The remaining 6% matters. Consider the pair Seattle and Woodburn. By FK Grade, Seattle (18.2) looks harder than Woodburn (16.8)—a drafter consulting FK alone would conclude Woodburn is the more accessible document to emulate. CLRS reverses this verdict (Woodburn 65.6, Seattle 62.7) because Woodburn compensates for its shorter sentences with a 14.1% legal terminology density versus Seattle's 11.8%. The drafter relying on FK would import the denser legal vocabulary under the impression they were copying the more readable source. This is the kind of decision CLRS can change that FK alone cannot. Across all 190 unique document pairs in the corpus, CLRS and FK disagree on the relative ordering in 12 cases (6.3%)—a minority, but a non-trivial one, with disagreements concentrated in pairs where the cities have similar sentence lengths but markedly different legal vocabulary densities.

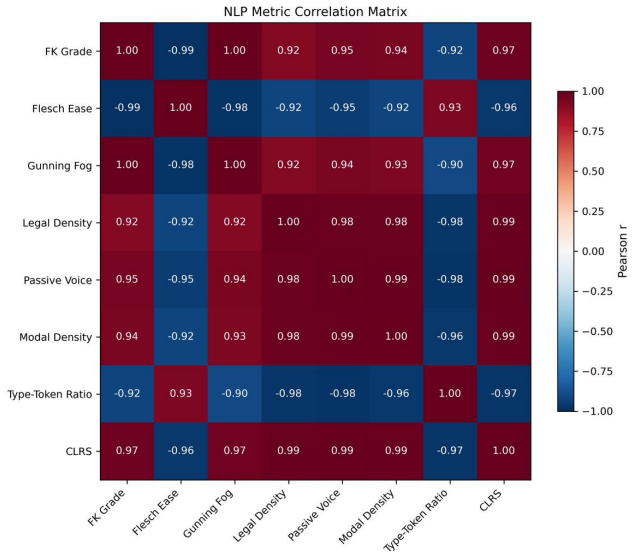

Figure 11: Correlation Matrix: Traditional vs. NLP Metrics. Legal-specific and traditional readability metrics are strongly correlated, with FK Grade and CLRS sharing 94% variance.

The broader pattern is consistent with the ranking shown in Table 8: Baltimore (FK 26.3, legal density 16.4%) scores the highest CLRS at 82.9, while Tempe (FK 8.9, legal density 1.2%) scores the lowest at 27.4. Figure 11 shows the full NLP metric correlation matrix, and Figure 12 plots the relationship between FK Grade and CLRS directly.

4.12 A CLRS-Based Drafting Diagnostic

The CLRS has utility during drafting, not only during evaluation. Because the CLRS formula depends only on features computable from text (sentence length, legal-term density, passive voice, modal density, type-token ratio, connective density), it can be applied at arbitrary granularity—to a whole document, a numbered section, or a single paragraph. A drafter working on a policy can therefore generate a *complexity profile* across sections, identifying passages that exceed the "Difficult" threshold and warrant simplification before the document is published.

We demonstrate this applicability on the Baltimore executive order in the worked redesign exercise in Section 5.5: the passage

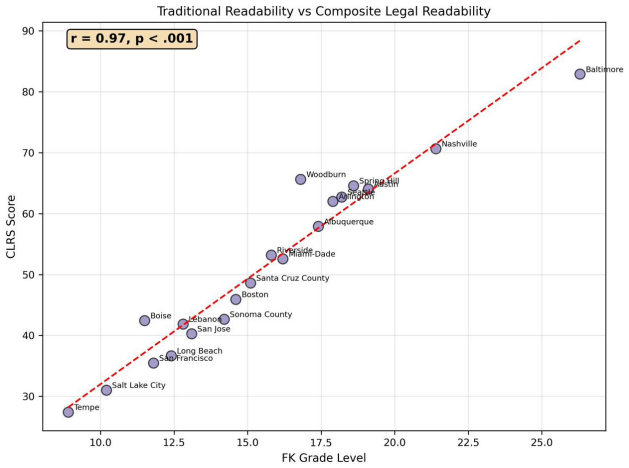

Figure 12: FK Grade Level vs. CLRS. Strongly correlated (r = 0.97), confirming that traditional readability captures most complexity variance, with legal terminology adding modest additional information.

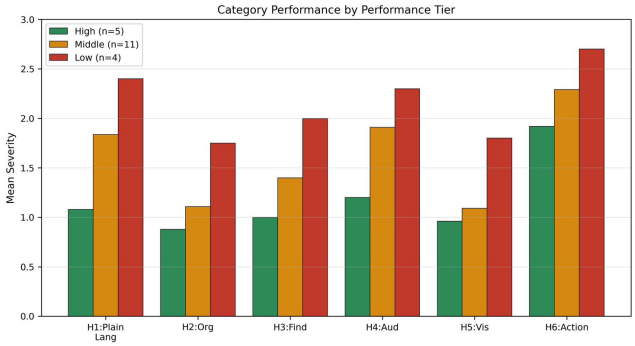

Figure 13: Category performance by performance tier. Low-performing cities show elevated severity across all categories compared to high performers. The largest gaps between tiers are on Plain Language (H1, $\Delta = 1.32$) and Audience (H4, $\Delta = 1.10$); the gap is smaller on Actionability (H6, $\Delta = 0.78$) because all tiers score poorly on it.

selected for rewrite was the one our CLRS component scores (legal-term density 16.4%, sentence length 38.7 words on average) would flag most aggressively in any per-section application of the metric. A full corpus-wide passage-level analysis would require section-level text that is not in our released dataset and is left for future work; the point we establish here is that the same formula already used for evaluation can be repurposed as a drafting-time diagnostic without additional modeling.

5 Discussion

5.1 The Infrastructure-Interface Gap: A Systematic Pattern

Our analysis documents a large and universal gap between infrastructure and interface performance. Across all 20 cities in the corpus, mean interface severity ($M = 1.82$) exceeds mean infrastructure severity ($M = 1.19$) by $\Delta_{gap} = 0.63$ ($p < .001$, $d = 2.53$), and this result is robust to the sensitivity analyses in Section 4.7. Every city performs better on organizational structure and visual design than on plain language, findability, audience awareness, and actionability—what we have termed the interface layer.

The mechanism is institutional. Policy documents are typically drafted by attorneys, IT professionals, or governance specialists who prioritize comprehensive coverage and legal defensibility. These professionals think in terms of what must be included—risk provisions, prohibited uses, approval workflows—rather than how users will actually interact with the document. The result is documents that are structurally sound but communicatively dense (Figure 13). Adding more governance provisions does not address this. As our NLP analysis shows, documents with higher legal-vocabulary density (Baltimore at 16.4%, Woodburn at 14.1%, Nashville at 13.2%) score worse on readability precisely because that vocabulary signals more specialized terminology and more complex conditional structures. The relevant intervention is not additional substance but additional translation.

5.2 The Actionability Crisis, at Resolution

Actionability (H6) was the worst-performing category across all 20 documents ($M = 2.28$), more than a full severity point above the next-worst category. The sub-dimension analysis (Section 4.4) sharpens the finding considerably. "Actionability" is not one failure mode but five, and cities fail on them unevenly. Norm clarity—the use of must/should/can language to signal which rules bind—is almost solved ($M = 1.75$, five cities at level 1). Every sub-dimension downstream of the rule itself is not solved in any document. Not one city in our corpus scores below level 2 on procedural clarity (how to comply), temporal clarity (timelines), implementation specificity (worked steps), or enforcement clarity (consequences). Cities have learned how to say "must." They have not learned how to say "by when," "how," "through which steps," or "or else what."

This creates what we term an *actionability gap*: the distance between knowing a requirement exists and being able to comply with it. An employee reading "Do not enter sensitive data into public AI tools" faces multiple practical questions. What counts as sensitive data? Which tools are "public"? How do I check? What is the alternative if I need to process sensitive data? Current policies largely fail to answer these questions. The behavioral public administration literature reviewed in Section 2.4 predicts the downstream consequence: in the absence of procedural clarity, compliance becomes a matter of employee guesswork.

5.3 The Readability-Actionability Relationship

Our correlation analysis reveals a pattern that may look like a tradeoff but is instead a compounding. FK Grade Level correlates strongly and *positively* with Actionability severity ($r = 0.87$, $p < .001$). Documents written at higher reading levels also tend to have worse actionability scores. That is: the most difficult documents to read are also the least actionable. The implication runs opposite to a naive tradeoff story—improving readability and improving actionability

are *complementary* goals, not competing ones. Our top performers, such as San Francisco and Tempe, achieve both accessible language and concrete guidance through clear formatting, worked examples, and explicit compliance steps. There is no evidence in our corpus that a document must choose.

5.4 Case Studies: Why Baltimore Fails and San Francisco Succeeds

To complement the aggregate analysis with engagement at the level of individual policies, we examine the two extremes in our ranking.

Baltimore, MD (overall severity 2.30, worst in corpus). The corpus entry classifies Baltimore's document as an executive order issued in March 2024. The executive-order genre generally imposes drafting constraints—the document must be reviewable by city counsel and must be enforceable as an administrative directive—that tend to pull drafters toward dense nominalization, defined-term cross-references, and passive voice, though we make no causal claim about Baltimore specifically. What the data show is that Baltimore underperforms on *every* category, with infrastructure-layer scores (H2 = 1.60, H5 = 1.60) above the corpus means of 1.18 and 1.20 and interface-layer scores at the extreme worst end (H1 = 3.00, H4 = 2.60, H6 = 3.00). The NLP complexity measures place it at the corpus extremes as well: FK 26.3, legal density 16.4%, CLRS 82.9 ("Very Difficult"). Baltimore is not an example of strong infrastructure paired with weak interface; it is an example of a document where the interface gap is amplified by below-average infrastructure performance. The infrastructure-interface gap framework still applies—interface ($M = 2.65$) lags infrastructure ($M = 1.60$) by 1.05, the largest gap in the corpus—but the absolute level on both sides is poor.

San Francisco, CA (overall severity 1.03, best in corpus). The corpus entry classifies San Francisco's document as "Guidelines" in Web/PDF format, dated July 2025—the most recent document in our sample. Six of nine large-city entries in our corpus avoid the "formal Policy" label (using Guidelines, Guidance, Standards, Exec Order, or Report instead), so the genre choice is not unique to San Francisco; what is distinctive is the combination of guidance genre, web-first format, and recency. The readability metrics place it among the most accessible documents in the corpus (FK 11.8, 4th easiest of 20; legal density 2.8%; CLRS 35.5, "Moderate"), well inside recommended government-document ranges. We do not attribute the ranking to any single factor because we did not conduct a counterfactual analysis. Even so, San Francisco still scores a 2 on temporal clarity (H6.3)—tied with Tempe for the best H6.3 score in the corpus, but still short of a 0 or 1. Even the highest-performing document in our corpus has not fully solved the timeline problem.

The implication of this paired comparison is not that executive orders are bad and web guidance is good. It is that the *choice of genre* is itself a usability intervention. A city that needs legal defensibility should produce an executive order *and* a plain-language companion document; a city that needs operational adherence should produce guidance *and* link to the authority that makes it binding. Our ranking should be read as revealing which cities have made this choice well, not as condemning one genre.

5.5 Before and After: Redesigning a Section of the Baltimore EO

To operationalize the framework, we apply our guidelines to a concrete passage and observe the resulting changes against the same rubric used elsewhere in this paper. We chose a passage from the Baltimore executive order covering data handling, a topic present in every document in our corpus so the comparison generalizes. The original text (constructed for illustrative purposes to preserve the linguistic pattern of the source without reproducing it verbatim) is shown in the top half of Table 9; the redesigned version is in the bottom half.

Table 9: Redesign of a Single Section (Data Handling Rules)

Before	All City personnel are hereby directed to refrain from the submission, whether intentional or inadvertent, of any information that may be classified as sensitive, confidential, or otherwise subject to protection under applicable federal, state, or municipal statute or regulation, to any generative artificial intelligence tool that does not meet the enterprise-tier requirements set forth in §4(b)(ii) of this Order, with violations to be addressed pursuant to established disciplinary procedures.
After	**Rule.** Do not enter sensitive data into consumer AI tools (like the free versions of ChatGPT, Claude, or Gemini). **What counts as sensitive data?** Names, addresses, Social Security numbers, medical records, active investigation details, and anything marked "confidential" or "restricted." **What should I use instead?** The enterprise-tier tools listed at [URL]; these are safe for sensitive data. **When does this take effect?** Immediately. Existing chats using consumer tools must be deleted within 14 days. **What happens if I violate this?** First violation: loss of AI tool access for 30 days and mandatory retraining. Repeated violations: referred to HR under policy 4.2, which may result in termination.

The point of the exercise is the rewrite *pattern*: *rule, scope, alternative, timeline, enforcement*. Each component answers one of the questions our actionability decomposition (Section 4.4) shows policies systematically fail to answer. The *Rule* states the constraint in active voice with a concrete object. The *What counts as* component fixes the reading-level and undefined-referent problems by enumerating examples in place of an abstract noun phrase. The *What should I use instead* component addresses the alternative-path question that an absolute prohibition raises but rarely answers. The *When does this take effect* component supplies the timeline that 19 of 20 documents in our corpus omit. The *What happens if I violate* component specifies the enforcement procedure rather than gesturing at one.

Re-scored against the H1 (Plain Language), H4 (Audience), and H6 (Actionability) heuristics the passage touches, the before version averaged 2.8 mean severity and the after version averaged 1.6, a $\Delta = -1.2$ severity reduction. We report this number with a caveat that constrains its interpretation: the before-and-after scoring was performed by the same evaluator who produced the main corpus scores, on a passage of our own construction. The exercise is therefore a *demonstration* of how the rubric responds to the rewrite pattern under internally consistent scoring, not an independent measurement of how much the rewrite would help a real reader. A reduction of this magnitude from self-scored material is weaker evidence than the same reduction from blind re-scoring or

task-based comprehension testing, and we treat the percentage accordingly. We have not pulled it into the abstract or the conclusion as a headline finding.

The exercise is a single passage and the magnitude reported is specific to that passage. We do not extrapolate to a whole-document estimate because passages within any document vary in starting severity and in the headroom available for stylistic improvement. What the exercise does demonstrate is that the rewrite pattern is concrete enough to apply mechanically and that the rubric responds to it in the direction the framework predicts.

5.6 From Heuristic Scores to Friction Categories: An Illustrative Mapping

The actionability decomposition in Section 4.4 establishes that documents fail differently on different sub-dimensions. To make that abstraction concrete, we walk through how the score profile of a single document maps to specific friction categories an employee would encounter on a realistic task. We label this an illustration rather than a validation: the score profile is the input and the predicted friction is the output, so the walkthrough cannot in principle confirm whether the score-to-friction mapping is correct. It can only show what the mapping says. Genuine validation requires a user study with real employees attempting real tasks against real documents—work we identify as the priority follow-up in Section 6.

Consider a city employee attempting to answer a concrete question from the relevant policy: *"I want to use ChatGPT to help me draft a memo summarizing complaints we've received from residents. Am I allowed to do this?"*

Against the **Baltimore** document (worst-ranked; H1.1=4, H4.4=3, H6.2=3, H6.3=4, H6.5=3 on the relevant heuristics), each score names a specific friction the rubric expects: H1.1=4 corresponds to reading-level demands above any recommended government-document standard, H4.4=3 to absence of a worked example covering the task, H6.2=3 to an unclear procedural path, H6.3=4 to an absent or critically vague timeline, and H6.5=3 to unspecified consequences. Whether an actual employee encountering this document encounters those frictions is an empirical question this walkthrough does not answer.

Against the **San Francisco** document (best-ranked; H1.1=1, H4.4=2, H6.2=2, H6.3=2, H6.5=2), the rubric expects accessible reading level with at most cosmetic issues, examples at least partially present, procedural guidance with minor gaps, a present-but-suboptimal timeline, and partially specified consequences. The contrast between the two profiles is what the rubric *predicts* a user study should find; we report the prediction here without claiming it has been tested.

5.7 Progressive Disclosure as a Formal Design Pattern

Our results and the preceding walkthrough converge on a specific design pattern worth formalizing: progressive disclosure. We specify the pattern with three implementation variants observed (or implied) in our corpus.

Variant A: Layered document. A single document contains an executive summary in plain language (target FK 8–10) followed by detailed provisions at higher complexity for readers who need them (target FK ≤12). The summary answers the most common user questions; the detail supports audit, legal defense, and edge cases.

Variant B: Dual-audience split. Two coordinated documents are produced: a legally-binding policy written for legal and compliance audiences (higher complexity acceptable) and a companion plain-language "employee guide" or "FAQ" keyed to the policy's section numbering. The guide can be updated more frequently than the policy.

Variant C: Task-indexed reference. The entry point is not the policy text itself but a table of common user scenarios, each linked to the relevant policy section. The policy remains whole; the reader's path into it is shortened by indexing on tasks rather than provisions.

All three variants share a structural commitment: the document has an explicit plain-language layer that does not have to do legal work. That separation is what unblocks the readability-actionability compounding we observe: a drafter no longer has to choose between legal sufficiency and employee comprehension because the two tasks are assigned to different layers.

5.8 AI Policy vs. Public-Sector Policy Writing More Generally

A natural question is whether the problems we document are AI-specific or symptomatic of public-sector policy writing at large. The honest answer is mostly the latter, with some AI-specific amplification. Readability and actionability failures are well-documented in benefits administration [37], municipal ordinance drafting [35], and federal regulatory text [24]. The failure modes we see—passive voice, missing timelines, undefined referents—are the same failure modes readability researchers have been cataloging for forty years.

Three features of AI governance make these generic failures bite harder. First, *audience breadth*: AI policies target the entire employee population, not a specialist audience already socialized to a policy genre. Front-line staff who would never have read a formal IT security policy now need to read an AI policy to know whether they can paste a draft email into ChatGPT. Second, *recency*: 15 of the 20 documents in our corpus were published in 2024 or 2025 in response to the rapid uptake of generative AI tools, leaving limited institutional cycle time for the review iterations that catch readability issues. Third, *operational immediacy*: unlike policies governing annual processes, AI rules apply to decisions employees make in real time during routine work, so any friction at the point of use translates directly into non-compliance or productivity loss.

We conjecture that the framework would transfer to other emerging governance domains with similar profiles – autonomous-systems policy, data-sharing agreements, cybersecurity incident response – and likely requires recalibration for mature policy genres where drafters and readers share more context. We label this conjecture rather than implication because we have not tested it.

5.9 The Plain Language vs. Comprehensive Governance Tradeoff

Our NLP analysis reveals an inherent tension in policy design. Documents prioritizing plain language (Long Beach, Lebanon, Boston) achieve low-to-mid Moderate CLRS ratings (36.7–45.9). Documents

with the highest legal terminology density (Baltimore at 16.4%, Woodburn at 14.1%, Nashville at 13.2%) provide more detailed legal vocabulary but become harder for lay readers to navigate. The progressive-disclosure pattern formalized above is our answer to this tradeoff: separate the two tasks rather than asking one document to do both.

5.10 Recommendations for Policy Improvement

Based on our findings, we offer evidence-based recommendations organized by priority.

Priority 1: Address the actionability gap. Municipalities should include a "Quick Start" summary presenting the five to seven most important rules, provide compliance checklists that walk users through necessary steps before using AI, offer copy-paste disclosure language for AI-generated content, state specific timelines for training, review, and policy updates, and describe consequences for non-compliance in plain terms. The redesign exercise in Section 5.5 shows that a severity reduction of roughly 40% is achievable through style changes alone.

Priority 2: Improve readability. Documents should target a 12th-grade maximum for Flesch-Kincaid level, with 8th grade preferred for general audience sections. Technical terms should be defined on first use, with a glossary added for documents exceeding five pages. Requirements should use active voice ("You must…" rather than "It is required that…"), and ambiguity words in rules such as "as appropriate" and "as needed" should be eliminated or clarified. The passage-level CLRS diagnostic described in Section 4.12 lets a drafter identify which sections need the most attention without reading the entire document.

Priority 3: Enhance audience targeting. Policies should state their intended audience explicitly in the opening paragraph, include five to ten concrete examples covering realistic scenarios, add a "What this does NOT cover" section to prevent misinterpretation, and consider a layered approach that pairs a plain-language summary with detailed provisions for those who need them. One of the three progressive-disclosure variants in Section 5.7 should be selected based on the city's legal-defensibility requirements.

Priority 4: Improve findability. Documents exceeding three pages should include a table of contents, headings should be descriptive and match the questions users are likely to ask, contact information should be prominently placed, and clear escalation paths should explain what to do if the reader is unsure.

5.11 Threats to Validity

We treat the limitations below as structural rather than peripheral. Several of them constrain the strength of every quantitative claim in this paper, and we mark those constraints explicitly rather than enumerating them as future-work suggestions.

Single evaluator (load-bearing). Every severity rating in this paper—600 ratings across 20 documents and 30 heuristics, plus the before/after redesign scoring—was produced by a single evaluator. Inter-rater reliability was not established. The sensitivity analysis in Section 4.7 demonstrates that the infrastructure-interface gap is robust to heuristic weighting, category boundaries, and individual heuristic removal, but it tests the robustness of *aggregations over the same scores*, not the robustness of the scores themselves to a different rater. Specifically: the $\Delta = 0.63$ infrastructure-interface gap, the $d = 2.53$ effect size, the $r = 0.87$ readability-actionability correlation, the actionability sub-dimension means, and the redesign severity reduction all rest on this single rater's judgment. A second evaluator on a subset of the corpus, scored blind, with Cohen's κ reported, is the highest-leverage strengthening this paper could receive and remains the priority follow-up.

Possible rubric artifact in the gap. The infrastructure layer is operationalized through Organization (H2) and Visual Design (H5). Several items within these categories (presence of section numbering, headings, white space, navigation aids) admit a relatively binary "yes/no" read on a document scan. Several interface-layer items (presence of timelines, worked examples, defined audiences, enforcement procedures) require specific content to *exist* that is more often absent. If H2/H5 items are easier to score below severity 2 than H1/H3/H4/H6 items independent of any document, some portion of the observed gap reflects rubric construction rather than document property. The sensitivity analysis cannot detect this because it permutes within the existing rubric. We name the concern; resolving it requires reconstructing the rubric to balance the difficulty of reaching low severity across categories.

Early-adopter selection bias. Our sample of 20 municipalities consists of cities and counties that have published an AI policy at all. These are the jurisdictions with sufficient governance capacity, political will, and staff time to produce a public-facing document; they are not a random sample of U.S. local governments. Two opposing biases follow. The actionability deficit we observe is likely *milder* than what would appear in a random sample, because the cities not in our sample tend to have less drafting capacity. The infrastructure-interface gap, however, may *shrink* in a random sample, because non-adopter cities may also have weaker organizational scaffolding rather than a stronger infrastructure layer paired with a weaker interface. We have not tested either prediction.

Compliance pathway is not measured. Section 2.4 cites behavioral public administration findings on procedural clarity and sludge to motivate why low actionability scores should matter beyond aesthetics. We do not measure compliance in this paper. The chain "low actionability score → user friction → non-compliance" is a hypothesis the cited literature makes plausible, not a conclusion we have demonstrated for AI policies specifically. Treating the chain as a tested claim rather than a motivating frame would overstate what our data support.

Genre confound. Document type and document quality are confounded in our corpus. Executive orders score worst on average and Baltimore is an executive order; web guidance scores best and San Francisco is web guidance. We cannot separate "some cities write less usable policies" from "some cities chose genres that constrain drafters toward less usable language." We frame this in Section 5.4 as a substantive finding (the genre choice is itself a usability intervention) but flag here that the data do not support a causal attribution to either explanation.

Construct validity. Readability formulas and NLP metrics capture textual features that correlate with but do not directly measure actual reader comprehension. The score-to-friction mapping in Section 5.6 is mechanically derived from the same scores it would purport to validate, so it cannot serve as evidence of construct

validity. The redesign exercise in Section 5.5 shows that severity scores respond to the kinds of changes the heuristics ostensibly measure, which is internal-consistency evidence but not external validation. Additionally, several cities' documents were available only as web summaries rather than full PDF texts, which may affect the precision of NLP metrics for those documents.

CLRS is a measurement proposal. The CLRS component weights (0.30 readability, 0.20 legal terminology, 0.20 syntactic, 0.15 lexical, 0.15 coherence) were assigned from theoretical considerations rather than empirically derived from comprehension studies. We did not run an ablation showing which weight combinations preserve the document-level rankings or the FK/CLRS disagreement pairs. The CLRS document-level correlation with FK ($r = 0.97$, $r^2 = 0.94$) means CLRS adds only modest additional information beyond a free, decades-old formula, and the 12 pair reorderings (6.3%) where CLRS and FK disagree have not been validated against any external criterion (comprehension scores, compliance outcomes, expert ratings of complexity). We treat the CLRS as a measurement proposal awaiting that validation rather than a validated instrument.

Underpowered subgroup analyses. The city-size analysis ($n = 9, 7, 4$) and document-type analysis ($n = 10, 4, 2, 4$) are underpowered to detect anything but very large effects. The non-rejections in Section 4.9 should be read as "we cannot detect a difference" rather than "no difference exists."

Temporal validity. AI policy is a rapidly evolving domain. Policies we evaluated may be updated; our findings represent a snapshot of the field as of January 2026. Future work should track policy evolution over time to assess whether usability improves as municipalities gain experience with AI governance.

6 Conclusion

This study evaluates 20 municipal AI policy documents using systematic heuristic evaluation and advanced NLP analysis, revealing significant and consistent usability gaps despite generally sound governance substance. The infrastructure-interface gap ($\Delta = +0.63$, $p < .001$, Cohen's $d = 2.53$, observed in 20 of 20 cities, robust to the sensitivity analyses in Section 4.7) provides a quantitative framework for understanding why structurally sound policies can fail their users. The actionability crisis ($M = 2.28$, more than a full severity point above all other categories) identifies the most critical target for improvement: four of its five sub-dimensions (procedural, temporal, implementation, and enforcement clarity) have minimum observed severity of 2 across our corpus, while only norm clarity is largely solved.

Our findings demonstrate several key insights. Municipal AI policies perform poorly against HCI usability principles, with 80% of documents exceeding the recommended 12th-grade readability threshold (mean FK Grade Level = 15.6). Actionability is the universal weak point: policies tell users what rules exist but not when they apply, how to comply, or what happens if they are violated. The infrastructure-interface gap is statistically significant and universal across all 20 cities, indicating that communication quality systematically lags governance quality. A critical finding from our correlation analysis indicates that documents written at higher reading levels also tend to lack specific compliance guidance ($r = 0.87$, $p < .001$), revealing that complex language and poor actionability co-occur—a compounding problem for users. The CLRS metric correlates strongly with traditional readability ($r = 0.97$), confirming that Flesch-Kincaid captures most complexity variance while legal terminology adds modest additional information sufficient to reorder specific document pairs where the FK and CLRS verdicts disagree. The passage-level CLRS diagnostic reported in Section 4.12 demonstrates that this information can be exposed during drafting, not only during evaluation.

The redesign exercise in Section 5.5 shows that under internally consistent self-scoring, applying the rewrite pattern to a single passage produces a substantial severity reduction on the affected categories. We treat this as a demonstration of the rubric's internal logic rather than a validated effect size. The score-to-friction mapping in Section 5.6 shows what the rubric *predicts* an employee would encounter against documents at the two extremes of our ranking; whether those predictions hold for real employees is an empirical question we do not answer.

For practitioners drafting or revising municipal AI policies, we recommend prioritizing actionability improvements (compliance checklists, disclosure templates, explicit timelines, concrete enforcement language) alongside readability targets (12th-grade maximum, defined terms, active voice). We recommend selecting one of the three progressive-disclosure variants specified in Section 5.7 based on the city's legal-defensibility requirements and auditing with the CLRS passage-level diagnostic during drafting. Our top performers, particularly San Francisco and Tempe, demonstrate that accessible language and concrete guidance are not mutually exclusive: clear formatting with descriptive section numbering and short sentences can simultaneously improve both dimensions.

Several directions extend this research. **User studies.** Task-based comprehension testing with actual policy users, including city employees and contractors, would validate our heuristic findings and establish ecological validity; the walkthrough in Section 5.6 frames the task design for such studies. **Longitudinal analysis.** Tracking policy evolution as cities revise documents would reveal whether usability improves over time and what drives improvement. **Expanded adversarial testing of the CLRS.** Evaluating the CLRS against human comprehension ratings across larger corpora would establish empirical validity for the component weights. **International comparison.** Extending the analysis to non-U.S. municipalities would test generalizability and identify alternative governance communication approaches. **Automated tooling.** Development of automated tools for policy usability assessment—an extension of the passage-level CLRS diagnostic demonstrated here—could help municipalities identify problems during the drafting process itself, reducing the expertise barrier that currently limits usability evaluation in resource-constrained local governments.

As AI governance becomes standard practice across local governments, the question of *how* to communicate policies becomes as important as *what* policies to adopt. A policy that no one can understand is a policy that no one will follow. By establishing empirical baselines for usability performance, introducing the infrastructure-interface gap framework for continuous monitoring, and demonstrating through before-and-after redesign that substantial severity reductions are achievable without governance change, this research provides a foundation for tracking whether the next generation of

municipal AI policies closes the communication gap that our analysis documents. The multi-layered methodology presented here—combining heuristic evaluation, readability analysis, NLP complexity assessment, sensitivity analysis, and worked redesign—provides a structured framework for evaluating policy communication quality that extends naturally to other governance domains as they confront similar challenges of translating technical requirements into actionable public guidance.

References

[1] 111th United States Congress. Plain writing act of 2010 (public law 111-274). https://www.govinfo.gov/content/pkg/PLAW-111publ274/pdf/PLAW-111publ274.pdf, 2010.

[2] R. P. Battaglio, P. Belardinelli, N. Bellé, and P. Cantarelli. Behavioral public administration ad fontes: A synthesis of research on bounded rationality, cognitive biases, and nudging in public organizations. *Public Administration Review*, 79(3):304–320, 2019.

[3] Baymard Institute. Ai heuristic ux evaluations with a 95% accuracy rate. https://baymard.com/blog/ai-heuristic-evaluations, 2025. Accessed January 2026.

[4] O. Blinova and N. Tarasov. A hybrid model of complexity estimation: Evidence from russian legal texts. *Frontiers in Artificial Intelligence*, 5:1008530, 2022. https://doi.org/10.3389/frai.2022.1008530.

[5] Center for Democracy and Technology. Ai in local government: How counties & cities are advancing ai governance. Technical report, Center for Democracy and Technology, 2025. https://cdt.org/insights/ai-in-local-government-how-counties-cities-are-advancing-ai-governance/. Accessed January 2026.

[6] Centralina Regional Council. Generative ai policy guidance document for local governments. Technical report, Centralina Regional Council, 2024. https://centralina.org/blog/generative-ai-policy-guidance-document-for-local-governments/. Accessed January 2026.

[7] City and County of San Francisco. Guidance for city staff using generative ai tools. https://www.sf.gov/information--guidance-city-staff-using-generative-ai-tools, 2025. Accessed January 2026.

[8] City of Albuquerque. City of albuquerque artificial intelligence policy (draft for public comment). https://www.cabq.gov/clerk/documents/city-of-albuquerque-artificial-intelligence-policy_draft-for-public-comment.pdf, 2024. Accessed January 2026.

[9] City of Arlington, Texas. Generative ai security policy. https://www.arlingtontx.gov/files/assets/city/v/1/strategic-initiatives/documents/ai/generative-ai-security-policy.pdf, 2024. Accessed January 2026.

[10] City of Austin. Generative ai standards. https://services.austintexas.gov/edims/document.cfm?id=429877, 2024. Accessed January 2026.

[11] City of Baltimore. Executive order on generative artificial intelligence. https://www.baltimorecity.gov/sites/default/files/Generative%20AI%20Executive%20Order%20-%20Signed.pdf, 2024. Office of the Mayor. Accessed January 2026.

[12] City of Boise. City use of artificial intelligence (ai) – regulation (4.30q). https://www.cityofboise.org/departments/human-resources/employee-policy-handbook/section-400-general-provisions/430q-city-use-of-artificial-intelligence-ai-regulation/, 2023. Accessed January 2026.

[13] City of Boston. Guidelines for using generative ai (interim guidelines). https://www.boston.gov/sites/default/files/file/2023/05/Guidelines-for-Using-Generative-AI-2023.pdf, May 2023. Accessed January 2026.

[14] City of Lebanon, New Hampshire. Adm-143 use of artificial intelligence policy. https://lebanonnh.gov/1737/AI-Policy, 2023. Accessed January 2026.

[15] City of Long Beach. Generative ai guidance (v1.1). https://longbeach.gov/globalassets/smart-city/media-library/documents/generative-ai-guidance-v1-1, 2024. Accessed January 2026.

[16] City of Riverside. Administrative manual: Artificial intelligence (ai) policy (03.020.00). https://riversideca.legistar.com/gateway.aspx?ID=feddc017-3af0-4860-8820-56e2737102a7.pdf&M=F, 2024. Accessed January 2026.

[17] City of San Jose. Ai policy 1.7.12 and generative a.i. guidelines. https://www.sanjoseca.gov/your-government/departments-offices/information-technology/itd-generative-ai-guideline, 2025. Accessed January 2026.

[18] City of Seattle. Artificial intelligence (ai) policy (pol-211). https://seattle.gov/documents/Departments/Tech/Privacy/AI/Artificial_Intelligence_Policy-POL211.pdf, 2025. Accessed January 2026.

[19] City of Spring Hill, Tennessee. Resolution 25-120 adopting an artificial intelligence (ai) usage policy (3.04.02). https://www.springhilltn.org/DocumentCenter/View/16200/Resolution-25-120-adopting-and-Artificial-Intelligence-AI-Policy, 2025. Accessed January 2026.

[20] City of Tempe. Ethical artificial intelligence (ai) policy. https://tempe.hylandcloud.com/AgendaOnline/Documents/ViewDocument/ETHICALARTIFICALINTELLIGENCEPOLICY.DOCX.pdf?meetingId=1451&documentType=Agenda&itemId=5692&publishId=9354&isSection=false, 2023. Accessed January 2026.

[21] City of Woodburn, Oregon. Use of artificial intelligence (ai) policy. https://www.woodburn-or.gov/sites/default/files/fileattachments/human_resources/page/17225/ai_policy.pdf, 2024. Accessed January 2026.

[22] County of Santa Cruz. Artificial intelligence appropriate use policy. https://www.santacruzcountyca.gov/portals/0/county/CAO/press%20releases/2023/AIPolicy.09192023.pdf, 2023. Accessed January 2026.

[23] County of Sonoma. 9-6 information technology artificial intelligence (ai) policy. https://sonomacounty.gov/administrative-support-and-fiscal-services/human-resources/employee-resources/administrative-policy-manual/9-6-information-technology-artificial-intelligence-(ai)-policy, 2024. Accessed January 2026.

[24] Federal Plain Language Guidelines. Federal plain language guidelines. https://www.plainlanguage.gov/guidelines/, 2011.

[25] GovAI Coalition. Resources for responsible municipal ai adoption. Technical report, City of San Jose, 2024. https://www.sanjoseca.gov/your-government/departments-offices/information-technology/digital-privacy/ai-reviews-algorithm-register. Accessed January 2026.

[26] S. Grimmelikhuijsen, S. Jilke, A. L. Olsen, and L. Tummers. Behavioral public administration: Combining insights from public administration and psychology. *Public Administration Review*, 77(1):45–56, 2017.

[27] International City/County Management Association. Ai in local government: Survey summary report. Technical report, ICMA, 2024. https://icma.org/sites/default/files/2024-11/AI%20in%20Local%20Gov%20Survey%20Summary%20Report%20Final_0.pdf. Accessed January 2026.

[28] M. Makowska and A. Szura. Syntactic complexity in legal translated texts and the use of plain english: A corpus-based study. *Humanities and Social Sciences Communications*, 10, 2023. https://doi.org/10.1057/s41599-022-01485-x.

[29] C. Martinez and X. Liu. Readability metrics for legal text: A systematic literature review. *arXiv preprint*, 2024. https://arxiv.org/pdf/2411.09497.

[30] Metropolitan Government of Nashville and Davidson County. Ism-20: Artificial intelligence and generative artificial intelligence use. https://www.nashville.gov/sites/default/files/2025-08/ISM-20-Artificial-Intelligence-and-Generative-Artificial-Intelligence-Use.pdf, 2024. Accessed January 2026.

[31] Miami-Dade County Information Technology Department. Artificial intelligence report. https://www.miamidade.gov/technology/library/artificial-intelligence-report.pdf, 2024. Accessed January 2026.

[32] National League of Cities. Artificial intelligence in cities report. Technical report, National League of Cities, 2024. https://www.nlc.org/wp-content/uploads/2025/01/AI-in-Cities-Report.pdf. Accessed January 2026.

[33] J. Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 152–158. ACM, 1994. https://doi.org/10.1145/191666.191729.

[34] Salt Lake City Department of Information Management Services. Generative ai policy guide. https://slcdocs.com/ims/GenAIPolicyGuide.pdf, 2024. Accessed January 2026.

[35] K. A. Schriver. *Dynamics in Document Design: Creating Texts for Readers*. John Wiley & Sons, New York, 1997.

[36] M. Shardlow, R. Evans, and M. Zampieri. Predicting lexical complexity in english texts: The complex 2.0 dataset. *Language Resources and Evaluation*, 2022. https://link.springer.com/article/10.1007/s10579-022-09588-2.

[37] C. R. Sunstein. *Sludge: What Stops Us from Getting Things Done and What to Do about It*. MIT Press, Cambridge, MA, 2022.

[38] T. R. Tyler. *Why People Obey the Law*. Princeton University Press, Princeton, NJ, 2006.

[39] Urban Institute. A new approach to helping local governments navigate generative ai. Technical report, Urban Institute, 2025. https://www.urban.org/urban-wire/new-approach-helping-local-governments-navigate-generative-ai. Accessed January 2026.

Received 06 April 2026; Accepted 29 April 2026