Scoring Local AI Policies: Applying NLP to Evaluate AI Governance Readiness

Katelynn Hernandez, Lawrence Wagner
Github Repository

Abstract

Local government entities are adopting new artificial intelligence (AI) systems to support under-staffed teams that provide various administrative services and make data-driven decisions. Considering the fast-paced adoption of AI systems and tools, developing regulatory frameworks and systematic techniques to assess the governance risk and compliance (GRC) of AI systems is an afterthought in the United States. However, that is not the case internationally; many AI regulations are published almost daily. This project attempts to address the need for usable Governance, Risk, and Compliance (GRC) assessment tools targeting AI by learning from our peers abroad. We propose an organized evaluation tool to examine AI policies in municipalities by scoring them based on semantic agreement with defined governance principles in four main governance focus areas: Govern, Map, Measure, and Manage.

By leveraging natural language processing (NLP) methods and models, high-level process evaluation of local government AI policies can be made possible. We used a combination of real and synthetic data to test the performance of the AI policy evaluation tool. By employing a BERT model fine-tuned for the semantic similarity analysis task to compare a user's AI policy with our golden standard corpus and a retrieval augmented generation pipeline for information extraction, we have the beginnings of a promising AI policy diagnostics tool.

Introduction

The implementation of artificial intelligence (AI) systems by local governments to automate housing allocation, public safety operations, and administrative processes has presented significant problems involving fairness, transparency, and accountability in governance. Though these technologies promise operating efficiencies, these are introduced with no corresponding policy structures able to address the risk to the system (David et al., 2024). Cities are under pressure to balance technological innovations with ethical considerations and thus need tools that translate abstract governance principles into actionable policy designs. Though ethical trade-offs like fairness, interpretability, and transparency are issues around which discussions on AI governance revolve, this study does not directly address those normative issues. Rather, it assesses the extent to which local government documents reflect governance standards by examining the proximity of language in the documents to existing standards.

When translated to local governments, governance frameworks with broader contexts in mind pose major constraints. The Organisation for Economic Cooperation and Development (OECD) has stated principles regarding the responsible use of AI. Still, these do not provide means to assess the presence of governance values within the language of policy (OECD, 2024). The National Institute of Standards and Technology (NIST) AI Risk Management Framework offers an organized structure for risk identification and risk mitigation throughout the lifecycle of AI. Yet, it does not consider the resource limitations and organizational gaps found in city halls (NIST, 2024). More research adds to an awareness of how local governments lack the technical skills and funds to apply these frameworks on the ground (Nonnecke et al., 2025).

An automated evaluation tool was developed to assess municipal AI policy documents across four governance categories: Govern, Map, Measure, and Manage. The system scores alignment against predetermined governance standards using a BERT model, and plain-language explanations are generated by a LLaMA language model referencing regulatory sources to describe scoring outcomes. Structured evaluation and easy interpretation enable the system to assist municipalities in turning governance principles into enforceable and audited policy language.

Background

Scholars and policy institutions engaged in policy making have created several frameworks to inform responsible AI integration into civic life with priority on ethical considerations like equity, explainability, and accountability in institutions. Though these frameworks are clear in theory, they are usually not accompanied by implementation methods specific to the day-to-day operations of local authorities. For example, the OECD presents high-level rules governing AI principles but does not provide any methods to measure the operation of these principles through city policies (OECD, 2024). The lack of linkage between general principles and actionable policies leaves cities without ways to assess their governmental strategies effectively.

The NIST AI Risk Management Framework presents an organized methodology to identify, quantify, and address AI-related risks throughout the development and deployment. Although its documentation and system evaluation focus offers a solid technical basis, the framework does not consider the budget limitations, personnel constraints, and jurisdictional heterogeneity characteristic of local environments (NIST, 2024). Smaller municipalities, for instance, may be unable to fund NIST's comprehensive risk assessment procedures, which are predicated on access to technical specialists. The disconnect reflects an overarching deficiency in applying national-level policies to localizable governing standards (Nonnecke et al., 2025).

Wider tools to measure readiness for AI include the AI Safety Index and confirm the lack of attention given to municipal needs. In that these tools measure risk preparedness on a national level but do not evaluate individual policy documents' quality or detail (Center for AI Safety, 2024), current methods are mostly centered on organizational compliance or national-level strategies and do not provide city and county governments with ways to enforce general guidance under local operating conditions (Smith et al., 2025). Such lack of attention is more concerning when municipalities are directly responsible for applying AI systems to high-consequence domains like housing and public safety.

Efforts to evaluate governance quality in policy documents have traditionally relied on qualitative reviews or high-level compliance checklists. These methods lack the consistency and scalability to systematically assess whether policy language aligns with core governance standards. While natural language processing (NLP) tools show promise, their design often prioritizes classification or summarization over structured policy evaluation. For example, BERT-based models excel at sentence similarity tasks but have rarely been adapted to handle the complexity and variability of municipal policy language (Keras Team, n.d.). This limitation underscores the need for domain-specific adaptations of NLP techniques.

Interpretability continues to be an ongoing difficulty in policy assessment. Retrieval-augmented generation (RAG) presents an approach to anchoring assessment outputs to the foundational material, adding transparency to the outcome. Yet current RAG uses are directed toward general-purpose question answering and not public-sector policy analysis (Warner et al., 2024). Few models pair semantic scoring with document-grounded explanations informed by local agencies' governance priorities. Such an absence denies policymakers the actionable insights necessary to adjust their strategies in instances needing to make granular trade-offs among ethical principles and effective implementation.

3. Methods

The evaluation system comprises three integrated components: document ingestion and preprocessing, sentence-level classification model, and retrieval-based explanation module. Subsequently, each component evaluates the alignment between policy language and known governance standards. Figure 1 depicts an overview of the complete pipeline.

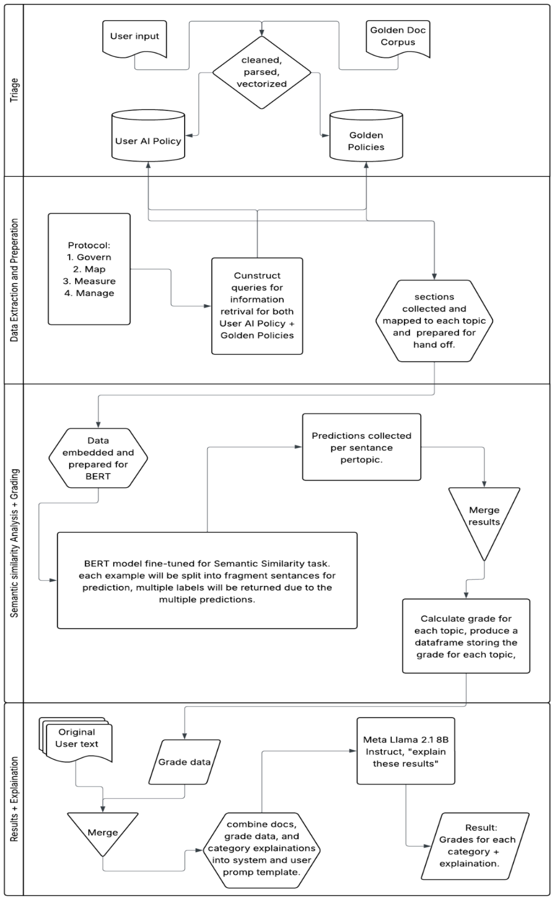
Triage
User input
Golden Doc Corpus
cleaned, parsed, vectorized
User AI Policy
Golden Policies
Data Extraction and Preperation
Protocol:
1. Govern
2. Map
3. Measure
4. Manage
Cunstruct queries for information retrival for both User AI Policy + Golden Policies
sections collected and mapped to each topic and prepared for hand off.
Semantic similarity Analysis + Grading
Data embedded and prepared for BERT
Predictions collected per sentance pertopic.
Merge results
BERT model fine-tuned for Semantic Similarity task. each example will be split into fragment sentances for prediction, multiple labels will be returned due to the multiple predictions.
Calculate grade for each topic, produce a dataframe storing the grade for each topic,
Results + Explaination
Original User text
Grade data
Meta Llama 2.1 8B Instruct, "explain these results"
Merge
combine docs, grade data, and category explainations into system and user promp template.
Result: Grades for each category + explaination.

Figure 1. AI Policy Evaluation Pipeline: Intake, Semantic Scoring, and Explanation Generation

3.1 Document Intake and Vector Encoding

The evaluation process starts with the intake of user-provided AI policy documents. Files uploaded are cleansed to eliminate formatting imperfections and cut into standardized blocks of text. In support of uniformity in format, documents can optionally be converted to markdown. The text thus processed is then inserted into a Chroma vector database with dense vector representations drawn from a pre-trained transformer model. After embedding, these fragments provide the foundation for subsequent testing and explanations. The explanation module employs a retrieval loop whereby the LLM assesses the relevance of fragments retrieved by the vector database to the query intent. If so identified, these fragments are employed to construct an explanation; if not, the system re-queries until a suitable context is retrieved. The testing is an exemplary instance of lightweight LLM reasoning akin to the 'theory of mind' capabilities outlined by (Zhang, 2024), whereby the model cleanses and reasons on contextual inputs before response generation.

3.2 Protocol Mapping and Prompt Engineering

To support structured analysis, policy evaluation is framed by the four governance categories presented above: Govern, Map, Measure, and Manage. Two prompts are aligned with each category to pull relevant material on diverse types of policies. Prompt development underwent an iterative exercise mapping NIST AI Risk Management Framework direction to NLP-conformant question form.

The reference corpus consists primarily of structured material from the NIST AI Risk Management Framework, to which some small sets of annotated instances from the AGORA dataset are added. Arranging this way enables the tool to evaluate to what extent the language employed in municipalities' policy documents conforms to expectations set by prevailing standards of governance practice.

3.3 Fragment Parsing and Preprocessing

Policy documents routinely surpass input lengths allowed by transformer models. In managing this, retrieved snippet fragments are split into overlapping segments employing the sliding window mechanism with 128-tokens-per-window and 64-tokens-per-overlap sizes chosen to maintain local context when crossing segment boundaries and to provide inputs suitable for sentence pair classification. Each fragment is aligned with its related evaluation prompt to form paired-up structured pairs compatible with the target model.

3.4 Semantic Similarity Classification

The evaluation pipeline employs a fine-tuned sentence transformer based on the BERT-base-uncased model to identify the relation of each policy sentence with its evaluation prompt. It is trained on a three-label task: entailment, contradiction, and neutral. Entailment encompasses support with the prompt, contradiction denotes conflict with the prompt, and neutral denotes ambiguous or missing alignment.

Both pairs of sentences are encoded and analyzed with cosine similarity to identify the most likely classification. The three labels are used to arrive at the final category of governance. Entailment and contradiction are assigned positive and negative weights, and neutral predictions are assigned minimal weights to show uncertainty without skewing the outcome.

3.5 Scoring Aggregation Logic

Entailment predictions are added to the category score, whereas contradiction predictions subtract. Neutral predictions are accorded little weight to dampen their effect on the final outcome and represent ambiguous or uncertain alignment. Two prompts are used to assess each governance category, with two fragments retrieved from each prompt. The pairs are labeled separately and are counted to arrive at a raw score by aggregating the label counts. Scores are subsequently normalized to obtain a grade based on percentages for any specific governance category.

3.6 Explanation Layer with Retrieval-Augmented Generation

A LLaMA-based retrieval-augmented generation (RAG) layer generates plain-language explanations for every evaluation to enhance transparency. Supporting excerpts are retrieved from the reference corpus to anchor every explanation to acknowledged governance language (Warner et al., 2024). Outputs can take the form of user policy quotes directly used in plain language alongside organized regulatory references and concise descriptions of how individual statements meet or are deficient in the evaluation standards. The explanation layer allows policymakers to interpret the findings and accordingly adjust their documents.

3.7 Human-in-the-Loop Validation

To complement the automated scoring process, both researchers on this project independently reviewed and manually graded a subset of policy documents. This human-in-the-loop (HITL) validation was a benchmark for the model's semantic label assignments and highlighted areas where the automated tool diverged from human interpretation. Discrepancies informed iterative refinements to the prompt design and scoring logic, strengthening the tool's reliability for practical use.

Evaluation Results

The automated evaluation tool analyzes policies produced by four language models: Claude, Perplexity, ChatGPT, and DeepSeek. All these policies were scored under four categories: Govern, Map, Measure, and Manage. The scores show different levels of compliance with governance standards, and there are areas of strength and gaps between policies and standards.

Claude Policies

The policies generated by Claude had the lowest governance score overall. The Govern category had a 0.0% score with no clear guidelines for oversight, accountability, and assigned responsibilities. While the Map category was strong descriptively, it was not scored because it did not evaluate well. The Measure category scored 66.6%, emphasizing bias and fairness but not concrete measures. There was a Manage score of 33.3%, with minimal procedures for AI tracking, inventory, and incident response. While ethical intention is expressed strongly at points, there is no actionable detail in most categories.

Perplexity Policies

The Perplexity-generated policies performed with mixed results. Govern and Map both obtained 33.3%, as there was minimal clarity regarding responsibilities, AI inventory, and classification processes. Measure obtained 80.0%, as there was a considerably improved understanding of system monitoring and performance. Manage had one of the highest at 85.7%, as a result of included audit procedures, monitoring in real-time, and permanent improvement. Risk management was strong, but the governance framework and visibility into AI system usage remained underdeveloped.

ChatGPT Documents

The policies generated by ChatGPT were balanced in specific categories but lacked substance in others. The Govern category received 50.0%, which confirmed the occurrence of governance language with limited procedural instruction. Map and Measure both received 83.3%, with elaborate portions on AI system mapping and measurement metrics. Nevertheless, Manage received only 16.7%, with minimal mention of incident response and mitigation measures. The document showed a reasonable basis for lifecycle writing and performance measurement but performed poorly at operationalizing risk management.

DeepSeek Documents

DeepSeek-generated policies had the best overall governance alignment. Govern earned a score of 78.6%, reflecting good structures for control and ethical alignment. Manage came second with a score of 81.3%, evidencing well-established procedures for mitigating risks, retraining, and response. Map scored 41.7%, evidencing a partial effort to describe AI system lifecycles but without specifying how those were implemented. While Measure was not thoroughly tested, mention of continuous risk management and considerations of tradeoffs appeared throughout the text.

Overall Assessment

DeepSeek-generated documents had the best overall governance alignment. Govern earned a score of 78.6%, reflecting good structures for control and ethical alignment. Manage came second with a score of 81.3%, evidencing well-established procedures for mitigating risks, retraining, and response. Map scored 41.7%, evidencing a partial effort to describe AI system lifecycles but without specification of how those were implemented. While Measure wasn't fully tested, mention of continuous risk management and considerations of tradeoffs appeared throughout the text.

Conclusion

This study introduces a systematic way to assess municipal AI policy documents using governance metrics derived from frameworks like the NIST AI Risk Framework and the OECD AI Principles. Integrating transformer-based scoring with prompt-based evaluation enables a uniform evaluation of how oversight, lifecycle mapping, performance monitoring, and risk management are addressed in local policies.

Despite a high-level governing language, the tool's application to several samples of municipal policy documents from the AGORA dataset demonstrated gaps in risk classification, lifecycle mapping, and performance monitoring. The findings highlight the disconnect between stated principles and enforceable policy details. Manual scoring by both researchers served as a human-in-the-loop validation method, reinforcing trust in model outputs and guiding improvements to prompt design and scoring logic. To further explore the model’s sensitivity to varied writing styles, we evaluated synthetic policy documents generated by large language models, revealing contrasting governance alignment strengths and confirming the tool’s flexibility across both real and synthetic content. The tool is adaptive and module-based, and expanding the corpus to more jurisdictions with greater diversity is suggested for future work.

Although not intended to drive ethical trade-offs, the tool offers an applied basis for enhancing policy structure and transparency. Policymakers can add these to refine AI governance through inventory procedures, audit cycles, and defined roles. The model built on an optimized BERT model does its finest work for well-structured documents and can overlook subtlety in casually formatted text. Refinement and increased datasets will enable wider application to environments outside its current setup.

Author Contributions

Lawrence Wagner—I came up with the idea and contributed to the research direction definition. I proposed the AGORA dataset and suggested aligning it with the NIST AI Risk Management Framework. I created the grading rubric based on the four main categories: Govern, Map, Measure, and Manage. I guided research for prominent AI governance frameworks, such as the NIST AI Risk Framework, OECD AI Principles, and AI Safety Act to ground our project in actual governance objectives. I wrote the research paper, including the background, related work, results, and evaluation conclusion. I refined the methods and abstract sections and co-developed and graded the synthetic datasets mimicking local AI policies. I assisted with initial experimentation with model creation through training and validation of a DistilBERT model for three-class classification. Though this was not included in the end project, it helped shape what we knew regarding classification issues. I prepared the last slides for presentations and assisted with bridging research to applications for AI governance. Finally, I created the PowerPoint presentation slides except the architecture slides.

Kat Hernandez: I worked on the literature review finding and researching methods and systems used to evaluate policies, holistic, impact, and effectiveness. Research on different NLP tasks to explore all possible options to implementation, BERT tasks and GPT tasks, as well as reasoning strategies for LLMs. I generated and collected data to build custom data sets for our project, I built and modified a total of 3 datasets; 1 dataset was synthetic AI policies, 1 dataset was of real AI regulation and policies from NIST, AGORA, and UC Berkeley documents, and 1 dataset was sentence pairs pulled from these same sources for semantic similarity. I tested out around 5-6 different models from huggingface, some were for embedding/vectorization, a few different BERT models, and some GPT models. I ultimately decided to go with the BERT-based-uncased tokenizer, Keras BERT-based-uncased model fine-tuned for semantic similarity, langchain GPT4ALL embedding model, and Llama 3.1 8B in the project. I built and implemented the document intake and preparation pipeline, vectorization of target data and storage in a Chroma database for retrieval. I also implemented the LLM reasoning strategy (theory of mind) for information retrieval, process and package data for hand off (manual). I also found the helper functions needed to process data to use the Keras BERT model (without the helper function it is not usable, DataGenerator()). I set up the workflow to use the data from the first hand off, implemented the data preparation to run through Keras BERT to predict labels. I implemented the grading process and score calculation for each category (govern, map, measure, manage). Package scored data for 2nd hand off (manual). Implemented the explainability module employing the Llama 3.1 model to explain the scores the user's policy received. I then ran the grading process for the "test docs", "deepseek docs", "claude docs", "perplexity docs", and the "chatgpt docs". I also created and I manage the repo for the project, uploading the 3 jupyter notebooks (data prep notebook, SSBERT notebook, and the LLM RAG notebook), the relevant data artifacts, and the paper. I helped write and edit the paper by doing the initial write up of the methods section and wrote the abstract section. I also created the architecture visual. For the powerpoint, I wrote the 2 slides on the architecture. I also managed the work on the project by delegating tasks and setting up regular meetings to check on progress and facilitate communication.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Goel, V., ... & Amodei, D. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. https://arxiv.org/abs/2212.08073

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 632–642). Association for Computational Linguistics. https://nlp.stanford.edu/projects/snli/

Center for AI Safety. (2024). AI Safety Index 2024. https://safetyindex.org/

David, A., Yigitcanlar, T., Desouza, K., Li, R. Y. M., Cheong, P. H., Mehmood, R., & Corchado, J. (2024). Understanding local government responsible AI strategy: An international municipal policy document analysis. Cities, 155, 105502. https://doi.org/10.1016/j.cities.2024.105502

Emerging Technology Observatory. (2024, December 16). AI Governance and Regulatory Archive (AGORA) (Version 1.11.0) [Dataset documentation]. https://eto.tech/dataset-docs/agora-dataset/

European Union. (2024). EU Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Keras Team. (n.d.). Semantic similarity with BERT. Keras. https://keras.io/examples/nlp/semantic_similarity_with_bert/

National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Nonnecke, B., Newman, J., & Pierson, S. (2025). Responsible AI in the public sector – WaTech & UC Berkeley report. CITRIS Policy Lab and Center for Long-Term Cybersecurity, UC Berkeley.

Organisation for Economic Co-operation and Development. (2024). OECD AI principles: Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Retzlaff, C. O., Das, S., Wayllace, C., Mousavi, P., Afshari, M., Yang, T., Saranti, A., Angerschmid, A., Taylor, M. E., & Holzinger, A. (2024). Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. Journal of Artificial Intelligence Research, 79, 359–415. https://doi.org/10.1613/jair.1.15348

Smith, G., Luka, N., Newman, J., Osborne, M., Nonnecke, B., Lattimore, B., & Mittelstadt, B. (2025). Responsible use of generative AI: A playbook for product managers and business leaders. Berkeley AI Research (BAIR) Responsible AI Initiative.

Stanford Institute for Human-Centered Artificial Intelligence. (2024). AI Index Report 2024. Stanford University. https://hai.stanford.edu/research/ai-index-2024

Warner, C., Chaffin, A., Clavié, B., Weller, O., Hallström, T., Taghadouini, S., ... & Poli, M. (2024). Modern BERT: A survey on recent trends in fine-tuning and prompting. arXiv. https://arxiv.org/abs/2412.13663

Zhang, T., He, W., Sun, M., Yang, Z., Tan, J., Lu, T., Yang, X., Tang, J., & Wang, W. (2024). Strategic reasoning in large language models. arXiv. https://arxiv.org/abs/2404.01230