AI-Powered Rules as Code

Government Briefing | April 23, 2025

GEORGETOWN UNIVERSITY

DBN Community Norms + Code of Conduct

- Being kind and respectful to each other
- Making space and taking space
- Assuming good will
- Staying curious beeckcenter.org/dbncode
- Being mindful and reflective
- Being open and honest
- Taking the best, leaving the rest

beeckcenter.org/dbncode

The Digital Benefits Network is dedicated to providing a harassment-free experience for everyone, We do not tolerate harassment of participants in any form.

If you have any questions or concerns, please email us at digitalbenefits@georgetown.edu.

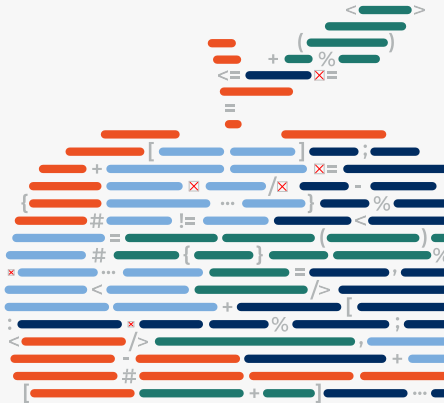
Agenda

Welcome

Research Briefing

- Highlights from our research publication using AI for RaC translation for SNAP and Medicaid (started as part of the Policy2Code Prototyping Challenge)

Open Discussion



GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK

Digital Benefits Network

Supports government in delivering public benefits services and technology that are accessible, effective, and equitable in order to ultimately increase economic opportunity.

SNAP | WIC | Medicaid/CHIP | TANF | Basic Income | UI | Child Care

digitalbenefitsnetwork.org

About the Massive Data Institute

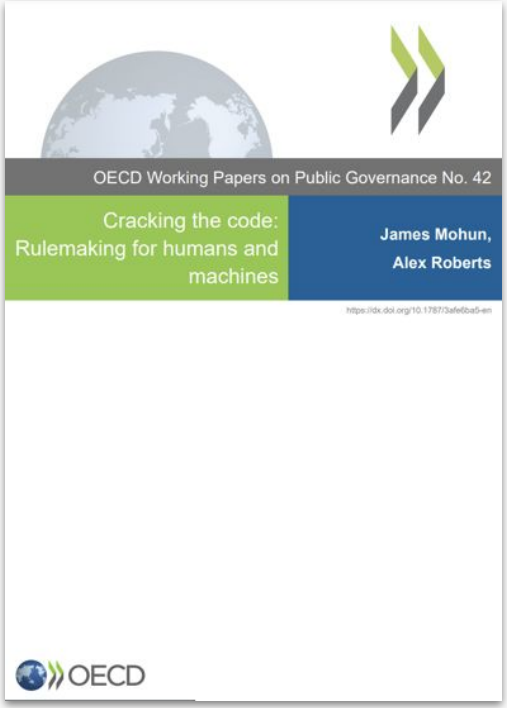
At Georgetown's McCourt School of Public Policy, the **Massive Data Institute (MDI)** is an interdisciplinary research institute that connects experts across computer science, data science, public health, public policy, and social science to tackle societal scale issues and impact public policy in ways that improves people's lives through responsible evidence-based research.

GEORGETOWN UNIVERSITY
McCourt School of Public Policy

MASSIVE DATA INSTITUTE

mdi.georgetown.edu

@MassiveData_GU



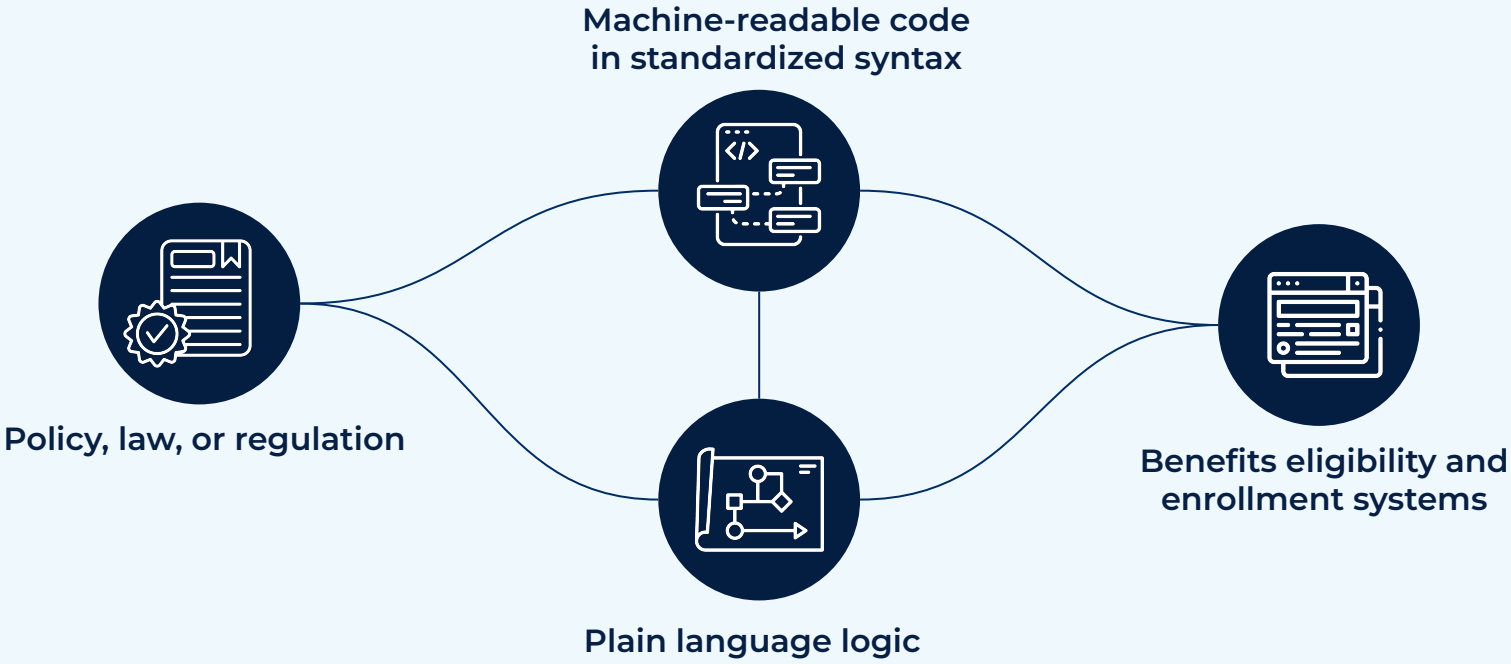
Rules as Code Definition

"an official version of rules (e.g., laws and regulations) in a machine-consumable form, which allows rules to be understood and actioned by computer systems in a consistent way."

- **Organisation for Economic Co-operation and Development**

beeckcenter.org/oecdRAC

Benefits Eligibility Rules as Code



Source: Beeck Center for Social Impact + Innovation
and the Massive Data Institute, Georgetown University, 2025

Our Rules as Code Work

Community — beeckcenter.org/dbncommunity

Research

Upcoming: State Eligibility & Enrollment Systems

2025: AI-Powered Rules as Code

2024: Cross-Sector Insights Report from our Rules as Code Community

2023: Exploring Rules Communication

2022: Benefit Eligibility Rules as Code

beeckcenter.org/dpolicy

Initiatives / Events

Quarterly: Ongoing roundtable sessions

2025: Program design for prototyping open syntax & code library

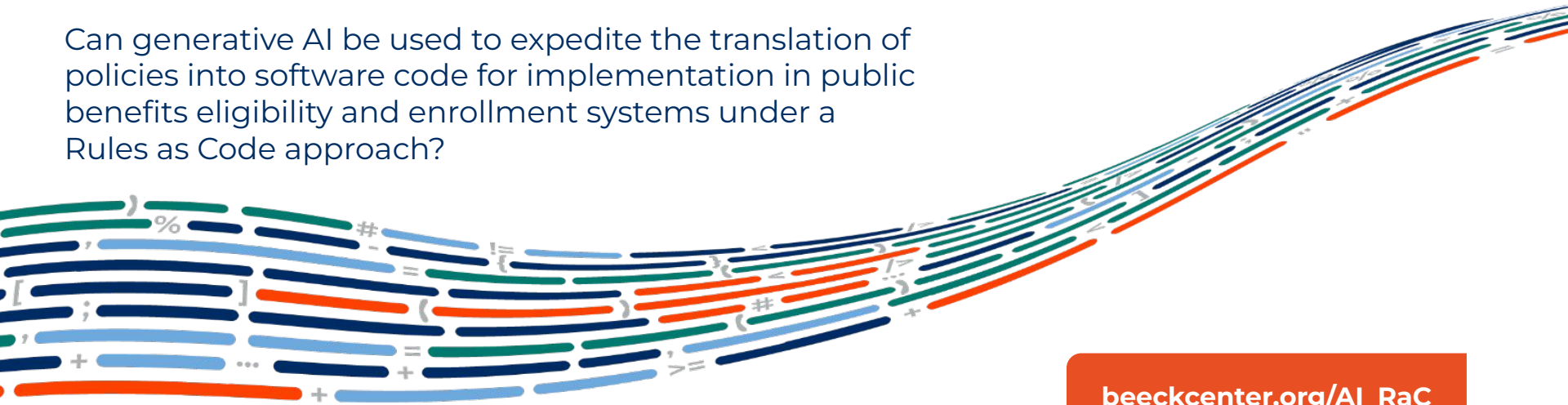
2024: Policy2Code Prototyping Challenge + Demo Day
beeckcenter.org/Policy2Code

2022: Rules as Code Demo Day

AI-Powered Rules as Code

Experiments with Public Benefits Policy

Can generative AI be used to expedite the translation of policies into software code for implementation in public benefits eligibility and enrollment systems under a Rules as Code approach?



beeckcenter.org/AI_RaC

GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK | GEORGETOWN UNIVERSITY McCourt School of Public Policy | MASSIVE DATA INSTITUTE

Policy2Code Prototyping Challenge

Summer meetups:
June - August 2024

12 teams

Demo Day:
September 2024

beeckcenter.org/Policy2Code

Source: Beeck Center for Social Impact + Innovation
and the Massive Data Institute, Georgetown University, 2025

GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK | GEORGETOWN UNIVERSITY McCourt School of Public Policy MASSIVE DATA INSTITUTE

Policy2Code Prototyping Challenge

Hoyas Lex Ad Codex

3 Georgetown Centers
Beeck Center

Massive Data Institute

Center for Security and Emerging Technology (CSET)

Faculty, students, staff collaboration

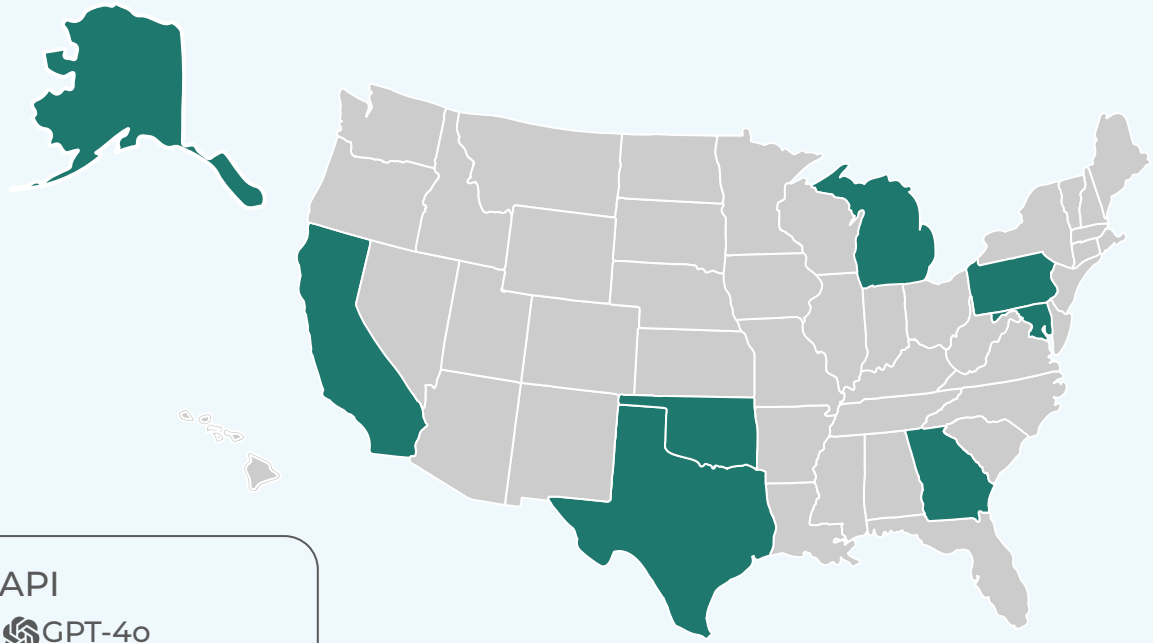
beeckcenter.org/Policy2Code

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

AI-Powered Rules as Code: Research Overview

We ran experiments with SNAP and Medicaid policies in seven states:

- Alaska
- California
- Georgia
- Michigan
- Oklahoma
- Pennsylvania
- Texas



We used the the following LLMs:

Chatbots
- ChatGPT
- Gemini

API
- GPT-4o

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Georgetown University | beeckcenter social impact + innovation | Digital Benefits Network | Georgetown University McCourt School of Public Policy | Massive Data Institute

AI-Powered Rules as Code: Research Overview

Experiment 1	Experiment 2	Experiment 3	Experiment 4
Asking eligibility questions to chatbots and API	**Asking eligibility questions to API guided by policy document**	**Prompting to write benefits eligibility rules using a template with and without policy documents**	**Prompting to generate software code for benefits eligibility**
How well can LLM chatbots answer general SNAP and Medicaid eligibility questions based on their training data and/or resources available on the internet? What factors affect their responses?	How well can an LLM generate accurate, complete, and logical summaries of benefits policy rules when provided official policy documents?	How well can an LLM extract machine-readable rules from unstructured policy documents in terms of output relevance and accuracy? How does the use of a structured rules template impact an LLM's ability to produce relevant and accurate output?	How effectively can an LLM generate software code to determine eligibility for public benefits programs?

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

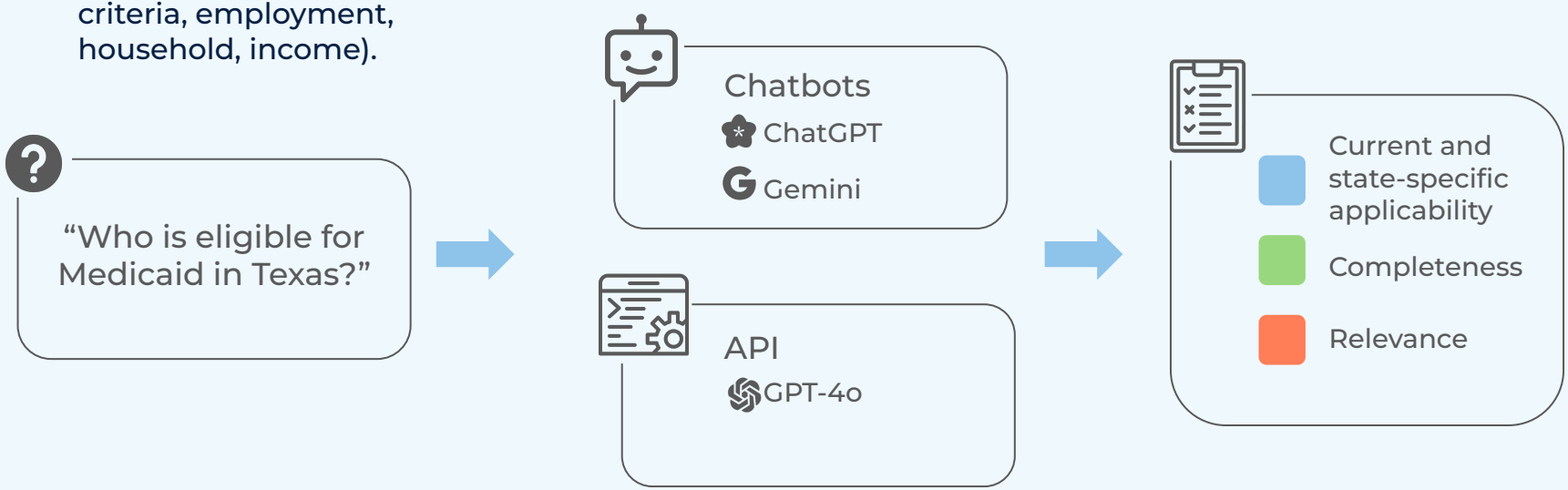
Experiment 1 Methodology

Ask eligibility questions and assess the outputs

1 **Generate prompts across different benefits scenarios and levels of specificity** (e.g. individual eligibility criteria, employment, household, income).

2 **Input prompts into web browser chatbots and API.**

3 **Evaluators score responses 1-5 points based on developed rubric.**

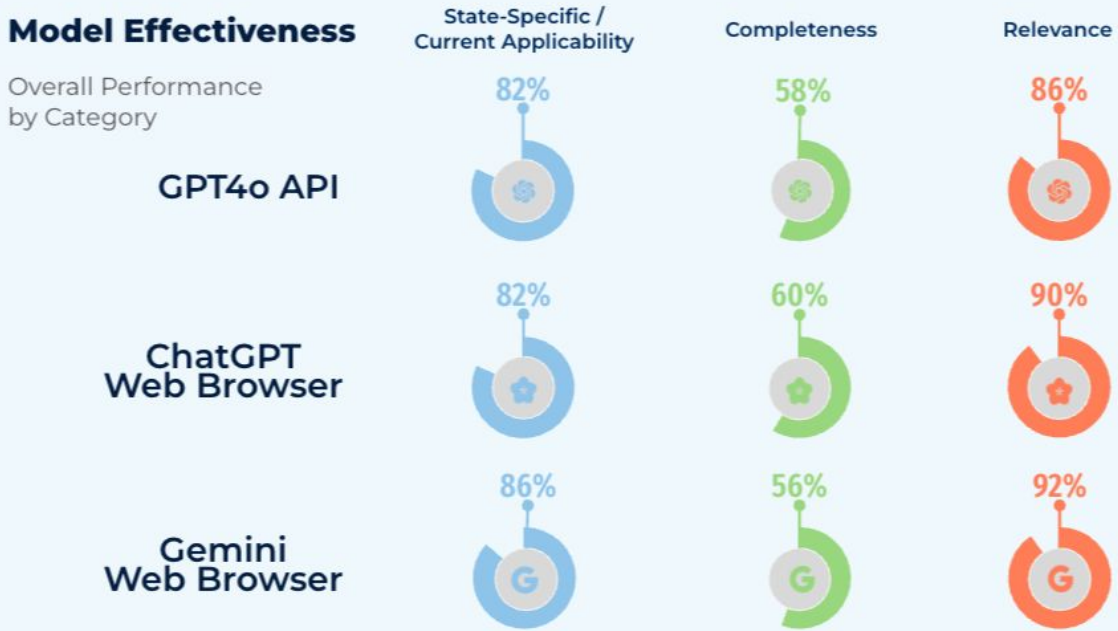


Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK | GEORGETOWN UNIVERSITY McCourt School of Public Policy | MASSIVE DATA INSTITUTE

Experiment 1 Results Summary

Ask eligibility questions and assess the outputs



- No major difference in the three models' performance.
- Perform reasonably well on state-specific /current and relevance, but big drop off for completeness

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 1 Results Summary

Ask chatbots eligibility questions and assess the outputs

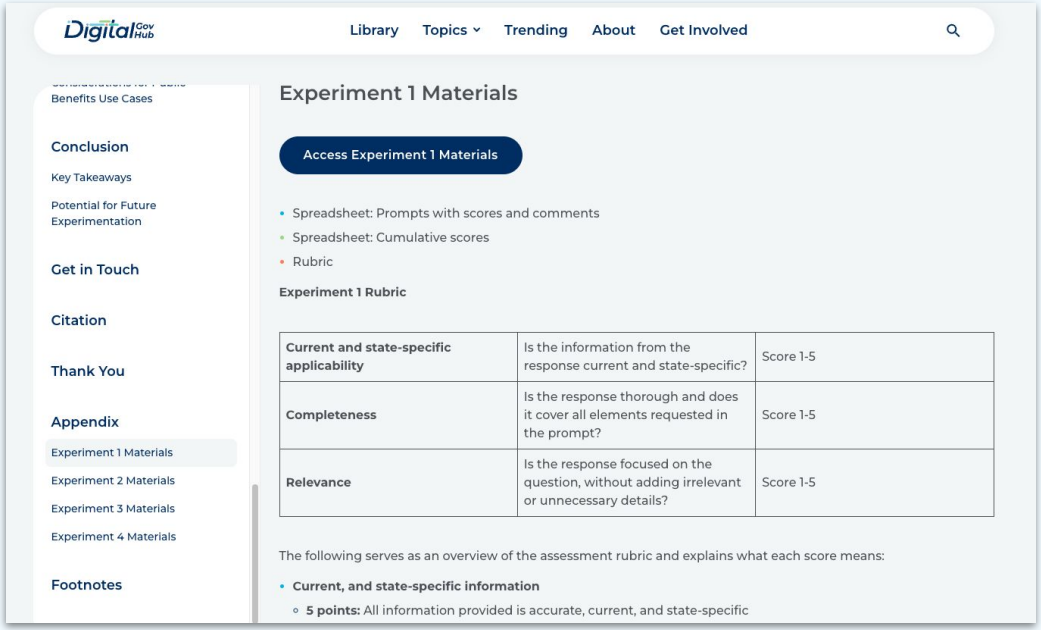
Average Scores By state and program

- Medicaid responses scored slightly higher than SNAP

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK | GEORGETOWN UNIVERSITY McCourt School of Public Policy | MASSIVE DATA INSTITUTE

Experiment 1 Materials



- Spreadsheet: Prompts with scores and comments
- Spreadsheet: Cumulative scores
- Rubric

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 2 Methodology

Collect policies, feed them to the LLM and assess the outputs

1 **Extract and consolidate policy documents.**

State Websites
Policy Manual PDFs

2 **Generate prompts for specific goals**
(e.g. plain language summaries, logic requests, eligibility criteria)

"Based on this document, what are the eligibility criteria for SNAP in Texas?"

3 **Build a RAG environment to capture knowledge about state specific policies for LLM to use.**

API

GPT-4o with RAG

Consolidated State Policy Document

4 **Evaluators score responses 1-5 points based on developed rubric.**

- Accuracy
- Completeness
- Relevance

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 2 Assessment of Policy Documents

1 **Extract and consolidate policy documents.**

State Websites
Policy Manual PDFs

Is your state program guide AI- ready?

Key Challenges:

- **Separate PDF documents** (Georgia)
- **Scanned documents make text unsearchable and unusable for LLMs** (California)
- **Inconsistent formats or interactive content that's hard to download** (Alaska, Pennsylvania)

Observations:

- **Option to download full PDF, easy access** (Michigan, Texas)
- **Separated or scanned non-readable documents** (Georgia, California)
- **Good navigation, but harder to extract as a single PDF** (Alaska, Pennsylvania)

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK | GEORGETOWN UNIVERSITY McCourt School of Public Policy MASSIVE DATA INSTITUTE

Experiment 2 Prompt Development and Response Assessments

2 **Generate prompts for specific goals**
(e.g. plain language summaries, logic requests, eligibility criteria)

? "Based on this document, what are the eligibility criteria for SNAP in Texas?"

State	Policy	Question on Topic	Prompt
PA	SNAP	"Plain language" summary	Please provide a plain language summary of the eligibility criteria for Medicaid in Pennsylvania found in the provided document.

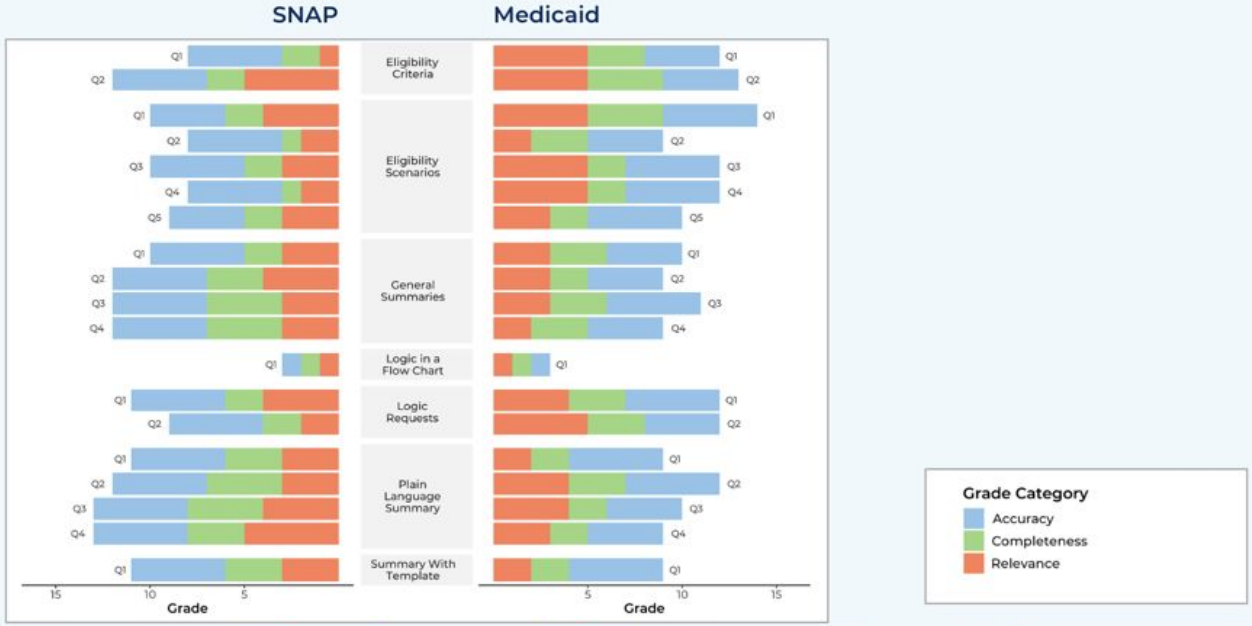
State	Policy	Question on Topic	Prompt
PA	SNAP	Eligibility Scenarios	I am a single adult who cannot work due to a medical condition in Pennsylvania. Based on the provided document, what would determine if I am eligible for SNAP?

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 2 Results Summary

Collect policies, feed them to the LLM and assess the outputs

Prompt Effectiveness for SNAP and Medicaid in Pennsylvania



- Generally returns accurate results; however, scores drop off significantly for completeness and relevance.

How did your state perform? See the report!

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK | GEORGETOWN UNIVERSITY McCourt School of Public Policy | MASSIVE DATA INSTITUTE

Experiment 2 Common Errors Found in Responses

Accuracy

Common errors:

- Misinterpreting criteria (like age or group eligibility)
- Confusing recent vs. still-applicable older policies.

Completeness

Common errors:

- Missing details for eligibility,
- Focus on admin information (timelines, verification)

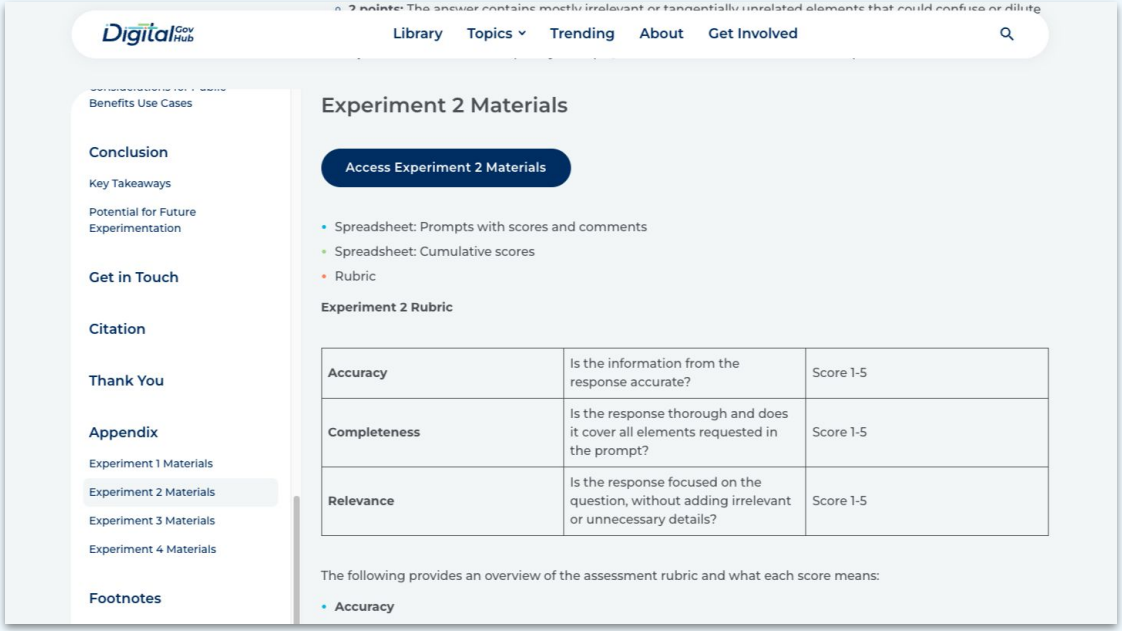
Relevance

Common errors:

- Off-Topic Answers
- Irrelevant pieces of Information

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 2 Materials



- Spreadsheet: Prompts with scores and comments
- Spreadsheet: Cumulative scores
- Rubric

Source: Beeck Center for Social Impact + Innovation
and the Massive Data Institute, Georgetown University, 2025

Experiment 3 Methodology

Rules generation

1. Develop rules template
2. Prompt LLM to generate rules under different conditions
3. Evaluate responses

```yaml
program: SNAP
state: "<STATE>"

eligibility:
  income_limits:
    gross: "<AMOUNT>"
    net: "<AMOUNT>"
  asset_limit: "<AMOUNT>"
  citizenship: ["<STATUS>"]
  work_requirements:
    age_range: ["<AGE>", "<AGE>"]
    hours: "<HOURS>"

benefits:
  max_allotment: "<AMOUNT>"
  standard_deduction: "<AMOUNT>"

certification:
  period: "<MONTHS>"

reporting:
  simplified: "<BOOLEAN>"
```

Simplified SNAP Rules Template

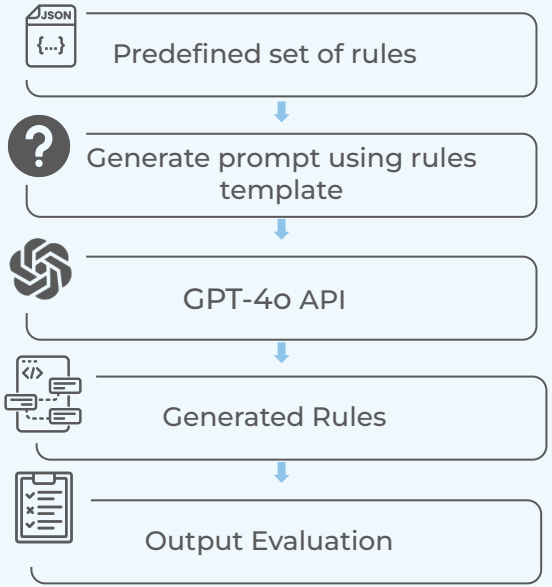
Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 3 Experimental Design

Develop rules template, prompt LLM to generate rules with and without policy documents

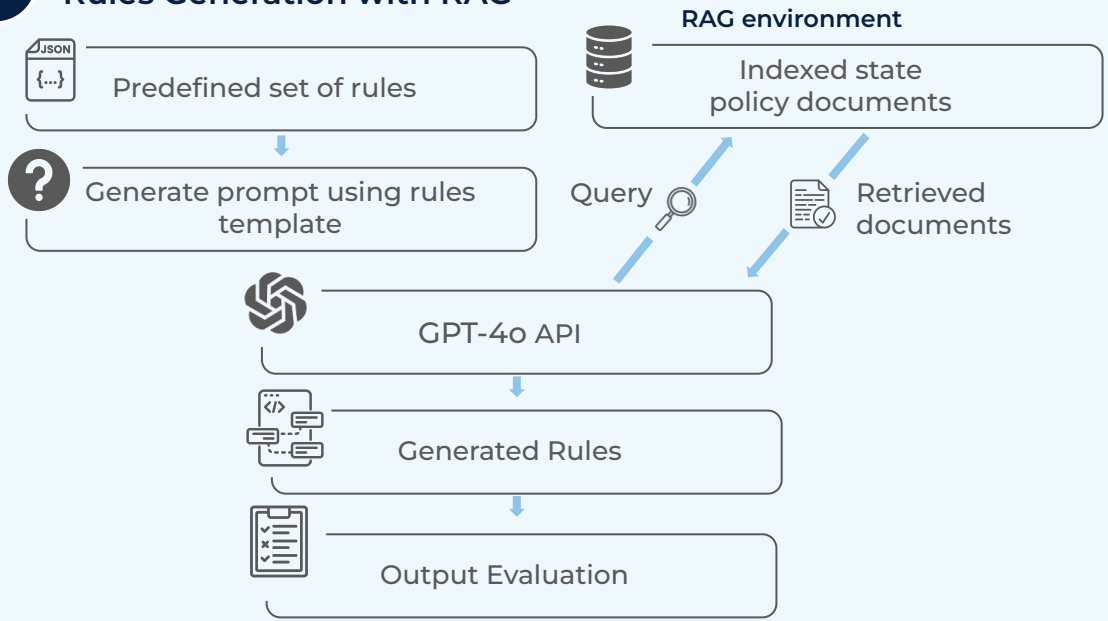
Approach

1 Template-based Rules Generation



Approach

2 Template-based Rules Generation with RAG



Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK | GEORGETOWN UNIVERSITY McCourt School of Public Policy | MASSIVE DATA INSTITUTE

Experiment 3 Results Summary

Develop rules template, prompt LLM to generate rules with and without policy documents

Average Rules Generation Accuracy

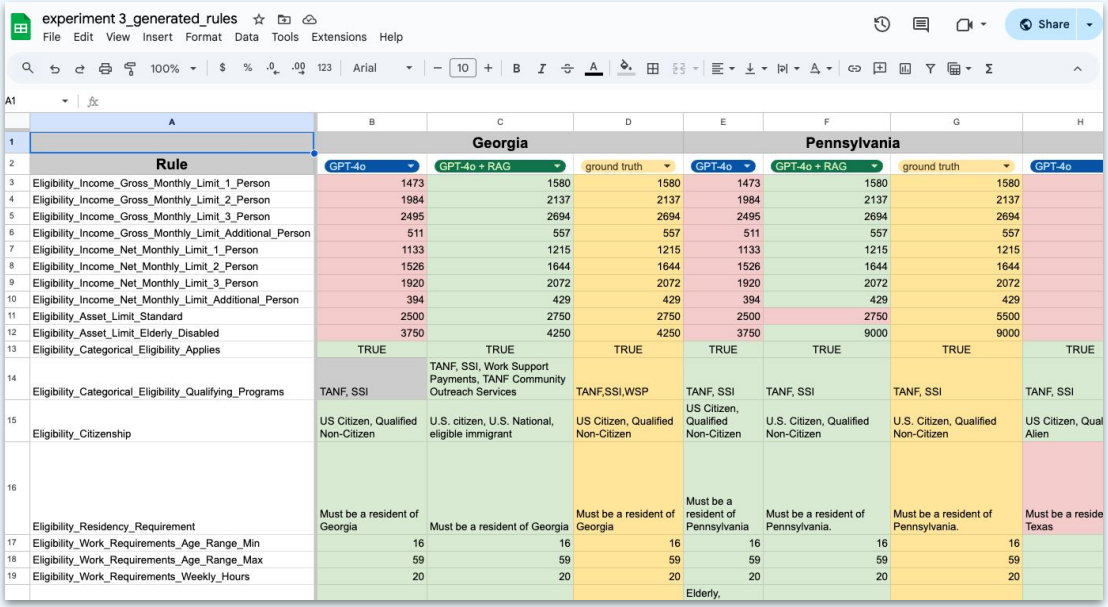
By state and category as percentages

	Georgia		Pennsylvania		Texas	
	LLM	LLM+RAG	LLM	LLM+RAG	LLM	LLM+RAG
Income	0	100	0	100	0	100
Asset Limit	0	100	0	50	0	100
Categorical Eligibility	75	100	100	100	100	100
Citizenship & Residency	100	100	100	100	0	100
Monthly Allotment	20	100	0	100	20	40
Work Requirements	100	100	100	100	100	100
Deductions	50	100	21	64	29	100
Certification Period	50	0	50	100	50	100

- Plain prompting yielded unreliable generation accuracy for state-specific rules.
- RAG enhanced the alignment of generated rules with policy documents.
  - Structured templates are essential for rules extraction.
  - Rules generation can provide a structure for code generation.

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 3 Materials



experiment 3_generated_rules

File Edit View Insert Format Data Tools Extensions Help

Share

	A	B	C	D	E	F	G	H
1		Georgia			Pennsylvania			
2	Rule	GPT-4o	GPT-4o + RAG	ground truth	GPT-4o	GPT-4o + RAG	ground truth	GPT-4o
3	Eligibility_Income_Gross_Monthly_Limit_1_Person	1473	1580	1580	1473	1580	1580	
4	Eligibility_Income_Gross_Monthly_Limit_2_Person	1984	2137	2137	1984	2137	2137	
5	Eligibility_Income_Gross_Monthly_Limit_3_Person	2495	2694	2694	2495	2694	2694	
6	Eligibility_Income_Gross_Monthly_Limit_Additional_Person	511	557	557	511	557	557	
7	Eligibility_Income_Net_Monthly_Limit_1_Person	1133	1215	1215	1133	1215	1215	
8	Eligibility_Income_Net_Monthly_Limit_2_Person	1526	1644	1644	1526	1644	1644	
9	Eligibility_Income_Net_Monthly_Limit_3_Person	1920	2072	2072	1920	2072	2072	
10	Eligibility_Income_Net_Monthly_Limit_Additional_Person	394	429	429	394	429	429	
11	Eligibility_Asset_Limit_Standard	2500	2750	2750	2500	2750	5500	
12	Eligibility_Asset_Limit_Elderly_Disabled	3750	4250	4250	3750	9000	9000	
13	Eligibility_Categorical_Eligibility_Applies	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
14	Eligibility_Categorical_Eligibility_Qualifying_Programs	TANF, SSI	TANF, SSI, Work Support Payments, TANF Community Outreach Services	TANF,SSI,WSP	TANF, SSI	TANF, SSI	TANF, SSI	TANF, SSI
15	Eligibility_Citizenship	US Citizen, Qualified Non-Citizen	U.S. citizen, U.S. National, eligible immigrant	US Citizen, Qualified Non-Citizen	US Citizen, Qualified Non-Citizen	U.S. Citizen, Qualified Non-Citizen	U.S. Citizen, Qualified Non-Citizen	US Citizen, Qual Alien
16	Eligibility_Residency_Requirement	Must be a resident of Georgia	Must be a resident of Georgia	Must be a resident of Georgia	Must be a resident of Pennsylvania	Must be a resident of Pennsylvania.	Must be a resident of Pennsylvania.	Must be a reside Texas
17	Eligibility_Work_Requirements_Age_Range_Min	16	16	16	16	16	16	
18	Eligibility_Work_Requirements_Age_Range_Max	59	59	59	59	59	59	
19	Eligibility_Work_Requirements_Weekly_Hours	20	20	20	20	20	20	
					Elderly,			

- Spreadsheet: Contains the generated rules from both the plain prompting and RAG approaches, compared against the verified SNAP Policy values in Georgia, Pennsylvania, and Texas.

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 4 Methodology

Code generation

1. Provide Policy
2. Prompt LLM to generate code under different conditions
3. Evaluate code

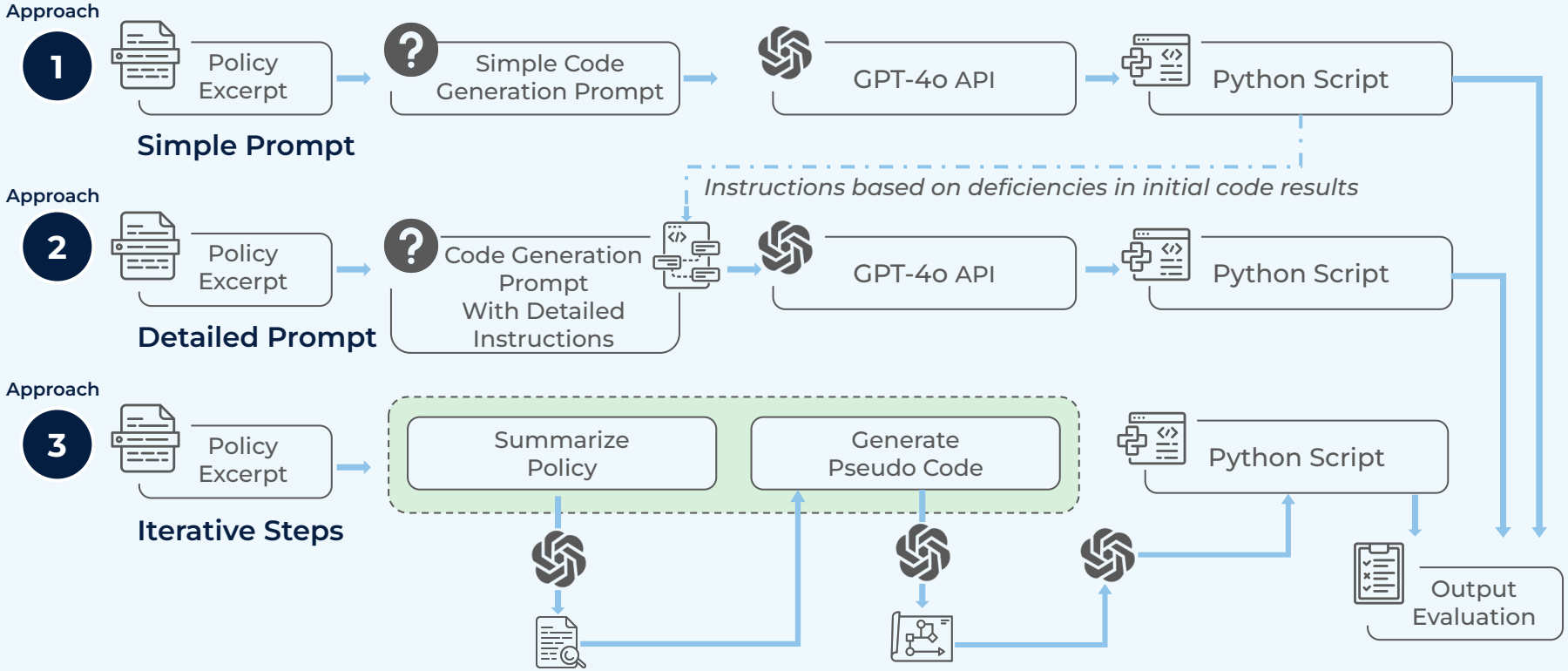
```
# Step 6: Documentation and Structure
# (This is done through comments in the code)

# Step 7: Rule Storage
# Save the decision-making rules in a JSON file
# [STRENGTH] Externalized rule storage shows understanding of maintainability
rules = {
    "exemptions": [
        {"age_less_than": 16},
        {"age_between": [16, 17], "head_of_au": False},
        {"age_greater_equal": 60},
        {"tanf_recipient": True},
        {"physically_unfit": True},
        {"mentally_unfit": True},
        {"pregnancy_status": True, "pregnancy_risk": True},
        {"chronic_homelessness": True},
        {"disability_status": True},
        {"student_status": True, "student_enrollment": True, "student_type": ["GED", "home_school", "high_school", "online_courses",
"post_secondary", "seminary", "trade_school"]},
        {"caretaker_child": True},
        {"caretaker_incapacitated": True},
        {"employment_hours_greater_equal": 30},
        {"employment_earnings_greater_equal": 7.25 * 30},
        {"vista_volunteer": True},
        {"migrant_seasonal_worker": True},
        {"drug_treatment_program": True},
        {"ucb_status": True},
        {"homeless_status": True},
        {"veteran_status": True},
        {"in_foster_care": True}
    ],
    "registrants": [
        {"age_between": [16, 59]}
    ]
```

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 4 Experimental Design

Generate code using simple and detailed LLM prompts and iterative steps



Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK | GEORGETOWN UNIVERSITY McCourt School of Public Policy | MASSIVE DATA INSTITUTE

Experiment 4 Results Summary

Generate code using simple and detailed LLM prompts and iterative steps

Performance of Designs Across Criteria

Criteria	Design 1 *Simple Prompt*	Design 2 *Detailed Prompt*	Design 3 *Iterative Prompts*
Variable Identification	Partial	Good	Good
Input Handling	Poor	Partially good	Poor
Output Correctness	Incorrect	Partially correct	Incorrect
Decision Making	Poor	Partially good	Poor (mechanical)
Logical Consistency	Poor	Improved	Partially consistent
Rule Coverage	Partially covered	Improved	Partially covered
Code Execution	Runs (unstable)	Runs (Improved)	Doesn't run

- At a high level, the detailed prompt design was the most successful.
- Summarized policy guidance for code generation reduces code errors.
- Modular design is particularly important in LLM workflows.

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Experiment 4 Materials

```
# Step 6: Documentation and Structure
# (This is done through comments in the code)

# Step 7: Rule Storage
# Save the decision-making rules in a JSON file
# [STRENGTH] Externalized rule storage shows understanding of maintainability
rules = {
    "exemptions": [
        {"age_less_than": 16},
        {"age_between": [16, 17], "head_of_au": False},
        {"age_greater_equal": 60},
        {"tanf_recipient": True},
        {"physically_unfit": True},
        {"mentally_unfit": True},
        {"pregnancy_status": True, "pregnancy_risk": True},
        {"chronic_homelessness": True},
        {"disability_status": True},
        {"student_status": True, "student_enrollment": True, "student_type": ["GED", "home_school", "high_school", "online_courses",
"post_secondary", "seminary", "trade_school"]},
        {"caretaker_child": True},
        {"caretaker_incapacitated": True},
        {"employment_hours_greater_equal": 30},
        {"employment_earnings_greater_equal": 7.25 * 30},
        {"vista_volunteer": True},
        {"migrant_seasonal_worker": True},
        {"drug_treatment_program": True},
        {"ucb_status": True},
        {"homeless_status": True},
        {"veteran_status": True},
        {"in_foster_care": True}
    ],
    "registrants": [
        {"age_between": [16, 59]}
    ]
```

- Prompts: Includes the system and user prompts used for code generation. For the iterative approach, there are three sets of prompts for each stage, as outlined in the experiment details.
- Generated Code: Contains the code results from each approach, with our manual annotations highlighting strengths [STRENGTH] and weaknesses [WEAKNESS].

Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Key Takeaways: Rules as Code Generation

- **LLMs can help support the Rules as Code pipeline.**
- **Humans in the loop:** Accuracy and equity considerations must outweigh efficiency in high-stakes benefits systems.
- **Authoritative sources:** It is possible to improve the performance of the benefits-related responses by pointing LLMs to authoritative sources like policy manuals.
- State governments can make it easier for LLMs to use their policies by making them digitally accessible.
- Asking LLMs to write policy code directly leads to inconsistent code quality.
- Code generation can be improved by using an LLM with RAG to generate machine-readable policy rules, but requires a manually-curated template.

Source: Beeck Center for Social Impact + Innovation
and the Massive Data Institute, Georgetown University, 2025

Key Takeaways: Impacts on People

- Mixed LLM results have a direct impact on people seeking or receiving benefits, **risking incorrect information** when they ask generative AI models questions about programs like SNAP and Medicaid.
- When AI models provide incorrect information, they often do so in a **confident tone,** which can mislead those without subject expertise.

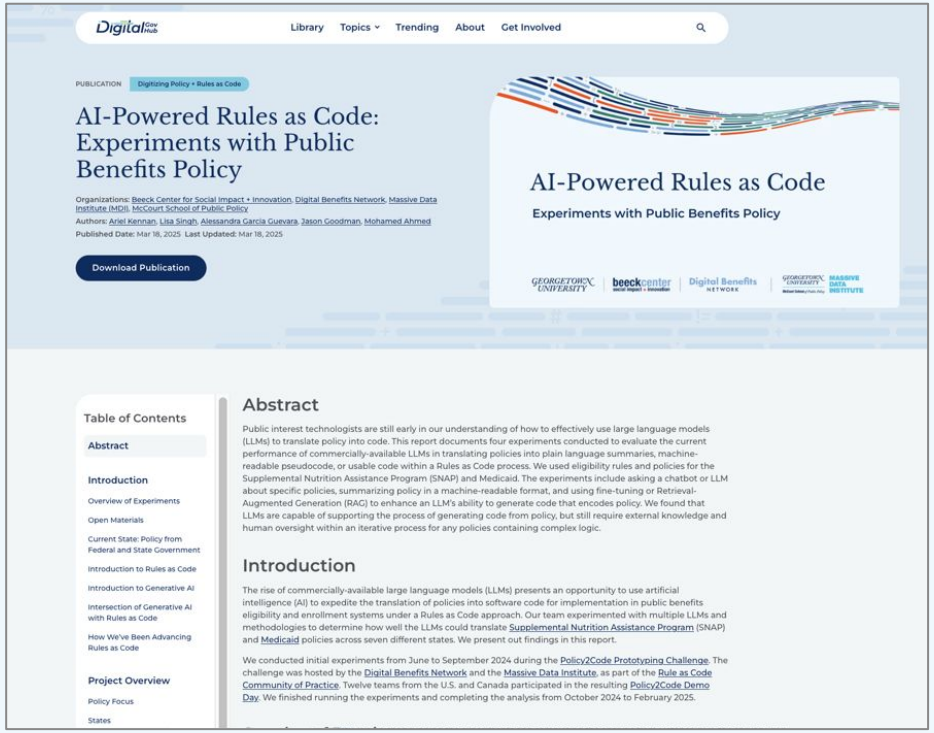
Source: Beeck Center for Social Impact + Innovation and the Massive Data Institute, Georgetown University, 2025

Future Experimentation

- Repeat methodologies for other states, programs, policies or LLMS.
- Repeat for different programming languages for outputs.
- Pair between policy experts and software engineers to program and evaluate.
- Explore specific program or policy LLM
- Compare rules against existing systems
- Explore extracting code from legacy systems.
- Compare code writing efforts to policy analysis efforts.
- Design a prototype toolkit to test different rules and code generation prompts for SNAP.

Source: Beeck Center for Social Impact + Innovation
and the Massive Data Institute, Georgetown University, 2025

AI-Powered Rules as Code



Full report

Summary + Key Takeaways

Rubrics
Prompts
Results
Scores
Code Samples

beeckcenter.org/AI_RaC

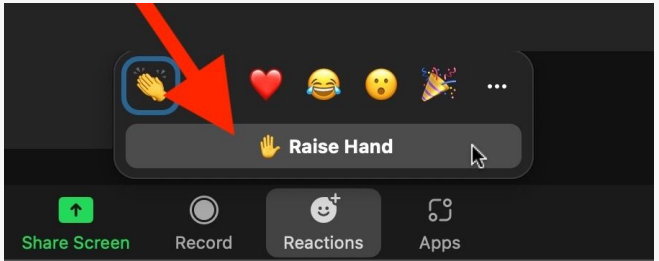
Source: Beeck Center for Social Impact + Innovation
and the Massive Data Institute, Georgetown University, 2025

GEORGETOWN UNIVERSITY | beeckcenter social impact + innovation | Digital Benefits NETWORK | GEORGETOWN UNIVERSITY McCourt School of Public Policy | MASSIVE DATA INSTITUTE

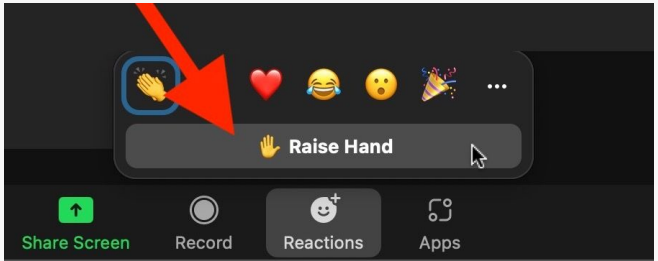
AI-Powered Rules as Code

Experiments with Public Benefits Policy

Questions?

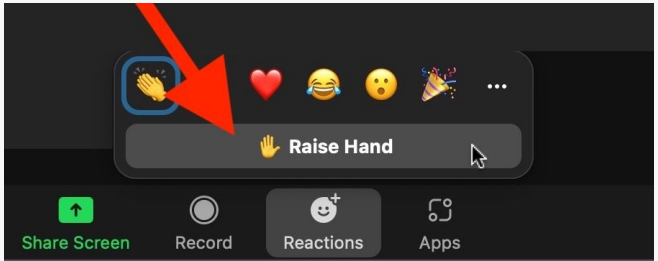


Open Discussion



Guide Questions

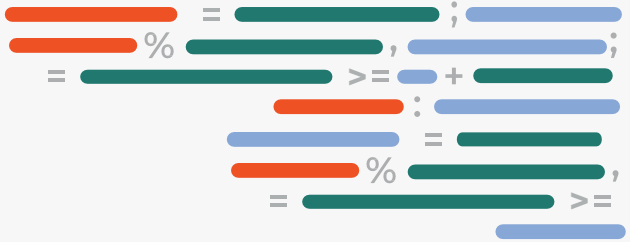
- How have you been using AI with policy and program implementation?
- Are there any new wins? New challenges?
- Do you have questions for this group?



Survey

Rate the Briefing

Thank you!



@digitalbenefits.bsky.social

Digital Benefits Network

rulesascode@georgetown.edu

Join our Rules as Code Community of Practice

beeckcenter.org/raccop