PolicyBench

Benchmarking tax-and-benefit reasoning in frontier language models

PolicyEngine

Table of contents

Source: Article Notebook

1 Abstract

PolicyBench evaluates whether frontier language models can estimate household tax and benefit outputs from household facts without tools. The benchmark covers the United States and the United Kingdom, uses nationally calibrated household scenarios, and scores outputs with a bounded 0-100 metric. In the frozen snapshot used for this manuscript (2026-04-04), the top US model is gemini-3.1-pro-preview at 73.3, the top UK model is grok-4.20 at 83.0, and the top shared global model is gemini-3.1-pro-preview at 78.0. In both countries, the lowest-scoring outputs are tax-base variables rather than benefits or credits. The live benchmark is available at https://policybench.org.

2 Introduction

Household tax-and-benefit estimation sits between arithmetic and policy discussion. Each case has numeric labels, but generating those labels requires filing status, household composition, income concepts, program thresholds, and jurisdiction-specific rules. PolicyBench measures whether models can map household records to those outputs.

Most current language-model evaluations do not test this mapping directly. General math benchmarks emphasize symbolic manipulation and exact final answers (Cobbe et al. 2021; Hendrycks et al. 2021), while generic question-answering benchmarks emphasize recall or instruction following. PolicyBench instead asks whether models can transform household facts into policy outputs without access to tools or simulators.

This matters for public-facing calculators, analyst workflows, and tax-preparation or screening systems. The benchmark is intended to separate verbal policy fluency from household-level quantitative prediction.

3 Related work

The most relevant prior work is tax and statutory reasoning. SARA frames statutory interpretation, including tax-relevant rule application, as a language-understanding problem (Holzenberger et al. 2021). LegalBench broadens this view and includes tax-oriented numeric tasks such as sara_numeric (Guha et al. 2023). RuleArena evaluates rule-guided reasoning in regulation-style settings (Kim et al. 2024), and TaxCalcBench evaluates frontier models on tax calculation from structured return-like inputs (Mahesh et al. 2025). Shanahan et al. report a similar split on the VITA test: models do better on tax knowledge questions than on open-ended calculation tasks (Shanahan et al. 2025).

PolicyBench also sits within a broader literature on quantitative reasoning benchmarks. Canonical math evaluations such as GSM8K and MATH normalized exact final-answer scoring for

numeric tasks (Cobbe et al. 2021; Hendrycks et al. 2021). That norm is less natural for continuous-valued policy outputs, where being within 1 percent is closer than being off by an order of magnitude. Recent work has questioned pure exact-match scoring in numeric QA and temporal reasoning, arguing for error-aware metrics instead (Zhou et al. 2025). PolicyBench follows that direction by combining exactness with near-miss thresholds rather than collapsing everything into exact match alone.

Finance-domain benchmarks provide another relevant precedent. FinQA and TAT-QA both show that realistic quantitative reasoning often requires combining structured numbers with domain knowledge rather than solving stylized school-math problems (Chen et al. 2021; Zhu et al. 2021). PolicyBench differs in that its reference outputs are generated by executable tax-benefit microsimulation rather than annotated document questions, but the underlying motivation is similar: domain-specific quantitative reasoning deserves dedicated evaluation.

Finally, PolicyBench depends on two infrastructure literatures that are not themselves LLM benchmarks. One is structured-output reliability, since the benchmark relies on multi-output JSON responses and parse coverage rather than free-form prose (Shorten et al. 2024). The other is tax-benefit microsimulation, where systems such as EUROMOD provide the methodological precedent for evaluating policy rules over household microdata (Sutherland et al. 2023). The UK side of PolicyBench also builds on PolicyEngine's recent work on survey-backed calibration and imputation (Woodruff and Ghenis 2026). There is also operational work on applying and evaluating AI systems in public-benefits settings, including caseworker-assist and SNAP-focused evaluations (Nava Labs 2026, 2025; ZenML LLMOps Database 2025). We are not aware of a public cross-model benchmark over household-level benefit outputs.

4 Benchmark design

PolicyBench asks models to predict all benchmark outputs for a household in a single structured response. One response per household reduces repeated prompt cost and keeps the task at the household-return level rather than turning it into a sequence of unrelated one-output calls.

The benchmark uses a bounded 0-100 score. For amount variables, the score averages four hit rates:

1. exact
2. within 1%
3. within 5%
4. within 10%

For binary outputs, the score is exact accuracy. This keeps 100 as the ceiling while still giving partial credit for near misses on amount outputs. PolicyBench also tracks mean absolute error and related diagnostics, but those are secondary to the bounded score. This choice preserves

exact-match comparability while avoiding the failure mode that recent numeric-evaluation papers have criticized (Zhou et al. 2025).

The benchmark distinguishes between benchmark runs and diagnostic runs. Benchmark runs are the canonical leaderboard artifacts. Diagnostic runs use smaller sidecar samples and may request optional explanations. Those explanation runs are used for qualitative analysis and are not part of the headline ranking.

PolicyBench also has an earlier archived tool-assisted pilot. That condition used a single untuned function-calling contract rather than a multi-step agent setup. The only tool was calculate_policy(household, variable, year), where household was a standard PolicyEngine-US household JSON object and the archived scenarios supplied only basic fields needed for those tests: person ages, employment income, and household state. The prompt was a short instruction to calculate one output, use the tool with the given variable name and year, and return the numeric result from the tool. The archived US run covers 4,200 predictions (100 households x 14 outputs x 3 models: Claude Opus, Claude Sonnet, and GPT-5.2), and all 4,200 predictions match the simulator reference values exactly. We do not mix that condition into the current leaderboard because it measures whether the model can route the request to the simulator and use the returned value, not whether it can estimate the output without tools.

5 Data and scenario construction

5.1 United States

The US benchmark is built from Enhanced CPS-derived households using PolicyEngine US. The sampled households are filtered to keep a single-tax-unit structure while retaining variation in filing status, household composition, and income sources. Prompts include nonzero promptable raw inputs across relevant entities rather than a hand-curated summary, so the models see many of the same facts the simulator receives.

The current US release evaluates 13 outputs spanning federal tax, credits, benefits, and state-tax quantities. These include federal adjusted gross income, tax before refundable credits, EITC, CTC, SNAP, SSI, and multiple state-tax outputs.

5.2 United Kingdom

The UK benchmark is built from a calibrated public transfer dataset scored through PolicyEngine UK. The current public build starts from a public export of benchmark-compatible households from PolicyEngine US Enhanced CPS, maps those records into UK-facing inputs, and recalibrates them to selected UK targets. This creates a public UK benchmark path without publishing restricted household microdata. The current UK release evaluates six outputs:

income tax, national insurance, child benefit, universal credit, pension credit, and personal independence payment.

The UK data path is more synthetic than the enhanced FRS pipeline, so the benchmark should not be read as a substitute for native UK microdata work. It supports the current public cross-country benchmark, but it is not equivalent to an enhanced-FRS-based benchmark (Sutherland et al. 2023; Woodruff and Ghenis 2026).

6 Results

6.1 United States leaderboard

The US leaderboard for the frozen manuscript snapshot is shown in Table 1. The top three models in that snapshot are gemini-3.1-pro-preview (73.4), grok-4.20 (71.7), and gemini-3-flash-preview (69.5).

Source: Article Notebook

Table 1: Top US benchmark models in the current release.

	Model	Score	Exact	Within 10%	Parsed	Total
0	gemini-3.1-pro-preview	73.4	67.6	80.4	12974	12974
1	grok-4.20	71.7	66.3	78.5	13000	13000
2	gemini-3-flash-preview	69.5	64.5	75.7	13000	13000
3	gpt-5.4	67.2	62.7	72.7	13000	13000
4	claude-opus-4.6	66.2	61.4	71.9	12558	12558

Source: Article Notebook

6.2 United Kingdom leaderboard

The UK leaderboard for the frozen manuscript snapshot is shown in Table 2. The top three models in that snapshot are grok-4.20 (83.0), gemini-3.1-pro-preview (82.6), and gemini-3-flash-preview (71.6).

Source: Article Notebook

Table 2: Top UK benchmark models in the current release.

	Model	Score	Exact	Within 10%	Parsed	Total
0	grok-4.20	83.0	76.0	90.0	6000	6000
1	gemini-3.1-pro-preview	82.6	74.7	90.5	6000	6000
2	gemini-3-flash-preview	71.6	64.4	79.5	6000	6000
3	gpt-5.4	68.9	62.7	76.1	6000	6000
4	gemini-3.1-flash-lite-preview	68.7	62.9	76.3	6000	6000

Source: Article Notebook

6.3 Global shared-model leaderboard

The global leaderboard averages US and UK scores for the models that have been run in both countries. In the frozen manuscript snapshot, the top three are gemini-3.1-pro-preview (78.0), grok-4.20 (77.3), and gemini-3-flash-preview (70.6). The global table summarizes cross-country performance; the country tables identify the outputs driving the differences.

Source: Article Notebook

Table 3: Global shared-model leaderboard across US and UK runs.

	Model	Score	Exact	Within 10%	Parsed	Total
0	gemini-3.1-pro-preview	78.0	71.1	85.4	18974	18974
1	grok-4.20	77.3	71.2	84.2	19000	19000
2	gemini-3-flash-preview	70.6	64.4	77.6	19000	19000
3	gpt-5.4	68.0	62.7	74.4	19000	19000
4	gemini-3.1-flash-lite-preview	67.3	62.2	73.9	19000	19000
5	claude-opus-4.6	67.1	61.7	73.8	18533	18558
6	claude-sonnet-4.6	66.4	61.9	72.1	18974	18974
7	gpt-5.4-mini	64.5	62.3	67.2	19000	19000
8	claude-haiku-4.5	63.2	60.9	66.5	19000	19000
9	gpt-5.4-nano	61.8	60.8	63.2	19000	19000

Source: Article Notebook

6.4 Hardest benchmark targets

The lowest-scoring US variables are tax-base and state-tax quantities rather than benefits. As shown in Table 4, household state income tax scores 31.5, state AGI scores 32.9, federal

income tax before refundable credits scores 33.2, and adjusted gross income scores 37.1. By comparison, EITC scores 86.3, SSI scores 93.1, and free school meals scores 94.1.

Source: Article Notebook

Table 4: Hardest US variables by bounded score.

	Variable	Score	Exact	Within 10%
4	household_state_income_tax	31.9	28.2	37.3
10	state_agi	34.0	20.6	47.5
5	income_tax_before_refundable_credits	34.0	30.3	39.9
0	adjusted_gross_income	39.6	21.3	57.7
11	state_income_tax_before_refundable_credits	40.4	36.4	46.1

Source: Article Notebook

The same pattern appears in the UK run. Income tax scores 32.2 and national insurance scores 51.9. Child benefit scores 77.3, universal credit scores 77.6, PIP scores 84.0, and pension credit scores 88.4.

Source: Article Notebook

Table 5: Hardest UK variables by bounded score.

	Variable	Score	Exact	Within 10%
1	income_tax	34.8	25.4	45.8
2	national_insurance	55.3	50.6	61.5
5	universal_credit	77.7	76.9	78.8
0	child_benefit	78.1	70.7	86.1
4	pip	84.9	79.4	90.8

Source: Article Notebook

7 Failure modes

The benchmark surfaces a few recurring failure patterns.

First, models miss positive tax quantities more often than zero cases. In the US, the four lowest-scoring variables are household state income tax, state AGI, federal income tax before refundable credits, and adjusted gross income. These outputs require the model to choose the right income concepts and exclusions before applying any final subtraction.

Second, the UK benchmark shows the same split between benefits and tax. Income tax and national insurance score below the four benefit outputs in the frozen manuscript snapshot. This suggests that the models handle broad program thresholds more reliably than detailed tax-liability calculations.

Third, structured-output reliability still affects the ranking. Some provider-specific gaps reflect tool or JSON contract reliability in addition to policy reasoning quality. For example, Claude Opus 4.6 parses 5,975 of 6,000 UK rows in the frozen manuscript snapshot. PolicyBench therefore tracks coverage and parse rates alongside accuracy (Shorten et al. 2024).

8 Limitations

PolicyBench is not a substitute for a production tax-and-benefit calculator. Several caveats matter:

- cost reporting is still reconstructed from provider token usage rather than invoice truth
- model rankings can reflect output-contract reliability as well as policy reasoning
- the public UK calibrated dataset is not equivalent to enhanced FRS quality
- benchmark success should not be interpreted as policy-advice readiness

The current paper is therefore an evaluation of model performance under a specific structured-output benchmark, not a general certification of tax or benefit competence.

9 Conclusion

PolicyBench shows a consistent pattern across both countries. Benefit outputs score above tax-base outputs, while the lowest-scoring variables are adjusted gross income, state-tax quantities, income tax, and national insurance. In the frozen manuscript snapshot, gemini-3.1-pro-preview is the top-scoring US model and grok-4.20 is the top-scoring UK model.

Next steps are to expand diagnostics, extend country coverage, and tighten the data and scoring pipelines so that the leaderboard is driven less by output-format reliability and more by policy reasoning.

Source: Article Notebook

Chen, Zhiyu, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Derek Langdon, Reham Moussa, et al. 2021. "FinQA: A Dataset of Numerical Reasoning over Financial Data." arXiv Preprint arXiv:2109.00122. https://arxiv.org/abs/2109.00122.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, et al. 2021. "Training Verifiers to Solve Math Word Problems." arXiv Preprint arXiv:2110.14168. https://arxiv.org/abs/2110.14168.

Guha, Neel, Julian Nyarko, Daniel E. Ho, et al. 2023. "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." arXiv Preprint arXiv:2308.11462. https://arxiv.org/abs/2308.11462.

Hendrycks, Dan, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. "Measuring Mathematical Problem Solving with the MATH Dataset." In NeurIPS Datasets and Benchmarks Track. https://arxiv.org/abs/2103.03874.

Holzenberger, Nils, Benjamin Van Durme, Sarah Lawsky, and Kyle Richardson. 2021. "Factoring Statutory Reasoning as Language Understanding Challenges." arXiv Preprint arXiv:2105.07903. https://arxiv.org/abs/2105.07903.

Kim, Chan, Jiseung Kim, Jihyung Choi, Juho Lee, Seongyeon Ryu, Sangki Yoon, et al. 2024. "RuleArena: A Benchmark for Rule-Guided Reasoning with LLMs in Real-World Scenarios." arXiv Preprint arXiv:2412.08972. https://arxiv.org/abs/2412.08972.

Mahesh, Meenakshi, Elliott Ash, Elsen Tan, and Aman Madaan. 2025. "TaxCalcBench: Evaluating Frontier Models on the Tax Calculation Task." arXiv Preprint arXiv:2507.16126. https://arxiv.org/abs/2507.16126.

Nava Labs. 2025. "Experimenting with AI-Powered Tools in Public Benefits." Case study. https://www.navapbc.com/case-studies/ai-tools-public-benefits.

———. 2026. "Evaluating a GenAI-Powered Assistive Chatbot for Caseworkers." Case study. https://www.navapbc.com/case-studies/evaluating-ai-assistive-chatbot-caseworkers.

Shanahan, Catherine, Emma McCarthy, Yan Zhao, Wesley H. Holliday, Jenny Zhao, Janet Ainsworth, and Harry Surden. 2025. "Performance of LLMs on VITA Test: Potential for AI-Assisted Tax Returns for Low Income Taxpayers." Artificial Intelligence and Law. https://doi.org/10.1007/s10506-025-09465-7.

Shorten, Connor, Charles Pierse, Thomas Benjamin Smith, Erika Cardenas, Akanksha Sharma, John Trengrove, and Bob van Luijt. 2024. "StructuredRAG: JSON Response Formatting with Large Language Models." arXiv Preprint arXiv:2408.11061. https://doi.org/10.48550/arXiv.2408.11061.

Sutherland, Holly, Francesco Figari, Ruth Hancock, Manos Matsaganis, Alari Paulus, Iva Tasseva, Alessandra Tumino, and Paola De Agostini. 2023. "EUROMOD: The European Union Tax-Benefit Microsimulation Model." International Journal of Microsimulation. https://microsimulation.pub/articles/00075.

Woodruff, Nikhil, and Max Ghenis. 2026. "Improving the Accuracy of UK Tax-Benefit Microsimulation with Machine Learning." PolicyEngine technical paper. https://policyengine.org/uk_data_enhancement.pdf.

ZenML LLMOps Database. 2025. "Building and Automating Comprehensive LLM Evaluation Framework for SNAP Benefits." Industry writeup. https://www.zenml.io/llmops-database/building-and-automating-comprehensive-llm-evaluation-framework-for-snap-benefits.

Zhou, Xinyue et al. 2025. "Time to Revisit Exact Match: Evaluating Temporal Question Answering with Numeric Error Metrics." arXiv Preprint arXiv:2509.16720. https://arxiv.org/abs/2509.16720.

Zhu, Fengbin, Wenqiang Lei, Youcheng Huang, Chao Wang, Shujie Zhang, Hongjie Wang, Vincent Chen, et al. 2021. "TAT-QA: A Question Answering Benchmark on a Hybrid

of Tabular and Textual Content in Finance." arXiv Preprint arXiv:2105.07624. https://arxiv.org/abs/2105.07624.