NBER WORKING PAPER SERIES

DO AI MODELS KNOW THE EVIDENCE ON SOCIAL PROTECTION? COMPARING LARGE LANGUAGE MODELS WITH POLICYMAKERS

Chenhao Tan

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2026

The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

© 2026 by Chenhao Tan. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Do AI Models Know the Evidence on Social Protection?
Comparing Large Language Models with Policymakers
Chenhao Tan
November 2026
JEL No. C88, H53, I38, O15

ABSTRACT

Policymakers increasingly consult large language models (LLMs) for policy advice. We ask whether these models hold beliefs about social protection that match the empirical evidence. We take the survey of 46 current and former ministers and senior civil servants from low- and middle-income countries in Olken et al. (2025), which covers ten topics from the best form of transfers to the effects of subsidized health insurance, and pose the same 19 items to 13 frontier models. An LLM judge grades each answer against the literature summary in that chapter and against the policymakers' aggregate answer. Evidence scores range from 98 to 71 out of 100, and 10 of 13 models score above the median policymaker (86). Models do better than policymakers on the items where policymakers were furthest from the evidence, such as demand for self-paid health insurance among informal workers, but share some of their errors, such as believing that subsidized health insurance improves health. Agreeing with policymakers is not the same as agreeing with the evidence: among models that always commit to an answer, those closer to policymakers score lower on the evidence.

Chenhao Tan
University of Chicago
chenhao@uchicago.edu

Acknowledgments. Claude (Anthropic) assisted with writing the evaluation code, running the analysis, and drafting the text.

I. INTRODUCTION

Research on social protection in developing countries has grown quickly. A recent review counts more than 300 papers (Banerjee et al., 2024), and a meta-analysis of unconditional cash transfers alone covers 115 studies (Crosta et al., 2025). Olken et al. (2025) ask whether this evidence has reached the people who design programs. They interview 46 high-level policymakers about ten topics and compare their answers with the literature, highlighting where beliefs match the evidence, where they diverge, and where the research is missing or ambiguous.

Policymakers and their staff now also ask LLMs for advice. If a model's beliefs differ from the evidence, that advice can spread the same errors widely; if they match the evidence, models could help close the gap that Olken et al. (2025) document. In this paper, we take the same approach as that chapter, but our respondents are 13 AI models rather than policymakers. We give each model the same questions, grade the answers against the same literature, and compare them with the policymakers' answers.

We make three points. First, most models match the evidence at least as well as the median policymaker, and the best ones do so on nearly every item. Second, models improve most on the items where policymakers were furthest from the evidence, but they share some policymaker errors, so the gains are not uniform across topics. Third, agreement with policymakers and agreement with the evidence are different things: the model that agrees most with policymakers has the lowest evidence score.

II. SURVEY AND QUESTION SET

Olken et al. (2025) surveyed 46 policymakers between April and July 2025, covering at least 18 countries. About 13 percent were current ministers, 30 percent former ministers, and the rest current or former senior civil servants. Respondents were asked to think about the context of their own countries.

The survey has ten topics. We convert it into 19 items: the question on the effects of cash transfers covers eight outcomes (consumption, food security, school attendance, health, high school graduation, spending on alcohol and tobacco, labor supply, and fertility), each of which we treat as an item, and we keep the chapter's side question on the trend in global extreme poverty. Items take four forms: rankings (four items), five-point scales (nine), single choices (four), and percentages (two).

For each item we record the policymakers' aggregate answer (mean ranks, shares by option, or the mean and median of a number) and the chapter's summary of the literature. We also assign a *gold strength*:

- **Clear** (10 items): the literature gives a settled answer, for example that cash transfers do not reduce adult labor supply (Banerjee et al., 2017; Baird et al., 2018).
- **Soft** (6 items): the literature points in a direction, but the exact answer is partly a judgment call, for example the full ranking of transfer forms.
- **None** (3 items): the literature gives no single answer (best targeting method, conditional versus unconditional transfers, and which programs voters like). These items are not part of the evidence score.

III. METHOD

Models. We query 13 models from 11 developers through OpenRouter, listed in Appendix Table A1. Each model receives each item once, with the instruction to answer from the perspective of a policymaker in a typical low- or middle-income country and to reply with a JSON object. For the payroll tax item, which asks about coverage "in your country," we substitute "a typical lower-middle-income country."

Declined answers. Some models refuse to choose, arguing that the answer depends on context. Because surveyed policymakers had to choose, we send one follow-up message in the same conversation asking the model to commit to the answer it considers best on balance. Both replies are recorded.

Grading. Claude Opus 5.5 grades each answer. The judge sees the item, the options, the policymakers' aggregate answer, the literature summary with the gold answer, and the model's reply. It returns (i) whether the model committed, declined, or gave no usable answer; (ii) an *evidence score* between 0 and 1 for committed answers; (iii) a *policymaker agreement* score between 0 and 1; (iv) on items where policymakers deviate from the evidence, whether the model sided with the evidence or with policymakers; and (v) a short rationale. The full rubric is in Appendix Appendix B.

Scoring declined answers. A declined answer receives an evidence score of 0 on clear items and 0.5 on soft items, and is not scored on items with no gold answer. It receives a policymaker agreement score of 0. A model's evidence score is the mean over the 16 items with a gold answer, and its policymaker agreement score is the mean over all 19 items, both scaled to 0 to 100. We report 95 percent intervals from a bootstrap over items.

Policymaker baseline. To place models relative to policymakers, we grade the median or modal policymaker answer to each item with the same judge and rubric. This baseline excludes the payroll tax item, which policymakers answered for their own countries.

IV. OVERVIEW OF RESULTS

A. OVERALL PERFORMANCE

Taken together, the models answer the survey in a way that is close to the evidence. Of the 247 answers (13 models times 19 items), 241 committed to an answer and 6 declined, all of them from a single model; none were unusable. The average evidence score across models is 89 out of 100 and the median is 92. Of the 208 answers to scored items, 74 percent received full credit and 3 percent received none, so most of the variation between models comes from a handful of items rather than from widespread errors.

The items fall into three groups. On six items every model received full credit: the effects of cash transfers on consumption, food security, school attendance, health, and high school graduation, and the trend in global extreme poverty. On a second group, models are mostly right but a few models repeat a concern that the evidence does not support: spending on temptation goods, labor supply, fertility, and the effect of social spending on votes. On a third group, most models are wrong or divided: the health effects of subsidized health insurance (mean score 64) and the metric for judging a cash transfer program (72).

Compared item by item with the median policymaker, the average model scores higher on 3 items, about the same on 9, and lower on 3. The three items where models score higher are those where the median policymaker was furthest from the evidence: demand for self-paid health insurance (94 against 31), the effect of social spending on votes (85 against 50), and the metric for judging a program (72 against 50). The three items where models score lower (temptation goods, labor supply, and fertility) are items where the median policymaker answer is correct but large minorities of policymakers were not; for example, half of the policymakers said cash transfers increase spending on alcohol and tobacco. The comparison with the median therefore understates how often individual policymakers depart from the evidence on these items.

Finally, agreeing with policymakers is not the same as agreeing with the evidence. Across all 13 models the correlation between the two scores is 0.18, but this is driven by the one model that often declines, which scores low on both. Among the models that always commit, the correlation is -0.44: models closer to policymakers tend to be further from the evidence.

Table 1. Agreement with the evidence and with policymakers

		Evidence score (0–100)			Policymaker	Sides with	
Model	Rank	All	95% interval	Committed	agreement	evidence	Declined
Claude Fable 5.2	1	98	[94, 100]	98	0.70	1.00 (7)	0
Kimi K3.1 Pro	2	97	[92, 100]	97	0.72	0.86 (7)	0
GLM 5.3	3	96	[92, 100]	96	0.74	0.86 (7)	0
Qwen3.8 Max	4	95	[88, 100]	95	0.78	0.88 (8)	0
MiMo V2.6 Pro	5	94	[86, 100]	94	0.75	0.71 (7)	0
DeepSeek V4 Pro	6	92	[83, 100]	92	0.73	0.71 (7)	0
GPT-6 Astra	7	92	[83, 100]	92	0.79	0.71 (7)	0
Grok 4.7	8	89	[78, 99]	89	0.71	0.57 (7)	0
Nemotron 3 Ultra	9	89	[78, 98]	89	0.79	0.75 (8)	0
Claude Opus 5.5	10	87	[72, 98]	87	0.66	0.71 (7)	0
Policymakers (median answer)		*86*	*[74, 97]*				
Gemini 4 Argon	11	82	[67, 94]	82	0.78	0.43 (7)	0
Gemini 4 Carbon	12	72	[54, 87]	83	0.57	0.50 (6)	6
Muse Wonder	13	71	[51, 88]	71	0.82	0.25 (8)	0

Notes: "All" includes partial credit for declined answers; "Committed" averages only items where the model chose an answer. Policymaker agreement (0 to 1) compares the model with the policymakers' aggregate answer. "Sides with evidence" is the share of items, among those where policymakers deviate from the evidence and the model committed, on which the model sided with the evidence (number of such items in parentheses). Declined counts items out of 19.

B. AGREEMENT WITH THE EVIDENCE

Figure 1 and Table 1 report evidence scores. Scores range from 98 (Claude Fable 5.2) to 71 (Muse Wonder). 10 of 13 models score above the policymaker baseline of 86. The top seven models score between 92 and 98 and their intervals overlap, so the order among them should not be read as a strict ranking.

C. AGREEMENT WITH POLICYMAKERS

Figure 2 ranks the same models by how closely they agree with the surveyed policymakers. Agreement scores range from 57 to 82, a narrower and lower range than the evidence scores. At the top, the order reverses Figure 1: Muse Wonder has the highest policymaker agreement (82) and the lowest evidence score, while Claude Fable 5.2 and Claude Opus 5.5 are among the three models that agree least with policymakers. Gemini 4 Carbon has the lowest agreement (57) because declined answers count as disagreement.

This pattern follows from how the survey was designed. On seven or eight of the 19 items, depending on the model's answers, the policymakers' aggregate answer differs from the evidence. On those items, a model that tracks the evidence must disagree with policymakers. The "Sides with evidence" column of Table 1 makes this explicit: on the items where policymakers deviate

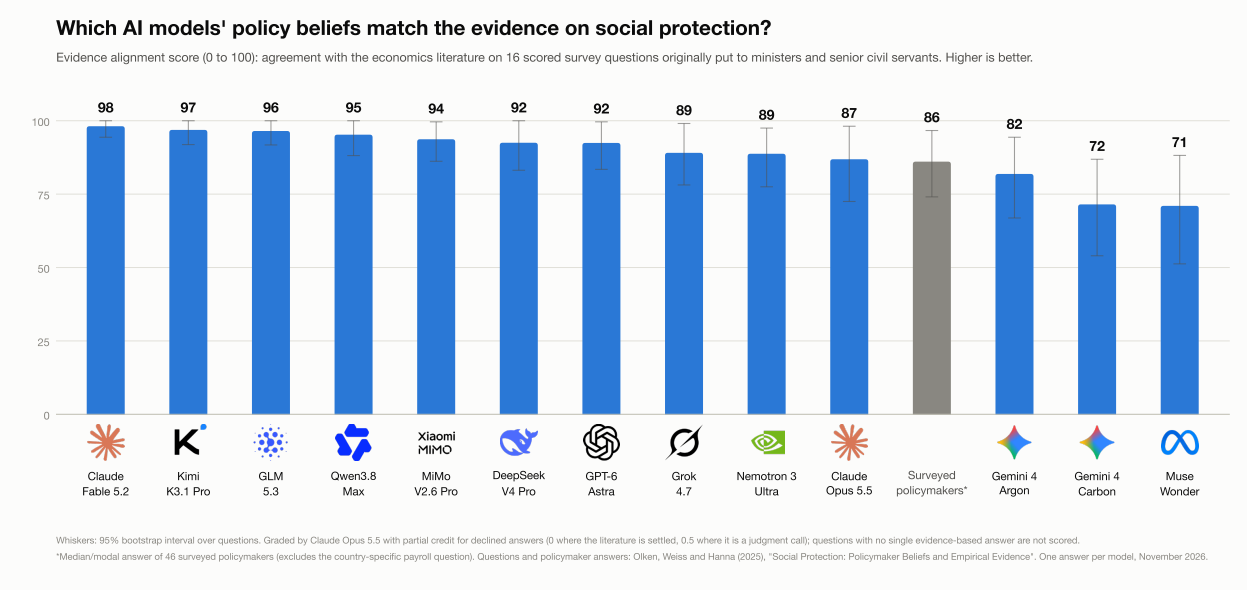


Figure 1. Agreement with the evidence, by model

Notes: Evidence score is the mean over 16 items with a gold answer, scaled to 0 to 100. Whiskers are 95 percent bootstrap intervals over items. The gray bar grades the median or modal policymaker answer from Olken et al. (2025) with the same judge and rubric (15 items).

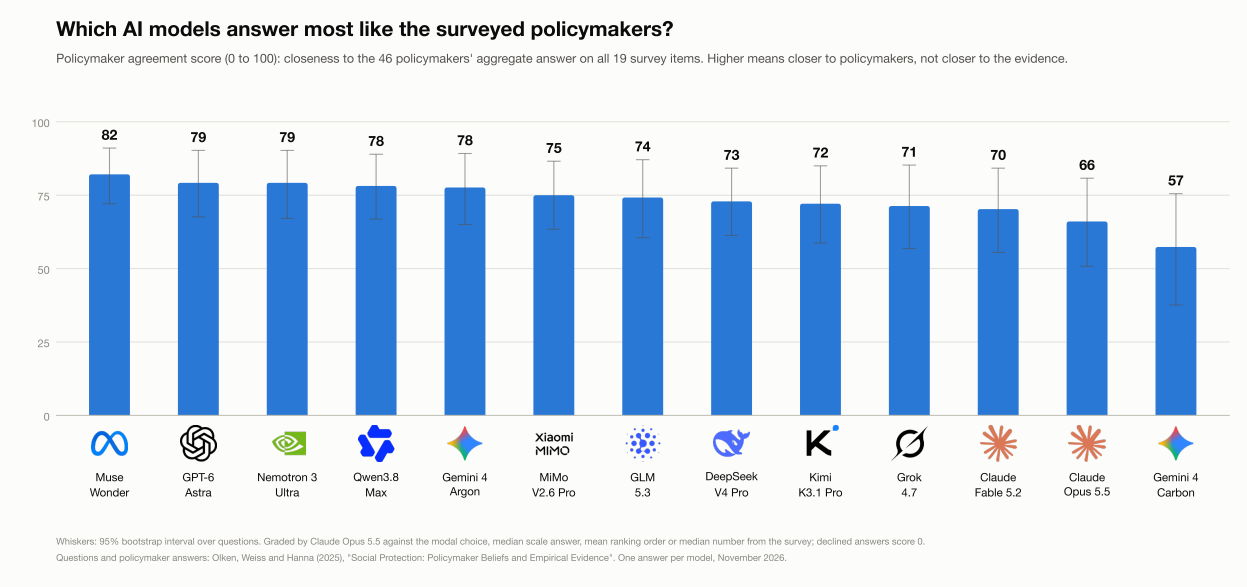


Figure 2. Agreement with surveyed policymakers, by model

Notes: Policymaker agreement score is the mean over all 19 items of the judge's rating of closeness to the policymakers' aggregate answer, scaled to 0 to 100. Declined answers score 0. Whiskers are 95 percent bootstrap intervals over items.

Table 2. Results by item

Item	Gold strength	Evidence score (0–100)		Policymaker agreement	Models declining
		Models (mean)	Policymakers		
Best form of transfer (rank)	soft	96	100	0.92	1
Cash transfers: consumption	clear	100	100	0.79	0
Cash transfers: food security	clear	100	100	0.81	0
Cash transfers: school attendance	clear	100	100	0.92	0
Cash transfers: health	clear	100	100	1.00	0
Cash transfers: high school graduation	clear	100	100	1.00	0
Cash transfers: alcohol, tobacco, drugs	clear	65	100	0.52	0
Cash transfers: labor supply	clear	85	100	0.88	1
Cash transfers: fertility	clear	88	100	0.91	0
Metric of program success (rank)	soft	72	50	0.67	0
Targeted or universal	soft	96	100	0.92	1
Trend in global extreme poverty	clear	100	100	1.00	0
Best targeting method (rank)	none	–	–	0.52	0
Conditional or unconditional	none	–	–	0.52	1
Payroll-tax insurance coverage (%)	soft	76	n.a.	0.64	0
Enrollment in self-paid health insurance (%)	clear	94	31	0.37	0
Effect of subsidized health insurance	soft	64	60	0.65	0
Effect of social spending on votes	soft	85	50	0.58	1
Most popular program with voters (rank)	none	–	–	0.34	1

Notes: "Models (mean)" averages the evidence score over the 13 models. "Policymakers" grades the median or modal policymaker answer. Items with gold strength "none" are not scored. The payroll tax item is not graded for policymakers because they answered for their own countries.

from the evidence, Claude Fable 5.2 sided with the evidence every time, and Muse Wonder did so on two of eight.

D. RESULTS BY ITEM

Table 2 reports results for each item. We discuss each topic in turn in Section V, following the order of Olken et al. (2025).

V. TEN SOCIAL PROTECTION POLICY QUESTIONS

A. WHAT FORM SHOULD TRANSFERS TAKE? CASH, IN-KIND TRANSFERS, AND PRICE SUBSIDIES

The chapter begins by asking about the form transfers should take. Policymakers were asked: "What is the best way to provide transfers to households?" and ranked four options from best to worst: cash, in-kind transfers, vouchers, and energy subsidies. Among the surveyed policymakers, cash was the top choice by far (mean rank 1.94), followed by vouchers (2.21), in-kind transfers (2.85), and energy subsidies (3.06).

The models agree. Twelve of the 13 models gave the same order as the policymakers' mean ranks, and every one of them put cash first and energy subsidies last. The models were more uniform than the policymakers: where the policymakers' mean ranks are close together, the models' ranks are not. This order is consistent with the evidence the chapter reviews. Head-to-head trials find similar consumption effects of cash, vouchers, and food, with in-kind transfers costing more than twice as much per outcome (Hidrobo et al., 2014); electronic vouchers in Indonesia delivered 45 percent more aid than in-kind rice because of lower leakage (Banerjee et al., 2023); and energy subsidies tend to be regressive (Burgess et al., 2020).

The one exception was Gemini 4 Carbon, which declined to rank the options even after our follow-up message, writing that "it is not possible to commit to a single definitive ranking." Its first reply nonetheless described energy subsidies as "highly regressive." As the chapter notes, what works "best" depends on the metric one uses, so this is the item where declining is easiest to defend; under our scoring it receives half credit.

B. HOW DO CASH TRANSFER PROGRAMS IMPACT HOUSEHOLDS?

Policymakers were asked: "How do cash transfer programs impact recipient households?" for eight outcomes. They reported positive effects on consumption (89 percent said it increases), food security (87 percent), school attendance (66 percent), health (65 percent), and high school graduation (53 percent). They also expressed concerns that are often raised against cash transfers: half said cash transfers increase spending on alcohol, tobacco, or drugs; 38 percent said they reduce labor supply, against 22 percent who said they increase it; and 30 percent said they increase fertility, against 11 percent who said they decrease it.

Positive outcomes. Every model said that all five positive outcomes increase. For consumption and food security, most models said "increases a lot" (9 and 10 of 13); for school attendance, health, and high school graduation, most said "increases a little." These answers match the meta-analytic evidence (Crosta et al., 2025) and are more uniform than the policymakers', a third or more of whom expected no effect on schooling and health.

Temptation goods. This is the outcome on which models disagree most. Seven models said cash transfers have no effect on spending on alcohol, tobacco, or drugs, in line with the review by Evans and Popova (2017). Three (Gemini 4 Argon, Gemini 4 Carbon, and Muse Wonder) said spending increases, as half of the policymakers did. Three others (Grok 4.7, Nemotron 3 Ultra, and Claude Opus 5.5) said spending decreases, which the evidence does not show either.

Labor supply. Eleven models said cash transfers have no effect on hours worked, in line with the reanalysis of seven randomized evaluations by Banerjee et al. (2017) and the review by Baird et al. (2018). Only Muse Wonder said labor supply decreases. Gemini 4 Carbon declined, writing that "the exact effect on hours worked depends heavily on who receives the transfer," even though its own answer stated that "broad fears of widespread dependency are largely unsupported by empirical studies." No model chose "increases a little," which recent meta-analytic evidence also supports (Crosta et al., 2025).

Fertility. Eleven models said cash transfers have no effect on the number of children; Gemini 4 Argon and Muse Wonder said fertility increases. The evidence finds no effect among adult recipients (Olken et al., 2025).

In short, the models largely reproduce the evidence on cash transfers, and the concerns that the chapter documents among policymakers appear in only a few models, mostly the same ones (Muse Wonder and the two Gemini models).

C. WHAT IS THE MOST IMPORTANT METRIC TO JUDGE WHETHER A CASH TRANSFER PROGRAM IS SUCCESSFUL?

Policymakers ranked seven metrics. Food security came first (mean rank 2.80), followed by targeting, that is, the percentage of the program that reached the poor (3.39). What households spend the money on, health, and employment came last. The chapter argues that, under standard assumptions, public finance theory implies that a dollar is a dollar: what matters is whether the transfer reaches households with a high marginal utility of income, so targeting and governance matter most and the composition of spending does not (Banerjee et al., 2024).

The models agree with one half of this argument. On average they ranked "what households spend the money on" last (mean rank 6.85, against 4.17 for policymakers), consistent with the idea that a dollar is a dollar. But only four models (Claude Fable 5.2, Gemini 4 Argon, Muse Wonder, and Qwen3.8 Max) put targeting first. Six put household consumption first and three put food security first. On average, models ranked targeting third (2.38), behind food security (2.08) and consumption (2.23). This item has the third-lowest mean evidence score of any item (72), although models still score above the median policymaker (50).

D. SHOULD CASH TRANSFERS BE TARGETED OR UNIVERSAL?

Policymakers were asked: "Should cash transfers be targeted to (aimed at) only the poorest households or universal?" No respondent chose universal transfers; 44 percent chose the bottom 30 percent and 27 percent the bottom 10 percent. The chapter notes that for a fixed budget, transfers

targeted with proxy-means tests deliver more welfare to the poor than universal transfers, at the cost of more horizontal inequity (Hanna and Olken, 2018).

Twelve models chose the bottom 30 percent, matching the policymakers' modal answer, and no model chose universal transfers. Gemini 4 Carbon declined, saying the choice "depends on value judgments."

The chapter's side question asked what has happened to extreme poverty over the last 20 years. Nearly three-quarters of policymakers answered, correctly, that the rate has almost halved, in contrast to the general public in high-income countries (Rosling et al., 2018). All 13 models answered correctly.

E. WHAT IS THE BEST WAY TO TARGET HOUSEHOLDS FOR GOVERNMENT ASSISTANCE?

Policymakers ranked eight targeting methods. "Create an application process that involves multiple verification steps" topped the list (mean rank 3.08), followed by categorical targeting (3.27), proxy-means tests (3.55), and geographic targeting (4.24). Community targeting, tax records, work requirements, and self-reported income came last.

The models' ranking is very different. All 13 models put proxy-means tests first, and nine gave the same top four: proxy-means tests, community-based targeting, geographic targeting, and categorical targeting. On average they ranked the multi-step application process sixth (5.85), the policymakers' first choice. The literature does not settle this question. Proxy-means tests do slightly better than community targeting at identifying households by consumption, while community targeting better matches households' own assessment of poverty (Alatas et al., 2012). A multi-step application process improved targeting in Indonesia (Alatas et al., 2016), but onerous applications can deter eligible households (Gupta, 2017). Models thus agree with each other far more than the evidence would justify, and they disagree with policymakers on the one method with direct experimental support for self-selection.

F. SHOULD CASH TRANSFERS BE CONDITIONAL OR UNCONDITIONAL?

Policymakers were asked: "Should poor households have to demonstrate that they are meeting certain conditions (like enrolling children in school or vaccinating children) to receive cash transfers?" Most (59 percent) said yes, 9 percent favored unconditional transfers, and about a third favored labeled transfers, in which households are told the intended use but benefits are not revoked.

The models lean the other way. Eight models chose labeled transfers, four (Gemini 4 Argon, Grok 4.7, Muse Wonder, and Nemotron 3 Ultra) chose conditional transfers, and Gemini 4 Carbon declined. Like the policymakers, no model chose fully unconditional transfers. The evidence

involves a tradeoff: conditions can raise schooling (Baird et al., 2011), but households that fail to meet conditions are often the poorest (Anagol et al., 2025), and a labeled transfer in Morocco reduced school dropout more than a conditional one (Benhassine et al., 2015). The models' preference for labeled transfers is consistent with this last study, but the question remains open.

G. CAN YOU PROVIDE SOCIAL INSURANCE THROUGH PAYROLL TAXES?

Policymakers were asked: "If you instituted a payroll tax in your country to provide unemployment insurance to workers, what fraction of the workforce would it cover?" The chapter compares their answers with formal employment rates in their countries and finds them somewhat predictive, with some respondents over- and some under-optimistic.

Because the models answered for a typical lower-middle-income country, we cannot compare them country by country. All models answered 15 or 20 percent, which reflects the idea that payroll-based insurance covers only formal workers. Mandating insurance for formal workers alone can also increase informality (Camacho et al., 2014).

H. WHAT IS THE DEMAND FOR HEALTH INSURANCE AMONG INFORMAL WORKERS?

Policymakers were asked what fraction of informal workers would enroll in a health insurance program in which they pay their own premiums at a break-even rate. On average, policymakers thought 48.3 percent would enroll; the median response was 37 percent. The chapter concludes that they overestimate demand: baseline enrollment is 8 percent in the Philippines (Capuno et al., 2016), 8 percent in Indonesia (Banerjee et al., 2021), and 4 percent in Vietnam (Wagstaff et al., 2016).

This is the item on which models improve most on policymakers. Twelve of 13 models answered between 5 and 15 percent: five said 10 percent, five said 15 percent, and Claude Fable 5.2 and Gemini 4 Argon said 8 and 5 percent. Only Muse Wonder (30 percent) was close to the policymakers. The mean model evidence score on this item is 94, against 31 for the median policymaker.

I. WHAT IS THE IMPACT OF FREE OR SUBSIDIZED HEALTH INSURANCE ON POOR HOUSEHOLDS?

Policymakers were asked how free or subsidized health insurance affects poor households. Nearly three-quarters (74 percent) said it improves both health and economic outcomes, 14 percent health only, and 9 percent economic outcomes only. The evidence supports the economic benefits, but the effects on health are more mixed: randomized studies find little effect on health outcomes (King et

al., 2009; Haushofer et al., 2020), while large reforms that also expanded the supply of care show mortality declines (Gruber et al., 2014, 2023).

Most models share the policymakers' view. Eight models said health insurance improves both outcomes. Four (Claude Fable 5.2, GLM 5.3, Grok 4.7, and Kimi K3.1 Pro) said it improves economic outcomes only, the answer most consistent with the experimental evidence. Claude Opus 5.5 said it improves neither, which is inconsistent with the evidence on financial protection. The mean model evidence score on this item is 64, close to that of the median policymaker (60), and the lowest of any scored item.

J. WHAT ARE THE EFFECTS OF SOCIAL PROTECTION SPENDING ON VOTERS? WHAT PROGRAMS DO VOTERS LIKE?

Policymakers were asked how social spending affects voting patterns. About half said it strongly increases voter support and about a third said it slightly increases it. The evidence is more muted: beneficiaries often become more supportive of the incumbent (Conover et al., 2020), but effects across the electorate are small or limited to competitive municipalities (Labonne, 2013; Imai et al., 2020). Nine models said spending slightly increases support, which matches the evidence better; three (Gemini 4 Argon, Grok 4.7, and Muse Wonder) said it strongly increases support, as most policymakers did; and Gemini 4 Carbon declined.

Policymakers then ranked six programs by popularity with voters. Targeted cash transfers came first (mean rank 2.10), followed by universal and conditional cash transfers; energy subsidies, health insurance, and in-kind transfers came last. The models' ranking nearly reverses this. Eight models put energy subsidies first and four put universal cash transfers first; on average, models ranked energy subsidies first (1.33) and targeted cash transfers fifth of six (5.00). The chapter treats this question as open. Energy subsidies are often politically popular (Burgess et al., 2020), although in Indonesia low-income households prefer targeted assistance to fuel subsidies except where leakage from transfer programs is high (Kyle, 2018). The gap between models and policymakers on this item is the largest in the survey, and it is a question on which both groups' beliefs could be tested directly.

VI. TWO KINDS OF FAILURE

The two lowest-scoring models fail differently. Muse Wonder commits on every item but answers three clear items incorrectly: it says cash transfers increase fertility and spending on alcohol and tobacco, and reduce work. These are the concerns that Olken et al. (2025) report among policymakers and that the evidence does not support, which explains why Muse Wonder agrees most with policymakers. Gemini 4 Carbon declines to choose on 6 of 19 items, even after the

follow-up message, arguing that the answer depends on context. On the items it answers, it scores 83, close to the middle of the distribution; under our scoring, its declines cost it about 11 points.

VII. LIMITATIONS

- **Sample size.** The evidence score averages 16 items and one answer per model, so differences under about 10 points are within sampling noise.
- **A single judge.** Claude Opus 5.5 grades all answers, including its own. It ranks 10th of 13 on the evidence score, so we see no sign of self-preference, but grading with more judges would make the scores more robust.
- **Gold answers.** For several items the chapter gives a direction rather than a single answer. Converting these into gold answers, and the partial credit for declined answers, involves judgment.
- **The policymaker baseline.** Using the median answer hides disagreement among policymakers. Half of them said cash transfers increase spending on alcohol and tobacco, but the median answer is "no effect," which scores 100. The baseline therefore overstates how well a typical policymaker matches the evidence.
- **Context.** Policymakers answered for their own countries; models answered for a typical low- or middle-income country.

VIII. CONCLUSION

In this paper, we posed the questions that Olken et al. (2025) asked of 46 policymakers to 13 AI models. Several facts stand out. Most models answer closer to the evidence than the median policymaker does, and the gains are largest on the items where policymakers were furthest from the evidence, such as demand for health insurance among informal workers. Models also share some of the policymakers' beliefs that the evidence does not support, most clearly on the health effects of subsidized health insurance and on targeting as the measure of a program's success. On the open questions, models agree with each other far more than the evidence would justify, and on which programs voters like, they reverse the policymakers' ranking.

These results suggest that models could be useful for communicating established findings, but that their confident answers on open questions should be treated with the same caution as anyone else's. Because models change quickly, the code and data behind this paper are designed to be rerun on new models.

REFERENCES

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias, "Targeting the Poor: Evidence from a Field Experiment in Indonesia," *American Economic Review*, 2012, *102* (4), 1206–1240.

_ , _ , _ , _ , Ririn Purnamasari, and Matthew Wai-Poi, "Self-Targeting: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, 2016, *124* (2), 371–427.

Anagol, Santosh, Thomas Fujiwara, and Martin Navarrete, "Conditioning Out the Poor? Consumption Inequality and the Design of Cash Transfer Programs," Working Paper 33841, National Bureau of Economic Research 2025.

Baird, Sarah, Craig McIntosh, and Berk Özler, "Cash or Condition? Evidence from a Cash Transfer Experiment," *The Quarterly Journal of Economics*, 2011, *126* (4), 1709–1753.

_ , David McKenzie, and Berk Özler, "The Effects of Cash Transfers on Adult Labor Market Outcomes," *IZA Journal of Development and Migration*, 2018, *8* (1), 22.

Banerjee, Abhijit, Amy Finkelstein, Rema Hanna, Benjamin A. Olken, Arianna Ornaghi, and Sudarno Sumarto, "The Challenges of Universal Health Insurance in Developing Countries: Experimental Evidence from Indonesia's National Health Insurance," *American Economic Review*, 2021, *111* (9), 3035–3063.

_ , Rema Hanna, Benjamin A. Olken, and Diana Sverdlin Lisker, "Social Protection in the Developing World," *Journal of Economic Literature*, 2024, *62* (4), 1349–1421.

_ , _ , _ , Elan Satriawan, and Sudarno Sumarto, "Electronic Food Vouchers: Evidence from an At-Scale Experiment in Indonesia," *American Economic Review*, 2023, *113* (2), 514–547.

_ , _ , Gabriel E. Kreindler, and Benjamin A. Olken, "Debunking the Stereotype of the Lazy Welfare Recipient: Evidence from Cash Transfer Programs," *The World Bank Research Observer*, 2017, *32* (2), 155–184.

Benhassine, Najy, Florencia Devoto, Esther Duflo, Pascaline Dupas, and Victor Pouliquen, "Turning a Shove into a Nudge? A "Labeled Cash Transfer" for Education," *American Economic Journal: Economic Policy*, 2015, *7* (3), 86–125.

Burgess, Robin, Michael Greenstone, Nicholas Ryan, and Anant Sudarshan, "The Consequences of Treating Electricity as a Right," *Journal of Economic Perspectives*, 2020, *34* (1), 145–169.

Camacho, Adriana, Emily Conover, and Alejandro Hoyos, "Effects of Colombia's Social Protection System on Workers' Choice between Formal and Informal Employment," *The World Bank Economic Review*, 2014, *28* (3), 446–466.

Capuno, Joseph J., Aleli D. Kraft, Stella Quimbo, Carlos R. Tan, and Adam Wagstaff, "Effects of Price, Information, and Transactions Cost Interventions to Raise Voluntary Enrollment in a Social Health Insurance Scheme: A Randomized Experiment in the Philippines," *Health Economics*, 2016, *25* (6), 650–662.

Conover, Emily, Román A. Zárate, Adriana Camacho, and Javier E. Baez, "Cash and Ballots: Conditional Transfers, Political Participation, and Voting Behavior," *Economic Development and Cultural Change*, 2020, *68* (2), 541–566.

Crosta, Tommaso, Dean Karlan, Finley Ong, Julius Rüschenpöhler, and Christopher Udry, "Unconditional Cash Transfers: A Bayesian Meta-Analysis of Randomized Evaluations in Low and Middle Income Countries," April 2025. Working paper.

Evans, David K. and Anna Popova, "Cash Transfers and Temptation Goods," *Economic Development and Cultural Change*, 2017, *65* (2), 189–221.

Gruber, Jonathan, Mengyun Lin, and Junjian Yi, "The Largest Insurance Program in History: Saving One Million Lives per Year in China," *Journal of Public Economics*, 2023, *226*, 104999.

_ **, Nathaniel Hendren, and Robert M. Townsend**, "The Great Equalizer: Health Care Access and Infant Mortality in Thailand," *American Economic Journal: Applied Economics*, 2014, *6* (1), 91–107.

Gupta, Sarika, "Perils of the Paperwork: The Impact of Information and Application Assistance on Welfare Program Take-Up in India," 2017. Job market paper, Harvard University.

Hanna, Rema and Benjamin A. Olken, "Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries," *Journal of Economic Perspectives*, 2018, *32* (4), 201–226.

Haushofer, Johannes, Matthieu Chemin, Chaning Jang, and Justin Abraham, "Economic and Psychological Effects of Health Insurance and Cash Transfers: Evidence from a Randomized Experiment in Kenya," *Journal of Development Economics*, 2020, *144*.

Hidrobo, Melissa, John Hoddinott, Amber Peterman, Amy Margolies, and Vanessa Moreira, "Cash, Food, or Vouchers? Evidence from a Randomized Experiment in Northern Ecuador," *Journal of Development Economics*, 2014, *107*, 144–156.

Imai, Kosuke, Gary King, and Carlos Velasco Rivera, "Do Nonpartisan Programmatic Policies Have Partisan Electoral Effects? Evidence from Two Large-Scale Experiments," *The Journal of Politics*, 2020, *82* (2), 714–730.

King, Gary, Emmanuela Gakidou, Kosuke Imai et al., "Public Policy for the Poor? A Randomised Assessment of the Mexican Universal Health Insurance Programme," *The Lancet*, 2009, *373* (9673), 1447–1454.

Kyle, Jordan, "Local Corruption and Popular Support for Fuel Subsidy Reform in Indonesia," *Comparative Political Studies*, 2018, *51* (11), 1472–1503.

Labonne, Julien, "The Local Electoral Impacts of Conditional Cash Transfers: Evidence from a Field Experiment," *Journal of Development Economics*, 2013, *104*, 73–88.

Olken, Benjamin A., Alexa Weiss, and Rema Hanna, "Social Protection: Policymaker Beliefs and Empirical Evidence," August 2025. Chapter draft.

Rosling, Hans, Ola Rosling, and Anna Rosling Rönnlund, *Factfulness: Ten Reasons We're Wrong about the World, and Why Things Are Better than You Think*, Flatiron Books, 2018.

Wagstaff, Adam, Ha Thi Hong Nguyen, Huyen Dao, and Sarah Bales, "Encouraging Health Insurance for the Informal Sector: A Cluster Randomized Experiment in Vietnam," *Health Economics*, 2016, *25* (6), 663–674.

Appendix A. Models

Table A1. Models evaluated

Model	Developer	OpenRouter identifier	Reasoning effort
Claude Fable 5.2	Anthropic	`anthropic/claude-fable-5.2`	high
Claude Opus 5.5	Anthropic	`anthropic/claude-opus-5.5`	high
GPT-6 Astra	OpenAI	`openai/gpt-6-astra`	default
Gemini 4 Argon	Google	`google/gemini-4-argon`	default
Gemini 4 Carbon	Google	`google/gemini-4-carbon`	default
Muse Wonder	Meta	`meta/muse-wonder`	default
Kimi K3.1 Pro	Moonshot AI	`moonshotai/kimi-k3.1-pro`	default
GLM 5.3	Z.ai	`z-ai/glm-5.3`	default
DeepSeek V4 Pro	DeepSeek	`deepseek/deepseek-v4-pro-0813`	default
MiMo V2.6 Pro	Xiaomi	`xiaomi/mimo-v2.6-pro`	default
Grok 4.7	xAI	`x-ai/grok-4.7`	default
Qwen3.8 Max	Alibaba	`qwen/qwen3.8-max-0902`	default
Nemotron 3 Ultra	NVIDIA	`nvidia/nemotron-3-ultra-550b-a55b`	default

Notes: All models were queried through OpenRouter in November 2026. "default" means no reasoning effort was set.

Appendix B. Grading Rubric

The judge receives the item, its options, the policymakers' aggregate answer, the literature summary with the gold answer, the authors' verdict on whether policymakers matched the evidence, and the model's reply (both replies if a follow-up was sent). It is instructed to grade only what the model committed to and not to reward hedging. It returns:

1. **Answer status**: committed (chose an option, a full ranking, or a number, possibly with caveats); declined (deliberately refused to choose); or invalid (empty, off-topic, or truncated).
2. **Evidence score** (0 to 1, committed answers only). Rankings receive credit for placing the options the literature favors near the top or bottom. Five-point scales receive 1 if the answer is in the gold set and 0.5 if it is one step away in the right direction. Numbers receive 1 within the gold range or within 5 points of it, falling linearly to 0 at 40 or more points away.
3. **Policymaker agreement** (0 to 1): closeness to the modal choice, mean ranking order, or mean and median number.
4. **Evidence gap direction**: on items where policymakers deviate from the evidence, whether the model sided with the evidence, with policymakers, or with neither.

5. **Rationale**: one to three sentences.