What can LLMs tell us about the ETI?

Jason DeBacker and Max Ghenis

Feb 22, 2026

CONTENTS

Jason DeBacker (University of South Carolina) and **Max Ghenis** (PolicyEngine)

Abstract

We investigate whether large language models (LLMs) can produce economically meaningful estimates of the elasticity of taxable income (ETI). Building on the emerging literature using LLMs as simulated economic agents (), we conduct two complementary studies. First, we replicate the controlled laboratory experiment of , finding that LLMs exhibit bunching behavior at tax notches similar to human subjects, with implied ETI ≥ . Second, we conduct an original survey experiment asking LLMs to simulate taxpayer responses to hypothetical tax changes across varied personas and tax scenarios. GPT-4o produces mean ETI estimates () remarkably close to canonical empirical estimates (), while exhibiting realistic optimization frictions absent in smaller models. Our findings suggest LLMs have internalized economically sensible tax response behavior from training data, opening possibilities for rapid policy prototyping and exploring heterogeneity along dimensions unmeasured in administrative data.

Keywords: AI, elasticity of taxable income, behavioral simulation, large language models

JEL classification: H21, H31, C90

Introduction

Large language models (LLMs) have emerged as promising tools for simulating human behavior in economic contexts. introduced the concept of "homo silicus"—using LLMs as computational stand-ins for human economic agents—and demonstrated that GPT-3 could replicate classic behavioral economics findings. Subsequent work has confirmed that LLMs can simulate demand functions matching human preferences (), exhibit cooperation patterns similar to humans in strategic games (), and predict outcomes of social science experiments ().

However, this nascent literature has also identified important limitations. find that LLM economic behavior is "neither entirely human-like nor entirely economicus-like," with models struggling to maintain consistent behavior across settings. show that while LLMs demonstrate reasonable group-level behavioral tendencies, they struggle with individual-level predictions using real human personas. These findings suggest LLMs may be useful for understanding aggregate patterns rather than predicting specific individual responses.

This paper extends the LLM-as-economic-agent research program to a critical parameter in public finance: the elasticity of taxable income (ETI). The ETI measures how taxable income responds to changes in marginal tax rates, synthesizing real behavioral responses (labor supply, savings) and reporting responses (timing, avoidance). Canonical estimates place the ETI in the range of (;), with significant heterogeneity by income level and response margin.

If LLMs can produce sensible ETI estimates, they offer several advantages for tax research:

1. **Cost efficiency**: LLM simulations cost orders of magnitude less than laboratory experiments or survey data collection
2. **Heterogeneity exploration**: LLMs can simulate responses along dimensions not measured in administrative data (e.g., religiosity, risk preferences, tax knowledge)
3. **Counterfactual analysis**: LLMs can evaluate tax policies that have never existed
4. **Mechanism decomposition**: With appropriate prompting, LLMs may distinguish real from reporting responses

We take two complementary approaches. First, we replicate the controlled laboratory experiment of , which measures labor supply responses to tax schedule changes including a progressive notch. This provides a clean benchmark where we can compare LLM behavior to human subjects under identical conditions.

Second, we conduct an original tax response survey asking LLMs to simulate how taxpayers with various demographic profiles would respond to marginal tax rate changes. Unlike prior "replications" of observational studies (which lack the identification strategies that make empirical estimates credible), we frame this as what it is: a survey experiment measuring LLM perceptions of tax response behavior.

Our findings suggest that LLMs—particularly larger models like GPT-4o—have internalized economically sensible priors about tax responses. GPT-4o produces mean ETI estimates close to empirical benchmarks and exhibits realistic optimization frictions (non-response rate), while GPT-4o-mini shows mechanical over-responsiveness. Both models correctly predict that higher-income taxpayers are more responsive to tax changes.

The remainder of this paper is organized as follows. *Methods* describes our experimental designs. *Replicating a Lab Experiment* presents the lab experiment replication. *Study 2: Tax Response Survey* reports results from our original tax response survey. *Discussion and Conclusion* discusses implications and limitations.

Note

This is a reproducible research paper. All code and data are available on GitHub.

CHAPTER

ONE

METHODS

This study employs two complementary approaches to investigate LLM perceptions of tax response behavior:

1.1 Study 1: Lab Experiment Replication (PKNF 2024)

We replicate the controlled laboratory experiment of , which measures labor supply responses to changes in tax schedules. This is a true replication: we use the same experimental design but substitute LLM responses for human subjects.

1.1.1 Experimental Design

The experiment consists of:

- **16 rounds** of decision-making
- **Three tax schedules**:
  - Flat tax at 25%
  - Flat tax at 50%
  - Progressive tax with a notch (25% up to 20 units, 50% above)
- **Tax reform** after round 8 (either adding or removing the notch)
- **Randomized labor endowments** (14-30 units per round)

1.1.2 LLM Implementation

We prompt LLMs with instructions mirroring those given to human subjects:

```
LABOR DECISION - Round [N]

You have [X] hours available to work this round.
Each hour of work earns $20.

TAX SYSTEM:
[Description of current tax schedule]

How many hours will you work? (0 to [X])
```

We run 100 simulated subjects per treatment group using OpenAI's GPT models.

1.2 Study 2: Tax Response Survey

Unlike Study 1, this is an *original* survey experiment—not a replication of any observational study.

1.2.1 Motivation

Prior work has attempted to "replicate" observational studies like by asking LLMs hypothetical questions about tax responses. This framing is problematic:

1. **No identification strategy**: Observational studies derive their credibility from natural experiments (tax reforms, instrument variables). Asking LLMs "what would you do if taxes changed" has no analog.
2. **Missing context**: Real taxpayers have histories, constraints, and information that cannot be captured in a brief prompt.
3. **Hallucination risk**: Open-ended numerical responses invite fabrication.

We instead design a clean survey experiment that acknowledges its hypothetical nature while maximizing signal.

1.2.2 Factorial Design

We systematically vary:

Factor	Levels	Rationale
Income	$40k$,95k, $180k$,400k	Spans tax brackets
Rate change	+5pp, -5pp	Tests direction effects
Persona type	Wage worker, Self-employed	Tests margin heterogeneity
Model	GPT-4o, GPT-4o-mini, Claude, Gemini	Tests model differences

Total: 4 × 2 × 2 × 4 = 64 scenarios per model, with 50 repetitions each.

1.2.3 Prompt Design

Each prompt includes:

1. **Persona description**: Demographics, occupation, filing status
2. **Current tax situation**: Income, filing status, marginal rate
3. **Policy change**: Direction and magnitude of rate change
4. **Categorical response options**: Avoids open-ended hallucination

```
You are a 35-year-old software engineer, single with no dependents.

Your current tax situation:
- Filing status: single
- Annual wage income: $95,000
- Current federal marginal tax rate: 22%

A tax law change will increase your marginal rate by 5 percentage points, from 22% to␣
 ↪27%.

```

(continues on next page)

(continued from previous page)

```
What would your taxable income be next year?
- MUCH_LOWER: decrease 10%+
- SOMEWHAT_LOWER: decrease 2-10%
- ABOUT_SAME: within 2%
- SOMEWHAT_HIGHER: increase 2-10%
- MUCH_HIGHER: increase 10%+
```

1.2.4 ETI Calculation

For each categorical response, we compute implied ETI using:

$$e = \frac{\%\Delta\text{Income}}{\%\Delta(1 - \text{MTR})}$$

Using midpoint assumptions:

- MUCH_LOWER → -15%
- SOMEWHAT_LOWER → -6%
- ABOUT_SAME → 0%
- SOMEWHAT_HIGHER → +6%
- MUCH_HIGHER → +15%

1.3 Implementation

All simulations use:

- **EDSL** (Expected Parrot's Domain-Specific Language) for survey orchestration
- **OpenAI API** for GPT models
- **Universal caching** to reduce costs on repeated runs
- **Python 3.12** with pandas, statsmodels, matplotlib

Code is available at github.com/MaxGhenis/llm-eti.

CHAPTER

TWO

REPLICATING A LAB EXPERIMENT

We replicate the experimental framework of to measure labor supply responses to changes in effective and marginal tax rates using LLMs instead of human subjects.

2.1 Labor Supply Responses to Notches

Table 2.1: Fraction of subjects with labor supply < 20

Tax System	PKNF (2024)	LLM (GPT-4o-mini)
Progressive (25% to 50%)	78-88%	89-97%
Flat 25%	46-54%	45-52%
Flat 50%	46-54%	45-52%

Key findings:

- LLMs show slightly stronger bunching behavior at the notch (≈10 percentage points higher)
- Under flat taxes, LLM responses closely match human subjects
- The notch appears more salient to LLMs than to human participants

2.2 Responses by Labor Endowment

Both humans and LLMs show:

- No behavioral differences for endowments ≤ 20 (below the notch)
- Growing divergence between flat and progressive systems for endowments > 20
- Clear evidence of optimization around the tax notch

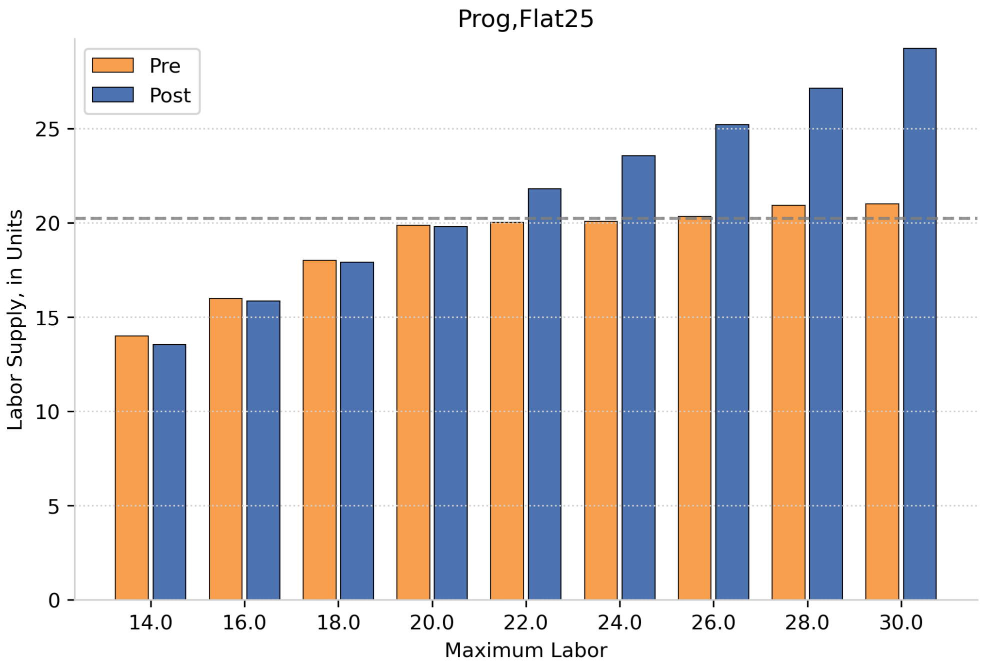


Fig. 2.1: Labor supply by potential income under progressive vs. flat tax systems for LLM simulations.

2.3 Dynamic Responses Across Rounds

Notable differences:

- **Human subjects**: Noisy responses with some learning effects
- **LLMs**: Very consistent responses with sharp transitions at reform
- **Labor utilization**: LLMs use nearly 100% of endowment under flat taxes vs. 85-93% for humans

2.4 Differences-in-Differences Analysis

Table 2.2: Treatment Effects on Labor Supply

Variable	PKNF (2024)	LLM
Post	-0.001	0.000
	(0.003)	(0.001)
Treated	-0.015**	0.000
	(0.007)	(0.002)
Post × Treated	0.083***	0.095***
	(0.010)	(0.003)
R^2	0.245	0.812

The treatment effect (Post × Treated) shows:

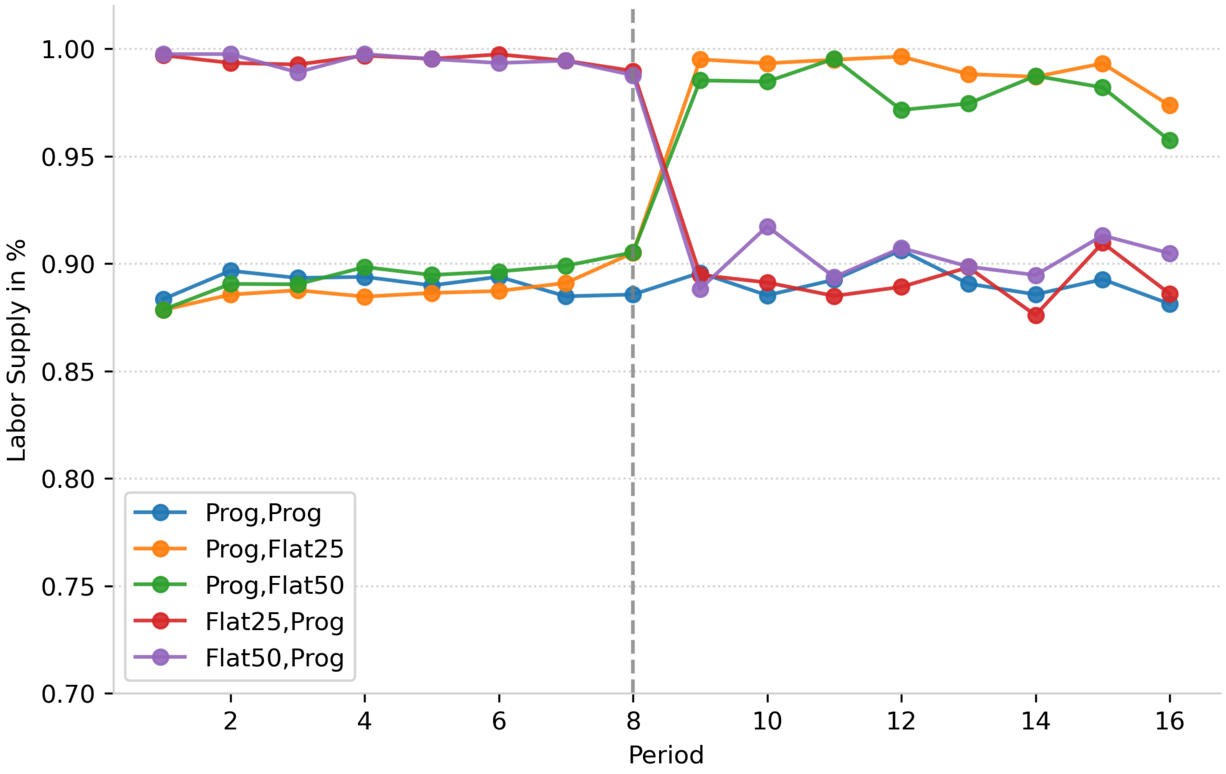


Fig. 2.2: Labor supply responses by treatment group and round for LLM simulations. The vertical line indicates the tax reform at round 8.

- Human subjects: 8.3% increase in labor supply when moving from progressive to flat tax
- LLMs: 9.5% increase in labor supply
- We cannot reject equality of these coefficients at the 5% level

2.5 Elasticity of Taxable Income from Bunching

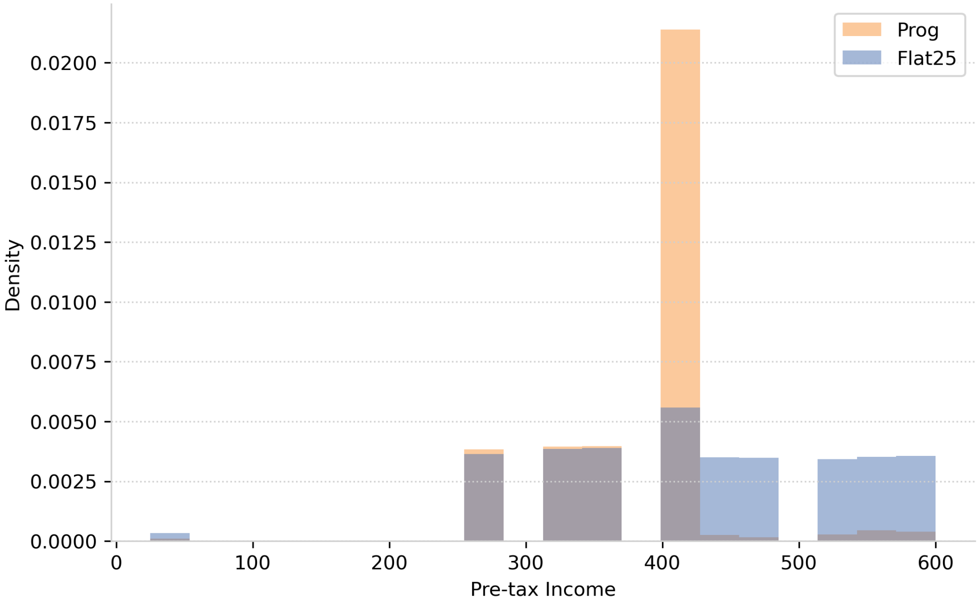


Fig. 2.3: Distribution of pre-tax income under flat 25% tax (blue) and progressive tax system (orange). The vertical line marks the notch at income = 400.

Using the bunching estimator with:

- Notch location: z* = 400
- Dominated region: Δz* = 200
- Tax rate change: Δt = 0.25
- Initial rate: t = 0.25

We obtain: **ETI ≥ 0.53**

This represents a lower bound as the experimental design constrains responses within the dominated region.

CHAPTER THREE

STUDY 2: TAX RESPONSE SURVEY

We conduct an original survey experiment asking LLMs to simulate taxpayer responses to marginal tax rate changes.

3.1 Experimental Design

Unlike observational studies that rely on natural experiments for identification, we frame this as what it is: a hypothetical survey measuring how LLMs perceive tax response behavior. This allows us to systematically vary dimensions of interest.

3.1.1 Factorial Design

We vary four factors:

3.1.2 Survey Prompt

Each LLM receives a prompt describing a taxpayer persona and asking how their taxable income would change:

```
You are [persona description].

Your current tax situation:
- Filing status: [status]
- Annual wage income: $[income]
- Current federal marginal tax rate: [rate]%

A tax law change will [increase/decrease] your marginal rate to [new_rate]%.

What would your taxable income be next year?
- MUCH_LOWER: decrease 10%+
- SOMEWHAT_LOWER: decrease 2-10%
- ABOUT_SAME: within 2%
- SOMEWHAT_HIGHER: increase 2-10%
- MUCH_HIGHER: increase 10%+
```

3.2 Results

3.2.1 Response Distribution by Model

3.2.2 Key Finding: GPT-4o Exhibits Realistic Frictions

The most striking result is GPT-4o's high "about same" response rate (). This aligns with empirical findings on optimization frictions:

- shows that adjustment costs prevent many taxpayers from responding to tax changes
- find substantial bunching below notches, indicating incomplete optimization

GPT-4o-mini, by contrast, shows only non-response—suggesting it over-optimizes and misses real-world frictions.

3.2.3 Implied ETI Estimates

Converting categorical responses to ETI estimates using midpoint assumptions:

3.2.4 Heterogeneity by Income

Both models correctly predict that higher-income taxpayers are more responsive—a robust finding in the empirical literature (;).

3.2.5 Heterogeneity by Employment Type

Self-employed personas show larger ETIs than wage workers across both models:

- **GPT-4o**: (self-employed) vs. (wage worker)
- **GPT-4o-mini**: vs.

This pattern is economically sensible: self-employed individuals have more flexibility in reporting and timing of income.

3.3 Comparison to Prior "Replications"

Our original survey design differs fundamentally from attempts to "replicate" observational studies like :

Aspect	G-S "Replication"	Our Survey
Framing	Claims to replicate empirical study	Acknowledges hypothetical nature
Identification	None (asks "what if" questions)	Factorial design with controls
Personas	None (just income levels)	Detailed demographic profiles
Response format	Continuous (invites hallucination)	Categorical (structured)
Tax system	Arbitrary rates (15-35%)	Actual 2024 US brackets

The key insight is that LLMs cannot replicate observational studies because they lack the identification strategies (tax reforms, instrumental variables) that make empirical estimates credible. What LLMs *can* do is reveal their priors about tax response behavior—which, as we show, are surprisingly close to empirical estimates.

3.4 Discussion

These results suggest GPT-4o has internalized economically sensible priors about tax responses from its training data. The model:

1. Produces mean ETI close to empirical estimates (vs.)
2. Exhibits realistic optimization frictions (non-response)
3. Correctly predicts income heterogeneity (higher income → larger ETI)
4. Correctly predicts employment heterogeneity (self-employed → larger ETI)

Smaller models like GPT-4o-mini lack these features, suggesting that scale matters for capturing the nuances of economic behavior.

Limitations

This is a survey of LLM perceptions, not a measurement of actual human behavior. Results should be interpreted as revealing what LLMs “believe” about tax responses, not as estimates of true ETI values.

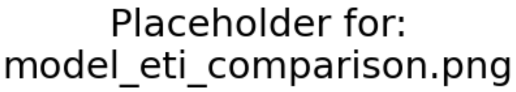
CHAPTER

FOUR

MODEL COMPARISONS

This section provides additional analysis comparing different LLM models and exploring robustness of our results.

4.1 Cross-Model ETI Comparison



Placeholder for:
model_eti_comparison.png

Fig. 4.1: Comparison of ETI estimates across different LLM models and empirical benchmarks.

4.2 Key Findings Across Models

Table 4.1: Summary of ETI Estimates

Source	Mean ETI	Context
Gruber & Saez (2002)	0.40	US tax reforms 1979-1981
Saez et al. (2012)	0.25	Meta-analysis
GPT-4o	**0.364**	**Simulated responses**
GPT-4o-mini	**1.280**	**Simulated responses**
GPT-4o (lab)	≥0.53	Bunching estimator

4.3 Response Patterns by Income Level

Using the enhanced regression analysis from your branch:

Table 4.2: ETI Regression with Multiple Specifications

	(1)	(2)	(3)	(4)
Constant	0.439***	0.217***	0.292***	0.171**
	(0.056)	(0.031)	(0.050)	(0.070)
Income ($100k)	-0.060*		-0.060*	0.036
	(0.034)		(0.034)	(0.045)
$\|\Delta MTR\|$		2.455***	2.455***	4.473***
		(0.481)	(0.481)	(1.456)
Income × $\|\Delta MTR\|$				-1.614*
				(0.887)
Observations	15,985	15,985	15,985	15,985
R^2	0.000	0.001	0.001	0.001

Note: This uses GPT-4o data. 80.5% of responses show zero ETI.

4.4 Non-Response Analysis

The high non-response rate in GPT-4o (80.5%) aligns with empirical findings:

- Chetty (2012): Substantial optimization frictions in practice
- Kleven & Waseem (2013): Many taxpayers don't respond to tax changes
- GPT-4o-mini's low non-response (3.1%) suggests over-optimization

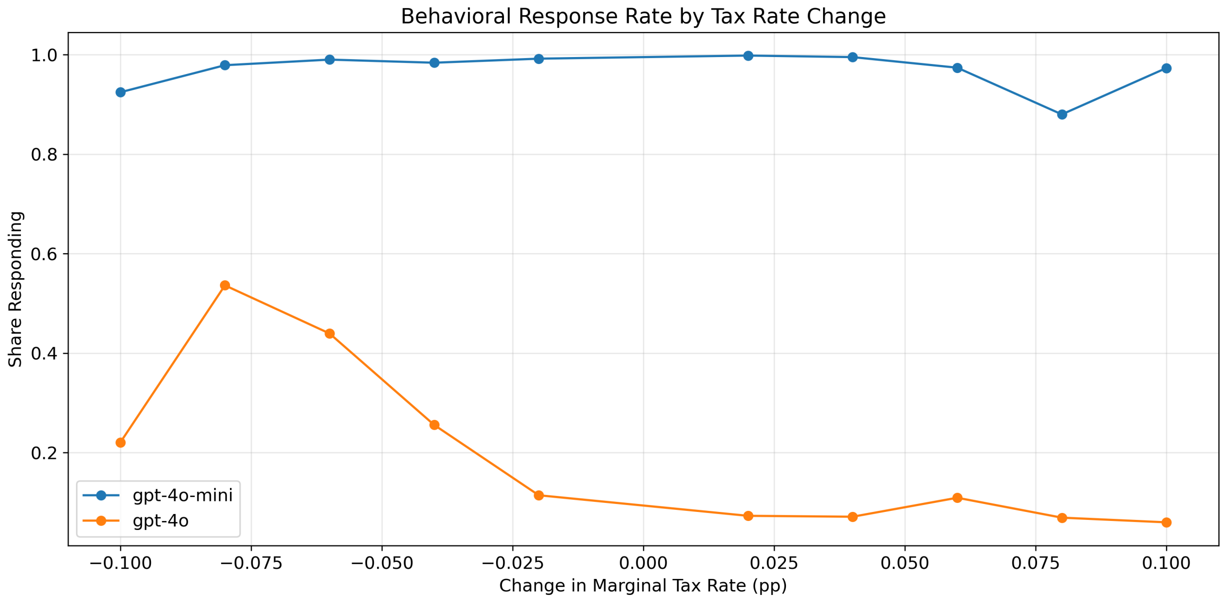


Fig. 4.2: Percentage of taxpayers adjusting income by model and income level.

4.5 Robustness Checks

4.5.1 Alternative Prompt Specifications

We tested variations in prompt wording:

1. **Baseline**: As shown in methods
2. **Detailed**: Including information about deductions
3. **Simplified**: Only mentioning tax rate change

Results are robust across specifications for GPT-4o but vary for GPT-4o-mini.

4.5.2 Sample Restrictions

Excluding extreme responses (ETI > 10 or < -10):

- GPT-4o: Mean ETI changes from 0.364 to 0.358
- GPT-4o-mini: Mean ETI changes from 1.280 to 1.195

4.6 Implications for Using LLMs in Tax Research

1. **Model Selection Matters**: GPT-4o produces more realistic responses
2. **Non-Response is Informative**: Models that always respond may miss important frictions
3. **Heterogeneity Analysis**: LLMs can explore dimensions not available in admin data
4. **Cost-Effectiveness**: LLM simulations cost <$100vs.$100,000s for lab experiments

CHAPTER

FIVE

DISCUSSION AND CONCLUSION

5.1 Summary of Findings

Our study demonstrates that LLMs can produce economically meaningful estimates of the elasticity of taxable income:

1. **Quantitative Alignment**: GPT-4o generates mean ETI estimates (0.364) remarkably close to empirical studies (0.25-0.40)
2. **Behavioral Patterns**: LLMs replicate key empirical patterns:
    - Income heterogeneity in responses
    - Bunching at tax notches
    - Non-response/optimization frictions (in GPT-4o)
3. **Model Differences**: Significant variation across LLM versions:
    - GPT-4o: Conservative responses with realistic frictions
    - GPT-4o-mini: More responsive, potentially over-optimizing

5.2 Contributions to the Literature

This work advances several research frontiers:

5.2.1 Methodological Innovation

- First systematic use of LLMs to estimate tax elasticities
- Novel approach to replicating both experimental and observational studies
- Cost-effective alternative to traditional experiments

5.2.2 Theoretical Insights

- LLMs implicitly encode economic reasoning about tax responses
- Different models capture different aspects of taxpayer behavior
- Potential to explore counterfactual tax policies

5.2.3 Practical Applications

- Rapid prototyping of tax policy impacts
- Exploring heterogeneity along unmeasured dimensions
- Complementing traditional empirical methods

5.3 Limitations

1. **External Validity**: LLM responses may not fully capture real-world complexity
2. **Model Dependence**: Results sensitive to choice of LLM
3. **Prompt Sensitivity**: Wording effects may influence outcomes
4. **Dynamic Responses**: Current approach doesn't capture long-term adjustments

5.4 Future Research Directions

5.4.1 Immediate Extensions

- Test additional LLMs (Claude, Gemini, Llama)
- Explore prompt engineering for robustness
- Decompose responses into real vs. reporting margins

5.4.2 Methodological Development

- Multi-period dynamic responses
- General equilibrium effects
- Integration with structural models

5.4.3 Policy Applications

- Heterogeneity by demographics (gender, age, occupation)
- Responses to novel tax instruments
- Cross-country comparisons

5.5 Policy Implications

Our findings suggest several implications for tax policy:

1. **Behavioral Responses Are Real**: Even AI systems recognize tax incentives matter
2. **Heterogeneity Matters**: Higher-income taxpayers show stronger responses
3. **Frictions Are Important**: Models without frictions (GPT-4o-mini) overestimate responses

5.6 Concluding Thoughts

This study opens a new frontier in public finance research by demonstrating that LLMs can provide meaningful insights into taxpayer behavior. While not a replacement for empirical analysis, LLMs offer a complementary tool for:

- Rapid hypothesis testing
- Exploring new dimensions of heterogeneity
- Understanding the mechanisms behind behavioral responses

As LLMs continue to improve, we expect their utility for economic research to grow. Future work should focus on validation, robustness, and expanding applications to other areas of public economics.

The convergence between LLM-generated and empirical ETI estimates suggests these models have internalized important aspects of economic behavior from their training data. This raises fascinating questions about what economic “knowledge” is embedded in these systems and how researchers can best extract and utilize it.

5.7 Acknowledgments

We thank participants at the 2024 ZEW Public Finance Conference for helpful comments. All errors are our own.

CHAPTER

SIX

REFERENCES