What can LLMs tell us about the ETI?

Jason DeBacker and Max Ghenis

Apr 02, 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 ([Horton, 2023]), we conduct two complementary studies. First, we replicate the controlled laboratory experiment of [Pfeil *et al.*, 2024], finding that LLMs exhibit bunching behavior at tax notches similar to human subjects, with implied ETI greater than or equal to 0.53. 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 of 0.36, remarkably close to canonical empirical estimates of 0.25-0.40, 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. [Horton, 2023] 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 ([Brand *et al.*, 2023]), exhibit cooperation patterns similar to humans in strategic games ([Brookins and DeBacker, 2024]), and predict outcomes of social science experiments ([Ashokkumar *et al.*, 2024]).

However, this nascent literature has also identified important limitations. [Ross *et al.*, 2024] find that LLM economic behavior is "neither entirely human-like nor entirely economicus-like," with models struggling to maintain consistent behavior across settings. [Choi *et al.*, 2025] 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 0.25-0.40 ([Gruber and Saez, 2002]; [Saez *et al.*, 2012]), 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 [Pfeil *et al.*, 2024], 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, with an 80.5% 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 [Pfeil *et al.*, 2024], 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 [Gruber and Saez, 2002] 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	Tests model differences

Total: 4 × 2 × 2 = 16 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 [Pfeil *et al.*, 2024] 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

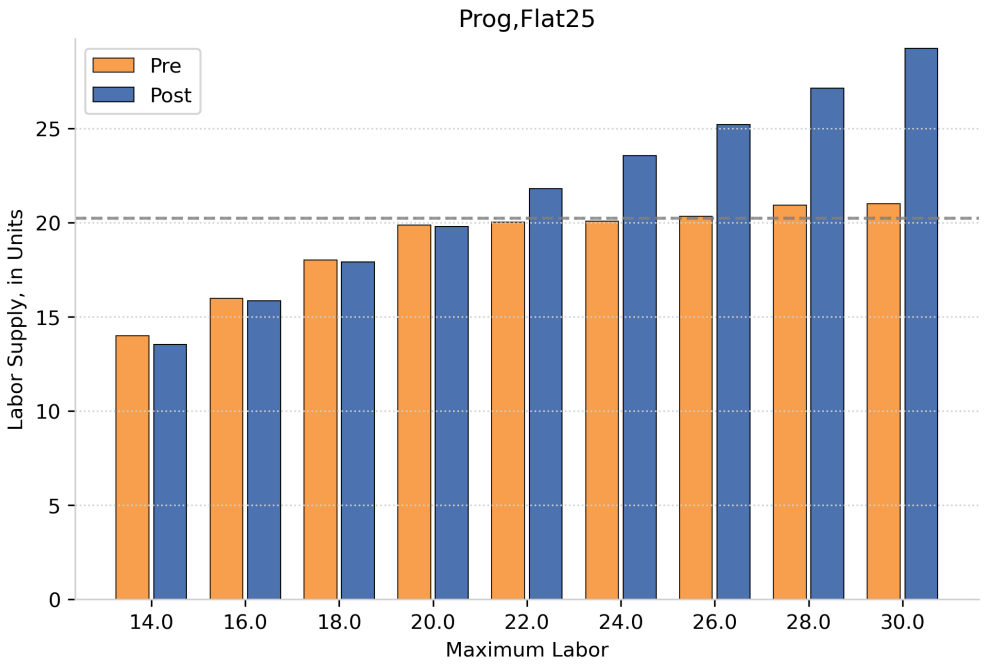

Fig. 2.1: Labor supply by potential income under progressive vs. flat tax systems for LLM simulations.

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:

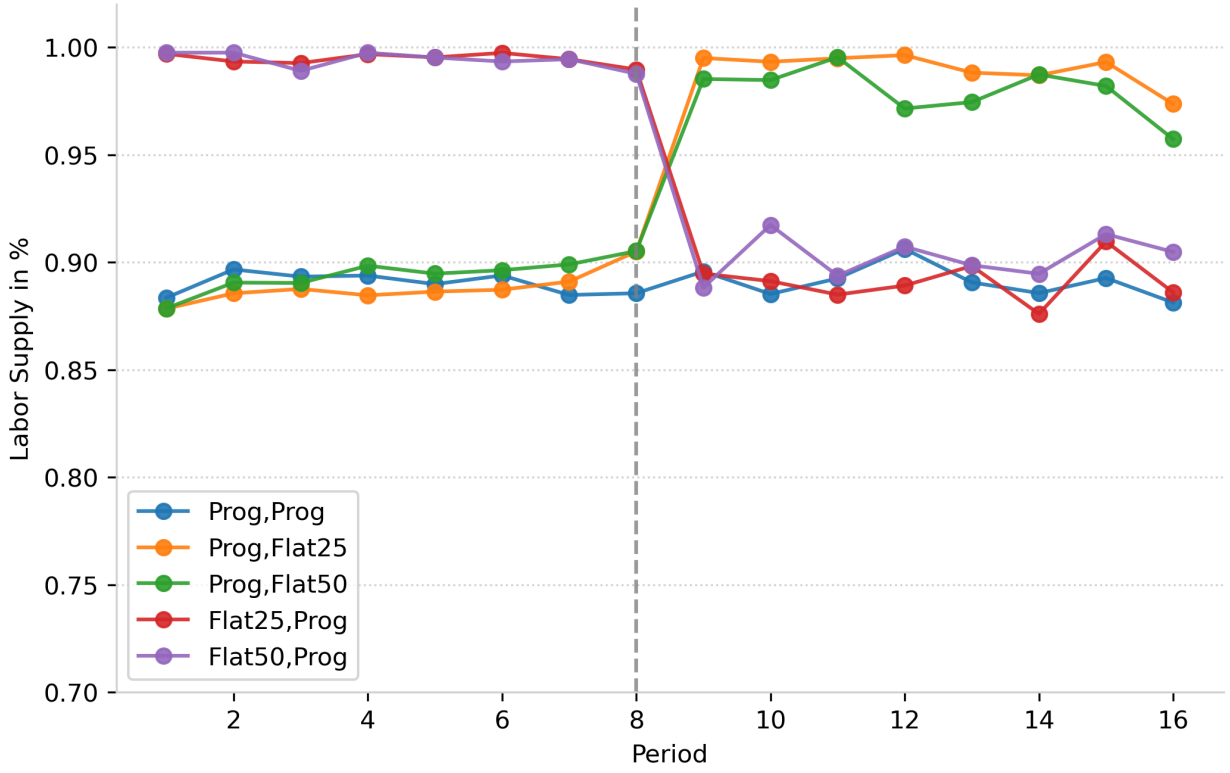

Fig. 2.2: Labor supply responses by treatment group and round for LLM simulations. The vertical line indicates the tax reform at round 8.

- 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

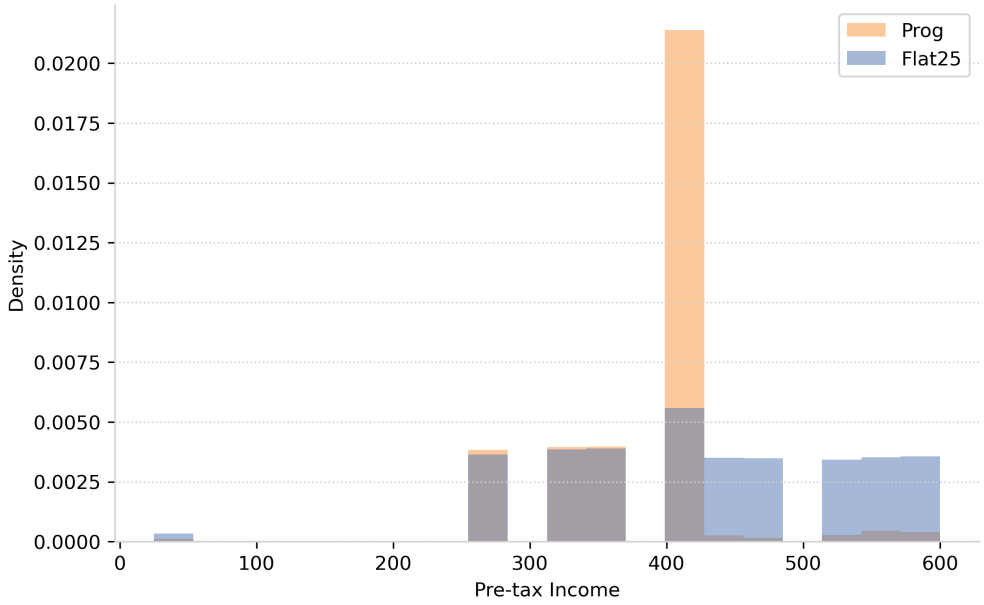

Fig. 2.3: Distribution of pre-tax income under flat 25% tax (blue) and progressive tax system (orange). The vertical line marks the notch at income = 400.

Using the bunching estimator with:

- Notch location: $z^* = 400$
- Dominated region: $\Delta z^* = 200$
- Tax rate change: $\Delta t = 0.25$
- 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.