Instrumental LLMs

Patrick Power

July 12, 2024

Abstract

In many applied microeconomic contexts, the underlying data is text - think Health Care, Education and Housing. Causal inference in this setting has typically proceeded by hand-selecting numerical representations of the text and estimating the corresponding conditional expectation function assuming that treatment or the instrument is locally randomly assigned. Recent developments in Natural Language Processing/AI though have an introduced alternative ways to produce causal estimates from text. In this paper we (1) clarify the general framework for using Large Language Models for causal inference and (2) highlight their relative strengths in the setting of Instrumental Variables.

1 Introduction

In many applied microeconomic contexts, the underlying data is text. Think Health Care, Education, and Housing. Causal inference in this setting has typically proceeded by hand-selecting numerical representations of the text and estimating the corresponding conditional expectation function, assuming the treatment or instrument is locally randomly assigned. This vector based approach to casual inference is so fundamental that it's often one of the first concepts emphasized in an Econometrics course.

> Paul Goldsmith-Pinkham writes in his first lecture– "We say that D_i is strongly ignorable conditional on a vector X_i."

With the development of Natural Language Processing models, though, a natural question to ask is what would it mean to instead condition directly on the underlying text? And as importantly, in what contexts would applied researchers prefer such an approach? We address these two questions, in this paper, in the style of the popular text *Mostly Harmless Econometrics* – meaning we prioritize the conceptual challenges that practitioners must think through rather than the large sample asymptotic properties of the estimators.

We do so first by extending the selection on observable assumption[1] to text. Formally, using terminology from *Topology* this allows researchers to speak of identification with respect to textual controls - controls such as Emotional well-being, Education, Health, Finances, Skills, Hope, Faith, Social Skills as in Evans et al. [2023]. More practically, though, it highlights how LLMs are that much more black-box than even feed-forward neural networks. That is as practitioners, we're placing no explicit restrictions on the learnt similarity between observations.

We then show that Large Language Models are potentially attractive in the context of instrumental variables with preferential treatment. That is, in the context when the offer for treatment is prioritized among the subset of people who are randomly eligible. As Figure 1 shows, low takeup rates can have an exponential effect on the sampling error of the IV estimator. Small first-stage estimates can lead to noisy IV estimates. In the context of IV with preferential treatment, though, LLM's are potentially more "efficient" than standard methods because of their representational abilities. They can incorporate information about how the treatment was prioritized making them more sample efficient.

At a high level, of course, perhaps we should have started by saying that there are a lot of good reasons for not wanting to use Large Language Models for causal inference. They are black-box estimators. They are potentially biased. They are expensive to fine-tune. They are harder to fit than linear models. They hallucinate. There is also an expanding way to construct causal estimates using LLMs – from using them for feature section, to

[1] This assumption underlies almost all popular identification strategies in applied econometrics, the paper provides a starting point for researchers interested in running diff-in-diff / instrumental variables / regression-discontinuity-designs on text. In difference-in-difference, we can think of the $\Delta\tilde{Y}_i \perp D_i | X_i$

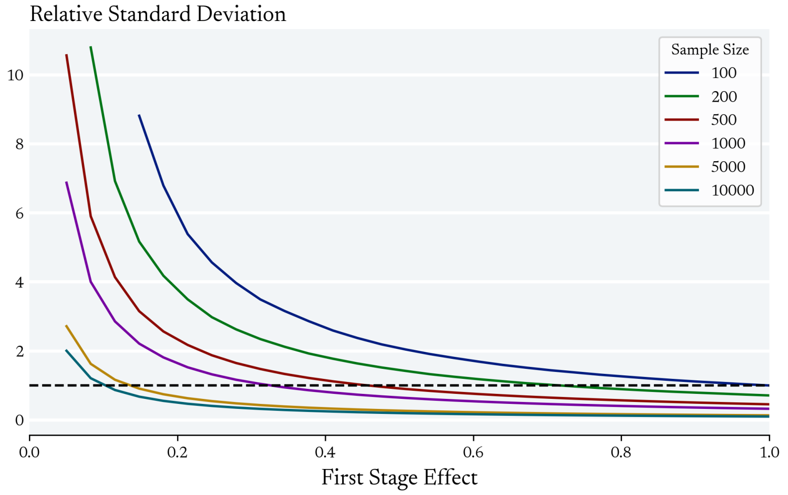


Figure 1: The standard deviation of an Instrumental Variable estimator exhibits exponential like decay with respect to the size of the first stage effect

in-context learning[2], to fine-tuning (see Figure), which makes an overview almost instantly out of date. The aim of this paper is to provide some insight about the relative tradeoffs bet

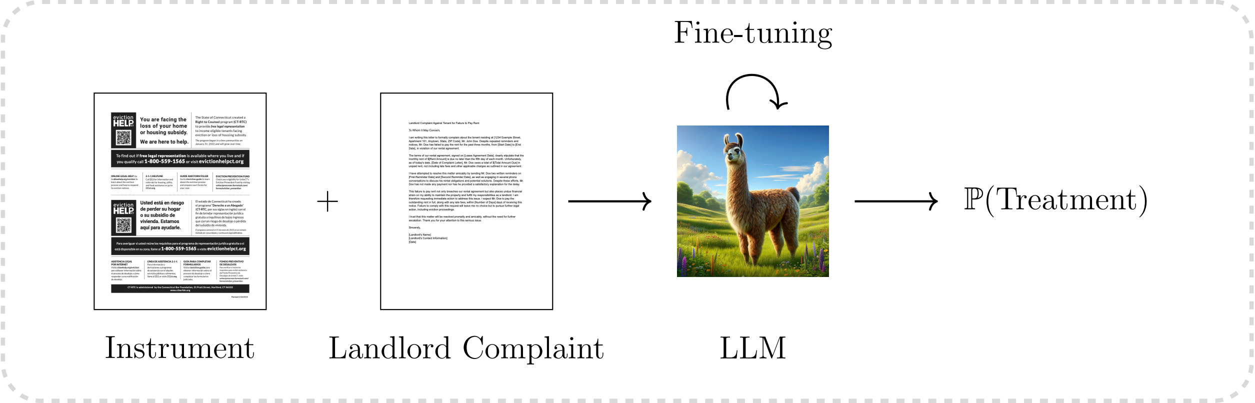


Figure 2: Illustrates how fine-tuning LLMs in the context of instrumental variables. First stage estimates are formed by passing the instrument (a text document) and the controls (another text document) as inputs to a language model which is fine-tuned to output the probability of treatment.

[2] where a frozen language takes the entire dataset in the prompt: $\mathcal{A} : \{(X_i, Z_i, D_i, Y_i)^n\} \rightarrow X \rightarrow Y$

2 Identification

2.1 The Fundamental Challenge of Language Models

The fundamental challenge in language models is that the **product topology** breaks down.[3] This is an overly mathematical way to capture the simple idea that context matters which was arguably best expressed by Firth [1957] who wrote, "You shall know a word by the company it keeps." We take this fantastic line and reinterpret it in the language of topology because topology is the coarsest mathematical construct for talking about casual identification.

A topology is formally a set of subsets of the set of interest that satisfies a series of conditions. For our purposes though, it is a way to encode similarity between elements of a set. For instance, if we think about the set of words, we can define a topology on this set which captures their semantic meaning. One element of the topology may then consist of all words which are related to sports. Another element could contain words related to statistics. And so on. Words which are in the same subsets are in some sense more similar.

We are all implicitly familiar with the standard topology – defined via the open intervals: $\mathcal{T} = \{(x, y),\ \forall y > x \in \mathcal{R}\}$.[4] One useful property of this topology is that is seamlessly extends to n-tuples of real numbers. That is a point $x := (x_1, x_2)$ is in an element of the topology $\mathcal{B} := (\mathcal{B}_1 \times \mathcal{B}_2)$ if $x_1 \in \mathcal{B}_1$ and $x_2 \in \mathcal{B}_2$. Which is to say that a point is similar to another point if it is similar across each component of the tuple.

Language doesn't cooperate with the product topology. Similar sentences have words at each position in the sentence that differ from each other in meaning. Therefore a topology defined on the set of words can not seemslessly scale to a topology on the set of sentences, the way a topology defined on the real numbers scales to a topology defined on vectors.

The recent success of large language models suggests that latently, these models have learned to construct a topology on sentences from an initial word level representations. Or as Radford et al. [2018] notes, they are "captur[ing] higher-level semantics."

2.2 Vector Based Identification

Economists tend to justify identification assumptions by arguing about local variation of the treatment. The typical assumption is that across individuals who share similar features, the treatment (or instrument) can be thought of as good as randomly assigned. That is, locally with respect to the controls there is no selection bias:

$$\mathbb{E}[\tilde{Y}_i(0)|X_i, D_i = 1] \approx \mathbb{E}[\tilde{Y}_i(0)|X_i, D_i = 0] \tag{1}$$

[3] Note: The product topology is the coarsest topology under which the projection function, $p_j : \Pi_{i \in J} V_i \to V_j$ is continuous - reference

[4] More precisely, the open intervals are a basis for the standard topology

In many areas of applied microeconomics, though, it's common for researchers to have access to the underlying documents. The entire causal inference pipeline, as reflected in Figure 3 then begins with these documents, which are mapped via the Encoder into a vector space, usually a hand-selected feature space. Then a Model, typically a linear model, but potentially a neural network, is fit to this vector space to estimate the corresponding Conditional Expectation Function.

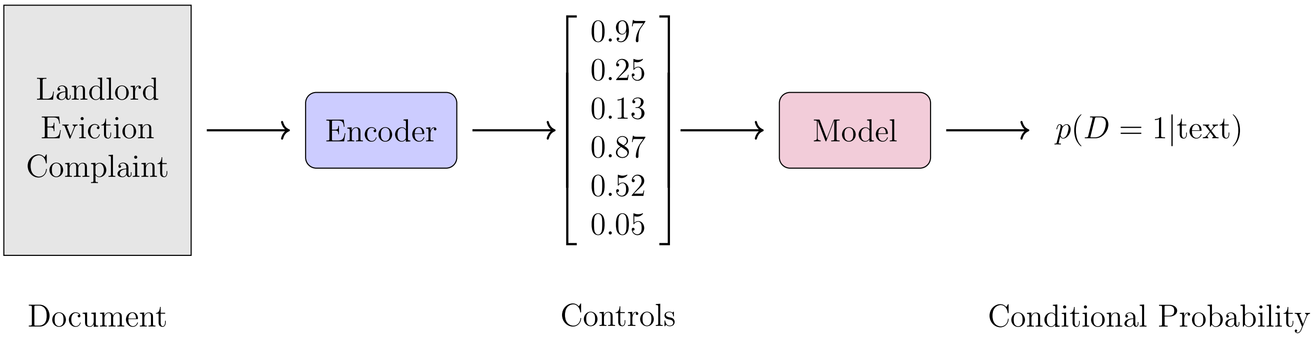


Figure 3: The Standard Pipeline

There are two key aspects in this setup. The first is that the topology on the underlying text can be defined such that the encoder is a continuous function. This allows us to reason about the conditional independence assumption in either the euclidean space defined by the vector representation of controls of more conceptually on the underlying space of text. The second key aspect is that the model is also continuous. Taken together (the composition of continuous functions is continuous), this highlights that the model exploits meaningful variation of the treatment variable. On in other words, the casual variation is preserved by the encoder and exploited by the model.

2.3 Identification with LLMs

Black-box Large Language models don't expose the encoder or the model separately. This means identification arguments cannot rely on a vector based structure. Everything must be defined or argued about with respect to the underlying textual space. Now, a practitioner may have a fuzzy idea of the underlying topology on the text space under which treatment is as good as randomly assigned but there's no way to pass this information to the language model. And even if there was, a continuous encoder would not preserve the conditional independence.

Therefore, the identification assumption with language models rests on the discrete topology. Or put another way, this means that we're not taking a stand or defining what is local. The argument at the population level is that across the exact same documents treatment is randomly assigned.

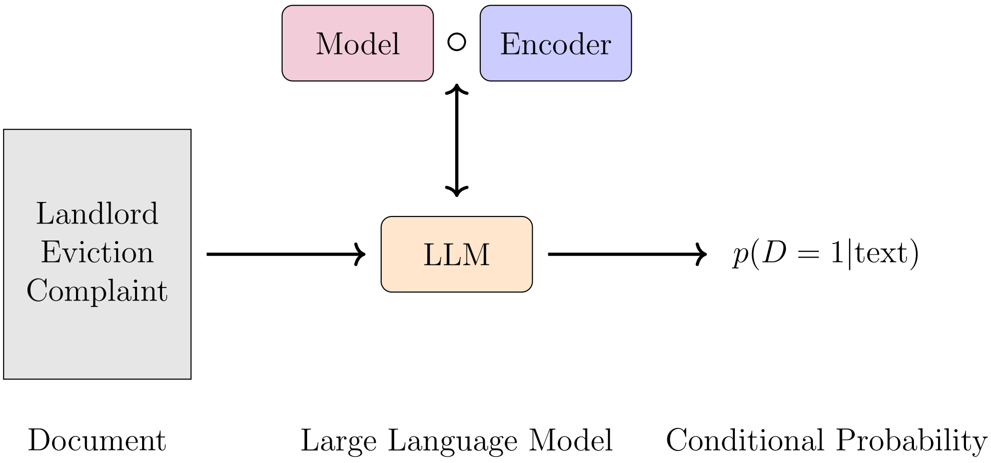


Figure 4: The Black Box Large Lange Model Pipeline

2.4 Formal Identification

We begin with the underlying probability space which is a tuple of the set containing all possible samples, a σ-algebra, and a probability measure. We subscript these terms by n to highlight their dependence on the sample size. Note this approach is different from others in the literature which emphasize an observed probability distribution and a counterfactual distribution and express the parameter of interest as a functional of the counterfactual distribution (Kennedy [2022]). With clustered data, which is a staple of empirical economics, the idea that we're sampling from the observed probability measure doesn't conceptually work.

$$\left(\Omega_n, \mathcal{F}_{\Omega_n}, \mathbb{P}_n\right) \tag{2}$$

On this space, we define the following three random variables of interest: Controls, Treatment, and Potential Outcomes, where for simplicity we assume that both treatment and outcomes are binary.[5]

$$X_i : \Omega_n \to \mathrm{L}^{\mathrm{V}} \tag{3}$$
$$D_i : \Omega_n \to \{0, 1\} \tag{4}$$
$$\tilde{Y}_i : \Omega_n \to \{0, 1\} \to \{0, 1\} \tag{5}$$

The key aspect is our setup is that the random variable X transforms the underlying probability space into a probability space defined over text - finite sequence of tokens of length L from a vocabulary of V tokens. On this space, we define the discrete topology $\mathcal{F}_{\mathrm{L}^{\mathrm{V}}}$ and assume that conditional on the text, treatment is as good as randomly assigned, which we express as follows.

$$\forall A, B, C \in \sigma(\tilde{Y}_i), \sigma(D_i), \sigma(X_i), \quad \int \mathbb{1}_{A \cap B} d\mathbb{P}_C = \int \mathbb{1}_A d\mathbb{P}_C \int \mathbb{1}_B d\mathbb{P}_C \tag{6}$$

Under this conditional independence assumption, the corresponding conditional expectation function has a casual interpretation. Note that the left hand side is integrated over the population probability space: $(\Omega, \mathcal{F}, \mathbb{P})$

$$\int_{\Omega} \mathbb{E}[\tilde{Y}_i(1) - \tilde{Y}_i(0)] d\mathbb{P} = \int_{\Omega_n} \mathbb{E}[Y_i | D_i = 1, X_i] - \mathbb{E}[Y_i | D_i = 0, X_i] d\mathbb{P}_n \tag{7}$$

3 Instrumental Variables

The typical instrumental variable setup is two stage least squares where we first predict treatment using a linear model of the controls (X) and the instrument (Z). Then in the second stage, fit a linear model to the outcome (Y) using the predicted treatment $(\hat{D})$ as a control.

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + \beta_2 X_i + \varepsilon_i, \quad \hat{D}_i = \hat{\gamma}_1 X_i + \hat{\gamma}_z Z_i \tag{8}$$

Under the Frish Waugh Lovell Theorem, this two step procedure can be understood as a single variable regression where we regress the outcome variable on a single residualized variable. The residual variable in this context is the difference between the predicted treatment based on the control and the instruments $(\hat{D}_i)$ and the predicted treatment based only on the controls $(\bar{\hat{D}}_i)$. This residualized term (which we can also capture via nonparametric methods) highlights the essence of the instrumental variable strategy: use only the local variation of the treatment variable generated by the instrument.

$$Y_i = \beta_1 \big(\hat{D}_i - \bar{\hat{D}}_i\big) + \eta_i \tag{9}$$

$$Y_i = \beta_1 \big(\mathbb{E}[D_i | X_i, Z_i] - \mathbb{E}[D_i | X_i]\big) + \eta_i \tag{10}$$

From this vantage point it's clear that flexible models are potentially more attractive in this context if they're better able to capture the local variation generated by the instrumental variable. Using real data, Figure 5 highlights how the structure imposed on the linear model can actually generated nonsensical variation: individuals offered the instrument have a negative signed residual.

For rest of the paper, we'll work with the nonparameteric residualized model, where we approximate the conditional expectation functions by fine-funning LLMs. As mentioned previously, in this setting both the instrument, the controls, and the treatment will be text.

$$Y_i = \beta_1 \big(\mathbb{E}[D_i | X_i, Z_i] - \mathbb{E}[D_i | X_i]\big) + \eta_i \tag{11}$$

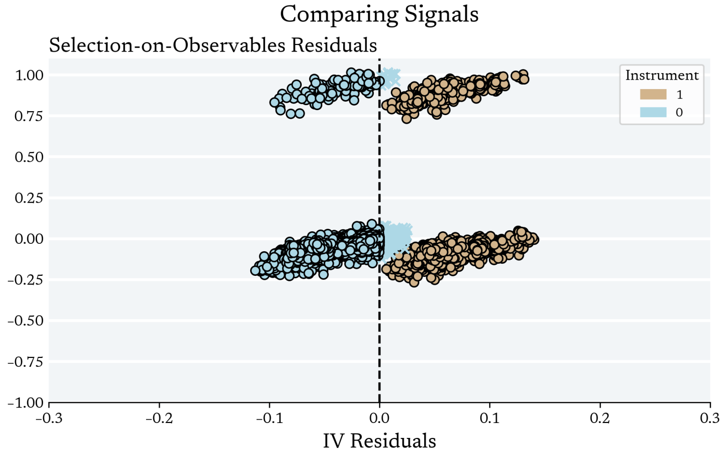


Figure 5: Scatter plot of the residual terms used in OLS and IV colored by the instrumental variable status.

$$\begin{aligned}\text{Residual}(Z=1,X) &= \mathbb{E}[D|Z=1,X]-\mathbb{E}[D|X]\\ &= \mathbb{E}[D|Z=1,X]-\big(\mathbb{P}(Z=1|X)\mathbb{E}[D|Z=1,X]\\ &\quad+(1-\mathbb{P}(Z=1|X)\mathbb{E}[D|Z=0,X]\big)\\ &= (1-\mathbb{P}(Z|X))\big(\mathbb{E}[D|Z=1,X]-\mathbb{E}[D|Z=0,X]\big)\\ &= \underbrace{(1-\mathbb{P}(Z|X))\big(\mathbb{P}(\text{Complier}|X)}_{\geq 0}\end{aligned}$$

$$\text{Residual}(Z=0,X) = \underbrace{-\mathbb{P}(Z=1|X)\big(\mathbb{P}(\text{Complier}|X)}_{\leq 0}$$

4 Efficiency

The primary motivation for including controls in the instrumental variable setup is for identification. We would like to have a sufficient set of controls such that the instrumental can be thought of as locally assigned.

$$\tilde{Y}_i \perp Z_i|X_i \qquad \tilde{D}_i \perp Z_i|X_i \tag{12}$$

A secondary reason for adding controls is that it can increase the precision of our estimate of the LATE. Intuitively, instrumental variables estimates treatment heterogeneity over a latent variable (the Compliers). One can rightly reason then that perhaps if the compliers are partially observed, we can leverage this information to do better. Using a saturated feature space, we'll show that adding controls which partially differentiate between compliers and non-compliers can (1) increase the variance of the predicted treatment, (2) has no impact

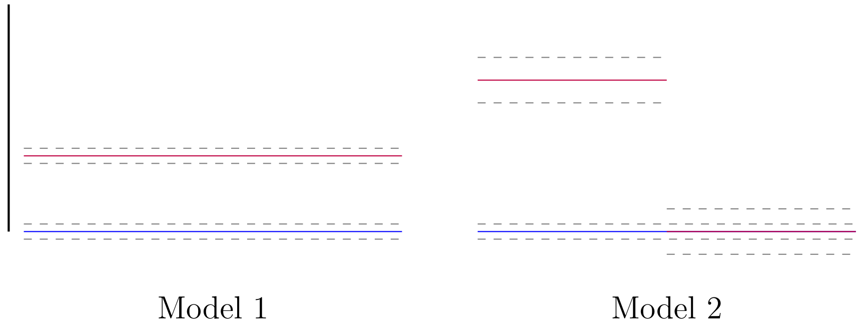


Figure 6: Caption

on the estimator of the First Stage but (3) reduces the variance of the LATE estimator. We'll then show via simulations that these results hold when we relax the saturated feature space assumption, and highlight the language models relative strength.

4.1 Saturated Feature Space

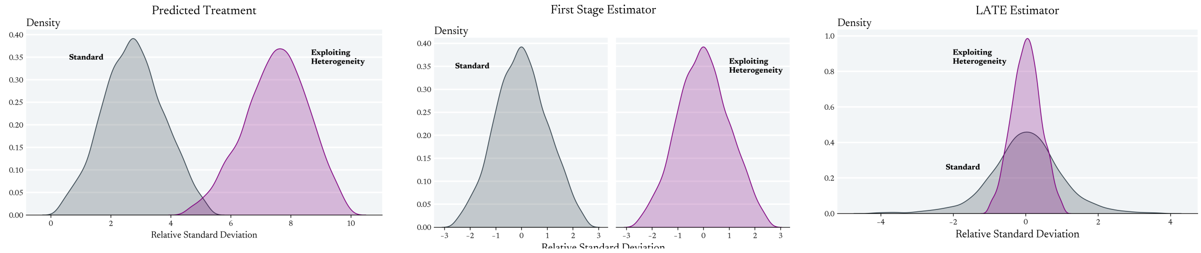
Consider the following setup. We' assume that the instrument is randomly assigned, and that we have a categorical variable X_i which is informative of who is a complier.

We then consider the following two estimators.

$$\hat{\mathbb{E}}[D_i|Z_i] \tag{13}$$

$$\hat{\mathbb{E}}[D_i|X_i, Z_i] \tag{14}$$

As figure 7a highlights, by conditioning on X, that variance of predictions increase. This isn't surprising. It's more of a sanity check on our simulation. What may be intitially suprising (although it won't be upon further reflection) is that this has no impact on the mean or standard deviation of the first stage estimator as captured in figure 7b.



(a) Predicted Treatment

(b) First Stage Estimator

(c) LATE Estimator

Figure 7

4.2 Simulation

The aim of these simulations is to demonstrate that language model – relative to feed-forward neural networks or linear models – *can* exploit characteristics which are associated with compliers. We generate synthetic observations by drawing numerical features: $x = [x_0, x_1, x_2, x_3, x_4]$. Using an Anthropic model, and the following prompt, we then map these numerical features into text, $x \mapsto t^*(x)$ so that we have a numerical representation of the features for the linear and feed-forward model and a textual representation for the LLM.

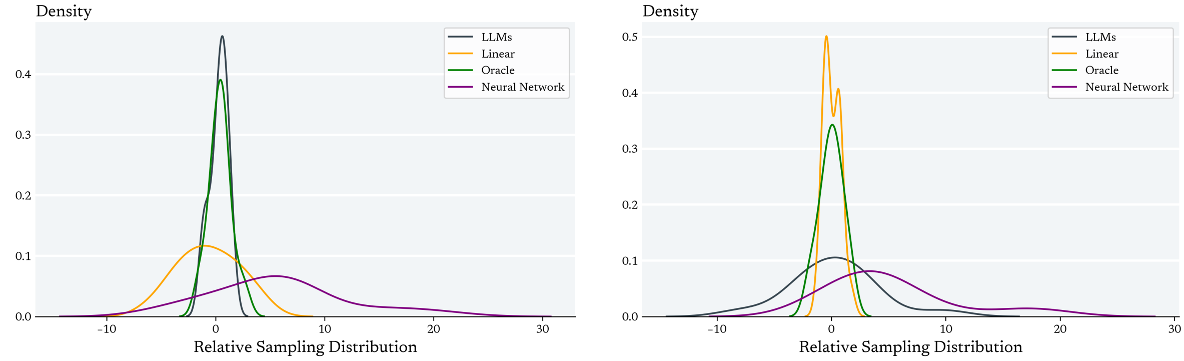
> **Prompt**
> Task: Write a paragraph description of a tenant in their {age_group} who is currently {overdue_phrase} {$}. Mention that they are in relatively {health}, live in a {living_situation}, have been living there for {months}, and have {pets}. Include some details about their {roomate_status} who {contribute_status} to the rent. Also mention somewhere that {additional_detail}

The instrumental variable is randomly assigned so we don't need to control for any features for identification reasons. A key design choice in this simulation is that (A) the first stage depends heavily on x_4 and (B) x_4 is not passed as input to the linear and feed-forward models. Or put another way, there is a "some information" in the text which is (A) highly indicative of who is a complier and (B) is not a feature that a researcher would have chosen apriori to select as a control variable. We then evaluate a fine-tuned LLM, a feed-forward neural network, a linear model with no controls, and a linear model which does control for contradicting what we said before but we label it the "oracle model" so it's not really a contradiction. We provide greater details of the simulation setup in the accompanying GitHub repository.

In figure 8, we show simulation results for two different first stage functions. Figure 8a shows the sampling distribution for the LATE effect when the treatment is completely determined by the interaction between the instrument and x_4. Since the LLM has a textual representation of this feature, and the linear and neural network models do not, the LLM is relatively more concentrated around the true parameter. In figure 8b, we simulate all of the control variables are uninformative about the takeup of the treatment variable. With no meaningful signal conveyed by the features, the LLM has greater dispersion relative to the linear model.

4.3 Leveraging Prompts

In certain policy contexts, the offer of a treatment is (conditionally) randomly assigned. But among those who receive the offer, more individuals would like to take up the offer than can be accommodated (excess demand!). As an example, consider again the roll-out of the



(a) Textual Complier Feature

(b) Noise

Figure 8: Sampling distributions under two different first stage functions

Right to Counsel across Connecticut. The offer of legal representation was made available in certain zip codes but not others. In zip codes where it was made available, demand for legal representation exceeded the availability of legal aid lawyers.[6] Legal aid lawyers therefore had to prioritize who to represent. According to one lawyer we talked with, tenants with vouchers and disabilities were often prioritized over other tenants.

The standard approach to leveraging this information would be to estimate the take-up rate of legal representation using a model that takes both the availability of legal aid as well as an indicator for a housing voucher and disability as inputs. To do so, though, we would first have to think through how what we what we might consider to be a disability. If we chose a coarse representation, when in fact certain disabilities are more prioritized than others, then our first stage estimate is less informative. On the other hand, if we choose a high dimensional representation of disabilities, then we potentailly introduce greater variance into our estimation.

$$D_i \approx g_\theta(\text{Offer}_i, \text{Disability}_i, \text{Housing Voucher}_i) + \varepsilon_i \tag{15}$$

Large Language Models are potentially attractive relative to standard methods because they don't require us to take a stand on how to represent a disability. They can potentially learn the appropriate representation directly from the data.

4.3.1 In practice

Whether they do so depends on a number of factors involving the sample size, the kind of text, the specific Large Language model, etc. But one aspect that we would like to highlight is the influence of the relative entropy of the target conditional expectation function.

$$H(Y_i|X_i) = -\sum_{x\in\mathcal{X}} P(x) \sum_{y\in Y} P(y|x) \log P(y|x) \tag{16}$$

[6] Based on conversations with legal aid lawyers across the state

Large Language models are massively overly parameterized. One hypothesis for their success is that that (1) they have their own latent topology, and (2) via-learning, they make this topology finer and finer, until the corresponding conditional expectation function can interpolate the data. More complex functions, require a finer-topology which is why we see training runs increase in length such as in Zhang et al. [2021].

5 Conclusion

The application of Large Language Models for causal inference draws one's attention to two import points. First, it makes us reflect on what it means to control for something. The second is that it forces us to understand that "'identification" is not sufficient - the question that we are truly interested in answering for the reader is why the estimate that we report in our papers has a causal estimate. The issue is that outside of a pure randomized control trial, there typically isn't a clean story for why we should interpret the estimate in a particular way.

References

Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*, 2021.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. In-context learning with long-context models: An in-depth exploration. *arXiv preprint arXiv:2405.00200*, 2024.

Vicki Boykis. What are embeddings, 2024. URL https://vickiboykis.com/what_are_embeddings/. Accessed: 2024-01-07.

Ryan Cotterell, Anej Svete, Clara Meister, Tianyu Liu, and Li Du. Formal aspects of language modeling. *arXiv preprint arXiv:2311.04329*, 2023.

Pedro Domingos. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.

William N Evans, Shawna Kolka, James X Sullivan, and Patrick S Turner. Fighting poverty one family at a time: Experimental evidence from an intervention with holistic, individualized, wrap-around services. Technical report, National Bureau of Economic Research, 2023.

John Rupert Firth. Ethnographic analysis and language with reference to malinowski's views. *Man and Culture: an evaluation of the work of Bronislaw Malinowski*, pages 93–118, 1957.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Graham Neubig. Introduction to NLP. YouTube, 2024.

Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. https://transformer-circuits.pub/2022/mech-interp-essay/index.html, 2023. An informal note on some intuitions related to Mechanistic Interpretability.

Omar Osanseviero. Understanding sentence embeddings. https://osanseviero.github.io/hackerllama/blog/posts/sentence_embeddings/, 2024. Accessed: 02-08-2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V Le. Inverse scaling can become u-shaped. *arXiv preprint arXiv:2211.02011*, 2022a.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022b.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L Griffiths. Recovering mental representations from large language models with markov chain monte carlo. *arXiv preprint arXiv:2401.16657*, 2024.

Bruno Zimmermann. Ictp diploma - topology. YouTube video, 2016. URL https://www.youtube.com/watch?v=28BluiBRdUk&t=1599s. Featured playlist of 20 videos. ICTP Mathematics.