Instrumental LLMs

Patrick Power

February 29, 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 fine-tuned large language models for causal inference and (2) highlight their relative strengths in the setting of IV with preferential treatment.

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 ubiquitous that it's often one of the first aspects emphasized in an Econometrics course: Paul Goldsmith-Pinkham writes in Definition Four of Lecture One of his applied PhD Econometrics course – "We say that D_i is strongly ignorable conditional on a vector X_i."

Recent developments in Natural Language Processing, though, have introduced an alternative way to produce causal estimates from text: fine-tuning Large Language Models. Figures 1 illustrates this setup in the context of instrumental variables. Taking the instrument (a text document) and the controls (another text document) as inputs, a fine-tuned language models learns to predict the probability of treatment. Averaging these results over a held out set of documents provides an estimate of the first stage effect.

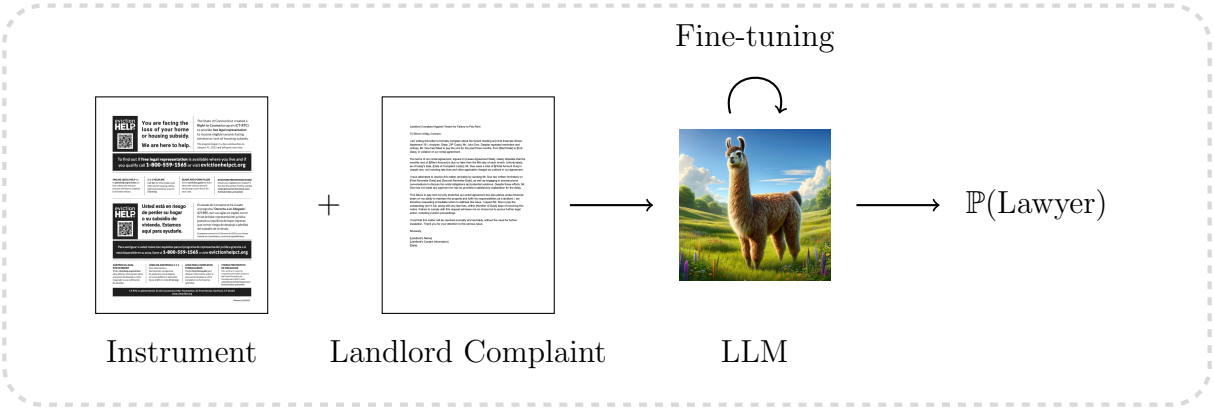


Figure 1: Fine-tuning Large Language Models: The instrument (a text document) and the controls (another text document) are the inputs to a language model which is fine-tuned to output the probability of treatment: whether a tenant has a lawyer

There are a number of issues that immediately come to mind, though, when considering this set-up. The first is how to think about identification. Economists tend develop identification strategies around local variations of the treatment. But with text, it's not clear what we mean by local. As Bengio et al. [2000] notes, language models work over discrete spaces.

The second issue is that it's not immediately evident what advantages fitting a fine-tuned large language model provides over the traditional hand-selected feature representation of the document. There hasn't been a wide adoption of flexible estimation methods like neural networks across the discipline because of lack of theory, sensitivity to hyperparameters, and lack of interpretability.[1] All of these issues apply to language models, and perhaps

[1]Reproducibility is potentially a concern as well as neural networks (even in Jax) are not guaranteed to produce the same result across difference accelerators

even more so.

The aim of this paper is to address these two questions. We do so in the following manner. First, we provide a conceptual understanding of causal inference with textual models within the potential outcome framework. We layout the mathematical framework for extending the selection on observable assumption to text. Second, we highlight the importance of generalization for causal inference. Using a residualized regression framework, we argue that better generalization generates meaning residual variation. This, we show, is particularly important in the context of instrumental variable models. And third, we highlight a particular context - instrumental variables with preferential treatment - where we believe fine-tuned language models have the opportunity to really shine. In contrast to "traditional" methods, language models can seamlessly incorporate partial information of the treatment assignment to more precisely estimate the first stage.

2 Mostly Harmless Large Language Models

In this section we overview a fundamental challenge of language models

2.1 The Fundamental Challenge of Language Models

The fundamental challenge in language models is that the **product topology** breaks down.[2] This is an overly mathematical way to capture a simple idea - 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, so it's good to get some practice thinking about it before we dive in deeper.

A topology associated with a set is a collection of subsets that satisfies the following conditions. The set and the empty set are in the topology. And, the topology is closed under arbitrary unions and under finite intersections.

A topology is a useful construct because it provides the necessary structure to begin to talk about the similarity of objects. If we think about the set of words, and define a topology on this set to capture their semantic meaning words, one element of the topology may 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 more similar.

Our first exposure to topology likely started with the real numbers. The standard topology associated with this set is defined via the open intervals: $\mathcal{T} = \{(x, y),\ \forall y > x \in \mathcal{R}\}$.[3] 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

[2]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

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

$x_1 \in \mathcal{B}_1$ 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.

Unfortunately, language doesn't cooperate with the product topology. Similar sentences have words at each place in the sentence that differ from each other in meaning. Therefore a topology defined on the set of words can not seemsly 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. That is, these models map a sequence of dense representation of words (together with the position number) into a dense representation of the sentence with the aim, as Radford et al. [2018] notes, of "captur[ing] higher-level semantics."

3 Identification

In this section we clarify the identification assumptions of causal inference with large language models.

3.1 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:

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

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 2 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.

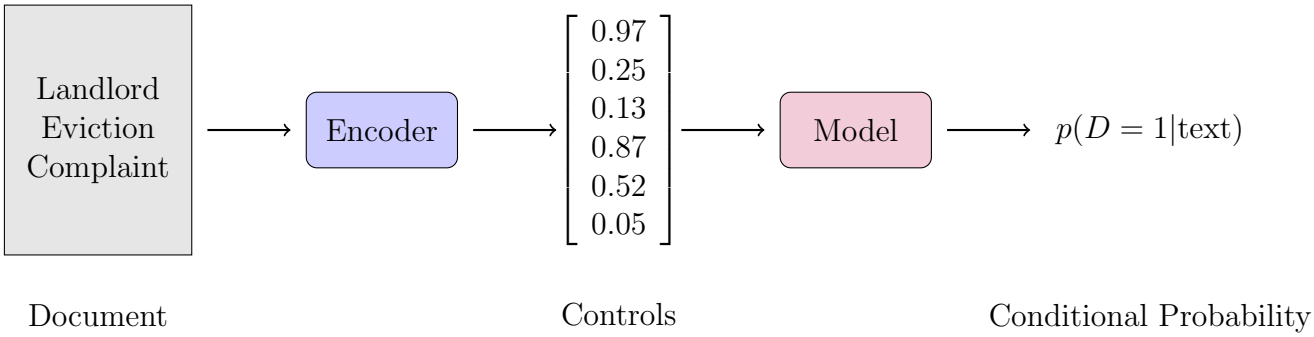


Figure 2: The Standard Pipeline

There are two key aspects in this setup. The first is that the encoder is a continuous function. When the encoder is continuous (i.e. it preserves the topology), we can 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 itself under the topology generated by the encoder. When researchers define a vector representation of the text documents, they are implicitly defining a topology on the underlying textual space.

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.

3.2 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, without knowing that the model is continuous, there's no guarantee that this topology would be preserved.

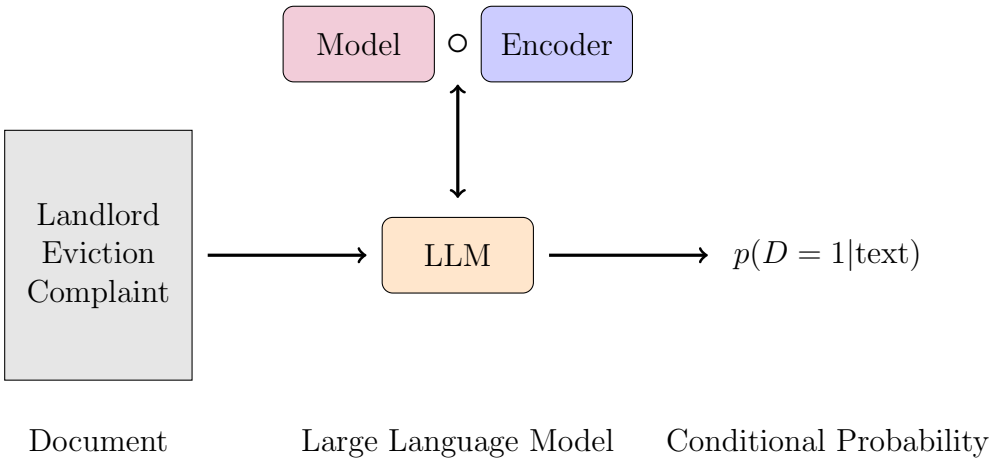


Figure 3: The Black Box Large Lange Model Pipeline

Therefore, the identification assumption with language models rests on the discrete topology. At the population level, we're not taking a stand or defining what is local. The argument is that across the exact same documents treatment is randomly assigned. Of course, in the finite sample this means that we are betting on the language's model's own representation of similarity. One can think that the language model has implicitly defined a topology and with respect to this topology, treatment is as good as randomly assigned.

Below we formalize this setup. We begin with the underlying probability space which is a tuple of the sample set (a 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.

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

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.

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

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: $(\Omega, \mathcal{F}, \mathbb{P})$

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

4 Framework

Having clarified the formal framework for causal inference with text based models, we'll restrict our focus for the rest of the paper to the residualizded approach to causal inference and in particular residualized instrumental variables.

4.1 Residualized Models

The typical approach is two stage least square where we first predict treatment using a linear model of the controls X and the instrument Z. And then in the second stage, fit a least squares model where the treatment variable has been replaced by its fitted counterpart.

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

Under the Frish Waugh Lovell Theorem, this two step procedure can be understood as regressing the outcome variable on a single residualized variable where $\bar{\hat{D}}_i$ is the fitted value of the first stage predictions based only on the controls.

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

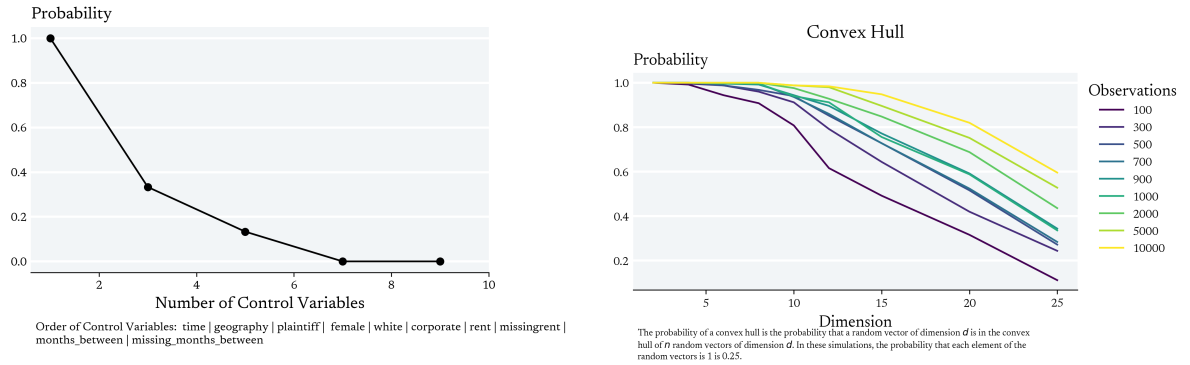
We can introduce a non-parametric equivalent of this setup by replacing the linear models with the associated conditional expectation functions. This residualized term highlights the essence of the instrumental variable strategy: we use only the local variation of the treatment variable which is generated by the instrument.

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

Having defined an instrumental variable approach in terms of conditional expectation functions, we can estimate parameters using text as the control and instrument by fine-tuning large language models as introduced before.

5 Motivation

In the finite sample, the success of these language models (as with all models) depends on their ability generalize (Balestriero et al. [2021]). Despite having within variation of the treatment at the population level, the need to generalize in the finite sample is clear – in a "typical" empirical settings, observations from the treated group lie outside the convex hull formed from observations in the control group (figure 4).



(a) Real Data

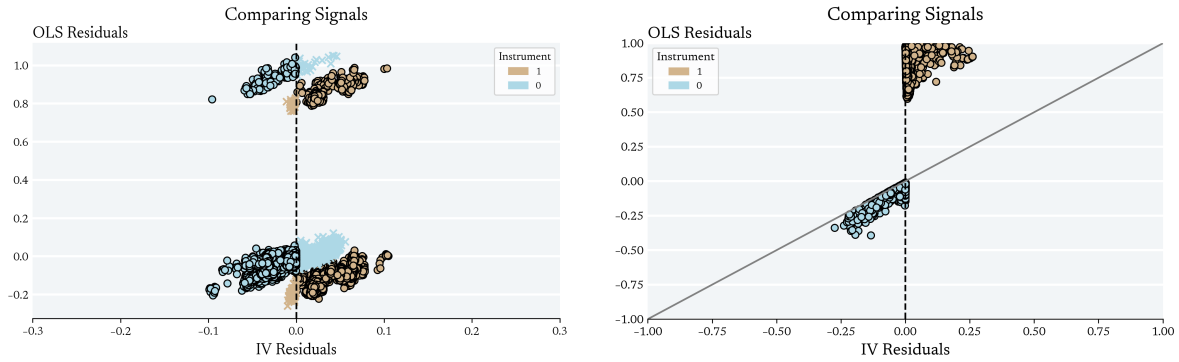
(b) Simulated Data

Figure 4: (Left) With a sample size greater than $10,000$ observations, with just a small number of controls, the probability that an observation is within the convex hull of the other observations diminishes significantly. (Right) Using simulated data, but focused on dimensions rather then control variables, the same pattern occurs.

One motivation for linear models then is that as kernel methods, we understand how they generalize because we've specified the similarity function between observations (the in-

ner product). Conversely neural networks, and large language models are black box because the similarity function is learnt during the training phase of the model (Domingos [2020]): $K(x, x') = \int_{c(t)} \nabla_\theta f_\theta(x)^T \nabla_\theta f_\theta(x') dt$.

One reason for not relying on linear models is that they can generalize poorly. As in the case of residualized IV (explained above), those with an instrument = 1 should have positive residuals while those with the instrument set to 0 should have negative residuals. As figure 5a illustrates, the linear model, by underfitting the data, doesn't necessarily satisfy this condition. Several individuals with the instrument set to 0 have positive residuals and vice-versus.[4] The non-linear language model, meanwhile has no problem satisfying the monotonicity assumption. It cleanly exploits the meaningful variation generated by the instrument.



(a) Linear Model

(b) Non-linear Model

Figure 5: Linear models underfit the data which can lead them to get the wrong sign of the residuals. Nonlinear models don't tend to underfit the data but of course are more likely to overfit the data (not shown).

In certain policy contexts we may have information about who is likely to take up the treatment. This is most likely in cases where treatment is a result of both an individual accepting the treatment, and the offer of the treatment being prioritized due to limited supply. For example, consider the context of the rollout of the Right to Counsel across Connecticut. Free legal representation was available in only certain zip codes (the instrument). However within these zip codes, given the limited supply of legal aid relative to the demand, tenants with vouchers and disabilities were often prioritized over other tenants (based on conversations with legal aid lawyers).

Language models can leverage this information to improve the precision of the first stage estimate. Mechanically we can do so by prepending each text observation with a description of how treatment is prioritized. As Figure 6 highlights fine-tuned language models with prompting can take advantage of this information – outperforming a standard large language model when the first stage noise is low – and yet is flexible enough to ignore this knowledge if it turns out not to be true – outperforming a fixed large language model

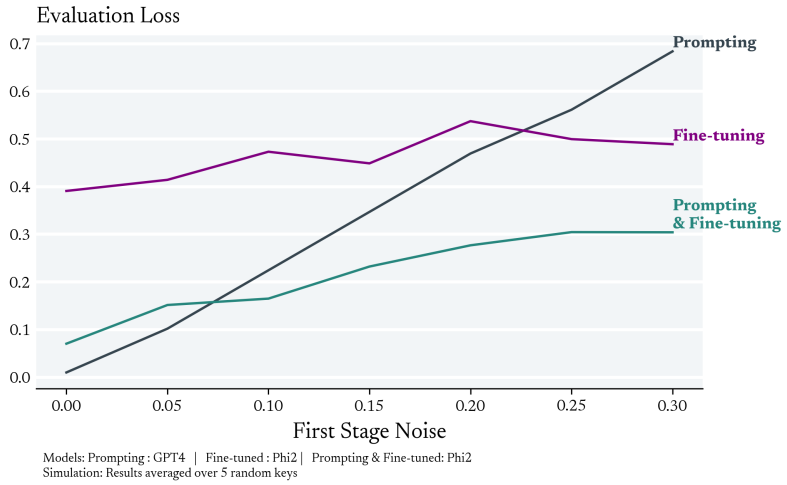


Figure 6: Optimal Evaluation Loss with Early Stopping

when the first stage noise is high.

6 Conclusion

In many contexts in applied microeconomics, conceptually, the most suitable set of controls is a textual document. One could imagine that if we were studying evictions and had access to the landlord's complaint against the tenant or were studying the Low Income Housing Tax Credit and were privy to the developers application, that if the treatment of interest was randomly assigned between semantically similar documents, we would have a great deal of confidence in our estimates.

Large Language models make this type of casal identification strategy feasible. In this paper we (1) clarify the conceptual framework of causal identification with text based models and (2) highlight their relative advantage when treatment is prioritized.

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.

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

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

[4]Using data from The Right to Counsel at Scale

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.

Graham Neubig. Introduction to NLP. YouTube, 2024.

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.

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.

7 Appendix

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

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