Mining Causality: AI-Assisted Search for Instrumental Variables*

Sukjin Han†

September 18, 2024

Abstract

The instrumental variables (IVs) method is a leading empirical strategy for causal inference. Finding IVs is a heuristic and creative process, and justifying its validity (especially exclusion restrictions) is largely rhetorical. We propose using large language models (LLMs) to search for new IVs through narratives and counterfactual reasoning, similar to how a human researcher would. The stark difference, however, is that LLMs can accelerate this process exponentially and explore an extremely large search space. We demonstrate how to construct prompts to search for potentially valid IVs. We argue that multi-step prompting is useful and role-playing prompts are suitable for mimicking the endogenous decisions of economic agents. We apply our method to three well-known examples in economics: returns to schooling, production functions, and peer effects. We then extend our strategy to finding (i) control variables in regression and difference-in-differences and (ii) running variables in regression discontinuity designs.

Keywords: Causal inference, instrumental variables, exclusion restrictions, artificial intelligence, large language models.

*I appreciate Guido Imbens for his thoughtful feedback and encouragement. I also thank Kevin Chen, Hide Ichimura and Jesse Shapiro for helpful discussions.

†School of Economics, University of Bristol. sukjin.han@gmail.com

1 Introduction

Endogeneity is the major obstacle in conducting causal inference in observational settings. Since the credibility revolution (Angrist and Pischke, 2010) and the causal revolution (Pearl, 2000), researchers in social science, statistics and other adjacent fields have developed various identification strategies to overcome endogeneity by restoring versions of natural experiments. A leading strategy is the instrumental variables (IVs) method. Over decades, researchers with their ingenuity have discovered IVs in various settings and justified their satisfaction of *exclusion restrictions* (e.g., IVs are conditionally exogenous of latent variables). With its various applicability, the IVs method has prevailed across all subfields of economics and beyond (e.g., Imbens and Angrist, 1994; Heckman and Vytlacil, 2005; Hernán and Robins, 2006).

Exclusion restrictions (or, more generally, exogeneity) are fundamentally untestable assumptions.[1] Often, in justifying them, researchers resort to *rhetorical* arguments specific to each setting. This non-statistical process follows the discovery of potential candidate IVs, which itself requires researchers' *counterfactual reasoning* and creativity—and sometimes luck. These elements all contribute to the heuristic processes employed by human researchers.

We demonstrate that large language models (LLMs) can facilitate the discovery of new IVs. Considering that narratives are the primary method of supporting IV exclusion, we believe that LLMs, with sophisticated language processing abilities, are well-suited to assist in the search for new valid IVs and justify them rhetorically, just as human researchers have for decades. The stark difference, however, is that LLMs can accelerate this process at an exponentially faster rate and explore an extremely large search space, to an extent that human researchers cannot match. It is now recognized that artificial intelligence (AI) shows remarkable performances in conducting systematic searches for hypotheses and refining the

[1] An exception is a favorable situation where one enjoys overidentifying restrictions. We discuss this point in our context below. Unlike the exclusion restriction, the IV relevance is testable from data (Stock and Yogo, 2005; Montiel Olea and Plagborg-Møller, 2021).

search (e.g., Jumper et al., 2021; Ludwig and Mullainathan, 2024). Furthermore, LLMs are argued to be capable of conducting counterfactual reasoning—or, perhaps more precisely, exploring alternative scenarios—which makes them a promising tool for causal inference.

There are at least three benefits to pursuing this AI-assisted approach to discovering IVs. First, researchers can conduct a systematic search at a speedy rate while adapting to the particularities of their settings. Second, the systematic search could increase the possibility of obtaining multiple IVs, which can then be used to formally (i.e., statistically) test their validity via overidentifying restrictions. Third, having a list of candidate IVs would increase the chances of finding actual data that contain IVs or guide the construction of such data.

We show how to construct prompts in a way that guides LLMs to search for candidates for valid IVs. The verbal translation of exclusion restrictions (among others) is used as the main component of the prompts. We propose a two-step approach in prompting that separates counterfactual statements of different complexities. We also propose to use role-playing prompts, arguing that they align with the very source of endogeneity, namely, agents' decisions.[2] By doing so, we endow LLMs with the perspective of agents, so that they mimic agents' endogenous decision-making. Role-playing prompts are also suitable for incorporating covariates.

To prove the proposed concept and assess the actual performance of an LLM, we conduct discovery exercises using OpenAI's ChatGPT-4 (GPT4), one of the leading LLMs, to find IVs in three well-known examples in empirical economics: returns to schooling, production functions, and peer effects. In all three examples, GPT4 produced a list of candidate IVs, some of which appear to be new in the literature and provided rationale for their validity. The list also contains IVs that are popularly used in the literature. Our preliminary assessment of the results suggests that the proposed method can work in practice. In the peer effect example, we also demonstrate that the proposed method can be effective in exploring relatively new topics for empirical research, which may in turn increase the possibility of

[2]For example, decisions of economic agents have been at the root of challenges for causal analyses in econometrics (e.g., Heckman, 1979; Manski, 1993).

finding novel IVs.

From a broader perspective, the proposal is to systematically "search for exogeneity." We extend the exercise to other causal inference methods: (i) searching for control variables in regression and difference-in-differences methods and (ii) searching for running variables in regression discontinuity designs. We construct relevant prompts and run them in well-known examples in the literature.

A list of candidate IVs or control variables produced as a result of the proposed method is not absolute. Rather, we hope that it serves as a valuable benchmark that inspires empirical researchers about which types of variables to consider and which domains to explore. The dialogue carried out with LLMs in the process can also help researchers solidify arguments or counterarguments for the validity of variables. After all, AI—like any machines—cannot be the ultimate authority (at least not yet). We believe a human researcher assisted by AI can choose research designs and conduct causal inference more effectively.

This paper contributes to a recent agenda in the social science literature on using AI to assist creative and heuristic parts of human research processes. In very interesting work, Ludwig and Mullainathan (2024) use generative models to systematically produce hypotheses that are comprehensible by humans in otherwise daunting settings. They make progress in research areas where the use of AI has been limited because, as they argue, establishing causal relationships in social science is an "open world" problem, unlike "closed world" problems in physical science.[3] In related work, Mullainathan and Rambachan (2024) use predictive (neural network) algorithms to recover old anomalies and discover new ones in economic theory models. Our paper does not attempt to generate hypotheses, although the new variables discovered implicitly maintain a range of hypotheses on their validity. LLMs has only very recently been used in social science research. Notably, Du et al. (2024) use fine-tuned LLMs (Meta's LLaMA in particular) to predict job transitions and understand career trajectories in labor economics. They show that the prediction accuracy remarkably

[3]The latter can be viewed as extremely difficult computation problems where machine learning makes significant progress; e.g., detecting new proteins using AlphaFold (Jumper et al., 2021) or advances in particle physics and cosmology using machine learning (Carleo et al., 2019).

outperforms those from traditional job transition economic models. Manning et al. (2024) propose to use LLMs to automate the entire process of social scientific research, from data generation to testing causal hypotheses. We employ LLMs in statistical causal inference by incorporating specific structure from econometric assumptions and allowing for human intervention in discovery processes. Overall, this paper contributes to posing philosophical questions on how social science researchers should conduct research in the AI era.

2 Notation and IV Assumptions

To formally state our discovery goal, let Y be the outcome of interest, D be the potentially endogenous treatment, $\mathcal{Z}_K \equiv \{Z_1, ..., Z_K\}$ be the list of IVs Z_k's with K being the desired number of IVs to discover, and X be the covariates. Let $Y(d, z_k)$ be the counterfactual outcome given (d, z_k). Let "$\perp$" denote statistical independence. We say Z_k is a valid IV if it satisfies the following two assumptions:

Assumption EX (Exclusion Restrictions)**.** *For any (d, z_k), (i) $Y(d, z_k) = Y(d)$ and (ii) $Y(d) \perp Z_k$ conditional on X.*

Assumption REL (Relevance)**.** *Conditional on X, the distribution of D given $Z_k = z_k$ is a nontrivial function of z_k.*

The goal of our exercise is to search for IVs that satisfy Assumptions EX and REL.[4] Suppressing X, Figure 1 depicts the causal direct acyclic graph (DAG) that implies EX and REL with $Y(d)$ being a transformation of latent confounders U. This diagram is useful in describing our procedures.

[4] One can consider a weaker version of the assumptions (i.e., mean independence and nonzero correlation). Although we do not believe our ultimate findings significantly differ from this relaxation, our prompts can reflect it.

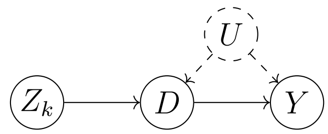

Figure 1: Causal DAG for a Validity IV (X suppressed)

3 Prompt Construction

We propose a two-step approach for IV discovery. In Step 1, we prompt an LLM to search for IVs that satisfy a verbal description of EX(i) and REL (i.e., $Z_k \rightarrow D \rightarrow Y$). In Step 2, we prompt the LLM to refine the search by selecting—among the IVs found in Step 1—those that satisfy a verbal description of EX(ii) (i.e., $Z_k \quad U$). In both steps, the prompts will involve counterfactual statements. In each step, we ask the LLM to provide rationale for its responses. This feature is useful for the user to understand the LLM's reasoning. Both steps are conducted in the same session so that Step 1's information is carried over to Step 2.[5] When submitting different queries, each two-step query should be conducted in an independent session to avoid interference across queries.

We propose a two-step approach for several reasons: First, LLMs are known to yield better performance when handing subtasks step-by-step, focusing on important details in interpreting the prompts and avoiding errors. Second, this approach creates more room for the user to inspect intermediate outputs, facilitating the evaluation of final outputs. In particular, Step 2 involves more complex counterfactual statements than Step 1, allowing the user (and the LLM) to apply varying degrees of attention when fine-tuning is needed. Third, intermediate outputs themselves can provide information and offer insights.

To simplify the exposition, we first demonstrate the prompt construction without introducing covariates in Sections 3.1–3.2 (in which case EX and REL should hold unconditionally). We then construct more realistic prompts with covariates in Section 3.3. The prompts presented in the paper can serve as a benchmark for more sophisticated prompts; we discuss them in Section 6.

[5]The feature of dynamic conversation in latest LLMs (including GPT4) enables us to do this.

3.1 Step 1: Prompts to Search for IVs

For Step 1, Prompt 1 is a role-playing prompt that queries the search for K_0 IVs ($|\mathcal{Z}_{K_0}| = K_0$) that satisfy verbal versions of EX(i) and REL (with no X). In all prompts below, each bracketed term represents a user input: `[treatment]` is the treatment D, `[agent]` is the economic agent whose decision is D, `[scenario]` is the specific setting of interest, `[outcome]` is the outcome Y, and `[K_0]` is the desired number of variables K_0. When prompting, we ask the LLM to play the role of `[agent]` to make a `[treatment]` decision in a hypothetical `[scenario]`. Examples of these inputs are given in Section 4.

Prompt 1 (Search for IVs).

```
you are [agent] who needs to make a [treatment] decision in
[scenario]. what are factors that can determine your decision
but do not directly affect your [outcome], except through
[treatment] (that is, factors that affect your [outcome]
only through [treatment])? list [K_0] factors that are
quantifiable. explain the answers.
```

It has been reported that LLMs—including GPT4—generate more tailored and unique responses when prompts are structured as role-plays.[6] In fact, in most scenarios, the explanatory variable D represents an economic agent's decision, which naturally facilitate role-playing. Additionally, role-playing prompts are more effective in guiding LLMs to respond as the relevant economic agent rather than as a researcher searching for IVs.[7] An abstract version of the prompt for comparison is presented in Appendix A.

There are at least two variants of Prompt 1 that may be useful in certain scenarios. First, instead of "`list [K_0] factors that are quantifiable`" one may simply write

[6]OpenAI Developer Forum: https://community.openai.com/t/make-chatgpt-better-for-roleplay-scenarios/344244

[7]Relatedly, we believe that it is not an issue if LLMs draw information from academic articles, given the extremely large search space. As long as LLMs identify IVs that human researchers assess as novel, we consider this acceptable.

"`list [K_0] factors.`" This may return candidates of IVs that are harder to measure but can inspire creative data collection (e.g., text or images). Second, one can expand Prompt 1 to be more specific about categorizing factors for relevant parties in a given setting. For example, in the schooling scenario (Section 4.1), we request separate lists for student factors and school factors. This approach can facilitate the user's evaluation of the results.

3.2 Step 2: Prompts to Refine the Search for IVs

Take the set of IVs, $\mathcal{Z}_{K_0} \equiv \{Z_1, ..., Z_{K_0}\}$, obtained by running Prompt 1 in Step 1. Next, for Step 2 in the same session, Prompt 2 is a role-playing prompt that queries the search for K IVs ($K \leq K_0$) within $\mathcal{Z}_{K_0}$ that satisfy a verbal version of EX(ii) (with no X). Below, `[confounders]` is the user input for unobserved confounders of concern and `[K]` is the user choice of K. In this prompt, we ask the LLM to continue playing the same role as in Prompt 1.

Prompt 2 (Refine IVs).

```
you are [agent] in [scenario], as previously described.  among
the [K_0] factors listed above, choose [K] factors that are
most likely to be unassociated with [confounders], which
determine your [outcome].  the chosen factors can still
influence your [treatment].  for each chosen factor, explain
the reasoning.
```

Unlike Prompt 1, this prompt contains a statement about variables typically unobserved to researchers, which may pose challenges. We believe that incorporating the researcher's prior knowledge on latent confounders helps simplify the overall search process and yield more desirable results.[8] For instance, in the schooling scenario, one can specify "`innate ability and personality and school quality.`" Alternatively, if

[8] This relates to *few-shot learning* discussed in Section 6.

the user prefers a more agnostic approach, they can list `[confounders]` as "`other possible factors`." Another option is to systematically search for possible unobserved confounders; see Section 5.1 for related prompting strategies. In Prompt 2, we use the term "`unassociated`." If the LLM ever captures the nuance of this word, it reflects the mean independence version of EX(ii), making the search easier. Interestingly, an alternative phasing such as "`choose [K] factors that are purely random`", which may seem a straightforward way to impose EX(ii), often fails to produce intended outputs.

There are useful variants of Prompt 2. First, one can omit `[K]` and instruct the LLM to "`choose all factors`" from $\mathcal{Z}_{K_0}$ that are likely to satisfy EX(ii), allowing the LLM to determine K independently; we apply this strategy in all examples later. Second, as a sanity check, one can direct the LLM to select elements in $\mathcal{Z}_{K_0}$ that *violate* EX(ii) in addition to those that satisfy it. This can be achieved by adding "`also choose factors that are, in contrast, associated with [confounders]`."[9]

3.3 Extension: Prompts to Search and Refine with Covariates

Typically, IVs are argued to be valid after conditioning on a list of covariates (as reflected in EX–REL). The IV discovery with covariates can be approached in at least two different ways. We can prompt the LLM to either (i) search for IVs conditional on predetermined covariates; or (ii) jointly search for IVs and covariates that satisfy EX and REL. We focus on option (i); option (ii) is discussed in Appendix A. Whenever covariates are searched, option (i) can be viewed as initiating an IV search in a new independent session with the searched covariates.

We construct a prompt that introduces the notion of conditioning variables; role-playing prompts are suitable for this purpose. Here, we only modify Prompt 2. Although REL also involves conditioning on X, we find that results are not sensitive to a relevant modification of Prompt 1. Prompt 2_x qualifies *both* `[agent]` and `[scenario]` by `[covariates]`,

[9] To gain further insights, the user can request explanations for factors that she identifies as valid IVs in initial set $\mathcal{Z}_{K_0}$ from Step 1, but which are somehow not included in the final set $\mathcal{Z}_K$ by the LLM.

the pre-determined user choice of covariates. It extends Prompt 2 by modifying the first sentence. Prompt 2_x is intended to be run after completing Prompt 1.

Prompt $\mathbf{2}_x$ (Refine IVs with Covariates)**.**

```
suppose you are [agent] in [scenario] with [covariates].
among the [K_0] factors listed above, choose [K] factors
that are most likely to be unassociated with [confounders],
which determine your [outcome].  the chosen factors can still
influence your [treatment].  for each chosen factor, explain
the reasoning.
```

The recommended approach for incorporating `[covariates]` is to assign specific values for the covariates. For instance, in the schooling scenario, one can write "`suppose you are an asian female high school student from california who considers attending a private college.`"[10] Alternatively, one can simply use terms like "`specific`" or "`particular`" along with the name of chosen covariates (e.g., "`suppose you are a high school student with specific gender, race, and regional origin who considers attending a college of specific type`").

4 Discovered IVs

Using Prompts 1 and 2_x described in the previous section, we aim to identify candidates for IVs in three examples: returns to schooling, production functions, and peer effects. These examples are chosen for their significance in the empirical economics literature (representing labor economics, industrial organization, and development economics, respectively). They commonly employ the IVs method as an empirical strategy. In this exercise, we seek to discover new IVs and verify well-known ones from the literature. The main purpose of this

[10] One can run multiple queries across different values of covariates for robustness, although this does not appear to be necessary in most cases unless extreme values are assigned in the initial run.

exercise is to evaluate the performance of LLMs in executing the proposed method and to demonstrate the practical applicability of the method. With the results produced, we hope to spark debates and inspire the discovery of new and better IVs.

The prompts we construct in each example slightly deviate from the templates of Prompts 1 and 2_x to better adapt to the scenario and enhance the flow of English language. For each example, we present results from the *initial single* run of the prompts without any curation or further refinement. Results across sessions are largely consistent, although they can vary substantially when different values of K_0 and K are chosen. By aggregating them across independent sessions, one may produce a more compelling set of IVs, as LLMs randomize responses with each run. We use GPT4 as our LLM.[11]

4.1 Returns to College

Suppose we are interested in estimating the causal effects of a college degree on earnings. The main latent confounders in this setting is unobserved individual and school characteristics (e.g., student ability and personality, school quality) that affect both the attendance decision and future earnings. To address this endogeneity and recover meaningful causal effects (e.g., local average treatment effects (Imbens and Angrist, 1994)), IVs such as distance to college and college tuition have been popularly used in the literature (Card, 1999).

The following is the prompts we use. We choose $K_0 = 40$ and let GPT4 choose K. We explicitly request separate lists for individual factors and school factors. Note that in Prompt 2_x-1, we repeat the requirement for EX(i) written in Prompt 1-1 to confirm.

[11] Although GPT4o is lighter and faster than GPT4, it is not fully functional in producing history dependent responses, which is key for our two-step discovery procedure.

Prompt 1-1 (Example: Returns to College).

```
you are a high school graduate.  you need to make a college
attendance decision.  what would be factors (factors of
schools and factors of yourself) that can determine your
decision but that do not directly affect your future earnings,
except through college attendance (that is, that affect your
earnings only through college attendance)?  list forty factors
that are quantifiable, twenty for school factors and twenty
for factors of yourself.  explain the answers.
```

Prompt 2_x-1 (Example: Returns to College).

```
suppose you are a student with family income $10K per year,
who is asian female from california, whose parents have
college education, who is catholic.  among the forty factors
listed above, choose all factors that are not associated
with your innate ability and personality and school quality,
which determine earnings.  only choose factors that affect
your earnings only through your college attendance decision.
create separate lists for school factors and factors of
yourself.  for each factor chosen, explain the reasoning.
```

Table 1 presents the results from a single session of running Prompts 1-1 and 2_x-1. It contains IVs suggested by GPT4 and GPT4's rationale for the suggestions. In the table, we find IVs that are already popular in the literature (e.g., #1, 2, 6, 14) as well as IVs that seem to be new (to our best knowledge) (e.g., #3, 5, 10, 11, 12, 13, 16, 17). The latter have potential to be valid, especially after being conditioned on additional covariates that are not considered in the prompt. Producing all these results took less than one minute in total. The rationale given by GPT4 can be elaborated further by requesting it in the same session, which we do not present here for brevity.

4.2 Production Functions

Suppose we are interested in estimating a production function that captures the causal relationship between inputs and outputs. The key identification challenge is that input decisions can be correlated with unobserved productivity shocks, which directly influence outputs. To address this, IVs such as input prices have been proposed in the literature (Griliches and Mairesse, 1995), which have been subsequently criticized (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015).

Here are the prompts we use. As before, we chose $K_0 = 40$ and let GPT4 choose K; we explicitly requested separate lists for market factors and firm and manager factors; in Prompt 2_x-1, we repeat the requirement for EX(i) written in Prompt 1-1 to confirm. Note that in Prompt 2_x-2, we use a loose description of covariates (unlike in Prompt 2_x-1 where we assigned specific values).

Prompt 1-2 (Example: Production Functions).

```
you are a manager at a manufacturing firm.  you need to make
a decision on how much labor and capital inputs to use to
produce outputs.  what would be factors (factors of markets
and economy and factors of yourself) that can determine
your decision but that do not directly affect your output
productions, except through the input choices (that is, that
affect your firm's outputs only through inputs)?  list forty
factors that are quantifiable, twenty for market factors and
twenty for managerial factors.  explain the answers.
```

#	IVs Suggested	Rationale Provided
1	Distance from home to college	Influences decisions based on familial support and potential living costs.
2	Personal savings for education	Affects the ability to afford college and choose schools without overwhelming debt.
3	Number of siblings attending college	Influences family financial resources and decisions about which college one can afford.
4	Parents' educational background	Even though parents have a college education, their influence on the decision may be more motivational rather than financial.
5	Risk aversion scale	Can determine the willingness to invest in higher-cost education or more distant schools.
6	Religious affiliations	Might seek institutions with strong religious communities, affecting the choice independently of potential earnings.
7	Personal career interests	Choosing a college based on specific programs aligned with career goals can affect earnings through the relevance and quality of education received.
8	Graduation rate	Provides an indication of the likelihood of completing a degree, which can influence future earnings by affecting the probability of graduation.
9	Retention rate	Reflects how well students are supported in their first year, influencing decisions to continue education.
10	Campus crime rates	A safety factor that can determine the desirability of attending.
11	Diversity of student body	Indicates an inclusive environment, which might be important for minority groups.
12	Availability of online courses	Offers flexibility in learning modes, which is crucial for students who may need to balance education with work or family commitments.
13	Campus facilities quality (gyms, labs, etc.)	Affects the daily student experience and could influence the decision to attend if facilities support the student's needs.
14	Percentage of students receiving financial aid	Critical for low-income students in deciding whether a college is financially accessible.
15	Proximity to urban centers	Access to broader opportunities can affect college choice and subsequent career opportunities through internships and networking.
16	Environmental sustainability rating	Might influence decision-making for environmentally conscious students.
17	Technology integration in classrooms	Reflects the school's adaptation to modern educational demands.

Table 1: Returns to College: Suggested IVs and Rationale for Validity

Notes: All IVs are discovered and explained by GPT4 from a single run of Prompts 1-1 and 2_x-1 with $K_0 = 40$ and K left unspecified. The first 7 rows are categorized by GPT4 to be student-related factors, and the next 10 rows to be school-related factors. The total running time was less than 1 minute.

Prompt 2_x-2 (Example: Production Functions).

```
suppose you are a manager at a firm with specific level of
capital intensity and specific scale of operations, which
has a specific market share in a specific industry.  among
the forty factors listed above, choose all factors that are
not influenced by productivity shocks of your firm, which
determine outputs.  only choose factors that affect your
outputs only through your input decisions.  create separate
lists for market factors and managerial factors.  for each
factor chosen, explain the reasoning.
```

Table 2 presents the results from a single session of running Prompts 1-2 and 2_x-2. It contains IVs suggested by GPT4 and GPT4's rationale. Interestingly, IVs that are suggested in the literature (i.e., input prices) are not chosen by GPT4 although they appear in the answer to Prompt 1-2 (not shown here for brevity).[12] This suggests that these IVs are not deemed by GPT4 to satisfy EX(ii), aligning with similar concerns in the literature (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015). However, GPT4 suggests IVs that may influence input prices (e.g., #1, 2, 3, 5, 6, 14), some of which can be arguably exogenous. There are a handful of other IVs suggested as market-related and managerial factors. Among the latter, there are variables related to long-term decisions of the firm, which are argued by GPT4 to not influence short-term productivity shocks. However, long-term decisions affect long-term outputs, which may or may not be relevant to the short-term outputs of concern. Overall, the explanations given by GPT4 are more detailed than those in Table 1, reflecting the random nature of the LLM's responses.

[12] This result was consistent over multiple runs.

#	IVs Suggested	Rationale Provided
1	Interest Rates	Influence the cost of borrowing, affecting decisions on how much capital to acquire or lease for production without altering current productivity levels.
2	Exchange Rates	Affect the price of imported inputs, which can change input costs and investment decisions in foreign capital, again without directly impacting the productivity of existing inputs.
3	Tariffs and Trade Policies	Alter the cost structures for imported and exported goods, impacting decisions on where to source inputs or sell outputs, but do not affect the productivity of the inputs themselves.
4	Regulatory Changes	Can necessitate changes in production processes or input materials, which may affect costs and sourcing decisions without directly impacting the productivity of the inputs once deployed.
5	Transportation Costs	Affect the total cost of inputs and can influence decisions regarding sourcing and logistics. These costs don't directly alter how effectively inputs are converted into outputs.
6	Real Estate Prices	Influence decisions about expanding or relocating production facilities, affecting capital expenditure and operational scale without altering current production efficiency.
7	Environmental Regulations	Stricter environmental regulations may force a firm to switch to greener, possibly more expensive inputs or production technologies, affecting input choices and costs but not directly modifying the productivity of any specific input set.
8	Patents and Intellectual Property Rights	Governed by legal systems and not directly influenced by a firm's internal productivity shocks.
9	Political Stability	Affect market predictability and investment decisions, impacting where and how much to invest in capital and labor, without changing how these inputs produce once acquired.
10	Global Economic Conditions	Influence market opportunities and risks associated with foreign investments, guiding input acquisition strategies without changing input productivity.

Table 2: (a) Production Functions: Suggested Candidates for IVs (Market Factors)

Notes: All IVs are discovered and explained by GPT4 from a single run of Prompts 1-2 and 2_x-2 with $K_0 = 40$ and K left unspecified. Panel (a) is categorized by GPT4 to be market-related factors, and Panel (b) to be managerial factors. The total running time was less than 1 minute.

#	IVs Suggested	Rationale Provided
10	Capital Availability	Determines its ability to invest in both labor and capital inputs, shaping the scale of operations and types of technology employed, without directly changing how productive these inputs are
11	Company's Risk Tolerance	Can decide the extent to which a firm is willing to invest in new, potentially more efficient, but riskier technologies or markets, affecting input decisions rather than the productivity of current inputs
12	Strategic Objectives	Long-term strategic objectives may dictate prioritizing certain types of inputs or production scales, influencing the firm's approach to markets and technology investments without affecting current input productivity.
13	Financial Health of the Company	The overall financial stability can limit or expand the firm's ability to procure and utilize inputs optimally, shaping how inputs are managed and financed rather than directly influencing their productivity
14	Compliance and Legal Considerations:	Driven by external legal requirements and internal ethics, not by short-term productivity
15	Corporate Social Responsibility (CSR) Initiatives	Strategic decisions about CSR are influenced by long-term planning and brand image considerations.

Table 2: (b) Production Functions: Suggested Candidates for IVs (Managerial Factors)

Notes: All IVs are discovered and explained by GPT4 from a single run of Prompts 1-2 and 2_x-2 with $K_0 = 40$ and K left unspecified. Panel (a) is categorized by GPT4 to be market-related factors, and Panel (b) to be managerial. The total running time was less than 1 minute.

4.3 Peer Effects

Suppose we are interested in the causal effects of peers on an individual's outcomes within a social network. We consider two well-known examples: (i) the effects of peer farmers on the adoption of new farming technologies (Foster and Rosenzweig, 1995; Conley and Udry, 2010); (ii) the effects of peers on teenage smoking (Gaviria and Raphael, 2001). In both examples, the main source of endogeneity is latent factors that determine the formation of network (e.g., latent homophily). To address this, the literature on peer effects sometimes uses friends of friends as IVs (Bramoullé et al., 2009; Angrist, 2014). In both examples, we construct prompts similar to the first two examples, except that we choose $K_0 = 20$.

4.3.1 Effects of Peer Farmers on New Technology Adoption

Here are the prompts. It is worth noting that, in this example, the role-playing is done from the peer's perspective, rather than from the perspective of the individual whose outcome is of concern.

Prompt 1-3-1 (Example: Peer Effects on Technology Adoption).

```
you are a farmer in a village in rural india.  you want to
influence your peer farmers in the same village to introduce
a new farming technologies that you introduced.  what would
be factors (factors of farming and village, and factors of
yourself) that can determine your influence on peers but
that do not directly affect your peers' technology adoption
decisions, except through your influence (that is, that affect
your peers' decisions only through your influence)?  list
twenty factors that are quantifiable.  explain your answer.
```

Prompt 2_x-3-1 (Example: Peer Effects on Technology Adoption).

```
suppose you are a 40 year old male farmer of a specific crop
in the village in rural india.  among the factors listed
above, which factors are not influenced by factors (e.g.,
similar background and preferences) that brought you and your
peers in the same neighborhood and social network from the
first place?  explain your reasoning.
```

Table 3 presents the results in a single session of running Prompts 1-3-1 and 2_x-3-1. It contains IVs suggested by GPT4 and GPT4's rationale. Interestingly, some IVs that are suggested in the literature (e.g., friends of friends) are not chosen by GPT4. It can be because GPT4 either views them as invalid or is incapable of identifying them. On the other hand, #10 seems to relate to the IV used in Conley and Udry (2010), which exploits variation in the presence of experienced farmers. Additionally, there are other IVs that seem to be new, notably #9.

4.3.2 Effects of Peer Teenagers on Smoking Behavior

Here are the prompts. Again, in this example, the role-playing is done from the peer's perspective. We choose a teenager in urban Indonesia for its relevance, given that the teenage smoking rate in Indonesia has been recently reported as one of the world's highest (Fithria et al., 2021). We consider a social media network in the scenario to illustrate the effectiveness of our approach in exploring relatively recent topics in the literature, thereby highlighting the potential to discover novel IVs.

#	IVs Suggested	Rationale Provided
1	Number of Training Sessions Attended	This reflects an individual's initiative to seek additional education and may vary widely among farmers with similar backgrounds due to personal choices, availability of time, or perception of the value of formal education
2	Number of Demonstrations Conducted	This is influenced by an individual's willingness to share knowledge and the resources they are willing or able to allocate for such activities, which might not be uniformly distributed even among neighbors
3	Increase in Yield After Technology Adoption	This factor depends significantly on how effectively an individual implements and adapts new technologies to their specific farming practices, which can differ greatly even among farmers growing the same crops
4	Access to Credit	Individual financial management skills, credit history, and relationships with lending institutions can influence this factor, which isn't necessarily shared among all peers in a community
5	Participation in Local Decision Making Bodies	Involvement in local governance or committees often depends on personal ambition, political skills, and community perception of an individual's leadership qualities, which can vary independently of shared initial factors
6	Awards or Recognitions Received	Achieving recognition for farming practices or community service is heavily dependent on individual effort and excellence and is not uniformly influenced by community-shared characteristics
7	Communications Skills Assessment	Communication effectiveness can be developed through personal experience, education, and innate ability, which vary among individuals regardless of their initial common background
8	Number of Articles or Case Studies Published	This is influenced by an individual's ability to write, research, and engage with broader audiences, skills that are not common to all individuals in a network
9	Engagement in Social Media or Technology Platforms	The level of engagement in modern communication and technology platforms depends on personal interest in technology, availability of resources, and individual attitudes toward innovation
10	Success Stories Documented and Shared	The ability and willingness to document and share success stories can be a personal choice influenced by an individual's perspective on self-promotion and communication

Table 3: Peer Effects on Technology Adoption: Suggested Candidates for IVs

Notes: All IVs are discovered and explained by GPT4 from a single run of Prompts 1-3-1 and 2_x-3-1 with $K_0 = 20$ and K left unspecified. The total running time was less than 1 minute.

Prompt 1-3-2 (Example: Peer Effects on Smoking).

```
you are a teenager in indonasia who smokes.  you want to
influence your peers in your social media network to smoke.
what would be factors (factors of social media, your school
and region, and factors of yourself) that can determine
your influence on peers but that do not directly affect
your peers' smoking decisions, except through your influence
(that is, that affect your peers' decisions only through
your influence)?  list twenty factors that are quantifiable.
explain your answer.
```

Prompt 2_x-3-2 (Example: Peer Effects on Smoking).

```
suppose you are a teenage boy in urban indonesia who goes
to high school and is from middle-income family.  among
the factors listed above, which factors are not influenced
by factors (e.g., similar background and preferences) that
brought you and your peers in the same social network from the
first place?  explain your reasoning.
```

Table 4 presents the results from a single session of running Prompts 1-3-2 and 2_x-3-2. It contains IVs suggested by GPT4 and GPT4's rationale. It is important to note that , given that the prompts are written from the perpective of peers, the variables in the table should be understood as factors influencing *peers* of the focal individual. Given that the setup incorporates modern elements such as social media, we identify many potentially new and interesting IVs, particularly from the social media category (i.e., #1, 2, 3, 4, 7). Interestingly, #7 can be viewed as a "friends of friends" IV.[13]

[13]Note that there were other "friends of friends" IVs that did not survive Step 2.

#	IVs Suggested	Rationale Provided
1	Engagement Rate	While the number of followers might be initially influenced by shared backgrounds, the engagement rate depends more on the content quality and how it resonates with the audience at any given time, rather than the reasons why the audience initially formed
2	Frequency of Posts	This is largely a personal choice and strategic decision, independent of peer background. It reflects personal commitment to the platform rather than shared interests or backgrounds
3	Content Reach	This metric depends on the algorithm and how engaging the content is, which can transcend the initial reasons for network formation
4	Number of Platforms Used	Choosing to diversify presence across various platforms is a strategic decision that doesn't necessarily correlate with shared backgrounds or interests
5	School Policies on Smoking	These are institutional decisions that apply to all students, regardless of the factors that brought them together
6	Local Regulations on Tobacco Advertising	Like school policies, these are determined by local government and affect all residents uniformly, irrespective of social network formations
7	Access to Smartphones Among Peers	This might vary widely even within similar socio-economic backgrounds due to individual family decisions or priorities
8	Visibility at School Events	Active participation and visibility in school events are personal choices and efforts that go beyond shared backgrounds, reflecting individual initiative
9	History of Disciplinary Actions at School	This is generally a result of personal behavior and choices rather than group influence
10	Academic Performance	Although there could be a correlation with socio-economic status, individual effort and capability play significant roles, making this somewhat independent of why peers might group together initially
11	Extracurricular Leadership Roles	Holding leadership positions is often based on personal qualities, skills, and choices rather than the shared preferences and backgrounds that might define a social network initially

Table 4: Peer Effects on Smoking: Suggested Candidates for IVs

Notes: All IVs are discovered and explained by GPT4 from a single run of Prompts 1-3-2 and 2_x-3-2 with $K_0 = 20$ and K left unspecified. The first four rows concern social media factors; the next three rows concern school and regional factors; and the last four rows concern personal factors. Given the perspective in the prompts, the variables should be understood as factors of *peers*. The total running time was less than 1 minute.

5 Variables Search in Other Causal Inference Methods

In this section, we demonstrate how prompting strategies similar to those for the IV discovery can be used to find (i) control variables under which treatments are conditionally independent (i.e., exogenous); (ii) control variables under which parallel trends are likely to hold in difference-in-differences; and (iii) running variables in regression discontinuity designs.

5.1 Conditional Independence

Using the same notation as in Section 2, consider a conditional independence (CI) assumption that assigns a more crucial role to the vector of control variables $X \equiv (X_1, ..., X_L)$:

Assumption CI. *For any d, $D \perp Y(d)|X$.*

Assumption CI is commonly introduced in causal inference settings, especially when combined with machine learning to estimate nuisance functions; e.g., debiased/double machine learning methods (Chernozhukov et al., 2024). More traditionally, this assumption is closely related to matching and propensity score matching techniques (Heckman et al., 1998). The mean independence version of CI (i.e., $E[Y(d)|D, X] = E[Y(d)|X]$) is relevant to regression methods.

We propose to use LLMs to systematically search for X that satisfies a verbal version of CI. The prompt writing is slightly simpler than that for IVs. In particular, we construct prompts that solicit the relationship between X and D (Step 1) and X and $Y(d)$ (Step 2). Therefore, *only* the second-step prompt involves a counterfactual statement. Let L_0 be the number of controls to be found in Step 1 ($L_0 \geq L$).[14]

[14]One may want to choose the value of L_0 to be larger than one would normally use for K_0 and leave L unspecified.

Prompt C1 (Search for Control Variables).

`you are [agent] who needs to make a [treatment] decision in [scenario]. what factors determine your decision? list [L_0] factors that are quantifiable. explain the answers.`

Prompt C2 (Refine Control Variables).

`among the [L_0] factors listed above, choose all factors that directly determine your [outcome], not only indirectly through [treatment]. the chosen factors can still influence your [treatment]. for each chosen factor, explain the reasoning.`

The prompts are constructed to search for confounders and need to be controlled for. Researchers sometimes mistakenly control for "colliders" and/or "mediators" (Pearl, 2000), which are intended to be exluced from the search. Note that Prompts C1–C2 can also be adapted to jointly search for covariates and latent confounders in the IV search. In this case, one can distinguish X from latent confounders by referring to the former as "`quantifiable`." Also, one may want to use the phrase "`demographic factors`" to refer to X, as they are common control variables in many empirical applications.

5.2 Difference in Differences

The difference-in-differences (DiD) method is popular in empirical research, partly due to the simplicity and intuitiveness of its main assumption, namely, the parallel trend assumption (stated below). However, this assumption is not directly testable and typically hard to justify (Ghanem et al., 2022; Rambachan and Roth, 2023). It is believed that conditioning on the right control variables can make this assumption more justifiable, which can motivate the search for such controls.

Assumption PT. $E[\Delta Y(0)|D, X] = E[\Delta Y(0)|X]$ *where* $\Delta Y(0) \equiv Y_{after}(0) - Y_{before}(0)$.

Assumption PT can be viewed as a mean independence version of CI, where the counterfactual outcome is replaced with the temporal difference of counterfactual (untreated) outcomes before and after the event. Therefore, Prompts C1–C2 can be directly used to search for X that satisfy a verbal version of PT. This can be done by inputting "`average temporal changes in [outcome_t] during the time of no [treatment]`" for `[outcome]` in Prompt C2, where `[outcome_t]` refers to Y_t for $t \in \{before, after\}$. The example of such prompts is constructed to revisit the classical empirical example, namely, the effects of minimum wage on the fast food industry's labor markets (Card and Krueger, 1994); see Appendix B.1 for the actual prompts. Table 5 contains the control variables suggested by GPT4, conditional on which the parallel trend is likely to hold, and GPT4's rationale. On the list, #3, 4, 7, 10, 11 are particularly interesting and #11 seems particularly novel. In the table, the first four rows (#1, 2, 3, 4) are chosen by GPT4 from an additional prompt that emphasizes the requirement with respect to $\Delta Y(0)$: "`be sure to choose all factors that do not determine the average wage level but only determine the temporal changes in average wages.`" Nonetheless, controls that satisfy the mean version of CI with the level, $Y_t(0)$ for $t \in \{before, after\}$, are also valid controls for PT.

5.3 Regression Discontinuity

Regression discontinuity designs (RDDs) are another well-known method for causal inference that closely relates to the IVs method (Lee and Lemieux, 2010).[15] The key for this method to work is to find a running variable (i.e., assignment variable) that satisfies the following:

Assumption RD. *There exists a variable R_j and a cutoff r_0 such that $D = 1$ if $R_j \geq r_0$ and $D = 0$ if $R_j < r_0$.*

One can use LLMs to systematically search for running variables $\{R_1, ..., R_J\}$ for a given D and Y of interest. We provide the example of prompts here. It is worth noting that, unlike in all the previous cases, none of the prompts below involve counterfactual statements.

[15] For example, the fuzzy RDD estimand can be viewed as the two-stage least squares estimand.

#	Control Variables Suggested	Rationale Provided
1	Inflation Rates	Higher inflation can decrease the real income of workers if wages don't adjust accordingly, making a case for introducing a minimum wage to maintain purchasing power.
2	Consumer Price Index (CPI)	A rising CPI without wage increases can reduce workers' real earnings, suggesting a minimum wage could help offset cost of living increases.
3	Job Vacancy Rates	Low vacancies could keep wages down due to ample labor supply; high rates may drive wages up, affecting the urgency and approach to minimum wage legislation.
4	Labor Productivity Growth	This affects temporary wage increases linked to efficiency improvements. Productivity gains can lead to bonus payments or higher wages temporarily but don't define starting wage rates.
5	Employment Growth Rates	Robust job growth in the sector might show the industry can support higher wages, while slow growth suggests caution in wage hikes.
6	Labor Force Participation Rate	High participation may suppress wages due to increased labor supply, supporting the need for a minimum wage to ensure fair earnings.
7	Union Membership Rates	Low unionization in fast food often means weaker wage bargaining, possibly justifying minimum wage laws to improve wage standards.
8	Turnover Rates	High turnover might indicate wage dissatisfaction, pointing to the need for a minimum wage to stabilize the workforce.
9	Corporate Profit Trends	If profits are high but wages are stagnant, it could suggest that the industry can afford to pay more, bolstering the argument for a minimum wage.
10	Economic Diversity Score	Dependency on low-wage sectors like fast food in less economically diverse areas might necessitate a minimum wage to ensure living standards.
11	Percentage of Workforce in Gig Economy	Increased gig work could pressure fast food employers to offer competitive wages, influencing when and how to implement minimum wage laws.

Table 5: Difference-in-Differences for Minimum Wage: Suggested Control Variables

Notes: All controls are discovered and explained by GPT4 from a single run of Prompts C1–C2, adapted to Assumption PT with $L_0 = 40$ and L left unspecified. Among them, the first four row are factors that are chosen from the additional emphatic prompt: "`be sure to choose all factors that do not determine the average wage level but only determine the temporal changes in average wages.`" The total running time was less than 1 minute.

Therefore, if LLMs outperform a traditional search for running variables, it would be due to their automated and comprehensive search behavior.[16] Similarly as above, we only specify initial J_0 and leave J unspecified.

> **Prompt R1** (Search for Running Variables).
>
> `you are [agent] who needs to make a [treatment] decision in [scenario]. what would be the possible criteria based on which your eligibility for [treatment] is determined? provide [J_0] of the most relevant criteria that are (1) quantifiable and (2) have specific cutoffs determining eligibility. explain the answers.`

> **Prompt R2** (Refine Running Variables).
>
> `among the [J_0] criteria listed above, choose all criteria that involve continuous or ordered measures and have precise cutoffs determining eligibility. also report the cutoff value for each criterion from verifiable sources only (ensuring no fabricated or hypothetical numbers are used). explain the answers.`

Note that when Prompt R2 is run on GPT4, it will engage in a series of automated web searches. The request for cutoff values may lead the LLM to provide hypothetical numbers as possibilities. When one wants to get the actual values from verifiable sources, it is important to explicitly state that, as we do above. We apply Prompts R1–R2 to a range of famous examples in the literature where RDDs are used as empirical strategies. Table 6 presents the results obtained by running the prompts, which are adapted to each specific context and country of the empirical example. In most cases, a handful of new possible running variables

[16] In further refining the candidates of running variables to ensure that RRD's continuity assumptions are satisfied, counterfactual prompting would be necessary; see Appendix B.2.

are suggested by GPT4 with specific cutoffs obtained from web sources. Except for one case (i.e., #5), GPT4 also identifies the running variables used in the literature.

6 Discussions

This paper merely serves as a starting point to pursue the proposed agenda. We discuss potential next steps. In constructing prompts for IVs, there are many possible ways for sophistication: First, one can consider using previously known IVs in the literature to guide LLMs to discover new ones. This can be done by adding textual demonstration of how Assumption EX–REL are satisfied with known IVs *before* starting the proposed prompts. This approach may evoke *few-shot learning* of LLMs (Brown, 2020), which can enhance their performances. Second, none of the results reported in the current paper are findings aggregated across sessions. To account for and potentially exploit the random aspects of LLMs, exploring the possibility of aggregation (e.g., taking the union or intersection of $\mathcal{Z}_K$'s across sessions) would be beneficial.

More broadly, we hope to further investigate the proposed approach in other empirical examples and other causal inference methods. Additionally, we can consider having a horse race among multiple LLMs or using an open-source LLM to fine-tune it (e.g., Du et al., 2024) for our purpose. For the former, a potential challenge is that the performance metric is hard to define in our context due to the lack of ground truth for valid IVs. In fact, this is the very reason we propose to use LLMs from the first place: for any IVs found by human researchers or the machine, there are only more compelling narratives or less compelling ones.[17] Finally, and related to the previous point, we hope to conduct a survey from empirical researchers to evaluate outputs obtained from the proposed methods.

[17]When data eventually come into play, overidentification tests can be a fruitful framework for the evaluation of LLMs.

#	Outcome(s) (Country)	Treatment(s)	Suggested Running Variable, Same as the Literature	Other Suggested Running Variables (Cutoffs for Eligibility)
1	Spending on schools, test scores (US)	State education aid	Relative average property values (Guryan, 2001)	- Percentage of low-income students (e.g., Equity Multiplier 2023-2024, above 70%) - Mobility rate (e.g., Equity Multiplier, above 25%) - Age (e.g., Transitional Kindergarten (TK) expansion 2023-24, 15th b-day by April 2) - Local Control Funding Formula (LCFF) (California)*
3	College enrollment (US)	Financial aid offer	SAT scores, GPA (Van der Klaauw, 2002)	- Expected family contribution (EFC) (e.g., the Pell Grant 2023-2024: below \$6,656)
4	Overall insurance coverage (US)	Medicaid eligibility	Age (Card and Shore-Sheppard, 2004)	- Federal Poverty Level (FPL) (e.g., Washington D.C.: below 215% and below 221% (family of 3); equiv. annual incomes below \$31,347 and \$54,940, reps.) - Household Size (e.g., Modified Adjusted Gross Income (MAGI) rules: expressed as % of FPL, adjusted by 5% FPL disregard)
5	Employment rates (Italy)	Job training program	Attitudinal test score (Battistin and Rettore, 2002)†	- Age (e.g., below 35; source: National Policies Platform) - Income: (e.g., below 60%; source: National Policies Platform) - Salary (e.g., EU Blue Card: above 3/2 of average Italian salary; sourse: ETIAS Italy)
6	Re-employment probability (UK)	Job search assistance, training, education	Age at end of unemployment spell (De Giorgi, 2005)	- Age (e.g., Jobseeker's Allowance (JSA): above 18, with exceptions for some 16 or 17; source: UK Rules) - Minimum Salary (e.g., Skilled Worker visa: above £38,700 or going rate for job type, whichever is higher; source: GOV.UK) - Residency Duration (e.g., JSA: above 3 months prior to claim, for new or returning UK nationals; source: UK Rules)

Table 6: Regression Discontinuity: Suggested Candidates for Running Variables

Notes: All running variables are discovered and explained by GPT4 from a single run of Prompts R1–R2, adapted to each context with $J_0 = 20$ and J left unspecified. All running variables used in the literature (Column 4) are also found by GPT4, except #5. The total running time for each row was less than 1 minute (even with an automated web search for Prompt R2). The sources indicated are given by GPT4 with links. $*$: A formula, not a running variable. $\dagger$: Not found by GPT4.

A Alternative Prompts for IV Search

A.1 Abstract Prompts with No Role-Playing

Here we present an example of abstract prompts where no role-playing is involved. Writing this version of prompt may be easier (especially to a trained social science researcher), but we believe that (i) it is not effective for either the user or the LLM in mimicking an economic agent's endogenous decision-making, and (ii) it may limit LLMs to soliciting information only from academic sources.

Prompt A1 (Search for IVs without Role-Playing).

```
what are factors that determine [treatment] of [agent] in
[scenario] but do not directly determine [agent]'s [outcome],
except through [treatment] (that is, factors that affect
[outcome] only through [treatment])? list [K_0] factors.
explain the answers.
```

Prompt A2 (Refine IVs without Role-Playing).

```
among the [K_0] factors listed above, choose [K] factors that
are most likely to be unassociated with [confounders] that
determine [outcome]. the chosen factors can still depend on
[treatment]. for each chosen factor, explain the reasoning.
```

A.2 Alternative Prompts with Covariates

Instead of using user-specified covariates in Prompt 2_x, an alternative way is to search for covariates in Step 1. Prompt 1_x below is designed to jointly search for (Z_k, X) that satisfy EX(i) and REL2:

Assumption REL2. *(i) The distribution of D given $(Z_k, X) = (z_k, x)$ is a nontrivial function of (z_k, x) and (ii) the distribution of $Y(d)$ given $X = x$ is a nontrivial function of x.*

Prompt 1_x (Search for IVs and Covariates).

```
you are [agent] who needs to make a [treatment] decision in
[setting]. what are factors that can determine your decision
but that do not directly affect your [outcome], except through
[treatment] (that is, factors that affect your [outcome] only
through [treatment])? list [K_0] factors. also, what are
your characteristics that directly influence [treatment] and
directly influence [outcome] (not just through [treatment])?
list [L_0] characteristics. explain the answers.
```

Instead of running Prompt 11 in Step 1, one can run Prompt 1_x to find an initial set of covariates and select a subset among them at her discretion to run Prompt 2_x in Step 2.

B Prompts for Section 5

B.1 Search for Control Variables in Difference-in-Differences

The following prompts are used to produce the results in Table 5 in Section 5.2. This example is motivated from Card and Krueger (1994), who explore the causal effects of minimum wage on labor market outcomes in the fast food industry.

Prompt C1-1 (Example: Minimum Wage).

```
you are the policymaker in the department of labor, deciding
whether to increase the minimum wage or not and to which state
to introduce this minimum wage law. what factors determine
your decision? list forty factors that are quantifiable.
explain the answers.
```

Prompt C2-1 (Example: Minimum Wage).

`among the forty factors listed above, choose all factors that directly determine the temporal changes in average wages at fast food restaurants, not only indirectly through the minimum wage law. the chosen factors can still determine your decision of introducing minimum wage law. for each chosen factor, explain the answer.`

B.2 Further Refining Search for Running Variables in Regression Discontinuity

An additional refinement prompt could follow Prompt R1–R2 in Steps 1–2, as detailed below. Note that this prompt involves a counterfactual statement due to `[confounders]`.

Prompt R3 (Further Refine Running Variables).

`among the criteria listed in the last answer above, choose all criteria that are difficult for you to manipulate. the chosen criteria should satisfy the following: [confounders, covariates] just below the cutoff and [confounders, covariates] just above the cutoff are not systematically different. explain the answers.`

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015): "Identification Properties of Recent Production Function Estimators," *Econometrica*, 83, 2411–2451. 4.2, 4.2

Angrist, J. D. (2014): "The Perils of Peer Effects," *Labour Economics*, 30, 98–108. 4.3

Angrist, J. D. and J.-S. Pischke (2010): "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *Journal of Economic Perspectives*, 24, 3–30. 1

Bramoullé, Y., H. Djebbari, and B. Fortin (2009): "Identification of Peer Effects through Social Networks," *Journal of Econometrics*, 150, 41–55. 4.3

Brown, T. B. (2020): "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*. 6

Card, D. (1999): "The causal effect of education on earnings," *Handbook of labor economics*, 3, 1801–1863. 4.1

Card, D. and A. B. Krueger (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–793. 5.2, B.1

Carleo, G., I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová (2019): "Machine learning and the physical sciences," *Reviews of Modern Physics*, 91, 045002. 3

Chernozhukov, V., C. Hansen, N. Kallus, M. Spindler, and V. Syrgkanis (2024): "Applied causal inference powered by ML and AI," *arXiv preprint arXiv:2403.02467*. 5.1

Conley, T. G. and C. R. Udry (2010): "Learning about a New Technology: Pineapple in Ghana," *American Economic Review*, 100, 35–69. 4.3, 4.3.1

Du, T., A. Kanodia, H. Brunborg, K. Vafa, and S. Athey (2024): "LABOR-LLM: Language-Based Occupational Representations with Large Language Models," *arXiv preprint arXiv:2406.17972*. 1, 6

Fithria, F., M. Adlim, S. R. Jannah, and T. Tahlil (2021): "Indonesian adolescents' perspectives on smoking habits: a qualitative study," *BMC Public Health*, 21, 1–8. 4.3.2

Foster, A. D. and M. R. Rosenzweig (1995): "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture," *Journal of Political Economy*, 103, 1176–1209. 4.3

Gaviria, A. and S. Raphael (2001): "School-Based Peer Effects and Juvenile Behavior," *The Review of Economics and Statistics*, 83, 257–268. 4.3

Ghanem, D., P. H. Sant'Anna, and K. Wüthrich (2022): "Selection and parallel trends," *arXiv preprint arXiv:2203.09001*. 5.2

Griliches, Z. and J. Mairesse (1995): "Production Functions: The Search for Identification," *NBER Working Paper No. w5067*. 4.2

Heckman, J. J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153–161. 2

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an econometric evaluation estimator," *The Review of Economic Studies*, 65, 261–294. 5.1

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation1," *Econometrica*, 73, 669–738. 1

Hernán, M. A. and J. M. Robins (2006): "Instruments for causal inference: an epidemiologist's dream?" *Epidemiology*, 17, 360–372. 1

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 1, 4.1

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. (2021): "Highly accurate protein structure prediction with AlphaFold," *Nature*, 596, 583–589. 1, 3

Lee, D. S. and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355. 5.3

Levinsohn, J. and A. Petrin (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 70, 317–341. 4.2, 4.2

Ludwig, J. and S. Mullainathan (2024): "Machine learning as a tool for hypothesis generation," *The Quarterly Journal of Economics*, 139, 751–827. 1

Manning, B. S., K. Zhu, and J. J. Horton (2024): "Automated social science: Language models as scientist and subjects," *National Bureau of Economic Research*. 1

Manski, C. F. (1993): "Identification of endogenous social effects: The reflection problem," *The Review of Economic Studies*, 60, 531–542. 2

Montiel Olea, J. L. and M. Plagborg-Møller (2021): "Local projection inference is simpler and more robust than you think," *Econometrica*, 89, 1789–1823. 1

Mullainathan, S. and A. Rambachan (2024): "From predictive algorithms to automatic generation of anomalies," *National Bureau of Economic Research*. 1

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297. 4.2, 4.2

Pearl, J. (2000): *Causality: Models, Reasoning, and Inference*, Cambridge, UK: Cambridge University Press. 1, 5.1

Rambachan, A. and J. Roth (2023): "A more credible approach to parallel trends," *Review of Economic Studies*, 90, 2555–2591. 5.2

Stock, J. H. and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. Andrews and J. H. Stock, Cambridge: Cambridge University Press, 80–108. 1