Mining Causality:

AI-Assisted Search for Instrumental Variables*

Sukjin Han†

October 22, 2025

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 dramatically accelerate this process and explore an extremely large search space. We demonstrate how to construct prompts to search for potentially valid IVs. We contend that multi-step and role-playing prompting strategies are effective for simulating the endogenous decision-making processes of economic agents or social actors and for navigating language models through the realm of real-world scenarios. We apply our method to four well-known examples in economics: returns to schooling, demand and supply, and peer effects. Expert surveys reveal that some of the discovered IVs in each domain appear both novel and likely valid.

†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), researchers in social science have developed various identification strategies to overcome endogeneity by restoring versions of quasi-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 social science and beyond (e.g., Imbens and Angrist, 1994; Blundell and Powell, 2003; Heckman and Vytlacil, 2005; Hernán and Robins, 2006).[1]

Exclusion restrictions are fundamentally untestable assumptions.[2] 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 done 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

---

[1]See Mogstad and Torgovitsky (2024) for a more recent survey.

[2]An exception is a favorable situation where one enjoys over-identifying restrictions. We discuss this point in our context below. Unlike the exclusion restriction, the IV relevance is testable from data (Olea and Pflueger, 2013; Stock and Yogo, 2005).

human researchers cannot match. It is now recognized that LLMs are capable of conducting counterfactual reasoning—or, perhaps more precisely, exploring alternative scenarios—which makes them a promising tool for causal inference (Jin et al., 2023a,b; Kiciman et al., 2023).

There are at least four benefits to pursuing this machine-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, interacting with these tools can inspire ideas for possible domains for novel IVs. Third, the systematic search could increase the possibility of obtaining multiple IVs, which would then enable formal (i.e., statistical) testing of their validity via over-identifying restrictions. Fourth, having a list of candidate IVs would increase the chances of finding actual data that contain IVs or guide the construction of such data, including the design of (natural) experiments to generate IVs.

We show how to construct prompts that guides LLMs to search for candidates for valid IVs. The text representation of exclusion restrictions (among others) is the main component of the prompts. We propose a multi-step approach in prompting that divides a discovery task into multiple subtasks, and thus separates counterfactual statements of different complexities. At the same time, we propose using role-playing prompts, arguing that they align with the very source of endogeneity, namely, agents' decisions.[3] By doing so, we equip LLMs with the perspective of agents, enabling them to mimic agents' endogenous decision-making processes and gather contextual information in realistic scenarios. Another benefit of multi-stage, role-playing prompts is that they help navigate language models through the realm of real-world scenarios, rather than anchor them within the narrow realm of academic discourses on IVs. Each stage's prompt focuses only on a portion of the IV assumptions, framed as an agent's real-world problem, thereby obscuring the IV-discovery purpose and maximizing the chance of finding novel IVs.[4]

To prove the proposed concept and illustrate the actual performance of an LLM, we

---

[3]Decisions of economic agents or social actors have been at the root of challenges for causal analyses (Heckman, 1979; Manski, 1993).

[4]Even if LLMs circumvent this safeguard and exhibit *memorization* from academic texts, we still find value in the procedure as long as the list of discovered IVs includes those that are recognized as new by researchers.

conduct discovery exercises using OpenAI's ChatGPT-4 (GPT4), one of the leading LLMs, to find IVs in four well-known examples in empirical economics: returns to schooling, demand and supply, and peer effects. In all four 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. 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 finding novel IVs. We then elicit expert evaluations through a survey, which reveal that some of the discovered IVs in each domain are novel and at least as valid as the most established IVs in the relevant literature.

From a broader perspective, the proposal is to systematically "search for exogeneity." In this spirit, the procedure can be extended to other causal inference methods, such as (i) searching for proper control variables in regression and difference-in-differences analyses and (ii) searching for new running variables in regression discontinuity designs. We construct relevant prompts, apply them to well-known examples in the literature, and successfully identify new variables.

A list of candidate IVs or control variables produced as a result of the proposed method is far from absolute. Rather, 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, artificial intelligence (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.

We contribute to a growing economics literature on using machine learning and AI to assist research processes. This agenda views these methods as not only data-processing and prediction tools but also as tools that can improve conventional research practices themselves (Mullainathan and Spiess, 2017; Athey and Imbens, 2019; Gentzkow et al., 2019; Dell, 2025;

Ludwig et al., 2025). Such practices include the creative and heuristic aspects of research, such as hypothesis generation and empirical design. The use of AI tools (e.g., self-supervised learning, deep learning, generative models) in these stages of scientific discovery has become increasingly common in the natural sciences (see the comprehensive survey in *Nature* by Wang et al., 2023), but remains less common in the social sciences. In interesting work, Ludwig and Mullainathan (2024) and Mullainathan and Rambachan (2024) use generative models to systematically produce hypotheses that are comprehensible by humans in otherwise daunting settings. While we do not explicitly seek to generate hypotheses, the variables we discover implicitly maintain a range of hypotheses about their validity. Regarding the use of foundation models such as LLMs in economic research, Du et al. (2024) use fine-tuned LLMs to predict job transitions and understand career trajectories in labor economics, and 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.

This paper also relates to the approach of using LLMs in causal discovery (see Wan et al., 2024, for a recent survey and references therein). However, the fundamental difference of our approach to this line of work is that we use LLMs to systematically discover variables with particular causal structure rather than using LLMs to find causal links among a *pre-determined* set of variables.

The paper is organized as follows. Section 2 states the IV assumptions, and Section 3 proposes the main idea of IV discovery along with the prompting strategies. Section 4 provides the examples of discovered IVs, and Section 5 contains the expert evaluation of them. Section 6 discusses the use of an adversarial LLM to review and refine the discovery process. Section 7 demonstrates that an alternative prompting strategy that discloses the IV-search purposes leads to poor results. Section 8 extends the approach to other causal inference settings. Section 9 concludes. The Appendix includes alternative prompting strategies for IV search, and additional prompts and findings in the IV discovery examples.
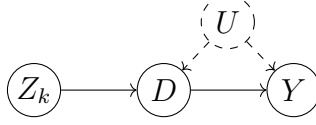
Figure 1: Causal DAG for a Valid IV (X suppressed)

2 IV Assumptions

We first 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). We say Z_k is a valid IV if it satisfies the following assumptions:

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

Assumption EX (Exclusion). *For any (d, z_k), $Y(d, z_k) = Y(d)$.*

Assumption IND (Independence). *For any d, $Y(d) \perp Z_k$ conditional on X.*

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

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 REL and EX (i.e., $Z_k \longrightarrow D \longrightarrow Y$). 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 IND (i.e., Z_k U).[5] In both steps, the prompts will

---

[5]In the Appendix, we present a three-step prompting that focuses on each of REL, EX and IND in each step.

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.

We propose a multi-step approach for several reasons: First, LLMs are known to yield better performance when handling subtasks step-by-step, focusing on important details in interpreting the prompts and avoiding errors (Wu et al., 2022). Second, this approach creates more room for the user to inspect intermediate 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. Finally, this approach—especially when each step is executed in a separate session—significantly reduces the likelihood that LLMs recognize the task as IV discovery and generate text from relevant academic sources.

Alongside the multi-step method, we propose a role-playing approach. When prompts are structured as role-plays, LLMs tend to gather better contextual information, exhibit improved reasoning performances, and generate more tailored and unique responses (Kong et al., 2023). 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.

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 (obtaining $\mathcal{Z}_{K_0}$) that satisfy verbal versions of REL and EX (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.

7

Prompt 1. *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.*

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 "list [K_0] factors" or even "list [K_0] factors that are hard to quantify." This would return candidates of IVs that are harder to measure but can inspire creative data collection (e.g., text, images, or other unstructured data). 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 and school factors. This approach can facilitate the 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, Prompt 2_x is a role-playing prompt that queries the search for K IVs (obtaining $\mathcal{Z}_K$, $K \leq K_0$) within $\mathcal{Z}_{K_0}$ that satisfy a verbal version of IND (with no X). Below, [confounders] is the user input for unobserved confounders of concern and [K] is the user choice of K. In addition, we introduces the notion of conditioning variables in the prompt to reflect covariates X in IND;[6] role-playing prompts are suitable for this purpose. 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], the pre-determined user choice of covariates.

Prompt 2$_x$. *suppose you are [agent] in [scenario] with [covariates]. among the [K_0] factors listed above, choose [K] factors that are most*

---

[6]The Appendix presents an alternative approach to introducing covariates, namely, jointly searching for IVs and covariates that satisfy REL–IND.

likely to be unassociated with [confounders], which determine your
[outcome]. the chosen factors can still influence your [treatment].
for each chosen factor, explain your 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. For instance, in the schooling scenario, one can specify "innate ability and personality and school quality."[7] In Prompt 2_x, we use the term "unassociated." If the LLM ever captures the nuance of this word, it reflects the mean independence version of IND, making the search easier.

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."[8] 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").

There are useful variants of Prompt 2_x. First, one can omit [K] and instruct the LLM to "choose all factors" from $\mathcal{Z}_{K_0}$ that are likely to satisfy IND, 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* IND in addition to those that satisfy it. This can be achieved by adding "also choose factors that are, in contrast, associated with [confounders]." To gain further insights, the user can request explanations for factors that she identifies as valid IVs in the 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. We apply the last approach to the application in Section 4.2.1.

---

[7]An alternative option is to systematically search for possible unobserved confounders.

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

4   Discovered IVs

Using Prompts 1 and 2_x, we aim to identify candidates for IVs in four well-known examples in economics: returns to schooling, demand and supply, and peer effects. These examples are chosen for their significance in the literature and commonly employ the IVs method as an empirical strategy. This exercise is to evaluate LLM performance and demonstrate the proposed method's practicality.

In all the examples, LLMs appear to discover new candidates for IVs and candidates that are related to well-known IVs in the literature. Also, many candidates demonstrate high levels of specificity to their context. We report the full results for the first two examples. For the last two, we only provide summaries and include the comprehensive results in the Appendix. 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 when different values of K_0 and K are chosen. In all the examples, we choose $K_0 = 40$ and let K be determined in the discovery process. We use GPT4 as our LLM.[9]

4.1   Returns to Education

Estimating the causal effects of educational attainment (e.g., college attendance, years of schooling) on earnings has been a significant objective in economics. 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 schooling decisions 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 schools, tuition fees, and compulsory schooling laws have been widely used in the literature (Card, 1999).

---

[9]We focus on GPT4 as the other versions of ChatGPT (e.g., GPT4o) produce similar findings.

4.1.1 Returns of College Attendance

As the first example, we focus on the returns to college attendance. The following is the prompts we use. We explicitly request separate lists for individual and school factors.

Prompt 1.1. *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. *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. create separate lists for school factors and factors of yourself. for each factor chosen, explain your reasoning.*

From running Prompts 1.1 and 2_x.1, the LLM produces candidates for valid IVs and its rationale for the suggestions (Table 1). The list includes IVs that are already established in the literature, such as *distance to college* (Card, 1995, 2001) and *family background* (Card, 1995; Ashenfelter and Zimmerman, 1997), as well as IVs that seem to be new (to our best knowledge): student-related factors as *personal savings for education* and various school characteristics such as *campus crime rates*, *diversity of student body*, *availability of online courses*, *campus housing capacity* and *proximity to urban centers*. The newly found variables have potential to be valid, especially after being conditioned on covariates. Producing all these results took less than one minute in total.

11

#	IVs Suggested	Rationale Provided
1	Distance from Home to College (Card (1995), 1955; Card (2001), 3134)	This is purely geographical and independent of personal attributes.
2	Personal Savings for Education	While this might be influenced by a student's ability to save, it's largely determined by external economic conditions and family support.
3	Number of Siblings Attending College (Griliches (1977), 2393)	This depends on family decisions and the educational pursuits of siblings, unrelated to the student's personal capabilities or characteristics.
4	Parents' Educational Background (Card (1995), 1955; Ashenfelter and Zimmerman (1997), 335)	A demographic factor that precedes the student's own choices and abilities.
5	Religious Affiliations	A personal and family cultural factor, not necessarily a direct influence on personal academic or career potential.
6	Campus Crime Rates	Safety measures and local crime statistics are external to the student's attributes.
7	Diversity of Student Body	Reflects the school's admissions policies and demographic outreach, independent of any one student's traits.
8	Library Size (Volumes Held)	A resource factor that's determined by the institution's funding and priorities.
9	Availability of Online Courses	Indicates the institution's commitment to flexible learning options, independent of student characteristics.
10	Campus Facilities Quality (Gyms, Labs, etc.)	Dependent on the school's investment in infrastructure.
11	Campus Housing Capacity	Dictated by the university's accommodations and facilities, not by student traits.
12	Proximity to Urban Centers	Geographic location of the school relative to cities.
13	Environmental Sustainability Rating	Reflects the institution's environmental policies and infrastructure.
14	Technology Integration in Classrooms	Indicates the level of modern technological adoption by the school.

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

Notes: IVs for college attendance are presented. 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 5 rows are categorized by GPT4 to be student-related factors, and the next 9 rows to be school-related factors. When the responses match the IVs described in the literature exactly, we include the corresponding references and their citation counts in parenthesis. The rationale given by LLM can be elaborated further by requesting it in the same session. The total running time was less than 1 minute.

4.1.2 Returns to Years of Education

As a second example, we consider the returns to years of schooling. We omit the prompts as they are similar as before, except that we impose the following role: "`you are a student beginning high school in the united states.  you need to make a decision on how many more years you will stay in school.`"

Table 2 presents the list of candidate IVs (together with the LLM's reasoning). Among them are instruments well documented in the literature, such as *state education laws* (Angrist and Krueger, 1991; Staiger and Stock, 1997) and *distance to college* (Card, 1995, 2001), alongside several that appear novel and interesting: regulatory and contextual factors such as *credits required for graduation, transportation availability, local education funding per student, local college enrollment rates*, and *local unemployment rates*.

4.2 Demand and Supply

4.2.1 Demand Estimation

Consider estimating demand for consumers in a given market. In estimating the effect of price on demand, the main concern is that price is endogenous as an equilibrium outcome. Researchers have used supply-side IVs that are excluded from the demand equation in a simultaneous system for supply and demand (Angrist et al., 2000) or motivated by structural models (Berry et al., 1995).

Below are the prompts we use, focusing on the setting of Angrist et al. (2000).[10]

Prompt 1.2. `you are a dealer at a fish market. you need to set the prices of fish. what would be factors that can determine your decision but that do not directly affect the customers' demand for fish, except through the price you set (that is, that affect the demand only through fish prices). list forty factors that are quantifiable. explain your answer.`

---

[10]Note that in Prompt 2.2, covariate information is omitted.

#	IVs Suggested	Rationale Provided
1	Credits Required for Graduation	This is a regulatory requirement and does not depend on your personal qualities or the inherent quality of your school but can determine the length of time you stay in school.
2	State Education Laws (Angrist and Krueger (1991), 3951; Staiger and Stock (1997), 10656)	Laws mandating education until a certain age or credits affect all students uniformly, irrespective of individual abilities or school quality.
3	Distance to Nearest College/University (Card (1995), 1955; Card (2001), 3134)	The geographical proximity to higher education institutions is an external factor that doesn't depend on your ability but can influence the decision to pursue further education based on accessibility.
4	Transportation Availability	The ease of getting to and from educational institutions is a logistical concern that affects your ability to attend and continue education, independent of personal academic qualities or school resources.
5	Local Education Funding per Student	This reflects the financial investment in education at a community or state level, impacting the resources available to you, which can influence educational continuation decisions, yet is not a reflection of personal ability.
6	Local College Enrollment Rates	Community trends in college enrollment can create a social environment that encourages further education, which is more about local culture and less about individual capacity.
7	Local Unemployment Rates	Higher local unemployment might incentivize staying in school longer as a way to defer entering a tough job market, a decision influenced by economic conditions rather than personal academic prowess.

Table 2: Returns to Years of Schooling: Suggested IVs and Rationale for Validity

Notes: IVs for years of schooling are presented. All IVs are discovered and explained by GPT4 from a single run of Prompts 1.1 and 2_x.1 (modified for this example) with $K_0 = 40$ and K left unspecified. When the responses match the IVs described in the literature exactly, we include the corresponding references and their citation counts in parenthesis. The rationale given by LLM can be elaborated further by requesting it in the same session. The total running time was less than 1 minute.

Prompt 2.2. `suppose you are a dealer at the fish market who is selling fish and setting its prices on a day of the week. among the factors listed above, choose all factors are not influenced by fish market conditions or customers' characteristics that determine demand for fish. for each factor chosen, explain your reasoning.`

Table 3 contains the candidate IVs from running Prompts 1.2 and 2.2. There are many *supply-side factors* (e.g., *fuel costs*), which are reasonable candidates for IVs. The *weather conditions* variable used as IVs in Angrist et al. (2000) appears in the list. Interestingly, another supply-side factor, *labor costs*, produced by Prompt 1.2 is not included in the final list produced by Prompt 2.2. When asked "`explain why you didn't include ''labor costs'' in the final list`," the LLM responded that "labor costs are somewhat flexible and can be adjusted in response to changes in market conditions and customer demand, making them more dynamic than some of the other factors listed."

4.2.2 Production Function Estimation

Consider 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).

In the prompts, we explicitly request separate lists for market factors and firm and manager factors. For this example and all examples below, the actual prompts are contained in the Appendix.

The LLM finds sixteen IV candidates (Table A1 in the Appendix). Notably, IVs commonly suggested in the literature (i.e., *input prices*) do not appear among the LLM's final selections, echoing similar concerns in the literature (Olley and Pakes, 1996; Levinsohn and

#	IVs Suggested	Rationale Provided
1	Fuel Costs	Costs for fuel typically depend on broader economic factors such as oil prices and are not directly affected by market demand for fish or specific customer characteristics.
2	Fishing Equipment Depreciation	The depreciation of fishing gear like nets and boats is largely a function of use and time, rather than market conditions or customer demand.
3	Weather Conditions (Angrist et al. (2000), 454)	Weather affects fishing operations directly but is not influenced by market conditions or the specific characteristics of customers.
4	Regulatory Costs	Costs associated with licensing and compliance with fishing regulations are dictated by regulatory bodies and do not fluctuate based on customer demand or market conditions.
5	Interest Rates	The rates charged on loans for purchasing boats or equipment are determined by financial markets and policy, rather than direct market conditions of the fish market or customer demands.
6	Insurance Costs	Premiums for insuring fishing equipment and operations are typically based on risk assessments and sector-wide data, rather than being directly influenced by day-to-day market conditions or characteristics of customers.
7	Utility Costs at Sales Points	The cost of utilities like electricity and water at sales points tends to be fixed or based on usage rates that are independent of market demand specifics.
8	Economic Conditions	Broader economic factors that affect overall spending and investment patterns influence operational costs but are not dictated by the fish market conditions or customer preferences.
9	Technological Advances	Investments in technology to improve fishing or sales operations are usually planned based on long-term business strategy and efficiency gains rather than immediate market conditions or specific customer demographics.
10	Government Subsidies	Subsidies to the fishing industry are determined by government policies, which are independent of daily market conditions or customer demands in the fish market.
11	Tariffs on Imports	Tariffs imposed on imported fish are a matter of international trade policy and do not change based on daily fluctuations in market demand or customers characteristics.
12	Employee Training Costs	Costs of training employees in handling and selling fish relate to operational efficiency, which are not directly influenced by the day's market conditions or customer demands.

Table 3: Demand Estimation: Suggested Candidates for IVs

Notes: IVs for price are presented. All IVs are discovered and explained by GPT4 from a single run of Prompts 1.2 and 2.2 with $K_0 = 40$ and K left unspecified. When the responses match the IVs described in the literature exactly, we include the corresponding references and their citation counts in parenthesis. The rationale given by LLM can be elaborated further by requesting it in the same session. The total running time was less than 1 minute.

Petrin, 2003; Ackerberg et al., 2015). Instead, the model suggests factors that may influence input prices, such as *interest rates*, *exchange rates*, *tariffs and trade policies*, *transportation costs*, *real estate prices*, and the firm's *financial health*, several of which could plausibly be exogenous. There are also other instruments that fall into regulatory factors (e.g., *environmental regulations*) and managerial factors (e.g., *risk tolerance*, *strategic objectives*). The model justifies these by arguing that such variables shape firms' input choices and long-term positioning but do not directly influence short-run productivity shocks. However, some of these long-term decisions may in fact affect long-term outputs, which may or may not be relevant to the short-run outcomes of concern.

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).

It is worth noting that, in these examples, we prompt the LLM to role-play from the peer's perspective, which is more convenient than prompting from that of the individual whose outcome is of concern.

4.3.1 Effects of Peer Farmers on New Technology Adoption

Interestingly, some IVs that are suggested in the literature (e.g., *friends of friends*) are not chosen by the LLM (Table A2 in the Appendix). It can be because the model either views them as invalid or is incapable of identifying them. On the other hand, the variable *success stories documented and shared* 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 *access to credit* and *engagement in social media or digital platforms.* Finally, the LLM fails to identify the IV used in Foster and Rosenzweig (1995), namely, *endowed land size.* The validity of this variable is justified in their paper by the specific historical backgrounds of the Indian villages studied. This implies that, for some IVs, providing LLMs with institutional details can be necessary; this can be implemented via LLM's system messages.

4.3.2 Effects of Peer Teenagers on Smoking Behavior

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.

It is important to note that, given that the prompts are written from the perspective 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 such as *engagement rate* and *frequency of posts* (Table A3 in the Appendix). Interestingly, *access to smartphones among peers* can be viewed as a "friends of friends" IV.[11]

5 Expert Evaluation of Discovered IVs

Although evaluating the validity of IVs is intrinsically elusive, we attempt to validate our approach through expert evaluation. To that end, we conducted a survey of experts in various fields of empirical economics—where IVs are heavily used—to elicit their evaluation of the IVs discovered by LLMs. We received evaluations from ten experts.

---

[11]There were other "friends of friends" IVs that are produced from Step 1 but did not survive Step 2.

For each of the six domains in Section 4, we collected two metrics through the survey: the IV validity score (1 = least valid, 5 = most valid) and the non-novelty score (1 = least likely to exist in the previous literature, 5 = most likely to exist). Lower non-novelty scores would correspond to *higher* IV novelty. The validity scores should be interpreted with care. First, the difficulty of finding valid IVs varies across domains, and thus the scores should be interpreted relative to a benchmark. We define the benchmark as the average validity of the most established IVs (non-novelty score = 5) in each domain. Second, although the average score may be lower, individual IVs found by LLMs may be viewed as novel and valid. Having these points in mind, for each domain, we report (i) the average validity score given non-novelty score $E[\text{validity}|\text{non-novelty} = k]$ for each $k = 1, \ldots, 5$ and (ii) the names of all IVs such that $E[\text{validity}|\text{non-novelty} = 1 \text{ or } 2, \text{ IV name}] \geq \text{benchmark}$, where benchmark $= E[\text{validity}|\text{non-novelty} = 5]$. IVs meeting this criterion can be viewed as relatively novel yet at least as valid as existing ones.

Tables 4–5 report the results. Overall, we observe the following: (i) No IV is perfect; even the most established IVs sometimes averaged a validity score below 3 on our scale.[12] (ii) Finding valid IVs is challenging for both LLMs and human researchers (the latter reflected in the established IVs). (iii) The LLM has the capability of identifying IVs that are viewed as both novel and more valid than the most established benchmarks. (iv) The LLM discovered the greatest number of valid IVs in the most novel domain (i.e., the example of peer effects on teenage smoking in social-media networks), which was also the most challenging domain with the lowest benchmark score.

6   Adversarial Large Language Models

One way to refine the responses of an LLM is to request another LLM to play the role of an adversary and review the responses produced by the first LLM, namely the defender LLM. In

---

[12]The literature recognizes that mildly invalid (or only plausibly exogenous) IVs are ubiquitous (Conley et al., 2012; Deaner, 2025).

Domain[†]	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$ (benchmark)
1	1.87	2.23	2.72	2.17	2.63
2	3.00	2.10	2.39	2.97	3.19
3	1.92	2.39	2.63	3.76	3.00
4	1.92	2.22	2.70	3.19	3.72
5	2.36	2.26	2.96	2.39	3.67
6	2.10	2.03	2.53	3.00	2.00

Table 4: Average Validity Across Novelty ($k=1$: most novel; $k=5$: least novel)

[†]Domain ID: 1. Returns to Schooling; 2. Returns to Years of Education; 3. Demand Estimation; 4. Production Functions; 5. Peer Effects (technology adoption); 6. Peer Effects (teenage smoking).

Domain[†] (Benchmark)	Novel IVs Suggested
1 (2.63)[††]	Campus Housing Capacity (+0.70)[†††]
	Technology Integration in Classrooms (+0.37)
	Campus Crime Rates (+0.03)
2 (3.19)	Credits Required for Graduation (+0.31)
	Transportation Availability (+0.31)
3 (3.00)	Fuel Costs (+1.00)
	Tariffs on Imports (+1.00)
	Technological Advances (+0.00)
	Government Subsidies (+0.00)
4 (3.72)	Exchange Rates (+0.28)
5 (3.67)	Access to Credit (+0.33)
6 (2.00)	Access to Smartphones Among Peers (+1.50)
	Content Reach (+1.33)
	Engagement Rate (+1.00)
	Local Regulations on Tobacco Advertising (+1.00)
	Number of Platforms Used (+0.40)
	Frequency of Posts (+0.33)
	Visibility at School Events (+0.25)
	School Policies on Smoking (+0.00)
	History of Disciplinary Actions at School (+0.00)
	Academic Performance (+0.00)

Table 5: IVs with $k=1$ or 2 Outperforming Benchmark ($k=1$: most novel; $k=5$: least novel)

[†]Domain ID: 1. Returns to Schooling; 2. Returns to Years of Education; 3. Demand Estimation; 4. Production Functions; 5. Peer Effects (technology adoption); 6. Peer Effects (teenage smoking). [††]The validity score of benchmark is in parenthesis. [†††]The improvement of validity score over benchmark is in parenthesis.

the adversarial stage, we fully disclose the IV discovery task and ask the adversarial LLM to provide counter-arguments, but without using econometric jargon. These counter-arguments are then given to the defender LLM, which is asked to refine the previous answers. Overall, this process produces more sophisticated responses. We find this to be a useful exercise, which mimics the mental process of a human researcher.

The following is an example prompt given to the adversarial LLM. In the prompt, `[the list of variables by Defender]` is the list provided the defender LLM, generated from running, for example, Prompts 1–2$_x$.

Prompt AD1. *you are a researcher who wants to find instrumental variables to estimate the effect of [treatment] on [outcome]. below is a list of candidate instrumental variables. for each variable in the list, provide arguments as to why it may not be a valid instrument: [the list of variables by Defender]*

Then, the counter-arguments generated from Prompt AD1 are presented to the defender LLM using the following prompt, which follows Prompts 1–2$_x$.

Prompt 3. *below are counter-arguments for each of your previous answers. based on these arguments, revise your selection and provide a list: [the arguments by Adversary]*

When running Prompt 3, it is important to encourage the defender to create a list, otherwise, it might become pessimistic and reject all the previous selections.

We apply this adversarial process to our examples in Section 4. Overall, the defender provides more sophisticated answers in the end. In the "Returns to College" example (Section 4.1.1), the variables related to *geographical proximity* or *transportation* have been retained. The other variables related to *school facilities* (e.g., *library size*, *technology integration in classrooms*) have been modified to factors that may better satisfy exogeneity (e.g., *the percentage of renewable energy used on campus, the number of green spaces on campus, campus*

21

medical facilities). Interestingly, many variables now appear to be weak IVs, which is sensible. In the "Returns to Years of Education" examples (Section 4.1.2), the *state laws* variable survived, but was renamed to the more descriptive *mandatory minimum years of schooling required by state.* Other variables were also further refined. *GED program availability* emerged. In the "Demand Estimation" example (Section 4.2.1), the *weather condition* variable was refined to *global climate patterns (e.g., El Niño)*, because the counter-argument was that "bad weather can also discourage customers from coming to the market," which is a compelling argument. Many other variables were refined to reflect *global and macroeconomic conditions.* Again, these variables suffer from being weak IVs or having less individual variation.

7   Comparison to Direct Approach

An alternative to the proposed prompting strategy (i.e., multi-stage, role-playing prompts) would be to explicitly mention the term "instrumental variables" and inform LLMs that the ultimate objective is to "search for instrumental variables." Additionally, one can verbally train LLMs to understand the IV assumptions before conducting the search. Unfortunately, we find that this direct approach tends to suffer from *memorization*, where LLMs primarily focus on sources from existing academic articles and textbooks relevant to IVs. One can easily verify that, when applying prompts from the direct approach to the examples in Section 4, GPT4 generates answers that are reminiscent of established IVs in the literature. Moreover, when additionally prompted with "`did you have sources for your previous responses?`," GPT4 responds by citing well-known references in each application that propose IVs. In contrast, when the same question is asked after the proposed procedure, GPT4 responds that it "didn't have specific sources." When pressed to provide sources from the web, in the returns to education application, for example, it cites academic references from behavioral and social sciences on students' decision-making, including qualitative and case studies.

8 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.

8.1 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 using 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 ($L_0 \geq L$). 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 factor chosen, explain your 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 excluded 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.

8.2   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).

The following prompts are used to produce the results in Table 6. 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 factor chosen, explain your reasoning.`

Table 6 contains the control variables suggested by the LLM, conditional on which the parallel trend is likely to hold, and the LLM'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 the LLM 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.

#	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 6: 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.

8.3 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).[13] 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. Therefore, if LLMs outperform a traditional search for running variables, it would be due to their automated and comprehensive search behavior. 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

---

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

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 A5 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 are suggested by GPT4 with specific cutoffs obtained from web sources. Except for one case (i.e., #4), GPT4 also identifies the running variables used in the literature.

9 Conclusions

This essay proposes the agenda to use LLMs to systematically search for variables in designing causal inference. It merely serves as a starting point, and there are potential next steps that can follow. 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 Assumptions REL–IND are satisfied with known IVs *before* starting the proposed prompts. This approach would evoke *few-shot learning* in LLMs (Brown et al., 2020), which can enhance their performances. This approach would also "orthogonalize" the search (Ludwig and Mullainathan, 2024) to focus on novel IVs. Second, the elaborated search can be directed toward finding IVs that are more policy-relevant (Imbens and Angrist, 1994; Heckman and Vytlacil, 2005) by specifying targeted policies in the prompt. Third, none of the results reported in the current essay are findings aggregated across sessions. To account for and potentially leverage the stochastic nature of LLMs' responses, exploring the possibility of aggregation (e.g., taking the union or intersection of $\mathcal{Z}_K$'s across sessions) would be beneficial.

One challenge of our approach—shared with other hypothesis-generation procedures—is that the performance metric is difficult to define due to the lack of ground truth for valid IVs. In fact, this is the very reason we propose using 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. We partly address this challenge through an expert survey in Section 5.

#	Outcome(s) (Country)	Treatment(s)	Suggested Running Variable (Same as the Literature)	Newly 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)*
2	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)
3	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)
4	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; source: ETIAS Italy)
5	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 A5: 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 #4. 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. †: Not found by GPT4.

In later stages, when data eventually come into play, over-identification tests can potentially be a fruitful framework for the evaluation of LLMs. More broadly, evaluating hypothesis-generating systems remains an interesting open question. We still believe the paper's proposal serves as a thought-provoking starting point. It would be interesting to apply the proposed variable-search approach to other empirical settings and causal inference methods.

A   Three-Step Prompts with System Messages

We present three-step prompts, where each step corresponds to one of REL, EX and IND. This version includes *system messages* that establish the scenario (including institutional details) and role of interest as context information and response formats before initiating the main prompts.

Prompt A0 (System Messages). `you are [agent] who needs to make a [treatment] decision in [scenario]. you will be given questions on your decisions and related outcomes. in answering the questions, always provide explanatic for your answers. each answer should be within five words and each explanation should be within twenty words. do you understand your role, the scenario you are in, and the instructions given to you?`

Prompt A1 (Search for Relevance). `what are factors that can determine your decision?  list [K_0] factors that are quantifiable.  explain the answers.`

Prompt A2 (Search for Exclusion). `among [K_0] factors listed above, what are factors that affect your [outcome] only through [treatment]?  explain the answers.`

Prompt A3 (Search for Independence). `among the 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 factor chosen, explain your reasoning.`

B   Alternative Prompts with Covariate Search

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 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 characterist`
`that directly influence [treatment] and directly influence [outcome]`
`(not just through [treatment])?  list [L_0] characteristics.  explain`
`the answers.`

Instead of running Prompt 1 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.

C   Prompts Used in Applications (Section 4)

C.1   Production Function Estimation

We choose $K_0 = 40$ and let GPT4 choose K; we explicitly request separate lists for market factors and firm and manager factors. Note that in Prompt 2$_x$.2.1, we use a loose description of covariates (unlike in Prompt 2$_x$.1 in Section 4.1 where we assign specific values).

Prompt 1.2.1 (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.`

Prompt 2$_x$.2.1 (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. create separate lists for market factors and managerial factors. for each factor chosen, explain your reasoning.*

C.2   Peer Effects

C.2.1   Effects of Peer Farmers on New Technology Adoption

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? for each factor chosen, explain your reasoning.*

C.2.2 Effects of Peer Teenagers on Smoking Behavior

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.

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? for each factor chosen, explain your reasoning.*

D Additional Tables for Discovered IVs (Section 4)

#	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 A1: (a) Production Functions: Suggested Candidates for IVs (Market Factors)

Notes: IVs for inputs are presented. All IVs are discovered and explained by GPT4 from a single run of Prompts 1.2.1 and 2_x.2.1 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 rationale given by LLM can be elaborated further by requesting it in the same session. The total running time was less than 1 minute.

#	IVs Suggested	Rationale Provided
11	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
12	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
13	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.
14	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
15	Compliance and Legal Considerations:	Driven by external legal requirements and internal ethics, not by short-term productivity
16	Corporate Social Responsibility (CSR) Initiatives	Strategic decisions about CSR are influenced by long-term planning and brand image considerations.

Table A1: (b) Production Functions: Suggested Candidates for IVs (Managerial Factors)

Notes: IVs for production inputs are presented. All IVs are discovered and explained by GPT4 from a single run of Prompts 1.2.1 and 2_x.2.1 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 rationale given by LLM can be elaborated further by requesting it in the same session. The total running time was less than 1 minute.

#	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 A2: Peer Effects on Technology Adoption: Suggested Candidates for IVs

Notes: IVs for peers' decisions are presented. Given the perspective in the prompts, the variables should be understood as factors of *peers*. IVs for All IVs are discovered and explained by GPT4 from a single run of Prompts 1.3.1 and $2_x.3.1$ with $K_0 = 40$ and K left unspecified. The rationale given by LLM can be elaborated further by requesting it in the same session. The total running time was less than 1 minute.

#	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 A3: Peer Effects on Smoking: Suggested Candidates for IVs

Notes: IVs for peers' decisions are presented. Given the perspective in the prompts, the variables should be understood as factors of *peers*. All IVs are discovered and explained by GPT4 from a single run of Prompts 1.3.2 and 2_x.3.2 with $K_0 = 40$ 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. The rationale given by LLM can be elaborated further by requesting it in the same session. The total running time was less than 1 minute.

References

ACKERBERG, D. A., K. CAVES, AND G. FRAZER (2015): "Identification Properties of Recent Production Function Estimators," *Econometrica*, 83, 2411–2451.

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

ANGRIST, J. D., K. GRADDY, AND G. W. IMBENS (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67, 499–527.

ANGRIST, J. D. AND A. B. KRUEGER (1991): "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, 106, 979–1014.

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.

ASHENFELTER, O. AND D. J. ZIMMERMAN (1997): "Estimates of the returns to schooling from sibling data: Fathers, sons, and brothers," *Review of Economics and Statistics*, 79, 1–9.

ATHEY, S. AND G. W. IMBENS (2019): "Machine Learning Methods that Economists Should Know About," *Annual Review of Economics*, 11, 685–725.

BATTISTIN, E. AND E. RETTORE (2002): "Testing for programme effects in a regression discontinuity design with imperfect compliance," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 165, 39–57.

BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

BLUNDELL, R. AND J. L. POWELL (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," *Econometric Society Monographs*, 36, 312–357.

BRAMOULLÉ, Y., H. DJEBBARI, AND B. FORTIN (2009): "Identification of Peer Effects through Social Networks," *Journal of Econometrics*, 150, 41–55.

BROWN, T. B. ET AL. (2020): "Language Models are Few-Shot Learners," *arXiv preprint arXiv:2005.14165*.

CARD, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. N. Christofides, E. K. Grant, and R. Swidinsky, Toronto: University of Toronto Press, 201–222.

——— (1999): "The Causal Effect of Education on Earnings," *Handbook of Labor Economics*, 3, 1801–1863.

——— (2001): "Estimating the return to schooling: Progress on some persistent econometric problems," *Econometrica*, 69, 1127–1160.

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.

CARD, D. AND L. D. SHORE-SHEPPARD (2004): "Using Discontinuous Eligibility Rules to Identify the Effects of the Federal Medicaid Expansions on Low-Income Children," *Review of Economics and Statistics*, 86, 752–766.

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*.

CONLEY, T. G., C. B. HANSEN, AND P. E. ROSSI (2012): "Plausibly exogenous," *Review of Economics and Statistics*, 94, 260–272.

CONLEY, T. G. AND C. R. UDRY (2010): "Learning about a New Technology: Pineapple in Ghana," *American Economic Review*, 100, 35–69.

DE GIORGI, G. (2005): "Long-term effects of a mandatory multistage program: The New Deal for young people in the UK," IFS Working Papers W05/08, Institute for Fiscal Studies, London.

DEANER, B. (2025): "The trade-off between flexibility and robustness in instrumental variables analysis," *American Economic Review*, 12.

DELL, M. (2025): "Deep learning for economists," *Journal of Economic Literature*, 63, 5–58.

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.*

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.

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.

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

GENTZKOW, M., B. KELLY, AND M. TADDY (2019): "Text as data," *Journal of Economic Literature*, 57, 535–574.

GHANEM, D., P. H. SANT'ANNA, AND K. WÜTHRICH (2022): "Selection and Parallel Trends," *arXiv preprint arXiv:2203.09001.*

GRILICHES, Z. (1977): "Estimating the returns to schooling: Some econometric problems," *Econometrica: Journal of the Econometric Society*, 1–22.

GRILICHES, Z. AND J. MAIRESSE (1995): "Production Functions: The Search for Identification," *NBER Working Paper No. w5067.*

GURYAN, J. (2001): "Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts," Working Paper 8269, National Bureau of Economic Research, Cambridge, MA.

HECKMAN, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

HECKMAN, J. J., H. ICHIMURA, AND P. TODD (1998): "Matching as an Econometric Evaluation Estimator," *The Review of Economic Studies*, 65, 261–294.

HECKMAN, J. J. AND E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation1," *Econometrica*, 73, 669–738.

HERNÁN, M. A. AND J. M. ROBINS (2006): "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology*, 17, 360–372.

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

JIN, Z., Y. CHEN, F. LEEB, L. GRESELE, O. KAMAL, Z. LYU, K. BLIN, F. GONZALEZ ADAUTO, M. KLEIMAN-WEINER, M. SACHAN, ET AL. (2023a): "Cladder: Assessing causal reasoning in language models," *Advances in Neural Information Processing Systems*, 36, 31038–31065.

JIN, Z., J. LIU, Z. LYU, S. POFF, M. SACHAN, R. MIHALCEA, M. DIAB, AND B. SCHÖLKOPF (2023b): "Can large language models infer causation from correlation?" *arXiv preprint arXiv:2306.05836*.

KICIMAN, E., R. NESS, A. SHARMA, AND C. TAN (2023): "Causal reasoning and large language models: Opening a new frontier for causality," *Transactions on Machine Learning Research*.

KONG, A., S. ZHAO, H. CHEN, Q. LI, Y. QIN, R. SUN, X. ZHOU, E. WANG, AND X. DONG (2023): "Better zero-shot reasoning with role-play prompting," *arXiv preprint arXiv:2308.07702*.

LEE, D. S. AND T. LEMIEUX (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

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

LUDWIG, J. AND S. MULLAINATHAN (2024): "Machine Learning as a Tool for Hypothesis Generation," *The Quarterly Journal of Economics*, 139, 751–827.

LUDWIG, J., S. MULLAINATHAN, AND A. RAMBACHAN (2025): "Large language models: An applied econometric framework," Tech. rep., National Bureau of Economic Research.

MANNING, B. S., K. ZHU, AND J. J. HORTON (2024): "Automated Social Science: Language Models as Scientist and Subjects," *National Bureau of Economic Research.*

MANSKI, C. F. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 60, 531–542.

MOGSTAD, M. AND A. TORGOVITSKY (2024): "Instrumental Variables with Unobserved Heterogeneity in Treatment Effects," *National Bureau of Economic Research.*

MULLAINATHAN, S. AND A. RAMBACHAN (2024): "From Predictive Algorithms to Automatic Generation of Anomalies," *National Bureau of Economic Research.*

MULLAINATHAN, S. AND J. SPIESS (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

OLEA, J. L. M. AND C. PFLUEGER (2013): "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 31, 358–369.

OLLEY, G. S. AND A. PAKES (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

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

RAMBACHAN, A. AND J. ROTH (2023): "A More Credible Approach to Parallel Trends," *Review of Economic Studies*, 90, 2555–2591.

STAIGER, D. AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

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.

VAN DER KLAAUW, W. (2002): "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," *International Economic Review*, 43, 1249–1287.

WAN, G., Y. WU, M. HU, Z. CHU, AND S. LI (2024): "Large Language Models for Causal Discovery: Current Landscape and Future Directions," *arXiv preprint arXiv:2402.11068*.

WANG, H., T. FU, Y. DU, W. GAO, K. HUANG, Z. LIU, P. CHANDAK, S. LIU, P. VAN KATWYK, A. DEAC, ET AL. (2023): "Scientific discovery in the age of artificial intelligence," *Nature*, 620, 47–60.

WU, T., M. TERRY, AND C. J. CAI (2022): "AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–22.