# Measuring Industrial Policy
## A Text-Based Approach*

Réka Juhász (UBC, NBER & CEPR)
Nathan Lane (LSE, CESifo)
Verónica C. Pérez (Boston University)

First draft: August, 2022
This draft: November, 2026

**Abstract**
Since the 18th century, policymakers have debated the merits of industrial policy (IP). Yet economists lack basic facts about its use, largely because it is hard to measure. We develop an approach that classifies policies by their stated goals, using supervised machine learning applied to the text of a global database of commercial policies (2010–2022) and show why a systematic approach to measurement matters. We find that industrial policy systematically targets sectors of established comparative advantage — reinforcing existing strengths rather than building new ones, contrary to the infant-industry logic that motivates much of the literature. This finding depends on measuring industrial policy correctly: policies pursuing other goals through the same instruments show no such targeting, and the tariff-based proxy long used in the literature misses the pattern as well. Measuring industrial policy by its goals, rather than the instruments it uses, changes what we conclude about it.

*JEL: O25,L52,C38*

*Keywords: Industrial Policy, Supervised Machine Learning, Natural Language Processing*

*We are deeply grateful to Emily Oehlsen, who contributed substantially to this project and to earlier versions of this paper as a co-author. Previously titled "The Who, What, When, and How of Industrial Policy: A Text-Based Approach." We thank Elliott Ash, David Byrne, Chiara Criscuolo, John Andrew Fitzgerald, Jason Garred, Penny Goldberg, Douglas Gollin, Amy Handlan, Gordon Hanson, Tarek Hassan, Giammario Impullitti, Myrto Kalouptsidi, Guy Lalanne, Colin McAuliffe, Phillip McCalman, Suresh Naidu, Emi Nakamura, Ashwin Narayan, Gaurav Nayyar, Nina Pavcnik, Tristan Reed, Dani Rodrik, Francesco Squintani, Chris Woodruff, and seminar audiences at American University, Applied Machine Learning, Economics, and Data Science (AMLEDS) Workshop, Berkeley, Brown, Chapman University, Columbia, Council of Economic Advisors (U.S.), Duke University, Econometric Society Summer Meetings, ETH Zurich, ETH Zurich–Monash–Warwick Textual Data Conference, Federal Reserve Bank of Atlanta Firm Dynamics and the Macroeconomy Conference, Harvard, the IGC-Stanford Firms, Development and Trade Conference, the International Monetary Fund, Michigan, MIT Sloan, National University of Singapore, NBER Summer Institute, New Thinking in Industrial, Innovation and Technology Policy Conference at Columbia, Pakistan Institute of Development Economics, STEG-CEPR-IGC Conference, UC-Irvine, University of Melbourne, University of Miami, University of Washington Comparative Politics, Warwick, the World Bank, and Yale for helpful comments and suggestions. We benefited from conversations with Gerard DiPippo, Ilaria Mazzocco, and scholars at the CSIS; the OECD Productivity, Innovation and Entrepreneurship team, and the UNIDO Research and Industrial Policy Advice Division. We especially thank Johannes Fritz for helpful discussions about the *Global Trade Alert* project, and the GTA team for sharing their data. Ida Brzezinska, Caran Niousha Etemad, Lottie Field, Mikhael Gaster, Diti Jain, Max Marczinek, Esha Vaze, Sarah Wappel, and Megan Yamoah provided outstanding research support. We thank the team of Columbia, Oxford, and UBC annotators for their excellent assistance. We gratefully acknowledge funding from ISERP, STEG/CEPR, SSHRC, and the Alfred P. Sloan Foundation. Large language models (Anthropic, OpenAI) were used to copy-edit the manuscript; all analysis, results, and claims are the authors' own.

# 1. Introduction

Since the rise of modern capitalism, social scientists have debated the role of industrial policy — intentional state intervention aimed at transforming the composition of economic activity (List, 1856; Taussig, 1914; Chang, 2002; Rodrik, 2008). Despite these longstanding debates, empirical research on industrial policy (IP) remains incomplete (Harrison and Rodríguez-Clare, 2010). Issues of measurement (Barwick, Kalouptsidi and Bin Zahur, 2024) and the lack of systematic data have been significant hurdles to understanding policy (e.g., World Trade Organization 2006). An enduring challenge is that the same policy instrument (for example, a tariff) can serve both as industrial policy and as a tool for other policy objectives (Harrison and Rodríguez-Clare, 2010; Lane, 2021). As industrial policy experiences a global resurgence, addressing these issues is crucial for understanding these policies.

In this paper, we develop a text-based method to measure industrial policy and use it to show that how industrial policy is measured determines what one concludes about it. Using natural language processing, we classify policies by their stated objectives—the intent to alter the composition of economic activity—directly from policy text, and apply this to the world's largest corpus of real-world commercial policy documents to construct measures of industrial policy at the country–sector–year level. We then put the measure to a test: we ask a question on which theories of industrial policy make opposing predictions, and compare our measure against the instrument-based proxy economists have long relied on. The two approaches give different answers: the instrument-based proxy points to a conclusion our measure shows to be wrong. The same test also shows that industrial policy is a distinct object: policies pursuing other goals through the very same instruments behave differently, along dimensions our measure was not built to capture.

To illustrate the conceptual challenge for measuring industrial policy, consider tariffs, where the problem is well understood and has been widely discussed. Although sometimes used to promote infant industries (Juhász and Steinwender, 2024), tariffs are also implemented with the goal of raising government revenue (Johnson, 1951; Balassa, 1989; Cagé and Gadenne, 2018), managing terms of trade (Broda, Limao and Weinstein, 2008), or catering to political interests (Goldberg and Maggi, 1999; Gawande, Krishna and Olarreaga, 2015). Thus, a dataset of tariffs is not a dataset of industrial policies.[1] The same holds for other policy instruments.

1. In their assessment of the literature, Harrison and Rodríguez-Clare (2010, p. 4065) note that a fundamental problem with cross-industry studies of tariffs and growth is that there is "no evidence to suggest that intervention for IP reasons in trade even exists." Earlier work sought to address this issue by distinguishing tariffs used for fiscal revenue generation from those employed for industry promotion purposes (Lehmann and O'Rourke, 2011). Likewise, Nunn and Trefler (2010) used the skill bias of tariffs as a proxy for tariffs aimed at industry development.

Our solution to this measurement challenge is based on the observation that the *text* of policy announcements often includes information about a policy maker's goals. Tables 1 – 2 (industrial policy goals and other policy goals, respectively) provide illustrative examples from our source data, emphasizing goals (in *italics*). To see why this type of information is useful for distinguishing IP, consider three import tariff policies. One wants "to stimulate innovation and strengthen the national IT sector" (Table 1, example 1). Another wants "to increase the revenue of the government" (Table 2, example 2). A third wants to reinstate import tariffs on food staples "which were removed during the period of rising food prices" (Table 2, example 4). Although these policies use the same policy instrument, they clearly have different objectives. Of the three policies, only the first constitutes industrial policy because of its explicit goal of wanting to change the composition of economic activity towards the IT sector.

In the first part of the paper, we develop a text-based approach to measure industrial policy. Using supervised machine learning, we identify industrial policies (2010–2022) by analyzing policy text from the Global Trade Alert (GTA) database (Evenett, 2019; Evenett and Fritz, 2020), the largest inventory of commercial policies.[2]

We summarize our method in four steps. First, we define industrial policy as deliberate government action aimed at altering the composition of the domestic economy to achieve a public goal. Our definition draws on historical, economic, and legal scholarship that defines industrial policy by its goals rather than its instruments (Johnson, 1982; Corden, 1980; Warwick, 2013; United States International Trade Commission, 1983; Kapczynski and Michaels, 2023). We classify policies by their stated, *de jure* goals: our measure deliberately includes policies whose announced aims may mask other motives (e.g., protectionism or patronage) or may never be realized, because whether industrial policy pursues and achieves its stated goals is precisely the question a systematic approach to measurement must confront. Second, we manually categorize, or "label," a subset of policy descriptions in our database using a team of human annotators. We show that humans agree on the concept of industrial policy. Third, we fine-tune BERT, a compact pre-trained language model, on these labeled data; the resulting classifier achieves 93% accuracy on held-out annotations. Finally, we deploy the fine-tuned model across our full corpus to construct text-based measures of industrial policy activity.

A disadvantage of language models is the difficulty of interpreting what they have learned: a model may reproduce patterns from its training data (Magar and Schwartz, 2022; Razeghi, Logan IV, Gardner and Singh, 2022) or exploit spurious correlations (McCoy, Pavlick and Linzen, 2019) rather than the concept it is meant to capture. We

2. Our approach follows similar efforts to measure monetary policy (Romer and Romer, 2004) and fiscal policy (Romer and Romer, 2010) shocks through qualitative assessment of policy documents. It also automates the manual data collection typically required for industrial policy case studies (e.g., Barwick, Kalouptsidi and Zahur (2019); Bai, Barwick, Cao and Li (2020); Lane (2020); Barwick, Kwon, Li and Zahur (2024a)).

Table 1: Industrial Policy Goals: Examples

| | Policy Description | Targeted Activity | Policy Instrument |
|---|---|---|---|
| 1 | Brazil increased import tariffs for various IT and telecommunication goods *to stimulate innovation and strengthen the national IT sector.* | IT and telecommunications | Import tariff |
| 2 | The Ministry of Industry and Information Technology released a policy [...] *to boost growth in the Chinese battery industry*, particularly for automobiles. | Batteries | State Loan |
| 3 | [...] the Ministry of Information Industry (MII) of the People's Republic of China (PRC) issued a Planning Release [...] The release [...] seeks to provide guidance on *maintaining and strengthening the PRC's position in the global ship-building industry.* | Shipbuilding | State Loan |
| 4 | The Ecuadorian Executive adopted Decree 675, increasing the percentage of bioethanol in regular fuel [...] *aiming to boost biofuel consumption and production while supporting local agriculture.* | Biofuels, agriculture | Price stabilization |
| 5 | The government of Egypt increased for USD 1,000, its flight subsidies for international charter flights [...] [the] *core scope [of this program] was to boost Egyptian tourism overall.* | Tourism | Production subsidy |
| 6 | The German Federal Government published the Artificial Intelligence (AI) Strategy, *aiming to increase competitiveness and secure responsible AI development in Germany and Europe.* | AI | State aid, unspecified |
| 7 | [...] government of Japan approved [...] supplementary subsidy *to support exports of goods identified under the agricultural export expansion strategy* [...]. | Selected agricultural products | Other export incentives |
| 8 | Nigeria's Federal Executive Council approved a new national automotive policy *to strengthen the automotive sector* and limit imports of used cars. | Automobiles | Import tariff |
| 9 | The South African Executive launched the Green Economy Accord [...] to *promote the development of the Green Economy.* | Green activities | State loan |
| 10 | [...] US Administration enacted the CHIPS and Science Act of 2022 *to boost American semiconductor manufacturing, research, workforce development, and advanced wireless communication technologies*. | Semiconductors | Multiple |

*Notes*: The table shows excerpts of policy descriptions and instrument types from the Global Trade Alert database, which we use in this study. The text that refers to the goals of the policy has been italicized by us.

therefore validate that our measure identifies industrial policy by its objective rather than by surface features of the text. Our main test is an out-of-sample case that postdates our training data entirely: the international sanctions imposed on Russia following its 2022 invasion of Ukraine. Sanctions are a stringent test because they share the surface features of industrial policy — state-imposed, sector-targeted interventions — while differing fundamentally in objective. Our model classifies 95% of them as

either pursuing other objectives or lacking sufficient information, avoiding a trap that a classifier reading vocabulary rather than intent would fall into. A conventional bag-of-words classifier, which performs comparably on held-out data, misclassifies three times as many sanctions as industrial policy — evidence that the value of a fine-tuned language model lies in generalization to novel policy language, not in-sample accuracy.

We also open the model up. Using mechanistic interpretability techniques, we show that fine-tuning organizes the model's internal representation around a single direction that separates industrial policy from other policy with 92% accuracy, that this direction causally drives the model's predictions, and that it places genuinely ambiguous policies near the boundary between the two. This structure comes from fine-tuning on our annotations rather than from scale: pre-trained BERT and a Llama model roughly seventy times larger reach only 85% and 79% separability. Scale is not a substitute for supervision for this task, and our results suggest little would be gained from it: DeBERTaV3, a newer and larger encoder from the same family, matches BERT's performance on the industrial policy class. A compact model also makes validation practical. The exercises we report can in principle be run on larger open-weight models, but their cost rises with model size; the leading proprietary models — GPT, Gemini, Claude — cannot be examined internally at all, since their weights are not public.

Our measure has also been validated externally: Goldberg, Juhász, Lane, Lo Forte and Thurk (2026) and Barwick et al. (2024a) report substantial overlap between our classifications and their own analyses of global semiconductor and electric vehicle policy.

With a validated measure in hand, we put it to a test. We focus on targeting — which sectors industrial policy flows to — because it is where theories of industrial policy make opposing predictions: infant-industry arguments hold that policy should build up sectors lacking a comparative advantage, while a second tradition holds that it should reinforce demonstrated strengths. Until now there has been little systematic evidence on which describes practice. We find that industrial policy systematically targets sectors of established comparative advantage. The pattern is strongest in high-income countries, where the best-performing sectors are about four percentage points more likely to receive new industrial policy than the weakest, and it holds within broad sectors at the product level. In no country group do we find the reverse pattern that infant-industry logic would predict.

We next ask whether a purpose-built measure was necessary to find this pattern. Three comparisons suggest it was. First, policies pursuing other goals — deployed through the very same instruments, in the same countries and years — show no comparable targeting, so the pattern is specific to industrial policy rather than a general feature of policymaking in rich countries. Second, a researcher using tariffs, which have

long served as the literature's proxy for industrial policy, would find no evidence that policy flows to a country's strongest sectors. The reason is that instruments are a poor proxy for intent, a fact visible even inside our own classifier: projecting held-out policies onto the fine-tuned model's industrial-policy direction, we find that instruments split across the boundary — roughly a third of tariffs and three-fifths of subsidies fall on the industrial-policy side. A measure built on instruments therefore mixes industrial policy with policies pursuing quite different objectives, and the targeting signal is lost. Third, the pattern sits uneasily with the empirical literature, which has concentrated on infant-industry episodes — cases in which policy built up sectors that did not yet hold a comparative advantage. Those cases turn out to be atypical of industrial policy as it is generally practiced, which only systematic measurement could have revealed.

We contribute to three strands of literature. First, we advance the empirical study of industrial policy. We introduce a measurement approach based on stated policy goals rather than policy instruments, applicable to other corpora and to other domains where intent, rather than instrument, defines the object of interest. We show that this choice has substantive consequences: instrument-based proxies, of the kind the literature has long relied on, deliver a different answer to the question of what industrial policy targets. Our work joins a growing body of research measuring industrial policy from text (Fang, Li and Lu, 2025; Ju, Li and Wei, 2026), reflecting a broader convergence on goal-based measurement of state intervention. It also complements recent measurement efforts using fiscal data (Criscuolo, Gonne, Kitazawa and Lalanne, 2022; DiPippo, Mazzocco and Kennedy, 2022).

Second, our approach speaks to a broader measurement problem in the trade policy literature. Goldberg and Pavcnik (2016) identify measurement as a central bottleneck to understanding the effects of trade policy, and distinguish between measuring policy and assessing its effects. The difficulty we address is distinct from missing data: commercial policies are recorded in detail, but the recorded attributes — the instrument, the affected sector, the implementing country — do not reveal the objective a policy pursues. Text allows this unobserved attribute to be recovered from records that already exist. Recovering it is a precondition for the questions the literature wants to ask: just as studies of tariff overhang (Beshkar, Bond and Rho, 2015) or the restrictiveness of non-tariff measures (Kee and Xie, 2024) require primary measures of the *de jure* policies whose bite they assess, evaluating whether industrial policy works requires first knowing which policies are industrial policy. Our approach joins related efforts to extract policy information from documents, including Estefania-Flores, Furceri, Hannan, Ostry and Rose (2024), who measure trade restrictiveness from IMF reports, and Teti (2020), who apply computational methods to improve tariff data.

Finally, we contribute to the literature applying natural language processing to economic measurement (D'Orazio, Landis, Palmer and Schrodt, 2014; Gentzkow, Kelly

and Taddy, 2019; Grimmer, Roberts and Stewart, 2022; Ash and Hansen, 2023). Text-based measures in economics have largely relied on dictionary and word-frequency methods (Baker, Bloom and Davis, 2016; Hassan, Hollander, Van Lent and Tahoun, 2019), which are transparent but cannot capture concepts defined by intent rather than vocabulary. We show that supervised classification can measure such concepts at scale, and — more importantly for empirical practice — that the resulting measure can be verified. Applying tools from the mechanistic interpretability literature — Activation Patching (Vig, Gehrmann, Belinkov, Qian, Nevo, Singer and Shieber, 2020; Meng, Bau, Andonian and Belinkov, 2022), and Concept Activation Vectors (Kim, Wattenberg, Gilmer, Cai, Wexler, Viegas and Sayres, 2018) — and following the direction-extraction and steering framework of Wang, Pérez, Schlesinger and Sun (2026), we examine the classifier's internal representations directly, establishing that it encodes industrial policy as an organized, causally active concept rather than a set of surface correlations. As economists increasingly use language models to construct data, methods for validating what those models have learned become as important as the measures themselves.

The paper proceeds as follows. Section 2 sets out our definition of industrial policy and explains why we measure policies by their stated goals. Section 3 describes the corpus of policy text. Section 4 builds the classifier, and Section 5 establishes that it identifies industrial policy by objective rather than by surface features of the text. Section 6 puts the measure to work: it documents that industrial policy targets sectors of established comparative advantage, and shows that this conclusion depends on measuring policies by their goals rather than their instruments. Section 7 concludes.

## 2. Measuring Industrial Policy: Definition and Approach

This section introduces our general framework for measuring industrial policy. We begin by presenting our formal definition of industrial policy. Next, we demonstrate how to operationalize this definition by systematically using policy text to distinguish industrial policies from other government actions. Finally, we illustrate our approach through concrete examples, emphasizing how we use policy objectives, rather than instrument type, to identify industrial policy activity.

### 2.A. Defining Industrial Policy

We define industrial policy as intentional government action aimed at altering the composition of a domestic economy to achieve a public goal. This definition builds on an extensive body of literature and captures three key dimensions of industrial policy.

First, industrial policy is a *political* action (Juhász and Lane, 2024) pursued by governments and implemented by states. This excludes actions taken by nongovernmental actors (e.g., NGOs), firms, and other private entities.

Second, industrial policy is *intentional*: governments seek to alter the composition of economic activity *to achieve specific goals*. Historically, industrial policy aimed to modernize economies and drive structural change, often by promoting manufacturing (Juhász and Steinwender, 2024). More recently, goals such as facilitating the net-zero transition, enhancing supply chain resilience, and achieving strategic autonomy in key sectors have gained prominence (Juhász, Lane and Rodrik, 2024b). As Chalmers Johnson explains, "[t]he very existence of industrial policy implies a strategic, or goal-oriented, approach to the economy" (Johnson, 1982, p. 19). This intentionality is a recurring theme across definitions in the literature, which describe industrial policy as "deliberate," "inherently. . . intentional and active," and directed at outcomes the state "views as important" or "perceives. . . to be efficient" (Warwick, 2013; Boonekamp, 1989; Dubnick and Holt, 1985; Kapczynski and Michaels, 2023; Krugman and Obstfeld, 1991; Chang, 1994; Pitelis, 2006).

Third, industrial policies exhibit *specificity*. Nearly all definitions emphasize their role in altering the structure of the economy (Warwick, 2013), and thus they favor certain activities over others (see Juhász et al. (2024b)). Most famously, industrial policies may target specific industries or sectors. However, many policies cut across traditional sectoral boundaries. For example, South Korea's 1960s export-led policies broadly promoted export *activity* rather than targeting specific sectors (Lane, 2021). Similarly, recent green industrial policies promote green economic activity across sectors, such as encouraging green technologies or supporting battery supply chains. Practically, this breadth means our dataset nests more targeted studies, such as semiconductor policy (Goldberg et al., 2026) or green automotive policy (Barwick, Kwon, Li, Wang and Zahur, 2024b).

### *2.B.* Our Approach: Stated Industrial Policy Goals

Our approach classifies policies by their stated goals — the public, *de jure* articulation of an industrial-policy objective. Stated goals are not mere rhetoric: nearly all industrial policy operates by eliciting private-sector responses (investment, entry, technology adoption). A policy whose intent is never communicated cannot coordinate the behavior it seeks to change. Stated goals are, moreover, the right object for *systematic* measurement, because they avoid conditioning the sample on precisely the characteristics that research on industrial policy seeks to evaluate: i) motives and ii) outcomes.

First, our approach — by design — does not assess the sincerity of industrial policy goals. Whether industrial policy in practice corrects market failures or instead provides cover for protectionism and patronage is one of the oldest debates in the field (e.g.,

Krueger, 1990). Crucially, that debate is empirical, and it can only be adjudicated on a sample that includes both sincere and insincere policies. Our de jure approach admits every policy that publicly claims an industrial-policy objective — sincere and insincere alike — precisely because they make the same claim, and because private actors respond to the announcement whatever the motive behind it. Whether a given policy's design, incidence, or implementation matches its stated goal thereby becomes a question researchers can ask of the data, rather than one settled at the measurement stage.

Second, our approach does not assess whether announced policies are implemented, binding, or ultimately successful.[3] As such, our data include failed and abandoned policies, which addresses a distinct challenge in the existing literature (e.g., Juhász and Steinwender, 2024): empirical evaluations of individual episodes may favor policies that were long-lasting or successful, yielding a sample of exceptional rather than representative episodes.

### 2.C. Examples and Application

Let us illustrate our approach to measurement. We begin with import tariffs, a domain in which measurement issues are well documented. Table 1 provides examples of industrial policies implemented through import tariffs. A Brazilian import tariff (example 1), levied on IT and telecommunication goods, aims to stimulate innovation and strengthen the national IT sector. Likewise, a Nigerian import tariff (example 8) levied on used automobiles aims to "strengthen the automotive sector." These examples demonstrate our definition of industrial policy in practice. The explicit goal of these policies is to alter the composition of economic activity in favor of specific sectors (here, IT or automobiles).

In contrast, Table 2 provides examples of import tariff policies with other, identifiable (non-industrial policy) goals. For instance, a policy from Pakistan raises import tariffs to "increase the revenue of the government" (example 2), while Ghana reintroduces tariffs on staple food items that were removed during a period of "rising food prices" (example 4). Meanwhile, Turkey eliminates import tariffs on prefabricated buildings "to meet the need for shelter caused by earthquakes" (example 6). These policies do not articulate industrial policy goals (i.e., to alter the composition of economic activity), but express other government objectives: increasing fiscal revenue, managing the price of essential goods, or responding to major shocks like natural disasters. Nor is the

3. This mirrors the standard sequence in the trade policy literature, which distinguishes between measuring policy and assessing its effects (Goldberg and Pavcnik, 2016). Studies of tariff overhang (Beshkar et al., 2015) or the material restrictiveness of non-tariff measures (Kee and Xie, 2024) quantify gaps between *de jure* policy and *de facto* outcomes, and are possible only because primary measures of the *de jure* objects exist.

pattern confined to tariffs: Tables 1–2 show governments using the same instruments — subsidies, loans, grants — to pursue both industrial policy and other objectives.

Some of these non-industrial policies are nonetheless selective, targeting specific goods such as food items or building materials, and thus influence the composition of domestic economic activity. The key distinction is that this compositional change is not the policy's objective. In summary, neither the selectivity of a policy nor the policy instrument itself is sufficient to identify industrial policy; only its stated goal is.

## 3. Data

We apply our definition to policy text from the Global Trade Alert (GTA) project, the most comprehensive global database on state commercial policy (Evenett 2019). Our framework, however, applies to any corpus of policy text; see Fang et al. (2025) and Ju et al. (2026) for a contemporary applications to Chinese and US policy documents, respectively. GTA employs international experts and combines an automated search process with manual expert verification to identify and document new state actions. The database is designed to capture policies that change the relative treatment of foreign versus domestic interests (Evenett and Fritz, 2020).

The policy descriptions from GTA's inventory—examples of which are shown in Tables 1 and 2—serve as the primary input for our supervised machine learning workflow. GTA provides standardized, English-language summaries for each policy announcement, written by in-house experts. We use the April 2023 version of this continuously updated dataset.[4]

*i. Inclusion and Credibility.* To be included in the GTA database, a state act must be (i) credible and (ii) materially impactful (Evenett and Fritz, 2020). A credible act is one that has been implemented or whose future implementation is guaranteed; this excludes statements of intent, such as those made in speeches. Although GTA refers to entries as policy "announcements," these are de jure state actions — verified and documented through official statements of the acting institution (Evenett and Fritz, 2020, p. 1). A materially impactful act is one that alters international commercial flows, whether in goods, services, investment, or labor. This credibility filter underpins our approach in Section 2.B: the stated goals we classify are attached to legally committed state acts, not political declarations.

*ii. Scope and Coverage.* GTA's coverage extends beyond the inventories of multilateral institutions such as UNCTAD and the WTO: as an independent organization, it does not rely on countries' self-reporting, instead scanning official government sources

4. This version, provided by the GTA, supersedes the initial July 2020 extract used in earlier versions of the paper.

Table 2: Other (Non-Industrial Policy) Goals: Examples

| | Policy Description (from GTA) | Targeted Activity | Policy Instrument |
|---|---|---|---|
| 1 | [...] Austrian export agency [...] opened a [...] credit line *to support companies in need of liquidity as a result of the economic destabilisation caused by the Russian invasion of Ukraine*. | All exporters | Trade finance |
| 2 | [Pakistan] Economic Coordination Committee approved several measures *to increase the revenue of the government*. An additional 1% duty has been imposed on all imported products except certain exempted items [...]. | Imports | Import tariff |
| 3 | [Finland] Temporary aid scheme [...] *to support primary agricultural production in the current financial and economic crisis* ('the Temporary Framework'). | Agriculture | Financial grant |
| 4 | [Ghana] [...] import duties on rice, wheat and cooking oil *which were removed during the period of rising food prices* in 2008 will be restored. | Staple food items | Import tariff |
| 5 | [Morocco] [...] Ministry of Agriculture provided 2.5 mil. quintals of barley to livestock producers at a subsidized price. [...] *to alleviate the repercussions of the dry winter season that led farmers to purchase imported grains at high prices*. | Livestock producers | Production subsidy |
| 6 | Turkey temporarily terminated the additional duties on prefabricated buildings [...] *to meet the need for shelter caused by earthquakes*. | Building materials | Import tariff |
| 7 | [...] European Investment Bank (EIB) and Banco Comercial Portugues SA signed an agreement worth EUR 75 million [...] for financing small and medium size projects [...] *impacted by forest fires in Portugal.* | SMEs | State loan |
| 8 | [...] UK government prevented Russian companies in aviation/space industry from UK-based insurance services *in response to the invasion of Ukraine by Russia.* | Russian aviation/space firms | Export ban |
| 9 | Vietnam Ministry of Industry and Trade reduced electricity price by 10%. *to ease business difficulties amid the COVID-19 pandemic.* | All | Production subsidy |
| 10 | [...] U.S. enacted American Rescue Plan Act of 2021 *to speed up recovery from COVID-19 pandemic effects*. | Various | Financial grant |

Notes: Policy descriptions (excerpts) and policy instruments from the Global Trade Alert. The text that refers to the goals of the policy have been italicized by us.

— ministry websites, agency portals, and official gazettes — with leads traced back to original documentation (Evenett, 2019). Importantly for our purposes, the database is not limited to classically protectionist measures: it records restrictive and liberalizing policies alike, including subsidies, FDI incentives, trade financing, and research and development policies, and covers services — such as the tourism and insurance policies in Tables 1–2 — as well as traded goods.

*iii. Relationship to Industrial Policies.* By design, GTA captures state acts that change the relative treatment of foreign versus domestic interests. Our measure therefore covers industrial policy within this universe: policies that tilt the playing field toward domestic activity fall squarely inside GTA surveillance, while industrial policy that never touches international commercial interests falls outside it. Two examples illustrate the boundary. Consider consumer subsidies, such as those used to promote electric vehicle adoption. GTA includes these when they incorporate local content requirements or other conditions favoring domestic interests — the US Inflation Reduction Act's tax credits for electric vehicles meeting local content requirements are a case in point.[5] Here, GTA's boundary and our definition align; non-discriminatory consumer subsidies, which aim to alter consumption rather than domestic production, fall outside both. By contrast, education and workforce development policies can serve as industrial policy instruments — the US CHIPS and Science Act funds workforce development and STEM education alongside production incentives.[6] While these qualify as industrial policy under our definition, GTA records only the act's production-targeted incentives, because workforce measures do not directly affect foreign commercial interests. Our indicators should therefore be read as measuring industrial policy that affects the relative treatment of foreign and domestic interests.

*iv. Quality of Coverage.* GTA relies on publicly available information — the paper trail of policies — so coverage may be better for larger countries and for countries with more transparent governments (Evenett, 2019). We assess this directly: in Appendix D, we compare GTA to an OECD dataset (OECD, 2024) constructed with a similar methodology but covering a narrower set of policies, and find that the two sources have broadly similar coverage of the policies they both track. Our main results are identified within country and year, so cross-country differences in coverage are differenced out by construction.

*v. Units, Variables, and Summary Statistics.* An observation in our data is a "measure" or "intervention." For comparability, we focus on national-level policy making. We exclude 1,371 subnational policies from the analysis.[7] We also drop data for two partial years, 2008 and 2023. We conduct our analysis on the 47,283 observations that remain after these two filtering steps. We report summary statistics in Appendix Table G.2.

5. See Global Trade Alert intervention 113854.
6. See the White House Fact Sheet on the CHIPS and Science Act; GTA records the act as intervention 66540.
7. The GTA's reporting of subnational (regional) policies seems incomplete. Less than three percent of the policies in the dataset are subnational, with only 30 countries reported as having *any* subnational policies. In related work (Goldberg et al., 2026), we find that subnational, provincial policies are typically absent for Chinese semiconductor industrial policy. Given these concerns, we exclude subnational policies from the analysis. See Appendix Table G.1 for a complete distribution of the implementation levels of the policies in the database.

Beyond the descriptions of policies, we will also use additional policy variables, or metadata, provided in the source data. These variables include the announcement date; the type of intervention (e.g., a tariff, state loans, etc.); level of implementation; implementing jurisdiction; Harmonized System (HS) 6-digit code of affected sectors; and whether there was firm-level scope tied to the intervention.

We take our definition to the data using supervised learning, which we turn to next.

## 4. Methodology

We construct measures of industrial policy using a three-stage supervised learning process, described in this section. First, we manually label a subset of data according to our formal definition of industrial policy (Section 2.A). Second, we use this labeled data to train a classification model. Third, we apply the trained classifier to predict instances of industrial policy across our database of approximately 47,000 policies. We use these predictions to generate flow-based measures of industrial policy at the country-sector-year level. Appendix B contains a detailed description.

### *4.A.* Labeling: Annotating Subsamples

We begin by hand-labeling a subset of policies based on our formal definition. A team of annotators codes policies using a standardized codebook that provides explicit criteria for determining whether policies meet our definition. The codebook instructs annotators to identify policy goals either through direct statements (e.g., "in order to boost domestic industry by making Egyptian cars more competitive") or implicit indicators (e.g., "China's 'Major Technical Equipment' policy grants tax-free imports to firms in certain sectors involved in the production of said equipment."). The codebook used for classification can be found here.

Research assistants (RAs) hand-label 2,932 policies—approximately 6% of the dataset's 47,283 observations—drawn as a random sample stratified by measure type.[8] Four RAs independently evaluate each policy, assigning one of three labels: "industrial policy," "not industrial policy," or "not enough information" (NEI). Following machine learning conventions, the NEI category captures cases where the text provides insufficient content about the target class. This proves particularly useful for sparse policy references (e.g., brief mentions of tariff-line changes), allowing our classifiers to focus more precisely on distinguishing between industrial policy goals and other policy objectives.

8. RAs were hired from Columbia University, the University of Oxford, and the University of British Columbia.

Labels are assigned through majority voting. 41% of annotated descriptions contain identifiable policy goals, with industrial policies accounting for 24% of hand-labeled cases. For the ambiguous cases where annotators were evenly split, co-author Réka Juhász provided expert adjudication. To understand the consistency of annotations we measure intercoder reliability using two standard metrics: Krippendorff's alpha and Conger's kappa. As detailed in Appendix A, our six rounds of annotation yielded values between 0.6 and 0.8, demonstrating both initial agreement and improved convergence over successive rounds as coders gained experience.

### *4.B.* Model Choice: Fine-Tuned BERT for Industrial Policy Classification

We fine-tune a BERT (Bidirectional Encoder Representations from Transformers) model (Devlin, Chang, Lee and Toutanova, 2019) on our annotated data to map policy documents into one of three categories: *industrial policy*, *not industrial policy*, or *not enough information*.[9] BERT is an encoder-only model: it applies bidirectional self-attention, so each token's representation is conditioned on the entire input sequence rather than only on preceding tokens. This differs from decoder-only architectures such as GPT, Claude, and Llama, which process text to predict the next token and are designed primarily for text generation. BERT processes documents as sequences of tokens, explicitly modeling the contextual relationships between words. As it is pre-trained on large corpora before fine-tuning, it captures contextual meaning that bag-of-words representations cannot. Fine-tuning then adapts these representations to our three-way classification task.

We use a fine-tuned, encoder-only model for two reasons. First, bidirectional attention suits classification tasks: Devlin et al. (2019) shows bidirectional attention improves BERT's performance for classification tasks, and BehnamGhader, Adlakha, Mosbach, Bahdanau, Chapados and Reddy (2024) show decoder-only models become better text encoders once bidirectional attention is enabled. Accordingly, fine-tuned compact encoders remain standard for supervised text classification (Warner, Chaffin, Clavié, Weller, Hallström, Taghadouini, Gallagher, Biswas, Ladhak, Aarsen, Cooper, Adams, Howard and Poli, 2024; Bucher and Martini, 2024; Edwards and Camacho-Collados, 2024). Second, our validation strategy requires a model we can both fine-tune and mechanistically inspect (Section 5.B), which BERT's well-studied and open-source architecture permits (Rogers, Kovaleva and Rumshisky, 2020).

Larger open-weight models such as Llama could themselves be fine-tuned for this task. Two considerations lead us to a compact encoder instead. First, our results give little reason to expect fine-tuning a larger model would help: in Section 5.B we show that our fine-tuned BERT, roughly seventy times smaller than Llama 3.1 8B, produces a more structured internal representation of industrial policy than the larger

9. Specifically, we fine-tune the `BERT-base-uncased` model (Hugging Face, 2025). The computational environment, GPU, and libraries are described in Appendix B.1.

pretrained model — indicating that fine-tuning, not raw capacity, is what this task requires. Nor does additional scale within the encoder family help: DeBERTaV3 matches BERT's performance on the industrial policy class (Appendix B.2). Second, model size determines what validation is feasible. Our interpretability exercises — layer sweeps, direction extraction, and steering — each require repeated passes over the corpus, and their cost rises steeply with model size; fine-tuning and interpreting a model of the size of Llama would multiply the computational cost for a benefit our evidence gives little reason to expect. On proprietary models whose weights are not released, these exercises cannot be run at all.

For model training, we randomly partition our labeled dataset into three stratified subsets: 65% for training, 20% for validation, and 15% for held-out testing. We use the training set to estimate model parameters and the validation set to fine-tune hyper-parameters via bayesian sampling over the learning rate, batch size, number of training epochs, weight decay, and warm up steps.[10] The test set is reserved for evaluating out-of-sample performance of the model. To address class imbalance, we apply standard oversampling during both training and validation.

### *4.C.* Model Performance

We show the strong predictive performance of our BERT model using standard metrics of model performance: precision, recall, accuracy, and F1 score, the latter of which equally weights precision and recall.[11] Recall, also known as the probability of detection, measures the model's ability to correctly identify true policy instances. Precision measures the probability that an instance identified as policy is, in fact, industrial policy.

Table 3 reports classification performance on held-out annotations. The model achieves an F1 score of 88% on the industrial policy class, and 93% accuracy across all three classes. Replicating the exercise with DeBERTaV3, a more recent and larger encoder, yields identical performance on the industrial policy class (Appendix B.2), suggesting that the ceiling on this task is set by how much policy descriptions reveal about their goals, not by the choice of model.

### *4.D.* Prediction: Constructing Text-Based Indices

We use the fine-tuned BERT model to predict instances of industrial policy language throughout our entire policy dataset ($\sim 47,000$ policies). For each observation in the dataset, our model gives an indicator denoting each policy's class (industrial policy goal,

10. The optimal hyperparameters identified for BERT model training are a learning rate of $4.7145 \times 10^{-5}$, a batch size of 8, 4 training epochs, and a weight decay of $2.0953 \times 10^{-2}$ and warm up steps 114.
11. Formally, Recall = $\frac{\text{TP}}{\text{TP+FN}}$ and Precision = $\frac{\text{TP}}{\text{TP+FP}}$. F1 is a weighted combination of each: $F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$. Accuracy refers to the overall share of correct predictions: Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$.

Table 3: Predictive Performance of Three-Class fine-tuned BERT model on Test Data

| Class/Metrics | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| IP Goal | 0.89 | 0.88 | 0.88 | 104 |
| No IP Goal | 0.90 | 0.93 | 0.92 | 76 |
| Not Enough Information | 0.95 | 0.94 | 0.94 | 260 |
| Accuracy | | | 0.93 | 440 |
| Macro Avg | 0.91 | 0.92 | 0.91 | 440 |
| Weighted Avg | 0.93 | 0.93 | 0.93 | 440 |

*Notes:* This table reports the predictive performance of our main three-class model on a held-out test sample of annotated data. We assess performance by comparing model predictions to human-coded labels using a labeled test set. Macro Average refers to the unweighted mean of metrics across the three classes; Weighted Averages are weighted by class size. Accuracy is calculated across all classes.

no industrial policy goal, or not enough information). Before putting our industrial policy data to work, we first present a series of validation tests of our model.

## 5. Model Validation

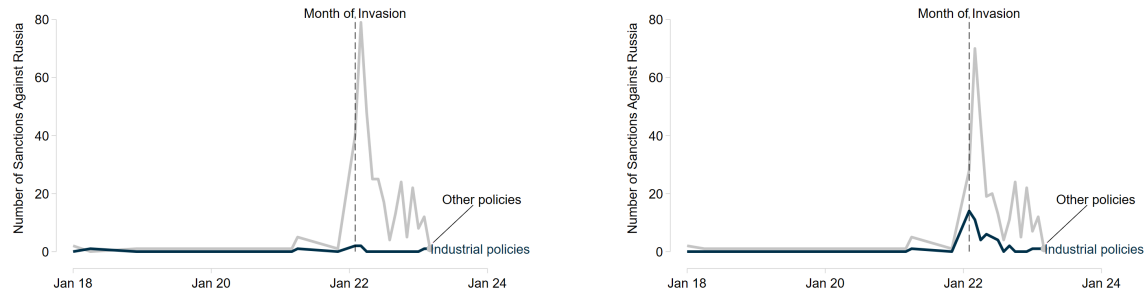
Having trained and deployed our classifier, we next assess whether it captures the concept of industrial policy, rather than a spurious or superficial pattern in the text. A natural concern with LLMs is that they are black boxes: a model can achieve high accuracy by exploiting spurious correlations (McCoy et al., 2019; Niven and Kao, 2019) or memorizing training data (Magar and Schwartz, 2022; Razeghi et al., 2022) rather than learning the concept of interest. We address this concern across three exercises.

First, we test the model's ability to generalize to unseen, out-of-sample policy events, using the 2022 Russian sanctions as a natural experiment. Second, we open the model itself, using mechanistic interpretability techniques. Third, we show that policy instruments alone are not good predictors of IP. Together, these exercises show that our classifier's strong held-out accuracy reflects a representation of industrial policy as a concept rather than of its surface features.

### *5.A.* Construct Validity: Objectives versus Surface Features

A measure of industrial policy is only useful if it classifies policies by their objective rather than by superficial resemblance. We test this using a demanding out-of-sample case: the international sanctions imposed on Russia following its February 2022 invasion

of Ukraine. Sanctions are a natural test of this distinction because they share the surface features of industrial policy — state-imposed trade restrictions targeting specific sectors — while differing fundamentally in objective: sanctions aim to exert geopolitical pressure, not to reshape the composition of domestic production. A classifier that reads surface vocabulary — the shared lexicon of targeted, state-imposed measures — would mistake them for industrial policy; one that has learned the underlying concept would not. The episode also lies entirely outside the model's training window: our codebook was developed in September 2021, before the invasion, and all annotated training policies were drawn from measures implemented through 2020. None of the sanctions regime that followed appears in the data the model was built on.




(a) Fine-tuned BERT (b) Logistic (bag-of-words) classifier

Figure 1: Classification of Russia-Related Sanctions: Fine-tuned BERT versus a Bag-of-Words Benchmark

*Notes:* The figure reports how each model classifies sanctions-related policies in the context of Russia's invasion of Ukraine. As of April 2023, 374 such measures had been published in the GTA. Panel (a) shows the fine-tuned BERT model, which classifies 95% as either pursuing other policy goals or lacking sufficient information to classify (NEI); it labels 5% as industrial policy. Panel (b) shows a logistic regression on TF-IDF-vectorized text (Appendix B.3), which labels 15% as industrial policy despite similar in-sample performance. The dashed line marks the February 2022 invasion.

We identify sanctions on Russia using two criteria: the affected countries include Russia, and the policy description contains at least one of "sanction," "invasion," "frozen," "aggression," or "illegal." Figure 1, Panel (a), plots the classification of these measures from 2018 through 2023; the dashed line marks the February 2022 invasion, after which sanctions spike sharply. Our model classifies 95% of Russian sanctions as either not industrial policy or lacking sufficient information to classify. It avoids the false-positive trap: despite the shared surface vocabulary, the model does not mistake geopolitical sanctions for industrial policy, indicating that it identifies industrial policy by objective rather than by the presence of targeted market interventions.

That this requires learning the concept, rather than its vocabulary, becomes clear when we repeat the exercise with a conventional classifier. We benchmark our fine-tuned

BERT model against a logistic regression on TF-IDF-vectorized text (Appendix B.3), a standard bag-of-words approach. On held-out test data the two have comparable performance — BERT's F1 score on the industrial-policy class is 0.88 against 0.87 for the tuned Lasso benchmark — so on the distribution the models were tuned for, the simpler method is almost as good. The case for a semantic model rests entirely on generalization to unseen policy language, and the sanctions provide exactly that test. Panel (b) of Figure 1 repeats the sanctions exercise with the logistic classifier: it labels 15% of the Russia-related measures as industrial policy, more than three times BERT's 5%. Sharing the surface vocabulary of industrial policy, sanctions are precisely the case a bag-of-words model conflates — and precisely the case a model that reads objectives does not. The value of the fine-tuned model, then, is not in-sample accuracy, where a simple classifier suffices, but in its ability to generalize to the novel language that systematic measurement will inevitably encounter.

### *5.B.* Internal Representation of Industrial Policy and Its Causal Role.

In this section, we open the black box of our LLM using mechanistic interpretability techniques. Following Wang et al. (2026), we ask four questions: whether the model has an internal "IP-ness axis" – a single direction in its internal space along which texts are ordered from clearly not industrial policy to clearly industrial policy; whether that axis actually drives the model's predictions; whether it behaves sensibly at the margin, placing ambiguous texts near the boundary; and whether it generalizes beyond our labeled data. Throughout, we use dataset $\mathcal{M}$, which combines our held-out test split with a separate set of human annotations not used at any stage of model training or validation.[12]

**The model learns an IP-ness axis.** We use activation patching (Vig et al., 2020; Meng et al., 2022) to identify where to look for a representation of *Industrial Policy*: we run a "No-IP" policy text through the model, but swap the internal representation of a matched "IP" text one layer at a time, and record where the swap moves the predicted probability of "IP" the most. The experiment shows that the effect is concentrated at layer 11, BERT's final layer.[13] This is not surprising on its own, the final layer is the closest to the model's decision. The more informative test is whether that layer contains an organized IP-ness axis. Figure 2 tests this. At each layer, we fit the mean-difference direction between IP and No-IP texts and measure how accurately its sign classifies held-out texts; the vertical axis reports the separability accuracy, the horizontal axis

12. No component of $\mathcal{M}$ was used to fit model weights. The total number of observations in $\mathcal{M}$ is $N = 956$. See appendix C for details on its construction and for the mathematical details underlying this section.
13. See Appendix Figure C.1. A complementary test based on Concept Activation Vectors (Kim et al., 2018), which measures the model's local sensitivity to the concept, also peaks at layer 11. See Appendix C.2.

runs from the first layer to the last.[14] The figure shows that at layer 11, there exists an "IP-ness axis" that separates the two classes cleanly: positions on this direction classify held-out "IP" and "No-IP" texts with 92.4% accuracy. Fine-tuning is central in the creation of this structure: repeating the exercise on pre-trained BERT and Llama 3.1 8B, neither of which saw our annotations, yields much weaker separability in the final layer: pretrained BERT has a separability of 84.7%, while the same metric for Llama in its final layer is 78.8%. The comparison is informative beyond the levels. The fine-tuned model's separability rises with depth and remains highest across its final layers, precisely where the model forms its decision. The pretrained models fluctuate across depth and never achieve the same level of separability. Fine-tuning, not scale, is therefore what organizes the industrial-policy concept: a small, task-specific encoder builds a cleaner representation of the concept than a much larger pretrained model, confirming that a fine-tuned BERT is sufficient for our purposes (Section 4.B).[15]

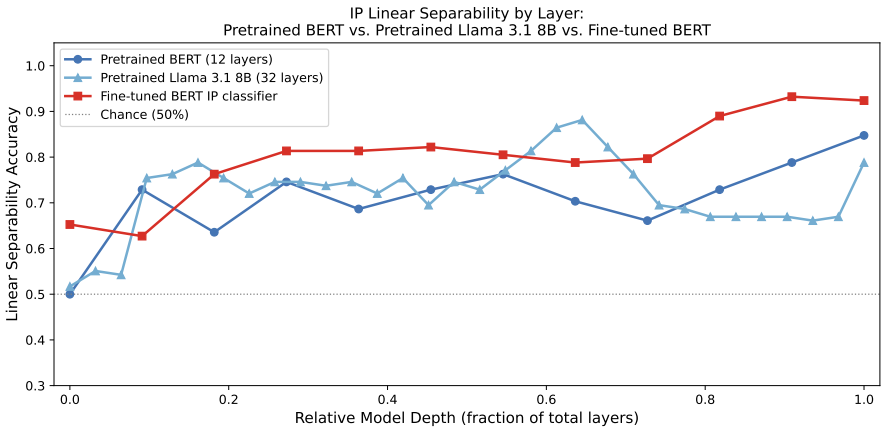



Figure 2: Linear Separability of the Industrial Policy Concept by Layer: PreTrained BERT vs Llama 3.1 8B vs. Fine-Tuned BERT

*Notes:* This figure reports linear separability by layer: at each layer, we fit a mean-difference direction on the extraction pairs and classify the evaluation pairs by the sign of their projection onto it, for the fine-tuned model, pretrained BERT (which never saw our annotations), and pretrained Llama 3.1 8B (last-token hidden states; layer depth normalized to a common 0–1 scale, since the models differ in depth: 12, 12, and 32 layers).

**The dimension drives the model's predictions.** Clean separation along the IP-ness axis is correlational: it shows the model *encodes* the concept, not that it *uses* it. To establish the latter, we run a dose-response experiment known as steering (Wang et al., 2026).

14. This separability accuracy is a probe of the internal representation, distinct from the model's own classification accuracy in Table 3.
15. Because the classification head is a linear function of the final-layer representation, high separability there after fine-tuning might be expected regardless of whether the underlying concept is meaningful. We address this concern below, where we show the same results replicate one layer upstream of the classification head.

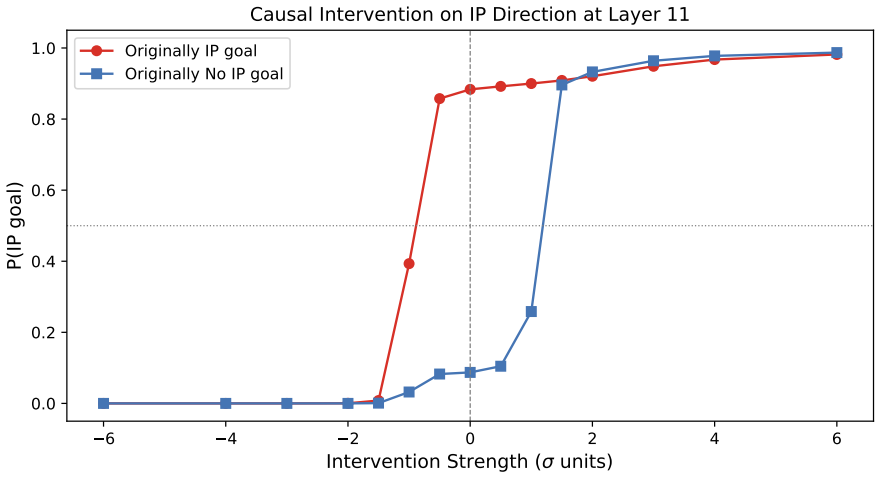



Figure 3: Steering experiment results: Causal Intervention on the IP Direction at Layer 11

Note: This figure reports the steering experiment described in Section 5.B. While the model processes a policy text, we artificially shift its layer-11 [CLS] representation along the unit-norm IP direction by $\alpha$ standard deviations, where the standard deviation is that of the texts' natural activations projected onto the direction; $\alpha = 0$ corresponds to the unmodified forward pass, and Appendix C details the construction of the direction and the intervention. Lines show the mean predicted probability of "IP goal" over the held-out evaluation pairs, separately for texts originally labeled as industrial policy (red) and not industrial policy (blue).

While the model reads a policy text, we artificially shift its layer-11 representation along the IP-ness axis by $\alpha$ standard deviations, and record the predicted probability of "IP". Figure 3 shows that the response is monotonic over $\alpha \in (-2, 2)$: pushing a text toward the "IP" end of the axis raises the probability the model classifies it as industrial policy, and pushing it away lowers it. The axis is therefore not a passive by-product of training but the dimension along which the model makes its decisions.[16]

**The IP-ness axis behaves sensibly at the margin.** An axis learned based on clear IP and No-IP example could fail at separating ambiguous cases. To test this we use our "NEI" (Not Enough Information) cases ($n_{NEI} = 578$) — texts our annotators found too ambiguous to label, and which played no role in constructing the axis. Figure 4 shows that they concentrate in a distinct intermediate region. While clear IP and No-IP texts project at average positions of +12 and −11, NEI texts average +3.[17] For some texts, ambiguity is a property of the data rather than noise among annotators, and should be

16. "Causal" here refers to the effect of the internal representation on the model's own output, not to causal inference about policies themselves.
17. The NEI cluster's positive offset is statistically significant (mean +2.9, s.e. 0.17; $p < 0.001$) but small relative to the axis's scale: the cluster lies roughly four times closer to the decision boundary than either class center.

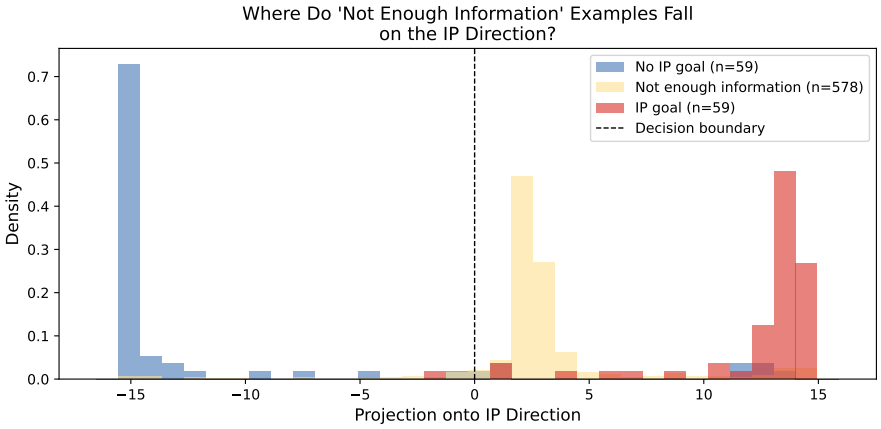



Figure 4: Where Do "Not Enough Information" Examples Fall on the IP Direction?

*Notes:* This figure shows the distribution of layer-11 projections onto the fine-tuned model's IP direction for the 59 evaluation pairs and the 578 held-out texts human-labeled "not enough information," which played no role in constructing the direction. The dashed line marks the decision boundary (projection = 0). Appendix C details the construction of the pairs, directions, and projections.

modeled as such (Pavlick and Kwiatkowski, 2019). Our third category preserves this ambiguity instead of averaging it away; the model's geometry independently ratifies the distinction, placing these texts in their own intermediate region.

**The dimension generalizes out of sample.** A remaining concern is that the axis reproduces patterns specific to our labeled data. We therefore project the Russian policy texts from Section 5.A onto the axis, restricting to the 337 measures announced after the invasion. This subset postdates our annotation codebook, was never seen by annotators, and played no role in constructing the axis. Figure 5 shows that the sanctions project strongly onto the "No-IP" side — the model's internal geometry, not just its final output, treats sanctions as distinct from industrial policy.

A final concern is that our experiments so far intervene at the model's last layer — the one that feeds directly into its classification decision. An intervention there has nothing downstream to test it, so it could move the model's predictions even if that layer told us little about what the model learned. To address this, Appendix C.6 repeats the analysis one layer earlier, where any intervention must still pass through a full layer of processing before reaching the decision. The results are the same: the layer separates industrial policy from other policies with high accuracy, steering along its IP direction flips the model's predictions, and the Russian sanctions land on the No-IP side. Moreover, the fine-tuned model's advantage over the pretrained models is even larger at this layer, 93% separability against 79% for pretrained BERT and 67% for Llama 3.1 8B at matched depth. Our conclusions therefore do not depend on having tested the model right at its decision point.

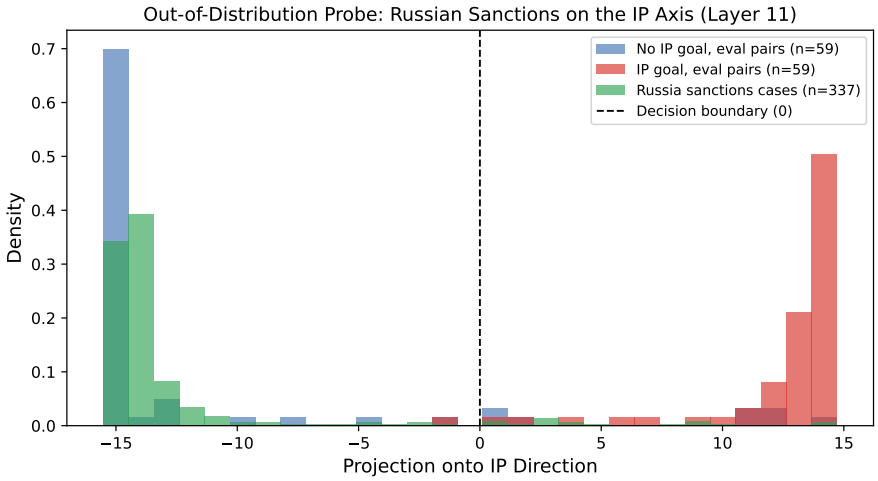



Figure 5: Projection of the Russian Sanctions on the IP-ness axis on layer 11

*Notes*: This figure projects the 337 post-invasion measures in our Russia-sanctions dataset (announced on or after February 24, 2022 — there are 374 total sanction policies, 337 post-invasion) onto the layer-11 IP direction, alongside the held-out evaluation pairs. Sanctions were selected without reference to the model and postdate the annotation codebook. 94.4% of sanctions project onto the No-IP side of the decision boundary (mean projection -12.7).

In summary, fine-tuning endows the network with an organized, low-dimensional representation of *Industrial Policy* that separates positives from negatives, places ambiguous cases near the boundary, causally controls predictions, and extends to events outside our training distribution.

### *5.C.* Instruments Do Not Identify Industrial Policy

A concern with any classifier is that it may exploit a shortcut rather than the concept: here, the worry is that the model learns to equate certain instruments — tariffs, subsidies — with industrial policy, rather than reading intent from the text. If it did, each instrument would project consistently to one side of the IP axis. It does not. We project a stratified sample of 300 policies per instrument category (MAST chapter), none seen by the model or annotators during training, onto the axis (Figure 6). Most instruments split across the boundary: among policies with identifiable goals, 34% of tariffs and 59% of subsidies fall on the industrial-policy side, and government procurement divides almost evenly (48%). Only instruments that are rarely industrial policy in any guise — contingent trade-protective measures, licensing and quotas — concentrate cleanly on the non-IP side. The model's representation of industrial policy is thus largely orthogonal to instrument type: it is reading goals, not instruments. Given the finding that policy instruments and industrial-policy status are largely orthogonal, any measure built from instrument type will misclassify a large share of policies. Whether this misclassification

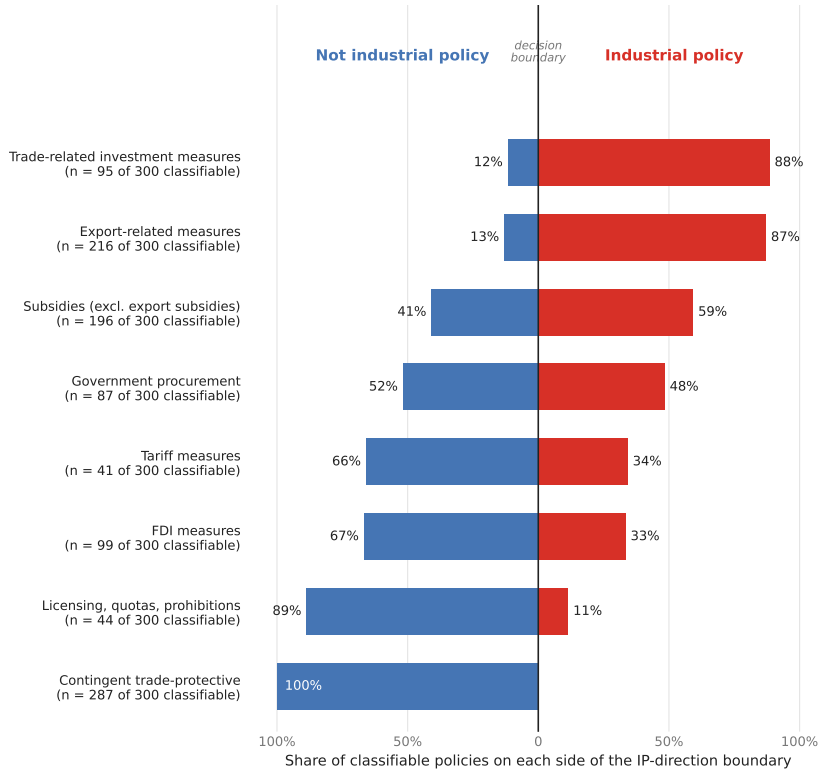


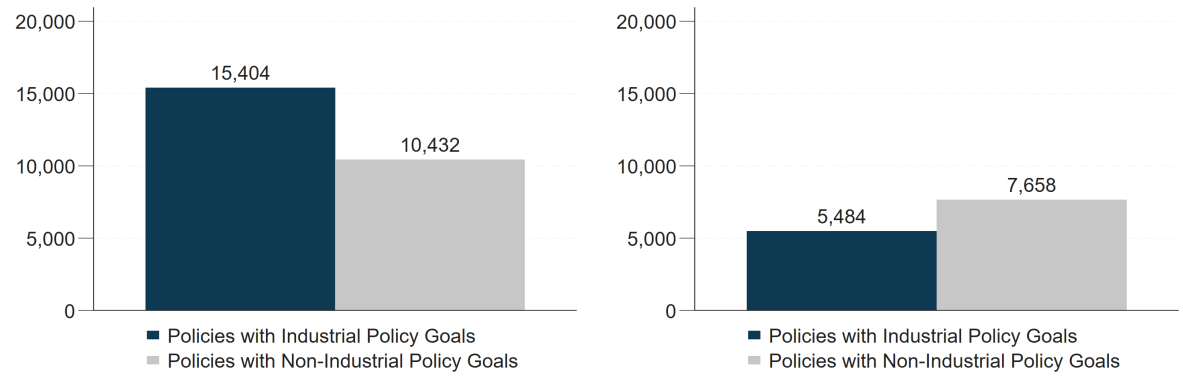

Figure 6: Share of policies on each side of the model's internal industrial-policy boundary, by policy instrument.

*Notes:* This figure shows, for each policy instrument, the share of policies falling on each side of the model's internal industrial-policy boundary. We draw a stratified random sample of 300 GTA measures per MAST chapter (2,400 total) from policies never included in the annotated data used for model training, validation, or testing, and project each measure's layer-11 [CLS] representation onto the IP direction of Section 5.B. Shares are computed among policies with an identifiable goal (n reported per row); remaining policies are classified as "not enough information." The dashed line marks the decision boundary (projection = 0).

actually matters — whether it changes what we conclude about industrial policy — is the question we take up next.

## 6. Industrial Policy in Practice: Measurement Matters

In this section, we show that measuring industrial policy accurately changes what we learn about it. How much this matters depends on how much industrial policy there is. If industrial policy were a small share of commercial policies, the choice of measure would be a minor concern. We therefore begin by documenting the scale of industrial policy. Two findings stand out: industrial policy accounts for a large share of commercial policy, and it is on the rise.




(a) Total Policies (b) National Policies

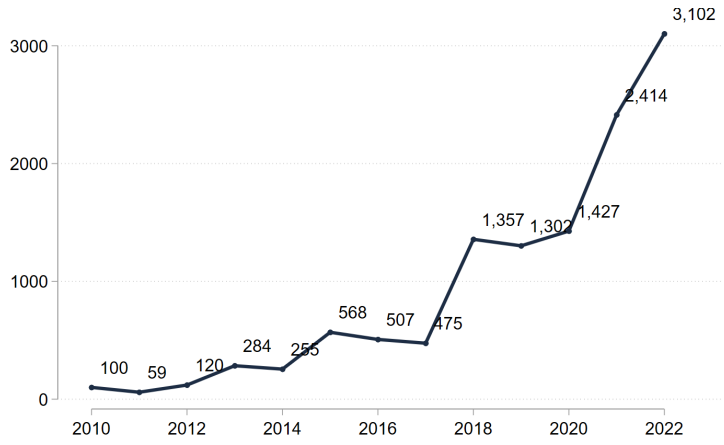
Figure 7: Model Classification of Policies with Identifiable Goals

*Notes:* This figure shows the results of the classification from our BERT model. This exercise excludes policies that did not have identifiable goals based on the BERT model. National policies are those not directed at a small set of firms.

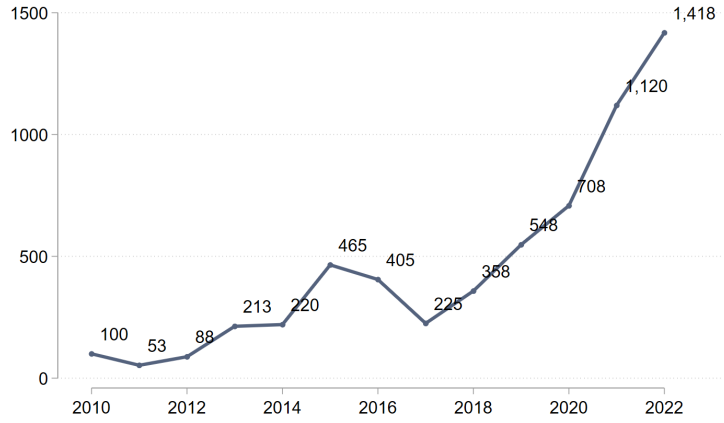
Figure 7 shows that between 42 and 60% of policies with identifiable goals are industrial policy.[18] Industrial policy is also on the rise. Figure 8 shows an expansion of industrial policy activity over the past decade on three counts, which differ in how they handle variation in the granularity of GTA's reporting. Panel (a) reports the raw count of industrial policies, which increased more than thirtyfold between 2010 and 2022. Panel (b) restricts attention to "national" policies — those not directed at a small set of firms — which increased fifteenfold, indicating that the rise is not simply the GTA recording policies at a finer level over time.[19] Panel (c) counts the public agencies — ministries, development banks, export-credit agencies, and the like — announcing at least one new industrial policy in a given year, which by 2022 was close to an order of magnitude higher than in 2010.[20] This rise is not an artifact of the GTA's expanding coverage: the *share* of industrial policy among all recorded policies has itself grown (Appendix Figure F.1), consistent with the widely noted resurgence of industrial policy in the 2010s (Stiglitz, Lin and Monga, 2013; Cherif and Hasanov, 2019).

Having established that industrial policy is a substantial and growing share of commercial policy, we turn to what it reveals — and to whether measuring it accurately, rather than proxying it, changes the answer. We make this case through the *targeting* of industrial policy: the sectors it flows to. Targeting is a useful lens because it is where theories of industrial policy disagree. The infant-industry tradition holds that policy

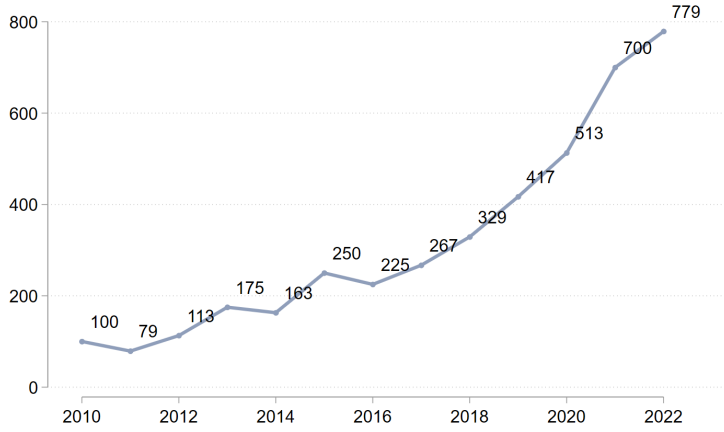
18. For this exercise, we exclude policies whose goals could not be identified (i.e., the "not enough information" group).
19. We use the term "national" policy to refer to policies that do not target only a small set of firms; this is distinct from the implementation level reported in Appendix Table G.1.
20. We extract implementing agencies from policy text; see Appendix E.




(a) All Industrial Policies




(b) National Industrial Policies




(c) Agencies Implementing Industrial Policies

Figure 8: Time Trend of Industrial Policy Activity

*Notes:* This figure shows the trend of industrial policy from 2010 to 2022. We follow GTA guidance and use only policies recorded by the GTA in the same year that they were announced for this exercise. This is due to the substantial backfilling of data that is a living dataset. By using only policies recorded in the same years as they were announced, we ensure the comparability of data across both more distant and recent years. The figure is presented as an index with the year 2010 set as the base year (indexed to 100). All subsequent values reflect changes relative to this baseline.

should target sectors where the country lacks a comparative advantage, while a second tradition holds that it should reinforce a country's demonstrated strengths (Hausmann and Rodrik, 2003; Lin and Chang, 2009). The two imply opposite patterns in the data, yet there is little systematic evidence on which describes industrial policy in practice.

We begin by establishing the targeting pattern using our measure, then probe it through three comparisons. First, we compare the targeting of industrial policy to that of policies pursuing other goals, often deployed through the very same instruments, and ask whether industrial policy is a distinct object. Second, we compare our measure of industrial policy to the tariff-based proxy economists have traditionally relied on, and ask whether the choice of measure changes the conclusion. Third, we compare the typical industrial policy to prominent cases the literature has studied, and ask whether those cases are representative.

### *6.A.* Industrial Policy Targets Comparative Advantage

We first establish the central empirical finding in this paper: based on our measure, industrial policy targets sectors in which a country already holds an established comparative advantage. To measure targeting we merge our data with trade flow data using the United Nations Commodity Trade Statistics Database (UN COMTRADE). Trade values are reported in USD, and we consider trade flows at the Harmonized System (HS) aggregate 2-digit and 6-digit levels. We use these data to construct revealed comparative advantage (RCA) (Balassa, 1965), which is a widely used metric for measuring export specialization. It is defined as $\text{RCA}_{kc} := \left(\frac{X_{kc}}{\sum_c X_{kc}}\right) \Big/ \left(\frac{\sum_k X_{kc}}{\sum_{c,k} X_{kc}}\right)$, where $X_{kc}$ denotes country $c$'s exports in industry $k$. When a country has a revealed comparative advantage in sector $k$ that is greater than one, it means the country is more specialized in exporting that sector than other countries on average.

We run linear probability model (LPM) regressions of the form

$$\text{IP}_{kct} = \alpha + \beta \text{RCA}_{kct} + \gamma_{ct} + \epsilon_{kct},$$

where $\text{IP}_{kct}$ is a binary indicator variable that takes the value of one if HS sector $k$ in country $c$ in year $t$ has at least one new industrial policy announcement, $\text{RCA}_{kct}$ is revealed comparative advantage, and $\epsilon_{kct}$ is the error term. All specifications include country-by-year effects $\gamma_{ct}$. We estimate the specification at the HS2 level for the years 2010-2022, using a sample of 175 countries reported in COMTRADE. Standard errors are clustered at the country level.

The fixed effects are central to our research design. They difference out every cross-country difference in how much policy a country undertakes and how completely the GTA records it — the coverage concerns of Section 3 — so that $\beta$ is identified purely from variation *within* a country and year.

Table 4: Regression of Industrial Policy Activity on RCA

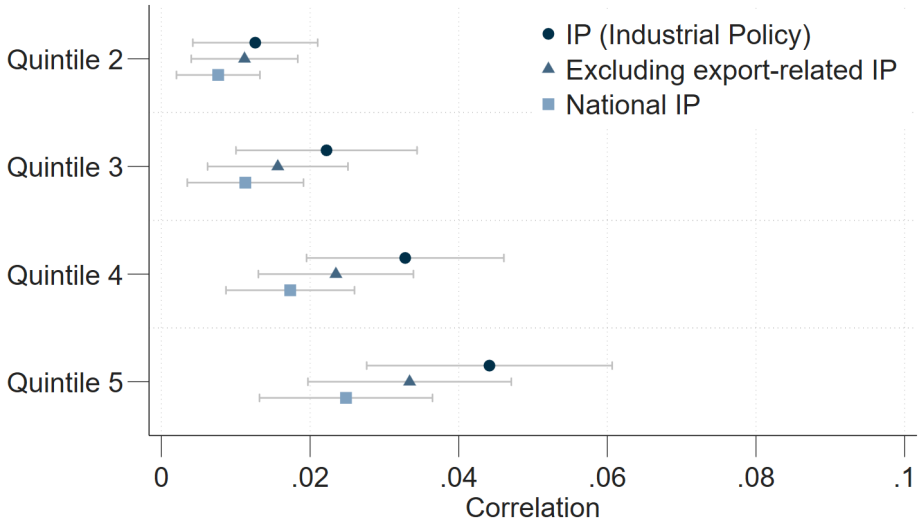
| Independent Variables | (1) IP = 1 | (2) IP = 1 | (3) IP = 1 | (4) IP = 1 |
|---|---|---|---|---|
| RCA > 1 | 0.01893*** | 0.00483** | | |
| | (0.00391) | (0.00190) | | |
| GDPpc > Median × RCA > 1 | | 0.02758*** | | |
| | | (0.00714) | | |
| ln(RCA) | | | 0.00193*** | 0.00065** |
| | | | (0.00037) | (0.00027) |
| GDPpc > Median × ln(RCA) | | | | 0.00273*** |
| | | | | (0.00077) |
| Observations | 199968 | 199968 | 180176 | 180176 |
| R-squared | 0.329 | 0.330 | 0.326 | 0.326 |
| Mean | 0.043 | 0.043 | 0.048 | 0.048 |
| # of Countries | 175 | 175 | 175 | 175 |

*Notes:* This table shows the relationship between industrial policy targeting and revealed comparative advantage. We regress an indicator of industrial policy at the country-year-HS2 level on revealed comparative advantage (RCA). IP takes the value of 1 if, for a given country-year pair, that sector (2-digit HS) received at least one industrial policy. RCA measures are created with trade data from COMTRADE. Standard errors are clustered by country. Asterisks denote statistical significance at the 1% (***), 5% (**), and 10% (*) levels. All columns include country-by-year fixed effects. Mean refers to the mean value of the dependent variable.

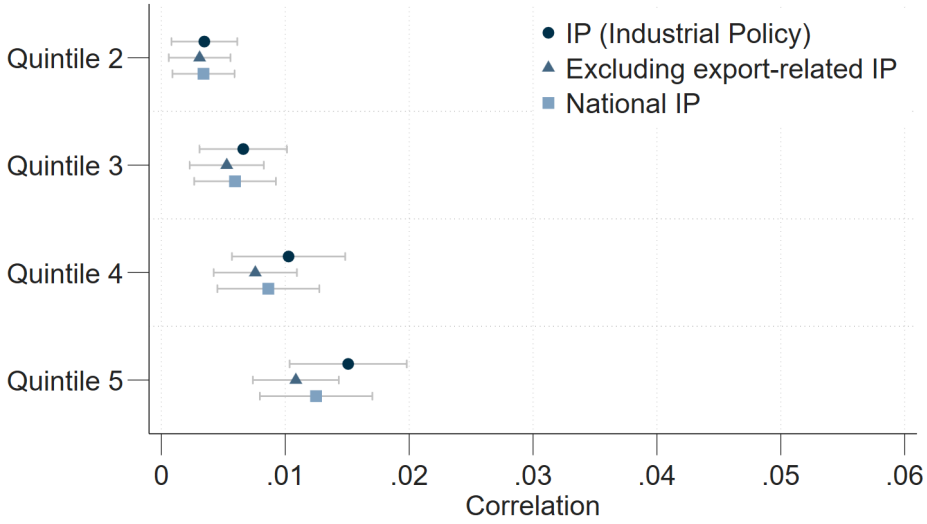
The fixed effects do not, however, address a distinct issue that arises *within* a country-year: the GTA records policies at different levels of granularity, enumerating every disbursement of some programs separately while recording others as a single policy. As the specification compares sectors within a country-year, this uneven granularity could bias which sectors appear most active. We therefore present results using two measures of industrial policy activity: raw counts, and "national" policies, which exclude firm-level measures.

Table 4 shows that sectors with higher RCA are more likely to receive new industrial policy. Sectors with an RCA above one are 1.9 percentage points more likely to be targeted in a given year (column 1) — a large effect relative to the base rate of 4.3%, and statistically significant. These results are qualitatively similar when using the continuous (log) RCA measure (column 3). Although this is a correlation, the timing of our measure aids interpretation. We observe the announcement of *new* policy in year $t$, while RCA in year $t$ reflects export specialization built up previously. A policy announced in a given year is unlikely to have produced the exports against which it is measured, making reverse causality an implausible explanation. It remains possible that current specialization reflects industrial policy in earlier periods — indeed, this is what the literature on successful industrial policy episodes would suggest. Our claim

is narrower: whatever produced a country's current export position, new industrial policy flows disproportionately toward the sectors where that position is strongest.




(a) HS2 level




(b) HS6 level

Figure 9: Industrial Policy and RCA in High-Income Countries

*Notes:* This figure plots the estimated $\beta_i$ coefficients from equation 1 and their corresponding 95% confidence intervals, for high-income countries (RCA quintiles relative to the country's own distribution; the omitted category is the lowest quintile). Panel (a) estimates the specification at the HS2 level; Panel (b) at the HS6 level, with country-year-HS2 fixed effects, so that identification comes from variation across HS6 products within a broad HS2 sector. In both panels, "IP (Industrial Policy)" takes the value of one if a country-sector-year receives at least one industrial policy; "Excluding export-related IP" excludes industrial policies deployed via export-related measures; "National IP" restricts to national (not firm-specific) industrial policies. High income refers to quintiles 4 and 5. All regressions cluster standard errors by country.

The results in Table 4 suggest potentially strong heterogeneity across the income distribution. The effect is much stronger for higher-income countries, shown by the interaction of RCA with an indicator variable for countries with above median income (columns 2 and 4). To further explore this heterogeneity, we split the baseline sample into different income quintiles and estimate the following specification:

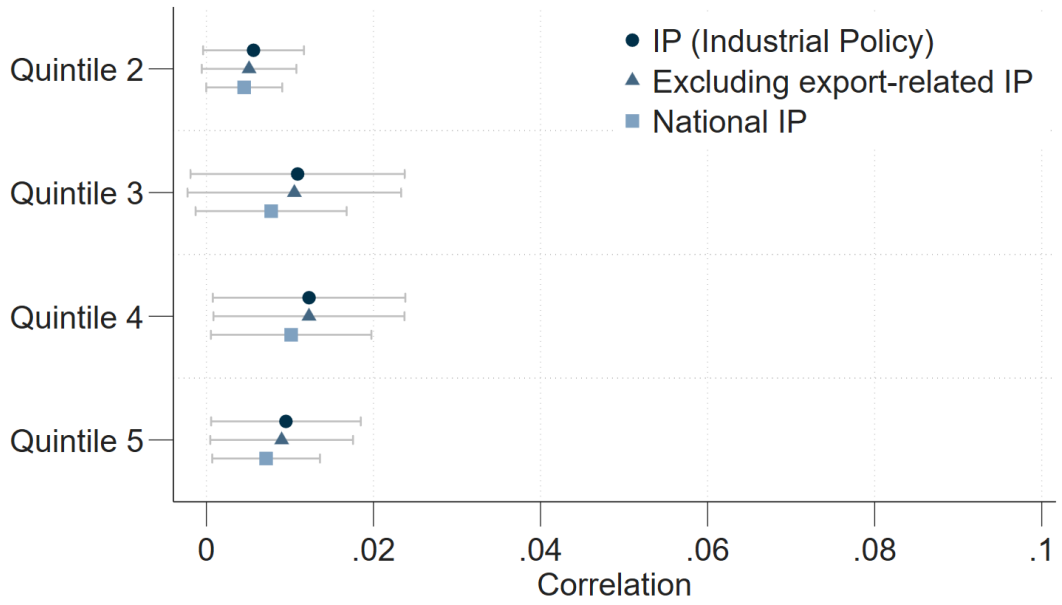
$$\text{IP}_{kct} = \alpha + \sum_{i=2}^{5} \beta_i \cdot \mathbb{1}\{\text{RCA}_{kct} = i\} + \delta_{ct} + \eta_{kct}, \tag{1}$$

Where $\text{IP}_{kct}$ is a binary indicator variable that takes the value of one if HS sector $k$ in country $c$ in year $t$ has at least one new industrial policy announcement, $\beta_i$ is the coefficient on an indicator variable that takes the value of one if $\text{RCA}_{kct}$ is in quintile $i$ of country $c$'s distribution of revealed comparative advantage in year $t$, $\delta_{ct}$ are country-year fixed effects, and $\eta_{kct}$ is the error term. The omitted category is the lowest RCA quintile, which implies that $\beta_i$ captures the difference in probability of industrial policy for RCA quintile $i$ relative to the lowest quintile of the RCA distribution.

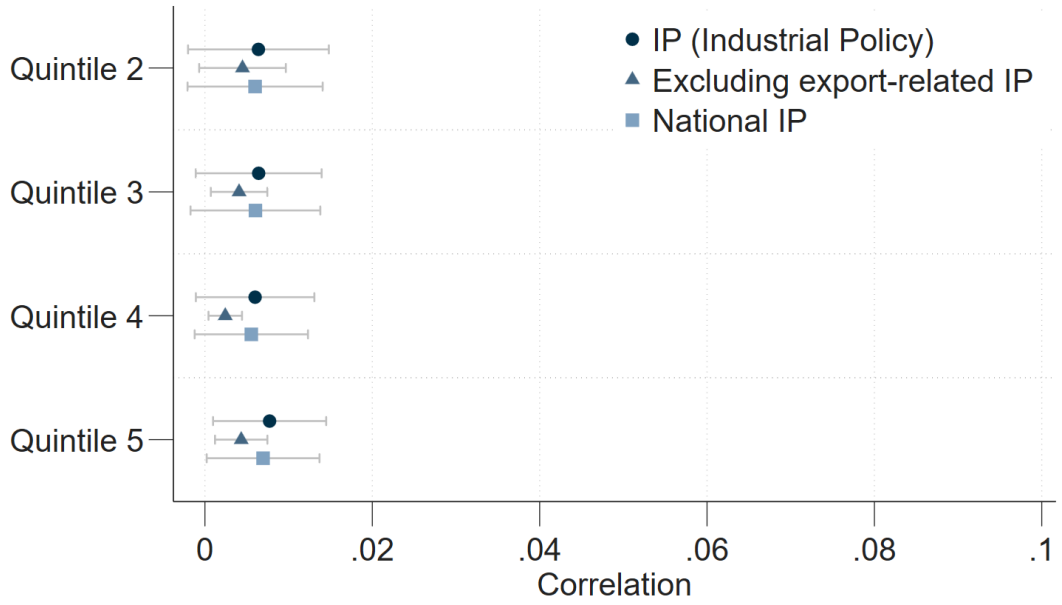
We run this specification separately for i) the two lowest income quintiles, ii) the middle income quintile, and iii) the two highest income quintiles. We split all countries in our data into income quintiles based on 2010 GDP per capita data from the World Bank. For high-income countries (Figure 9 panel a), we see a strong increasing monotonic relationship between a sector's position in the RCA distribution and the probability of receiving a new industrial policy. In high-income countries, the best-performing sectors (quintile 5) are about 4 percentage points more likely to receive a new industrial policy than the worst-performing sectors (quintile 1). The results are robust to using only national policies, and to dropping all export-related policies so it is not an artifact of export-promotion instruments mechanically following exports.

The relationship holds under disaggregation. Panel (b) re-estimates it at the HS6 product level with country–year–HS2 fixed effects, so that identification comes only from variation across products within a broad sector. The monotonic gradient survives: even within an HS2 sector, high-income countries direct industrial policy toward their strongest individual products. This is consistent with theories in which industrial policy builds on demonstrated comparative advantage rather than defying it, and with the idea that *maintaining* an edge at the technological frontier requires sustained support, as documented for advanced semiconductor manufacturing in frontier economies (OECD, 2019; Goldberg et al., 2026).

In comparison to high-income countries, Figure 10 shows a weaker relationship for low- and middle-income countries. Coefficients are smaller in magnitude and do not rise monotonically across the RCA distribution, though in some specifications the highest RCA quintiles remain positive and significant. In no income group, however,




(a) Middle-Income Countries




(b) Low-Income Countries

Figure 10: Sectoral Industrial Policy Activity by Income Levels and Revealed Comparative Advantage

*Notes:*This figure plots the estimated $\beta_i$ coefficients from equation 1 and their corresponding 95% confidence intervals. All regressions include country-year fixed effects and cluster standard errors by country. The dependent variables are as follows: "IP (Industrial Policy)" takes the value of one if a country-HS2 sector-year receives at least one industrial policy; "Excluding export-related IP" excludes industrial policies deployed via export-related measures; "National IP" takes the value of one if a country-HS2 sector-year receives at least one national (not firm-specific) industrial policy. The omitted category is the lowest quintile of the RCA distribution. Middle-income refers to quintile 3. Low-income refers to quintiles 1 and 2.

do we find the *opposite* pattern — industrial policy directed away from a country's areas of comparative advantage and toward its weakest sectors, as infant-industry theories, in which policy cultivates industries that lack comparative advantage, would predict. The difference across the income distribution is thus one of degree, not direction: where targeting is detectable, it follows comparative advantage rather than defying it.

We cannot distinguish whether this weaker relationship reflects genuinely less systematic targeting in lower-income countries or the thinner policy record available for them. Policy descriptions from these countries are considerably less informative about goals: the share of policies we cannot classify exceeds 60 percent in the bottom income quintiles, against roughly 20 percent in the top (Appendix Figure F.2). Where classification is noisier, a targeting gradient is harder to detect even if present. The comparisons that follow therefore focus on high-income countries, where the pattern is estimated precisely enough to serve as a benchmark against which alternative measures can be judged.

What explains this pattern is a separate question. In practice, targeting of industrial policy need not conform to any economic theory in which policy targets market failures. Equally, targeting may be driven by the political incentives of policymakers (e.g., Juhász and Lane, 2024). Better understanding the role of large, politically influential "superstar" exporting firms that can single-handedly shape a country's revealed comparative advantage (e.g., Freund and Pierola, 2015) is a question our measure now makes it possible to ask, and one we leave to future work.

In summary, our measure shows that industrial policy targets sectors in which countries already hold an established international presence, a pattern detected most strongly in high-income economies. In the next section, we subject this result to three comparisons that make the paper's central argument: getting the measurement of industrial policy right is what allows the pattern to be seen at all.

### *6.B.* The Targeting Result: Three Comparisons

We now turn to the second part of argument: getting measurement right matters for what we learn about industrial policy. We conduct three comparisons.

The first comparison asks whether industrial policy is a distinct object, or whether the targeting pattern is a general feature of government policy in rich countries. We re-estimate equation 1 with non-industrial policies — those our model assigns other goals — as the dependent variable. These policies are drawn from the same countries, years, and instruments as industrial policy; they differ only in their objective. Figure 11a shows they exhibit no comparable targeting: coefficients are shrunk toward zero relative to industrial policy, and there is no monotonic gradient in RCA. Sort government interventions by goal, then, and only industrial policy tracks comparative advantage; the same instruments deployed for other purposes do not. The distinction our measure

draws is thus not merely one of stated goals — which it assigns by construction — but of behavior: industrial and non-industrial policies target sectors differently, along a dimension the measure was never trained to capture. We read this as evidence that industrial policy is a distinct object.

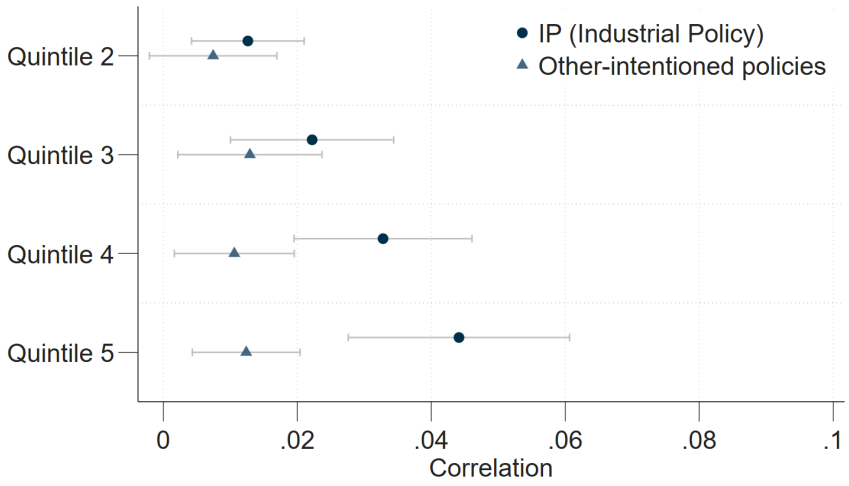
The second comparison asks whether standard proxies for industrial policy yield the same answer. We focus on tariffs, which have a long tradition in the literature as a proxy for industrial policy. As before, we re-estimate equation 1 with the flow of new tariffs as the dependent variable, consistent with our other flow-based measures. Figure 11b shows that new tariffs do not target high-RCA sectors in the way industrial policy does. The relationship is far weaker and non-monotonic: in high-income countries, the highest-RCA quintile is no more likely to receive a new tariff than the lowest, and while some intermediate quintiles differ from zero, the coefficients are small and show no systematic gradient in RCA. A researcher relying on the tariff proxy would therefore find no evidence that industrial policy flows to a country's strongest sectors — unlike the clear, monotonic pattern the goal-based measure reveals.

This is consistent with our result in Section 5.C that only about a third of tariffs with identifiable goals are industrial policy. A tariff-based measure conflates industrial policy with a larger set of tariffs that pursue other goals, and the two do not target sectors in the same way. This is why measurement matters: the standard proxy yields the wrong answer to a fundamental question — what does industrial policy target? — precisely because it cannot disentangle industrial policy from other policy.

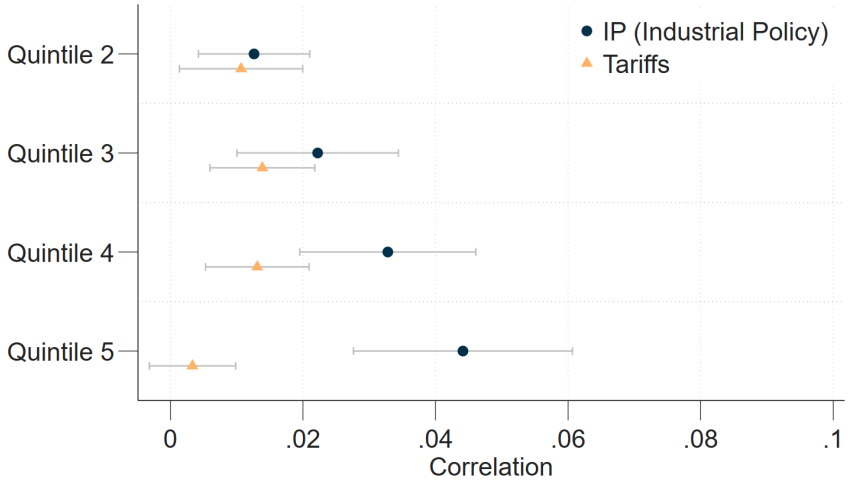
The third comparison considers what the targeting pattern implies for the emerging literature that empirically evaluates prominent episodes of industrial policy (e.g., Juhász, 2018; Hanlon, 2020; Barwick et al., 2019; Lane, 2021; Juhász, Squicciarini and Voigtländer, 2024a; Bai et al., 2020). These studies, several of them our own, examine cases in which policy built up sectors that did *not* yet hold a comparative advantage — the infant-industry pattern, which would appear in our data as a *negative* relationship between industrial policy and RCA.

Strikingly, we do not find a pattern consistent with systematic targeting of infant industries in *any* country group (which would entail a negative relationship between industrial policy and RCA). The episodes at the center of the empirical literature are thus informative about a particular kind of industrial policy: the deliberate cultivation of nascent sectors against a country's existing strengths. But, according to our data, this is not what most industrial policy, in most countries, actually does. Because our measure captures industrial policy as a whole — every policy that states an industrial-policy goal, not only the prominent episodes — it reveals that the canonical cases, valuable as they are, are atypical of the policy as most countries practice it, so that what they teach us need not generalize to industrial policy at large.

These comparisons together show the power of a systematic approach to measurement. Industrial policies target comparative advantage. The same instruments deployed for other purposes do not, showing that industrial policy is a distinct object. The tariff proxy misses this pattern entirely because it mixes industrial policies and policies with other goals. Finally, prominent cases studied previously in the literature are atypical in that they target infant industries rather than established strengths. Instruments have recurred throughout as a poor proxy for industrial policy. We close by turning to instruments directly — characterizing those through which industrial policy is deployed, and showing how little instrument choice reveals about industrial policy.




(a) Industrial and Other-Intentioned Policy Activity




(b) Policy Activity and Tariffs

Figure 11: Sectoral Industrial Policy Activity, non-Industrial Policy Activity and Revealed Comparative Advantage in High Income Countries

*Notes:* This figure plots the estimated $\beta_i$ coefficients from equation 1 and their corresponding 95% confidence intervals. All regressions include country-year fixed effects and cluster standard errors by country. In Panel A) the dependent variables are as follows: "IP (Industrial Policy)" takes the value of one if a country-HS2 sector-year receives at least one industrial policy; "Other-intentioned policies" takes the value of one when a country-HS2 sector-year receives at least one other-intentioned policy. In Panel B) "IP (Industrial Policy)" takes the value of one if a country-HS2 sector-year receives at least one industrial policy; "Tariffs" takes the value of one when a country-HS2 sector-year is affected by at least one tariff policy. The omitted category is the lowest quintile of the RCA distribution. High-income refers to quintiles 4 and 5.

### *6.C.* Contemporary industrial policy favors export promotion and subsidies, not import tariffs.

Contemporary industrial policy is overwhelmingly implemented using subsidies and export promotion, not tariffs. Figure 12 shows that subsidies and export-related

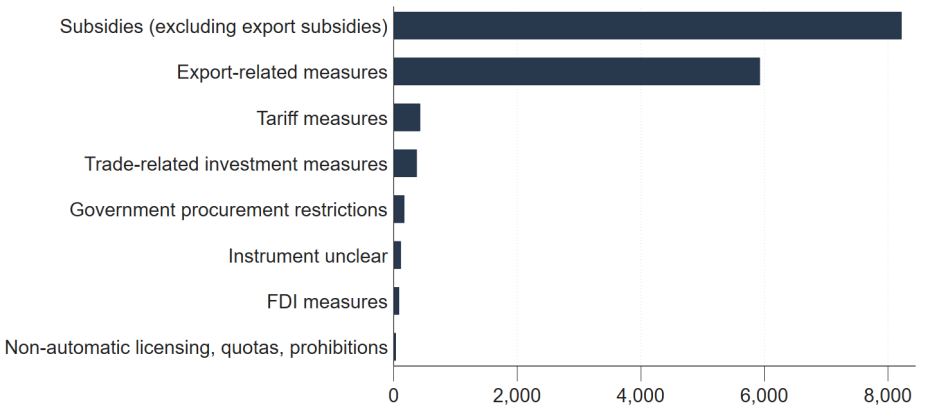
measures dominate industrial policy activity, regardless of how it is measured. Strikingly, these instruments are nearly an order of magnitude more common than import tariffs. This holds using national policies and implementing-agency counts as well as raw counts (Panels b and c), so it is not an artifact of some countries reporting subsidies firm-by-firm.[21]

This feature of industrial policy explains why the tariff proxy failed in Section 6.B, and it does so on two margins. Tariffs are *incomplete* as a measure of industrial policy: the large majority of industrial policy is not deployed through them. Tariffs are also *misleading*: most tariffs are not industrial policy at all (Figure F.3). A measure built on tariffs therefore both misses most of what industrial policy is and mistakes for it a great deal that is not.

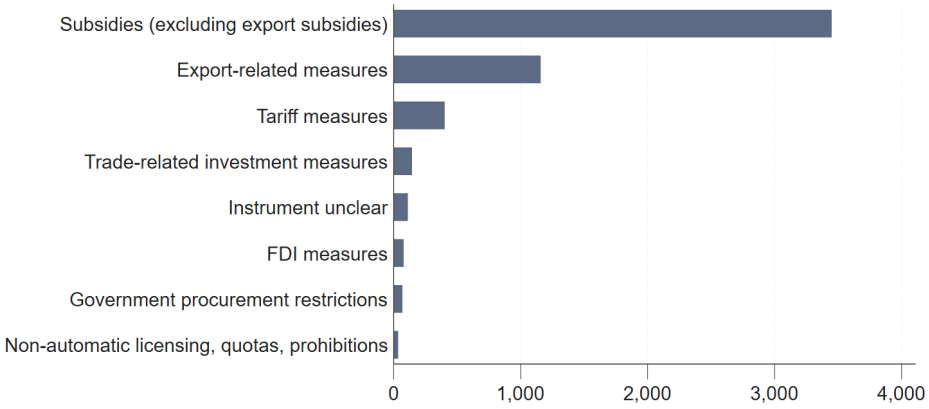
The composition of industrial policy also speaks to a common conflation of industrial policy with protectionism. The recent revival of industrial policy is frequently read as a turn toward trade barriers and economic nationalism. Our data suggest otherwise: the typical industrial policy is not protectionist. Much of it operates by providing finance — trade credit, state loans, and guarantees — to support domestic firms and boost their competitiveness abroad, rather than by restricting imports.[22] While tariffs may have been the characteristic instrument of industrial policy in earlier eras, they are not its characteristic instrument today.

21. Implementing-agency counts collapse agencies' industrial policy activity in a year to a single observation by policy instrument. We construct the implementing-agency measure by extracting institutions from policy text using a large language model with manual validation; see Appendix E and Appendix Table G.3 for examples.

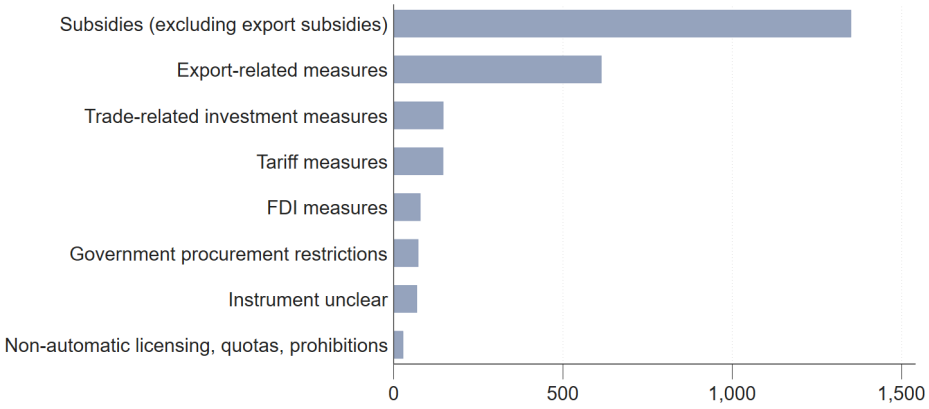
22. Appendix Figure F.4 reports finer instrument categories: export-related industrial policy operates mainly through trade financing rather than tax-based export incentives or (WTO-restricted) export subsidies, and subsidy-based industrial policy through state loans, grants, and guarantees rather than direct production subsidies.




(a) All Policies




(b) National Policies




(c) Implementing Agency-Policy Instrument Pairs

Figure 12: The Instruments of Industrial Policy

*Notes:* The charts show the top eight most-used policy instruments by all the measures. Below eight, IP activity is too small to display clearly. The most-used policy instruments by each measure of IP activity are the same, and the excluded policy instruments are the same for each measure of IP activity, too. The excluded MAST chapter codes are: Price-control measures, Migration measures, Capital control measures, Finance measures, Intellectual property, Contingent trade-protective measures. For Panel (c), Implementing Agency–Policy Instrument Pairs, we calculate the number of agencies implementing at least one industrial policy via each policy instrument in each year, and then sum these counts across years.

## 7. Conclusion

In this paper, we introduce a new approach to measuring industrial policy, addressing a longstanding measurement challenge in the literature. Given that the same instrument can serve industrial policy or any number of other objectives, measures built on instruments cannot distinguish them. We therefore classify policies by their stated goals, using the text of policy announcements, and validate that our classifier identifies industrial policy by objective rather than by surface features of the text.

Applying the measure, we find that industrial policy systematically targets sectors in which a country already holds an established comparative advantage — reinforcing existing strengths rather than building new ones. The pattern is strongest in high-income economies, and in no country group do we find the reverse. This conclusion depends on how industrial policy is measured: policies pursuing other goals through the same instruments show no comparable targeting, and the tariff-based proxy long used in the literature does not recover the pattern. It also sits uneasily with an empirical literature that has focused on infant-industry episodes, which turn out to be atypical of industrial policy as it is generally practiced.

Our findings present many avenues for future research. We highlight two. The first concerns motives. As we argue in Section 2.B, we deliberately measure policy as it is publicly stated, including policies whose announced aims may mask other objectives. We view this as a strength rather than a limitation of our approach: with a systematic measure of stated goals in hand, one can now compare what governments say they are doing to the design, incidence, and implementation of what they do, and ask when and where the two diverge. Whether industrial policy corrects market failures or supplies cover for protectionism and patronage is among the oldest debates in the field; it is also an empirical question, and answering it requires precisely the kind of measure we provide.

The second concerns the targeting pattern itself. We document that industrial policy follows comparative advantage; we do not explain why. The pattern is consistent with theories in which policy builds on demonstrated strengths, but equally with an account in which established exporters are simply better placed to obtain government support. Distinguishing these accounts is work that our measures now make feasible.

More broadly, as economists increasingly turn to language models to construct data, how such measures are validated will matter as much as the measures themselves. Our results show that measurement choices are not innocuous: the same underlying policies, sorted by instrument rather than by goal, support a different conclusion about what industrial policy does. Systematic measurement is what makes the difference visible.

# References


Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in "Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining" 2019, pp. 2623–2631.

Artstein, Ron, "Inter-Annotator Agreement," in Nancy Ide and James Pustejovsky, eds., *Handbook of Linguistic Annotation*, Dordrecht: Springer, 2017, pp. 297–313.

Ash, Elliott and Stephen Hansen, "Text Algorithms in Economics," *Annual Review of Economics*, 2023, *15*, 659–688.

Bai, Jie, Panle Jia Barwick, Shengmao Cao, and Shanjun Li, "Quid Pro Quo, Knowledge Spillover, and Industrial Quality Upgrading: Evidence from the Chinese Auto Industry," Technical Report, National Bureau of Economic Research 2020.

Baker, Scott R, Nicholas Bloom, and Steven J Davis, "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, November 2016, *131* (4), 1593–1636.

Balassa, Bela, "Trade Liberalisation and "Revealed" Comparative Advantage," *The Manchester School*, May 1965, *33* (2), 99–123.

_ , "Tariff Policy and Taxation in Developing Countries," Technical Report, World Bank, Washington, DC 1989.

Barwick, Panle Jia, Hyuk soo Kwon, Shanjun Li, and Nahim Zahur, "Drive Down the Cost: Learning by Doing and Government Policies in the Global EV Battery Industry," Technical Report, National Bureau of Economic Research 2024.

_ , Hyuk-Soo Kwon, Shanjun Li, Yucheng Wang, and Nahim B Zahur, "Industrial Policies and Innovation: Evidence from the Global Automobile Industry," Working Paper 33138, National Bureau of Economic Research November 2024.

_ , Myrto Kalouptsidi, and Nahim Bin Zahur, "China's Industrial Policy: An Empirical Evaluation," Technical Report, National Bureau of Economic Research 2019.

_ , _ , and Nahim Bin Zahur, "Industrial Policy: Lessons from Shipbuilding," *Journal of Economic Perspectives*, 2024, *38* (4), 55–80.

BehnamGhader, Parishad, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy, "LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders," 2024.

Bergstra, James, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl, "Algorithms for Hyper-Parameter Optimization," in "Advances in Neural Information Processing Systems," Vol. 24 2011.

Beshkar, Mostafa, Eric W Bond, and Youngwoo Rho, "Tariff Binding and Overhang: Theory and Evidence," *Journal of International Economics*, 2015, *97* (1), 1–13.

Boonekamp, Clemens, "Industrial Policies of Industrial Countries," *Finance & Development*, March 1989, *26*, 14–17.

Broda, Christian, Nuno Limao, and David E Weinstein, "Optimal Tariffs and Market Power: The Evidence," *American Economic Review*, 2008, *98* (5), 2032–2065.

Bucher, Martin Juan José and Marco Martini, "Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification," 2024.

Cagé, Julia and Lucie Gadenne, "Tax Revenues and the Fiscal Cost of Trade Liberalization, 1792–2006," *Explorations in Economic History*, 2018, *70*, 1–24.

Chang, Ha-Joon, *The Political Economy of Industrial Policy*, London: Palgrave Macmillan, 1994.

__ , *Kicking Away the Ladder: Development Strategy in Historical Perspective*, London: Anthem Press, 2002.

Cherif, Reda and Fuad Hasanov, "The Return of the Policy That Shall Not be Named: Principles of Industrial Policy," Technical Report 2019/074, International Monetary Fund, Washington, DC 2019.

Corden, Warner M., "Relationships Between Macro-economic and Industrial Policies," *World Economy*, 1980, *3* (2), 167–184.

Criscuolo, Chiara, Nicolas Gonne, Kohei Kitazawa, and Guy Lalanne, "An Industrial Policy Framework for OECD Countries: Old Debates, New Perspectives," Technical Report, OECD, Paris 2022.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)" 2019, pp. 4171–4186.

DiPippo, Gerard, Ilaria Mazzocco, and Scott Kennedy, "Red Ink: Estimating Chinese Industrial Policy Spending in Comparative Perspective," Technical Report, Center for Strategic and Internatianl Studies, Washington, DC 2022.

D'Orazio, Vito, Steven T Landis, Glenn Palmer, and Philip Schrodt, "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines," *Political Analysis*, 2014, *22* (2), 224–242.

Dubnick, Mel and Lynne Holt, "Industrial Policy and the States," *Publius*, 1985, *15* (1), 113–129.

Edwards, Aleksandra and Jose Camacho-Collados, "Language Models for Text Classification: Is In-Context Learning Enough?," 2024.

Estefania-Flores, Julia, Davide Furceri, Swarnali A Hannan, Jonathan D Ostry, and Andrew K Rose, "A Measurement of Aggregate Trade Restrictions and Their Economic Effects," *The World Bank Economic Review*, October 2024.

Evenett, Simon J, "Protectionism, State Discrimination, and International Business since the Onset of the Global Financial Crisis," *Journal of International Business Policy*, 2019, *2* (1), 9–36.

__ and Johannes Fritz, "The Global Trade Alert Database Handbook," Technical Report, Global Trade Alert 2020.

Fang, Hanming, Ming Li, and Guangli Lu, "Decoding China's industrial policies," Technical Report, National Bureau of Economic Research 2025.

Field, Lottie, "The Political Economy of Industrial Development Organisations: Are They Run by Politicians or Bureaucrats?," Discussion Paper, University of Oxford 2024.

Freund, Caroline and Martha Denisse Pierola, "Export Superstars," *The Review of Economics and Statistics*, 12 2015, *97* (5), 1023–1032.

Gawande, Kishore, Pravin Krishna, and Marcelo Olarreaga, "A Political-Economic Account of Global Tariffs," *Economics & Politics*, 2015, *27* (2), 204–233.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text As Data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Goldberg, Pinelopi K and Giovanni Maggi, "Protection for Sale: An Empirical Investigation," *The American Economic Review*, December 1999, *89* (5), 1135–1155.

__ and Nina Pavcnik, "The Effects of Trade Policy," in Kyle Bagwell and Robert Staiger, eds., *Handbook of Commercial Policy*, Vol. 1A, Amsterdam: Elsevier B.V. and North-Holland, 2016, chapter 3, pp. 161–206.

_ , Réka Juhász, Nathan J Lane, Giulia Lo Forte, and Jeff Thurk, "Industrial Policy in the Global Semiconductor Sector," Technical Report, National Bureau of Economic Research September 2026.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart, *Text As Data: A New Framework for Machine Learning and the Social Sciences*, Princeton, NJ: Princeton University Press, 2022.

Hanlon, W Walker, "The persistent effect of temporary input cost advantages in shipbuilding, 1850 to 1911," *Journal of the European Economic Association*, 2020, *18* (6), 3173–3209.

Harrison, Ann and Andrés Rodríguez-Clare, "Trade, Foreign Investment, and Industrial Policy for Developing Countries," in Dani Rodrik and Mark Rosenzweig, eds., *Handbooks in Economics*, Vol. 5, Elsevier, 2010, chapter 63, pp. 4039–4214.

Hassan, Tarek A, Stephen Hollander, Laurence Van Lent, and Ahmed Tahoun, "Firm-Level Political Risk: Measurement and Effects," *Quarterly Journal of Economics*, 2019, *134* (4), 2135–2202.

Hausmann, Ricardo and Dani Rodrik, "Economic development as self-discovery," *Journal of development Economics*, 2003, *72* (2), 603–633.

He, Pengcheng, Jianfeng Gao, and Weizhu Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," 2023.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd, "spaCy: Industrial-Strength Natural Language Processing in Python," 2020.

Hugging Face, "BERT Base Uncased Model," `https://huggingface.co/google-bert/bert-base-uncased` 2025. Accessed: 2025-01-10.

Johnson, Chalmers, *MITI and the Japanese Miracle: The Growth of Industrial Policy : 1925-1975*, Stanford, CA: Stanford University Press, 1982.

Johnson, Harry G, "Optimum Welfare and Maximum Revenue Tariffs," *The Review of Economic Studies*, August 1951, *19* (1), 28–35.

Ju, Jiandong, Yuankun Li, and Shang-Jin Wei, "The United States as an Active Industrial Policy Nation," Technical Report, National Bureau of Economic Research 2026.

Juhász, Réka, "Temporary Protection and Technology Adoption: Evidence from the Napoleonic Blockade," *American Economic Review*, 2018, *108* (11), 3339–3376.

_ and Claudia Steinwender, "Industrial Policy and the Great Divergence," *Annual Review of Economics*, 2024, *16*.

_ and Nathan Lane, "The Political Economy of Industrial Policy," *Journal of Economic Perspectives*, 2024, *38* (4), 27–54.

_ , Mara P Squicciarini, and Nico Voigtländer, "Technology adoption and productivity growth: Evidence from industrialization in France," *Journal of Political Economy*, 2024, *132* (10), 3215–3259.

_ , Nathan Lane, and Dani Rodrik, "The New Economics of Industrial Policy," *Annual Review of Economics*, 2024, *16*.

Kapczynski, Amy and Joel Michaels, "Administering a Democratic Industrial Policy," *Harvard Law and Policy Review*, 2023, *18* (2), 279–344.

Kee, Hiau Looi and Enze Xie, "Trade Policies Mix and Match: Theory, Evidence and the EU-Sino Electric Vehicle Disputes," Technical Report, World Bank Group, Washington, DC 2024.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in Jennifer Dy and Andreas Krause, eds., *Proceedings of the 35th*

*International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research* PMLR 10–15 Jul 2018, pp. 2668–2677.

Krippendorff, Klaus, *Content Analysis: An Introduction to Its Methdology*, 2nd ed., Thousand Oaks: Sage Publications Inc, 2004.

Krueger, Anne O., "Government Failures in Development," *The Journal of Economic Perspectives*, 1990, *4* (3), 9–23.

Krugman, Paul R. and Maurice Obstfeld, *International Economics: Theory and Policy*, 2nd ed., New York, NY: Harper Collins, 1991.

Lane, Nathan, "The New Empirics of Industrial Policy," *Journal of Industry, Competition and Trade*, 2020, *1* (2), 1–26.

_ , "Manufacturing Revolutions: Industrial Policy and Industrialization in South Korea," Technical Report, Centre for the Study of African Economies 2021.

Lehmann, Sibylle H and Kevin H O'Rourke, "The Structure of Protection and Growth in the Late Nineteenth Century," *The Review of Economics and Statistics*, August 2011, *93* (2), 606–616.

Lin, Justin Yifu and Ha-Joon Chang, "Should Industrial Policy in Developing Countries Conform to Comparative Advantage or Defy it? A Debate Between Justin Lin and Ha-Joon Chang," *Development Policy Review*, 2009, *27* (5), 483–502.

List, Frederick, *The National System of Political Economy*, Philadelphia, PA: J. B. Lippincott & Company, 1856.

Magar, Inbal and Roy Schwartz, "Data Contamination: From Memorization to Exploitation," 2022.

McCoy, R. Thomas, Ellie Pavlick, and Tal Linzen, "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference," in Anna Korhonen, David Traum, and Lluís Màrquez, eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics Florence, Italy July 2019, pp. 3428–3448.

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov, "Locating and Editing Factual Associations in GPT," in "Advances in Neural Information Processing Systems," Vol. 35 2022, pp. 17359–17372.

Niven, Timothy and Hung-Yu Kao, "Probing Neural Network Comprehension of Natural Language Arguments," in Anna Korhonen, David Traum, and Lluís Màrquez, eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics Florence, Italy July 2019, pp. 4658–4664.

Nunn, Nathan and Daniel Trefler, "The Structure of Tariffs and Long-Term Growth," *American Economic Journal: Macroeconomics*, 2010, *2* (4), 158–194.

Observatory of Economic Complexity, "Lithium oxide and hydroxide," `https://oec.world/en/profile/hs/lithium-oxide-and-hydroxide` 2025. Accessed: 2025-05-16.

OECD, "Measuring Distortions in International Markets: The Semiconductor Value Chain," Technical Report 234, OECD Publishing, Paris 2019.

_ , *OECD Inventory of Export Restrictions on Industrial Raw Materials 2024: Monitoring the use of export restrictions amid market and policy tensions*, Paris: OECD Publishing, 2024.

Passonneau, Rebecca J and Bob Carpenter, "The Benefits of a Model of Annotation," *Transactions of the Association for Computational Linguistics*, 2014, *2*, 311–326.

Pavlick, Ellie and Tom Kwiatkowski, "Inherent Disagreements in Human Textual Inferences," *Transactions of the Association for Computational Linguistics*, 2019, *7*, 677–694.

Pitelis, Christos N, "Industrial Policy: Perspectives, Experience, Issues," in Patrizio Bianchi and Sandrine Labory, eds., *International Handbook on Industrial Policy*, Cheltenham, UK: Edward Elgar Publishing, 2006, pp. 435–450.

Razeghi, Yasaman, Robert L Logan IV, Matt Gardner, and Sameer Singh, "Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning," in Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics Abu Dhabi, United Arab Emirates December 2022, pp. 840–854.

Reidsma, Dennis and Jean Carletta, "Reliability Measurement without Limits," *Computational Linguistics*, September 2008, *34* (3), 319–326.

Rodrik, Dani, *One Economics, Many Recipes: Globalization, Institutions, and Economic Growth*, Princeton, NJ: Princeton University Press, 2008.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky, "A Primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, 2020, *8*, 842–866.

Romer, Christina D and David H Romer, "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 2004, *94* (4), 1055–1084.

__ and __ , "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks," *American Economic Review*, 2010, *100* (3), 763–801.

Stiglitz, Joseph E., Justin Yifu Lin, and Célestin Monga, "Introduction: The Rejuvenation of Industrial Policy," in Joseph E. Stiglitz, Justin Yifu Lin, and Célestin Monga, eds., *The Industrial Policy Revolution I: The Role of Government Beyond Ideology*, New York, NY: Palgrave Macmillan, 2013, pp. 1–15.

Taussig, Frank W, *The Tariff History of the United States*, New York, NY: GP Putnam's Sons, 1914.

Teti, Feodora A, "30 Years of Trade Policy: Evidence from 5.7 Billion Tariffs," Technical Report, ifo July 2020.

United States International Trade Commission, "Foreign Industrial Targeting and Its Effects on U.S. Industries Phase 1: Japan.," Technical Report, Washington, DC 1983.

Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber, "Investigating Gender Bias in Language Models Using Causal Mediation Analysis," in "Advances in Neural Information Processing Systems," Vol. 33 2020, pp. 12388–12401.

Wang, R., V.C. Pérez, C. Schlesinger, and L. Sun, "Localization and Steering of Economic Uncertainty in Large Language Models.," Technical Report 2026.

Warner, Benjamin, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli, "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference," 2024.

Warwick, Ken, "Beyond Industrial Policy: Emerging Issues and New Trends," Technical Report, OECD, Paris 2013.

World Trade Organization, "World Trade Report 2006: Exploring the Links between the Subsidies, Trade and the WTO," 2006.

# Online Appendix

## Measuring Industrial Policy: A Text-Based Approach

Réka Juhász Nathan Lane Verónica C. Pérez

## A. Labeling: Annotator Agreement

Our classifier relies on labeled (annotated) data, which we use to train, validate, and test our classifiers. This appendix describes how we validate that human coders can consistently identify goals based on our definition and codebook.

### *A.1.* Inter-coder reliability metrics

We find agreement between annotators in coding industrial policy classes (i.e., industrial policy, policies with other goals, and cases of not enough information). We assessed this consistency using two standard metrics, Krippendorff's alpha and Conger's kappa. Both metrics accommodate more than two coders. A value of 1 indicates perfect agreement, 0 indicates agreement no better than chance, and values below 0 indicate systematic disagreement.

Traditional content analysis considers Krippendorff's alpha values of 0.67-0.8 tolerable and above 0.8 high quality (Krippendorff, 2004). However, recent research questions these thresholds for machine learning applications, as intercoder reliability measures can be misleading (Reidsma and Carletta, 2008) or inapplicable. This is because if the source of disagreement is due to random noise, machine learning can tolerate data with lower agreement (Artstein, 2017; Passonneau and Carpenter, 2014).

However, if the disagreement is systematic, even reliability measures with values 0.80 and above will provide an unwanted pattern for the machine to detect (Reidsma and Carletta, 2008). Therefore, we treat these metrics as general guidance, considering our measures roughly reliable, particularly in annotation rounds 2-4, where Krippendorff's alpha approaches 0.8.

## B. Model: Core LLM (BERT) and Benchmark (Logistic)

This technical appendix describes the models used in our analysis and describes their parameters, tuning procedures, and training workflows. Our primary model is BERT (Devlin et al., 2019) (Bidirectional Encoder Representations from Transformers), a deep neural network pre-trained on large-scale natural language corpora.

In our application, as in many others, we fine-tune the pre-trained BERT model on a task-specific dataset to perform our custom classification task. To benchmark performance in a more transparent and interpretable manner, we compare BERT to a regularized logistic regression classifier. For this logistic classifier, documents are represented using unigrams and bigrams, which are vectorized via term frequency–inverse document frequency (TF–IDF). We use the best-performing variant of logistic regression, which uses $L_1$ regularization.

### *B.1.* Fine-Tuned BERT Model

*1. BERT: Model Overview.* For our BERT-based classifier, we use the `bert-base-uncased` model (Hugging Face, 2025), which we then fine-tuned for our specific three-class classification task. We fine-tune and evaluate our BERT model using the annotated data splits: $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$ (training, validation, and test sets, respectively). Before oversampling, these splits consist of 1,905 samples in $\mathcal{D}_{\text{train}}$, 587 in $\mathcal{D}_{\text{val}}$, and 440 in $\mathcal{D}_{\text{test}}$ (2,932 total); to address class imbalance we oversample $\mathcal{D}_{\text{train}}$ to 3,384 samples for training, with labels assigned based on our predefined annotation procedure.

Note that the training process for a complex language model like BERT is not deterministic. The process involved in creating a fine-tuned BERT model involves randomness. Even with global seeds, randomness comes from differences in Python libraries, GPU processing, data shuffling, and differences in initialization weights. Thus, while the outputs of the classifier may have deterministic behavior, some aspects of training introduce randomness.

*2. BERT: Hyperparameter Tuning and Training.* Before training our BERT model, we first performed hyperparameter tuning using the training and validation sets ($\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$). We then followed standard practice by training the final model on the combined $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ split using the optimal hyperparameters.

All experiments—hyperparameter tuning and final training—were conducted on an NVIDIA RTX 4090 GPU, using PyTorch version 2.5.1, and Hugging Face Transformers version 5.8.0. To ensure reproducibility, we set a global random seed across Python's random module, NumPy, and PyTorch (including CUDA); individual Optuna trials also used derived seeds. Mixed-precision training with bfloat16 was employed for computational efficiency. Data loading was optimized using multiple worker threads and pinned memory. We used the `bert-base-uncased` tokenizer throughout.

The hyperparameters varied during optimization included learning rate, batch size, number of training epochs, weight decay, and warmup steps; see Appendix Table B.1. We used the standard `AdamW` optimizer with a tuned warm-up schedule (see Table B.1) and the default hidden dropout probability optimized for BERT (0.1).

Table B.1: Hyperparameter Search Space for BERT Model Tuning

| Hyperparameter | Range / Values | Sampling Strategy / Type | Best Parameter |
|---|---|---|---|
| Learning Rate | [1e-5, 8e-5] | Log-uniform | $4.7145 \times 10^{-5}$ |
| Batch Size | {4, 8, 16, 32} | Categorical | 8 |
| Number of Train Epochs | [2, 5] | Integer (Uniform) | 4 |
| Weight Decay | [1e-4, 1.1e-1] | Log-uniform | $2.0953 \times 10^{-2}$ |
| Warmup Steps | [0, 200] | Integer (Uniform) | 114 |

Hyperparameter optimization was conducted using an implementation of the Bayesian Tree-structured Parzen Estimator (TPE) algorithm (Bergstra, Bardenet, Bengio and Kégl, 2011; Akiba, Sano, Yanase, Ohta and Koyama, 2019). The TPE sampler was configured to maximize the F1-score on the validation set $\mathcal{D}_{\text{val}}$, with early stopping implemented via a median pruner to reduce runtime.

We executed 100 trials over the hyperparameter space defined in Appendix Table B.1. Large neural networks like BERT can exhibit variance in performance across runs, even with the same hyperparameter configuration. For this reason, each trial was replicated three times ($N = 3$), and the mean F1-score was used to evaluate each configuration.

The final model was trained using the best-performing hyperparameter set, applied to the combined $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ dataset. The selected hyperparameter values for our model are presented in Appendix Table B.1.

We assess the performance of the final model on the held-out test set $\mathcal{D}_{\text{test}}$. For completeness, Appendix Table G.4 reports predictive performance for the three-class BERT model on the test set ($\mathcal{D}_{\text{test}}$), as well as on the validation ($\mathcal{D}_{\text{val}}$) and training ($\mathcal{D}_{\text{train}}$) splits. This table is an expanded version of the main table shown in the text.

### *B.2.* BERT vs. DeBERTa-v3

To assess whether our results depend on BERT's specific implementation, we replicate the classification exercise with DeBERTa-v3 (He, Gao and Chen, 2023), a more recent encoder architecture that improves on BERT's pre-training objective and attention mechanism. We fine-tune `deberta-v3-base` on the identical training and validation splits, with the same oversampling procedure and hyperparameter-tuning protocol as the BERT model (Appendix B.1), and evaluate on the same held-out test set.

Appendix Table B.2 reports test-set performance for both models side by side. The two architectures perform identically on the industrial-policy class (F1 of 0.88 for both) and differ only marginally on the remaining categories (F1 differences of at most 0.03). Overall accuracy is 92.5% for BERT and 93.0% for DeBERTa-v3. Our measure therefore does not depend on the choice of encoder within the BERT family.

Table B.2: Test-Set Performance: Fine-Tuned BERT vs. DeBERTa-v3

| | BERT (`bert-base-uncased`) | | | DeBERTa-v3 (`deberta-v3-base`) | | |
|---|---|---|---|---|---|---|
| Class | Precision | Recall | F1 | Precision | Recall | F1 |
| IP Goal | 0.89 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 |
| No IP Goal | 0.90 | 0.93 | 0.92 | 0.94 | 0.95 | 0.94 |
| Not Enough Information | 0.95 | 0.94 | 0.94 | 0.95 | 0.94 | 0.95 |
| Accuracy | | | 0.93 | | | 0.93 |
| Macro Avg | 0.91 | 0.92 | 0.91 | 0.92 | 0.93 | 0.92 |
| Weighted Avg | 0.93 | 0.93 | 0.93 | 0.93 | 0.93 | 0.93 |

*Notes:* Test-set performance of the fine-tuned BERT and DeBERTa-v3 classifiers on the 440 held-out test policies (104 IP Goal, 76 No IP Goal, 260 Not Enough Information). Both models are trained on identical data splits with the same oversampling and tuning protocol. The BERT column reproduces Table 3.

## *B.3.* Logistic Classifier Benchmark

We use a three-class logistic regression classifier, based on TF–IDF vectorization, as a transparent benchmark throughout the paper. Specifically, we report results using the logistic model with $L_1$ regularization. This benchmark model uses unigrams and bigrams as input features and applies Lasso ($L_1$) regularization with a strength of $C = 0.09$ (equivalently, a regularization strength of $\lambda = 11.1$).

Given our setting and large vocabulary, the "more sparse" $L_1$-penalized model outperforms logistic regressions using Ridge ($L_2$), and performs slightly better than Elastic Net. We describe the full pipeline, model selection process, and resulting performance below.

*1. Logistic: Text Processing, Tokenization, and Vectorization.* Text is processed and represented differently in our baseline logistic classifier compared to the BERT-based model. In the logistic pipeline, each policy description is converted into a numerical array using standard Term Frequency–Inverse Document Frequency (TF–IDF) representations, incorporating both unigrams and bigrams.

Text is first tokenized and cleaned using the Python-based NLP library spaCy (version 3.8.3).[1] The preprocessing pipeline follows standard conventions to reduce noise, normalize linguistic structure, and enhance classifier performance. This involves several steps: tokens identified by spaCy as punctuation or number-like are removed; the remaining tokens are lemmatized (i.e., converted to their base morphological form), transformed to lowercase, and filtered through a stopword list. This stop list combines spaCy's default English stopwords with domain-specific terms, including units, to remove non-informative content. Where possible, we use spaCy's ready-made libraries

1. See Honnibal, Montani, Van Landeghem and Boyd (2020).

for our cleaning and filtering (e.g., stop word lists and lemmatizers). Tokens that become empty during preprocessing are discarded, and the pipeline delivers a cleaned and lemmatized sequence for vectorization.

Processed text is then vectorized using a TF–IDF vectorizer from the `Scikit-learn` library. The vectorizer is configured with max = 95% to exclude overly common terms and min = 2 (minimum 2 observations) to remove rare ones. Sublinear term frequency scaling is applied, along with the default $L_2$ normalization.

Following vectorization, a standard scaler is applied to further normalize the feature representation for logistic classification.[2] This step is particularly important for regularized models such as Lasso and Elastic Net (with Saga solvers), which are sensitive to the magnitude of input features. Scaling ensures that all features contribute proportionately to the regularization penalty, preventing features with larger inherent values from disproportionately influencing model coefficients. This results in more stable performance and more interpretable model outputs.

*2. Logistic: Hyperparameter Tuning, Training, and Model Selection.* We select the tuned logistic model—Ridge ($L_2$), Lasso ($L_1$), and Elastic Net—which performs the best out of sample. Below, we describe the hyperparameter tuning process, the training, and the out-of-sample performance of each.

Table B.3: Logistic Variants and Hyperparameter Tuning

| Logistic Model | Hyperparameter Space | Solver | Scaled Text | Best F1 | Best Hyperparameter |
|---|---|---|---|---|---|
| Ridge ($L_2$) | $C \in [.25, 2]$ | lbfgs | No | 0.8403 | $C = 0.5$ |
| Ridge ($L_2$) Scaled | $C \in [.05, 1]$ | saga | Yes | 0.8231 | $C = 0.25$ |
| Lasso ($L_1$) | $C \in [.075, 2]$ | saga | Yes | 0.8513 | $C = 0.09$ |
| Elastic Net | $C \in [.05, 1]$ | saga | Yes | 0.8536 | $C = 0.075$ |
| | $L_1$ ratio $\in [.5, .99]$ | | | | $\frac{L_1}{L_2} = 0.95$ |

*Notes:* All F1-scores are macro-averaged and computed as the mean across all $k$-folds of the validation set. Scaling refers to the application of a standard scaler to text features after TF–IDF vectorization; means are not subtracted to preserve sparsity in the TF–IDF representation. The solver indicates the optimization algorithm used to estimate the logistic regression parameters: `lbfgs` (Limited-memory Broyden–Fletcher–Goldfarb–Shanno), a quasi-Newton method effective for $L_2$ regularization; and `saga`, a stochastic optimization algorithm suitable for $L_1$ and Elastic Net penalties.

We use a standard grid search algorithm to optimize the hyperparameters of our logistic regression models implemented using Scikit-learn's `GridSearchCV`. For each variant, we select the logistic classifier that yields the highest macro-averaged F1-score—average F1 for validation sets across folds—during the tuning process.

2. The `Scikit-learn` standard scaler scales features to unit variance by dividing by their standard deviation. To preserve sparsity in the TF–IDF representation, the mean is not subtracted.

We began by identifying key hyperparameters for each model, such as the regularization strength $C$ and the L1 ratio (for Elastic Net), and defined a discrete set of candidate values $\Lambda_i$ for each. The resulting grid $\mathcal{G}$ comprised all possible combinations of these values.[3]

For every hyperparameter combination $h \in \mathcal{G}$, we trained a complete logistic model pipeline. This pipeline included text preprocessing, TF–IDF vectorization to produce features $X$, optional feature scaling, and logistic regression using the classifier $g_h(X'; w)$ parameterized by $h$.

Each candidate model $M_h$ was evaluated using a separate, fixed validation set $\mathcal{D}_{\text{val}}$, while the training set $\mathcal{D}_{\text{train}}$ was used solely to estimate model parameters $w$. We built on standard $k$-fold cross-validation by applying a predefined split: each $M_h$ was trained on the full $\mathcal{D}_{\text{train}}$ to estimate $\hat{w}_h$, and its performance was evaluated using the macro-averaged F1-score, denoted $\text{F1}_{\text{macro}}(M_h(\mathcal{D}_{\text{val}}))$. Internally, `GridSearchCV` applied a three-fold split of $\mathcal{D}_{\text{val}}$ during this evaluation.

We then selected the hyperparameter configuration $h^*$ that maximized validation performance.[4] This grid search and model selection procedure was repeated independently for each regularized logistic regression variant. The resulting best model $M_{h^*}$, trained using the optimal configuration $h^*$, was retained for final evaluation.

Finally, we chose among the tuned variants by their performance on the test set $\mathcal{D}_{\text{test}}$, reported in Appendix Table B.4; the figures in that table are therefore post-selection test performance rather than an untouched final evaluation. Specifically, each optimized variant from the grid search (Appendix Table B.3) was retrained on the combined $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ split and evaluated on the held-out test set $\mathcal{D}_{\text{test}}$. Appendix Table B.4 reports the out-of-sample performance and shows that the Lasso model performs best. The tuned Lasso model with $C = 0.09$ outperforms both Ridge and Elastic Net, although the margins are small: Elastic Net performed best on the validation folds (Appendix Table B.3) and trails Lasso by 0.2 percentage points of test accuracy.

3. The Cartesian product $\Lambda_1 \times \cdots \times \Lambda_k$.
4. Formally, the optimization problem:

$$h^* = \arg\max_{h \in \mathcal{G}} \text{F1}_{\text{macro}}\left(M_h(\mathcal{D}_{\text{val}})\right).$$

Table B.4: Predictive Performance of Different Logistic Models

| Logistic Model | Class | Precision | Recall | F1-score |
|---|---|---|---|---|
| Ridge | Industrial Policy | 0.820 | 0.875 | 0.847 |
| | No Industrial Policy | 0.889 | 0.842 | 0.865 |
| | Not Enough Information | 0.922 | 0.912 | 0.917 |
| | Accuracy | | | 0.891 |
| | Macro Avg | 0.877 | 0.876 | 0.876 |
| | Model Average | 0.892 | 0.891 | 0.891 |
| Lasso | Industrial Policy | 0.867 | 0.875 | 0.871 |
| | No Industrial Policy | 0.971 | 0.882 | 0.924 |
| | Not Enough Information | 0.921 | 0.942 | 0.932 |
| | Accuracy | | | 0.916 |
| | Macro Avg | 0.920 | 0.900 | 0.909 |
| | Model Average | 0.917 | 0.916 | 0.916 |
| Elastic Net | Industrial Policy | 0.858 | 0.875 | 0.867 |
| | No Industrial Policy | 0.971 | 0.882 | 0.924 |
| | Not Enough Information | 0.921 | 0.938 | 0.930 |
| | Accuracy | | | 0.914 |
| | Macro Avg | 0.917 | 0.898 | 0.907 |
| | Model Average | 0.915 | 0.914 | 0.914 |

*Notes:* This table presents the predictive performance of the logistic models on the test data split. Each model is trained using optimal hyperparameters on the train/validation splits. For each model, we detail standard classification metrics—Precision, Recall, and F1-Score—for each of the three classes. Accuracy is given for the model's predictions across all classes. The classes are 'IP Goal', 'No IP Goal', and 'Not Enough Information'. Additionally, the table shows the Macro Average and Weighted Average for these metrics across classes.

| Component | IP intention | Not IP | NEI | Total |
|---|---|---|---|---|
| Held-out test split | 104 | 76 | 260 | 440 |
| Human-labeled annotations | 108 | 90 | 318 | 516 |
| **Total** | **212** | **166** | **578** | **956** |

Table C.1: Distribution of labeled policy text in Mechanistic Interpretability Set $\mathcal{M}$

## C. BERT Internal Validation

This section outlines our workflow for understanding the internal representation of the "industrial policy" concept in our LLM. First, we explain in more detail our use of activation patching to localize the concept. Second we explain how we use Concept Activation Vectors (TCAV). For both workflows we use the original test dataset from our model ($N = 440$) and a subset of human annotations that were never seen by the model ($N = 516$). We refer to this combined data as our "mechanistic validation" dataset $\mathcal{M}$. Table C.1 shows the distribution by class of our data.

### *C.1.* Activation Patching and Separability

First, we validate that our model learned the concept of *industrial policy* rather than spurious patterns. We follow Wang et al. (2026) in using contrastive pairs and mean-difference direction extraction.

*Contrastive pairs.* From the non-NEI texts of $\mathcal{M}$ (212 IP, 166 Not IP; Table C.1), we randomly pair each "Not IP" text with a distinct "IP" text, without replacement,[5] yielding 166 pairs. Pairs are split 65/35 into an *extraction* set (107 pairs), used to construct directions, and a held-out *evaluation* set (59 pairs), used for all reported accuracy figures; no text appears in both splits.

*Activation patching.* For each extraction pair, we run the "Not IP" text through the model and replace its $[CLS]$ hidden state[6] one layer at a time with that of the paired "IP" text, recording the change in the predicted probability of "IP." Figure C.1 shows the effect is concentrated at layer 11 (mean $\Delta P = 0.79$); swaps at layers 0–9 leave predictions unchanged for the most part.

*Direction extraction and separability.* At the target layer, we compute the IP direction $d$ as the difference between the mean $[CLS]$ state of IP texts and of Not IP texts in the extraction set, normalized to unit norm. A text's position on the IP-ness axis is the inner product of its layer-11 $[CLS]$ state with $d$; texts with positive projections are classified

5. We include a fixed seed for reproducibility
6. The $[CLS]$ token is placed at the start of each tokenized input sequence. By the time the input passes through all transformer layers, the $[CLS]$ token's final hidden state has *attended* to every other token in the sequence, aggregating the contextual meaning of the entire input text.

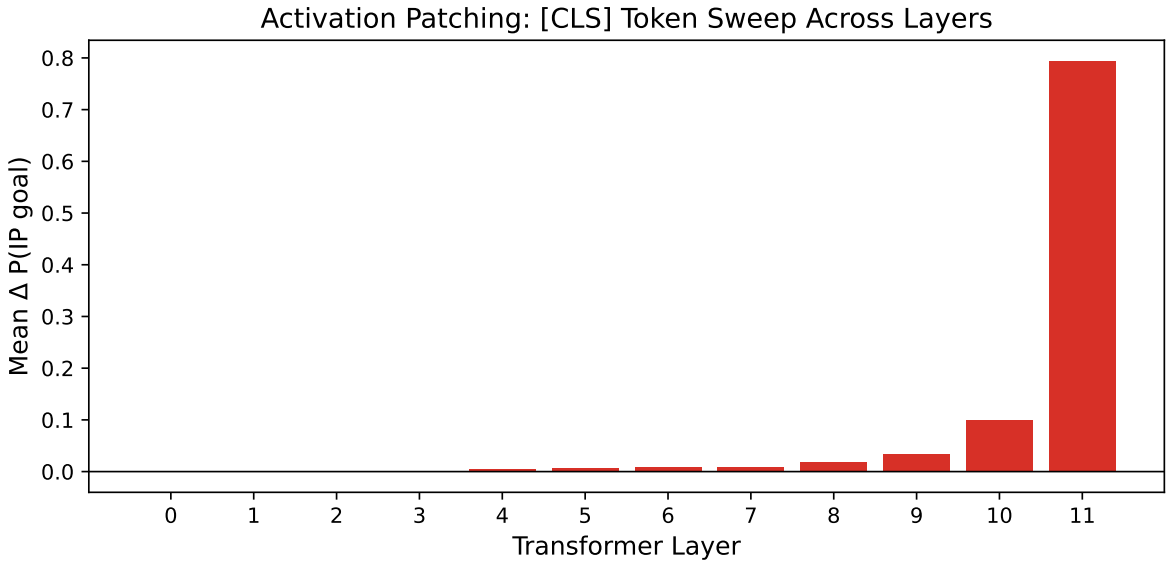



Figure C.1: Activation Patching Experiment Results

*Note:* This figures reports the activation-patching experiment: while the model processes a "Not IP" policy text, we replace its [CLS] hidden state at one layer at a time with that of a matched "IP" text and record the resulting change in the predicted probability of "IP goal," averaged over the extraction pairs.

IP. On the held-out evaluation pairs, this zero-threshold rule classifies 92.4% of texts correctly (98.3% of IP texts, 86.4% of Not IP texts). See Figure C.2.

*Pretrained comparisons.* We repeat the identical procedure on two models that never saw our annotations. For pretrained `bert-base-uncased`, the exercise is exactly the same: we extract per-layer $[CLS]$ states for the same extraction and evaluation pairs, fit the mean-difference direction on the extraction set at each layer, and evaluate zero-threshold accuracy on the evaluation set. Separability reaches 84.7% at its final layer. For pretrained Llama 3.1 8B, one adjustment is required: Llama is a decoder-only model with causal attention and has no $[CLS]$ token, so we use the hidden state at the *last* token position of each text — the only position that, under causal attention, has attended to the entire input, making it the analogue of BERT's $[CLS]$ summary. We extract these last-token states at each of Llama's 32 transformer layers (truncating inputs to the same 512-token maximum) and apply the same extraction-set direction fitting and evaluation-set scoring. Separability peaks at 88.1% at intermediate depth (layer 20 of 32) and declines to 78.8% by the final layer. Because the models differ in depth (12 versus 32 layers), Figure 2 plots each model's per-layer accuracy against relative depth (layer index divided by total layers); where a single Llama layer must be compared to a specific BERT layer, we use the layer at matching relative depth. Figure C.2 summarizes separability for the three models: the pretrained models at their peak layer, and the

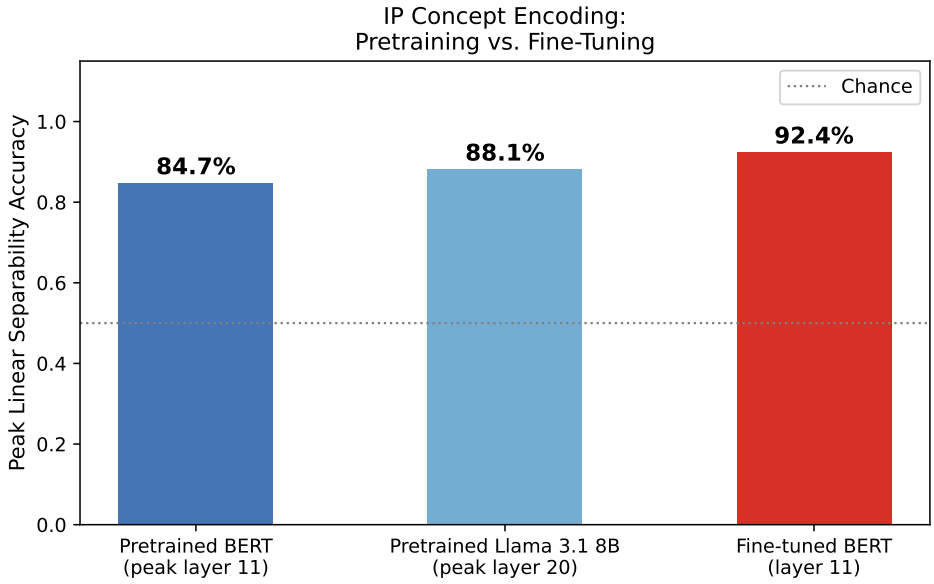



Figure C.2: IP Concept Encoding: Pre-trained models (peak layer) vs. Fine-Tuning (final layer)

*Notes:* Bars for the two pretrained models report their peak linear separability accuracy across layers (Pretrained BERT: layer 11; Pretrained Llama 3.1 8B: layer 20). The bar for fine-tuned BERT reports its final layer (layer 11), which is the layer used in the main analysis; its across-layer maximum is 93.2% at layer 10, see Appendix C.6 and Figure C.4.

fine-tuned model at its final layer, the layer used throughout the main analysis. The fine-tuned model's across-layer maximum is 93.2% at layer 10 (Appendix C.6).

## C.2. TCAV (Concept Activation Vectors)

This section describes our application of Testing with Concept Activation Vectors (Kim et al., 2018), which quantifies the sensitivity of the model's predictions to a human-defined concept, in our case texts that contains explicit industrial-policy intention.

**Concept, contrast, and evaluation sets.** From the labeled texts of $\mathcal{M}$, we define the concept set as texts human-labeled "IP intention" and the contrast set as texts labeled "Not IP" (ties among annotators are dropped); 20% of texts per class are held out as the evaluation set on which sensitivities are measured. As a generic baseline we draw a pool of 8,910 English Wikipedia sentences (`sentence-transformers/wikipedia-en-sentences`; sentences of 40–400 characters and at least seven words), used to form 50 non-overlapping random groups of 162 sentences each.[7] We run two analyses: *concept vs. random*, which contrasts the IP-intention texts with each of the 50 Wikipedia groups, and *concept vs. contrast*, which contrasts 50 independently resampled, equal-sized pairs of subsets, each containing 50 IP-intention texts and 50 Other-Intention texts. Sampling is without replacement within each subset, although texts may reappear across different runs.

7. Which results in using 8100 out of 8910 sentences drawn

The second analysis asks whether the model is sensitive to industrial-policy intent specifically, holding fixed that both sides are policy language.

**Concept activation vectors.** For each of BERT's 12 transformer layers (and the pooler output), we extract $[CLS]$-pooled activations for the concept and negative examples, reduce dimensionality by PCA (retaining enough principal components to explain 50% of activation variance), and fit a regularized linear classifier ($C = 10^{-5}$, assessed by 6-fold cross-validation) separating concept from negative activations. The CAV $v_C^\ell$ is the vector orthogonal to the separating hyperplane, the direction of the concept in the model's activation space at layer $\ell$. Fitted CAVs separate the two sides with mean accuracies of 0.92-0.94 at the final layers.

**TCAV scores.** Each output class is evaluated on a common held-out evaluation set $\mathcal{X}_{\text{eval}}$, containing texts from all three annotation categories. A common evaluation set allows us to compare how the same movement toward IP intention affects the IP, No-IP, and NEI outputs, as opposed to a class-conditional TCAV. For concept $C$, output class $k$, and layer $\ell$, the TCAV score is the fraction of evaluation texts that responds positively to local movement along the learned concept direction, increasing the class-k logit:

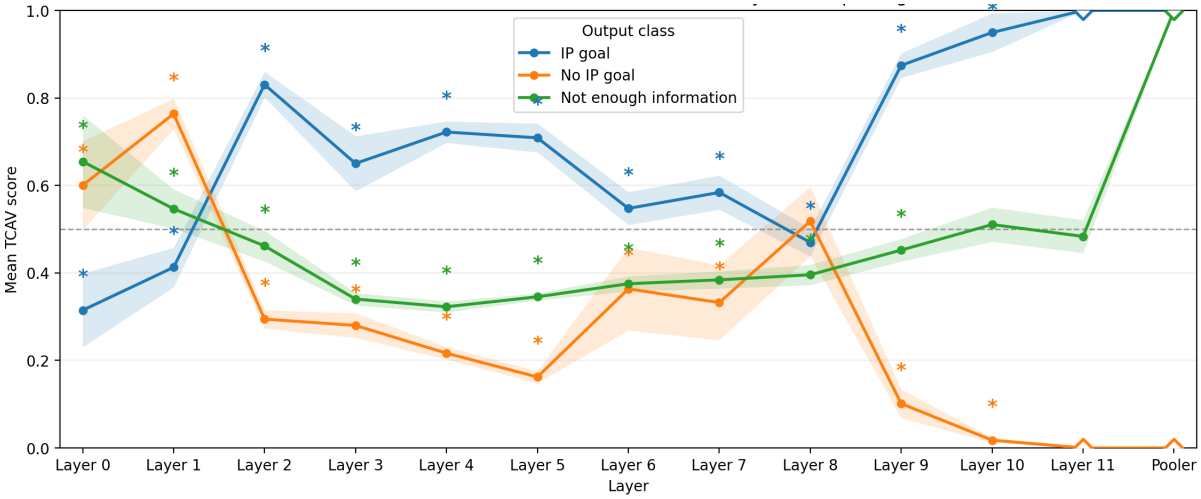
$$\text{TCAV}_{C,k,\ell} = \frac{1}{|\mathcal{X}_{\text{eval}}|} \sum_{x \in \mathcal{X}_{\text{eval}}} \mathbf{1}\left[\nabla h_k(f_\ell(x)) \cdot v_C^\ell > 0\right].$$

where $h_k$ maps layer-$\ell$ activations to the class-$k$ logit. A score of 0.5 indicates no systematic directional tendency under the sign-count metric: equal shares of evaluation texts have positive and non-positive directional derivatives. Because each analysis is repeated over 50 negative (or resampled) groups, we obtain a distribution of 50 scores per layer, class, and analysis; we test each distribution against 0.5 and apply a Bonferroni correction separately within each analysis across the non-degenerate layer-by-output tests ($\alpha = 0.05$).

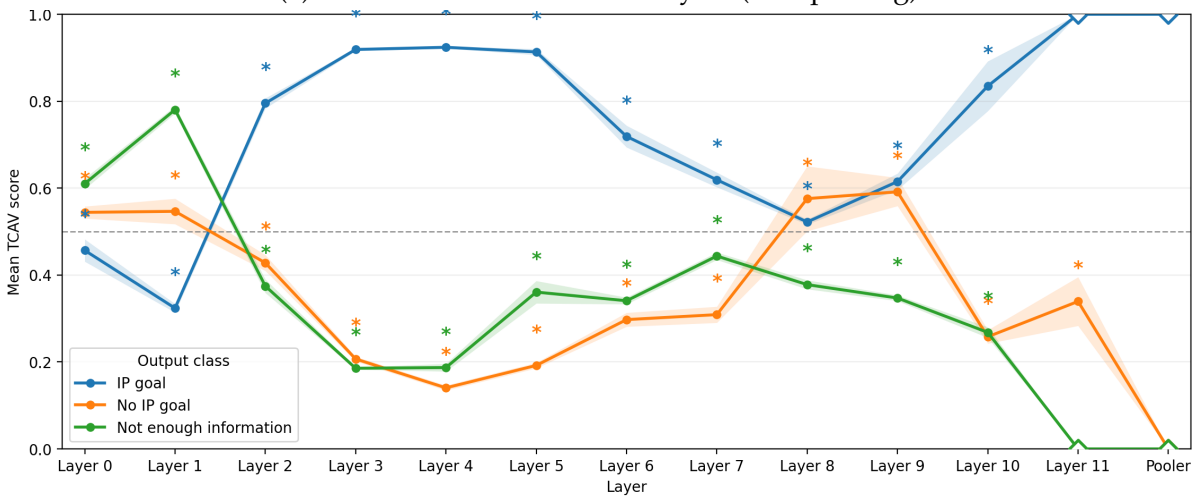
**Implementation.** The analysis uses the frozen fine-tuned classifier (weights not updated), PyTorch and Hugging Face Transformers, a fixed seed (42), maximum sequence length 512, and was run on an NVIDIA RTX 4090.

**Results.** Figure C.3 shows the results of TCAV across layers. At the final layer, Panel a shows that for every evaluation text, the local directional derivative is positive for the IP logit and non-positive for the No-IP logit — TCAV scores of 1.00 and 0.00 in all 50 comparisons, leaving no sampling variance to test. The score for the NEI class is 0.48, statistically indistinguishable from the neutral value of 0.5 (Bonferroni-corrected $p = 0.11$): the concept moves predictions between the two substantive classes and is uninformative for the insufficient-information class. One layer earlier, scores are 0.95 (IP) and 0.02 (No-IP), both significant (Bonferroni-corrected $p < 10^{-49}$), and sensitivity declines toward the middle of the network. The concept-vs-random analysis (Panel b) gives the same picture at the final layer (1.00 for IP in all 50 comparisons), although its

results for the No-IP and NEI outputs differ from the policy-language contrast. The directional effect on the IP output is maximally consistent at the final layer, consistent with the activation-patching results of Appendix C.1.




(a) IP vs Not IP: TCAV across layers (CLS pooling)




(b) IP vs Wikipedia random: TCAV across layers (CLS pooling)

Figure C.3: TCAV across layers

*Notes:* Lines show mean TCAV scores across repeated runs (concept-contrast runs in Panel (a), concept-random runs in Panel (b)); shaded ribbons show ±1 standard deviation across runs. The dashed line marks the neutral TCAV value of 0.5. Hollow diamonds indicate class-layer combinations with zero variation. Asterisks indicate significance after Bonferroni correction.

## *C.3.* Steering

The steering experiment of Section 5.B intervenes on the model's forward pass: while the model processes a policy text, we replace its layer-11 $[CLS]$ state $h$ with $h + \alpha\,\sigma_{11}\,d$, where $d$ is the unit-norm IP direction of Appendix C.1, $\sigma_{11} = 13.5$ is the standard deviation of the extraction texts' projections onto $d$, and $\alpha$ is the intervention

strength reported on the horizontal axis of Figure 3. Because IP and No-IP texts occupy opposite spaces of the axis, $\sigma_{11}$ is of the same order as the separation between the two classes, so $|\alpha| \approx 1$–2 corresponds to moving a text across the gap between them.

We evaluate $\alpha \in \{-6, -4, -3, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 3, 4, 6\}$ on the 59 held-out evaluation pairs. Predicted probabilities respond monotonically at every step. Texts originally labeled "No IP" cross $P(\text{IP}) = 0.5$ between $\alpha = 1$ and $\alpha = 1.5$ and saturate by $\alpha = 2$; texts originally labeled "IP" flip by $\alpha = -1.5$. At $\alpha = +3$, 94.9% of No-IP texts are classified as industrial policy; at $\alpha = -3$, 100% of IP texts are classified as not industrial policy.

### *C.4.* Projection Of NEI

We project all 578 texts of $\mathcal{M}$ labeled "Not Enough Information" onto the layer-11 IP direction; these texts played no role in constructing the direction. Their mean position is +2.9 (s.e. 0.17), between the evaluation-set class means of +12.4 (IP) and −11.3 (No-IP), and roughly four times closer to the decision boundary than either class. 93.9% of NEI texts project weakly positive. The positive offset is statistically significant ($t = 17.2$) but small relative to the axis's scale.

### *C.5.* Projection Of Russian Sanctions

We project the Russia-sanctions measures of Section 5.A onto the layer-11 IP direction. Because this is an out-of-sample probe, we restrict this exercise to the 337 of 374 measures announced on or after February 24, 2022, which postdate the annotation codebook and all training data. 94.4% project onto the No-IP side of the boundary (mean −12.7, s.d. 4.9) — that is, sanctions land where the geometry places clear non-industrial policies, not merely on the non-IP side of the softmax output. The axis and the model's classification agree perfectly on this set: among the 321 sanctions the model classifies as IP or No-IP (7 and 314, respectively; 16 are classified NEI), the sign of the projection matches the predicted class in 100% of cases.

### *C.6.* Interventions on Layer 10

Interventions at layer 11 face a mechanical objection: nothing downstream can test them. This section repeats the full workflow at layer 10, one layer upstream of the classification head, extracting a layer-10 direction from the extraction pairs by the same procedure as Appendix C.1.

**Separability.** The layer-10 direction classifies the held-out evaluation texts with 93.2% accuracy, against 78.8% for pretrained BERT at its layer 10 and 66.9% for pretrained Llama 3.1 8B at matched relative depth (layer 28 of 32); see Appendix Figure C.4. The fine-tuned model's advantage over the pretrained models is larger at this layer than at the final one.

**Steering.** Steering along the layer-10 direction (strength measured in standard deviations of the layer-10 projection distribution, as in Appendix C.3) produces the same monotonic response as at layer 11, with complete label reversals at larger strengths: No-IP texts cross $P(\text{IP}) = 0.5$ between $\alpha = 3$ and $\alpha = 4$, against 1–1.5 at layer 11 (Appendix Figure C.5). Note that an intervention in layer-10 must survive an entire transformer layer before reaching the classification head, and the final layer re-aggregates information from the surrounding (unmodified) tokens, which dilutes the intervention, however, the response curve is shifted, not absent.

We must also reconcile the patching and steering results. A patch is a fixed intervention: it displaces the $[CLS]$ state by the natural distance between a matched pair of texts, on the order of one standard deviation along the axis. An intervention of this size exceeds layer 11's flip threshold but falls below layer 10's, which is why the patching sweep of Appendix Figure C.1 concentrates at the final layer even though a causal pathway exists from layer 10 onward. Consistent with dilution rather than absence, a $[CLS]$ patch at layers 0–9 recovers essentially none of the clean run's layer-11 projection (at most 6%), a patch at layer 10 recovers 24%, and only the layer-11 patch recovers it fully (Appendix Figure C.6). This pattern is exactly what BERT's architecture would lead us to expect. Through the middle of the network, the information relevant to a document-level judgment is distributed across the token positions, and each layer's attention rebuilds the $[CLS]$ state from those surrounding tokens, so overwriting the $[CLS]$ state alone at an early layer changes little, because the next layer simply re-aggregates from the unmodified tokens. Only in the final layers the $[CLS]$ state itself becomes the carrier of the industrial-policy signal, which is where interventions on it are able to stick.

**Sanctions.** Projecting the 337 post-invasion sanctions onto the layer-10 direction places 94.7% on the No-IP side (Appendix Figure C.7), nearly identical to the layer-11 result of Appendix C.5.

In conclusion, the layer-10 representation passes the same tests as layer-11. Our results reflect the network's learned representation of industrial policy, not proximity to the classification head.

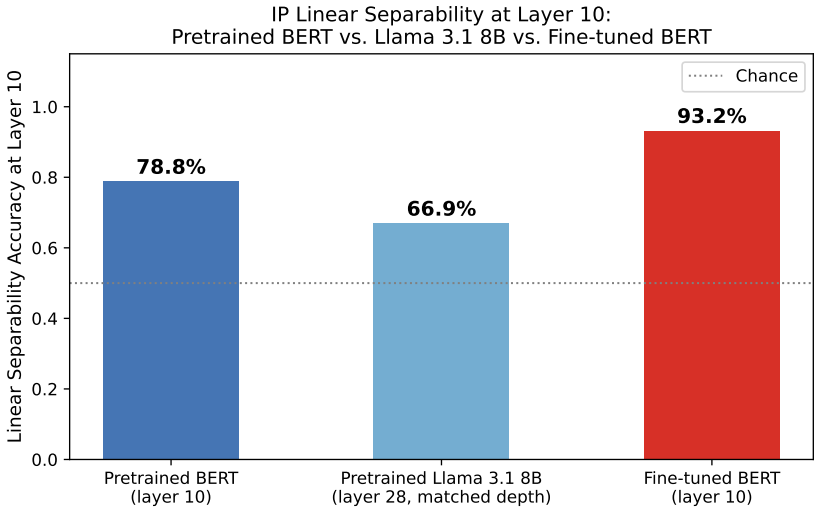



Figure C.4: Linear separability at layer 10: pretrained BERT vs. Llama 3.1 8B vs. fine-tuned BERT

*Notes:* Zero-threshold projection accuracy on the held-out evaluation pairs using each model's layer-10 mean-difference direction (Llama 3.1 8B evaluated at layer 28 of 32, the matching relative depth). Directions are fit on the extraction pairs of the respective model; neither pretrained model saw our annotations. The dotted line marks chance (50%).

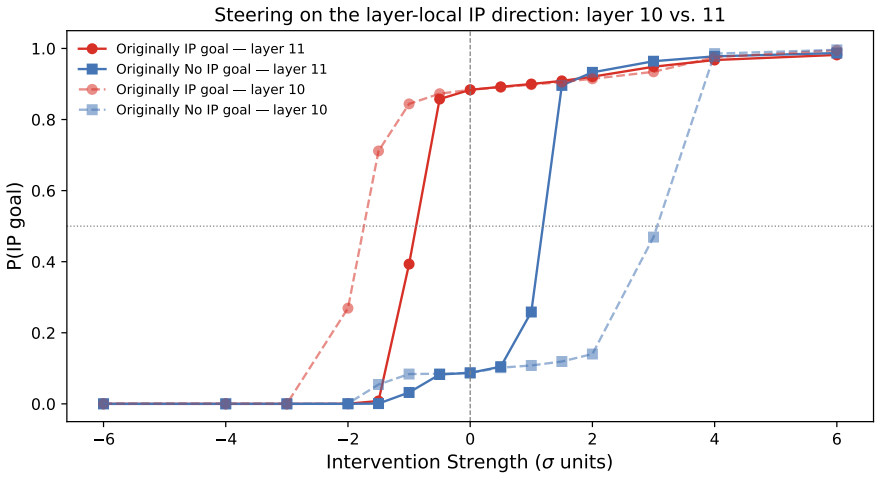



Figure C.5: Steering on the layer-local IP direction: layer 10 vs. layer 11

*Notes:* This figure repeats the steering experiment of Figure 3 at layer 10 (dashed lines), using the layer-10 local IP direction, overlaid on the layer-11 results (solid lines). Intervention strength $\alpha$ is measured in standard deviations of each layer's own natural projection distribution, making the two layers directly comparable. Lines show the mean predicted probability of "IP goal" over the 59 held-out evaluation pairs, by original label. Layer-10 interventions must propagate through a full subsequent transformer layer; they produce the same monotonic dose-response, with complete flips at roughly twice the intervention strength.

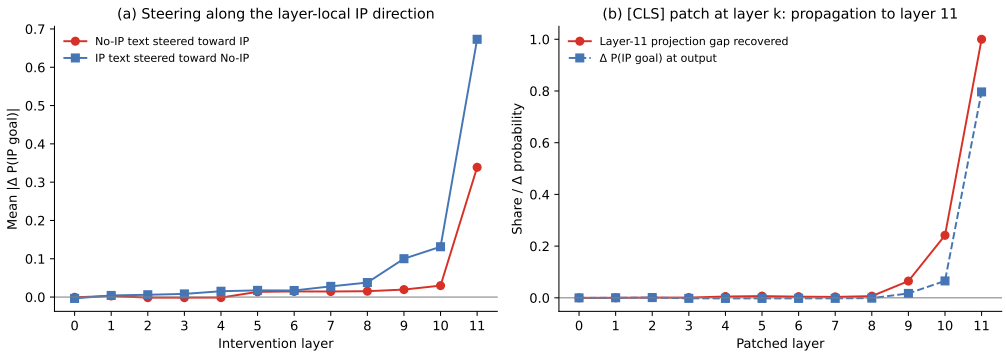



Figure C.6: Layer-wise causal diagnostics: steering by layer and patch propagation

*Notes:* Panel (a) steers the $[CLS]$ state at each layer along that layer's IP direction and reports the mean change in $P$(IP goal) on the evaluation pairs. Panel (b) patches the $[CLS]$ state at layer $k$ of a No-IP forward pass with the state of a matched IP text and reports the share of the clean-versus-corrupt gap in the layer-11 projection recovered by the patch (red) and the change in the output probability (blue). Early-layer patches recover none of the layer-11 projection: they are not overridden by later processing — the industrial-policy signal is simply not yet carried by the $[CLS]$ token, which the model populates from surrounding tokens in its final two layers.

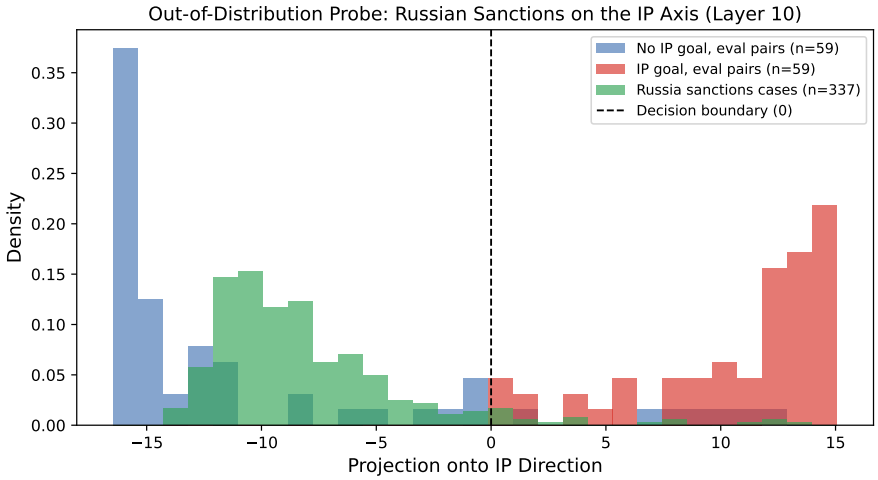



Figure C.7: Projection of the Russian sanctions on the layer-10 IP direction

*Notes:* This figure repeats the sanctions projection of Figure 5 using the layer-10 IP direction. Of the 337 post-invasion sanctions, 94.7% project onto the No-IP side of the decision boundary, indicating that the separation of sanctions from industrial policy is present in the model's geometry before the final readout layer.

## D. Validating the GTA using OECD data on Export Restrictions on Industrial Raw Materials

This section contains a detailed description of the GTA data validation exercise referenced in Section 3.[8] We first describe the OECD dataset. Next, we explain our hand-matching protocol. Then, we discuss the findings.

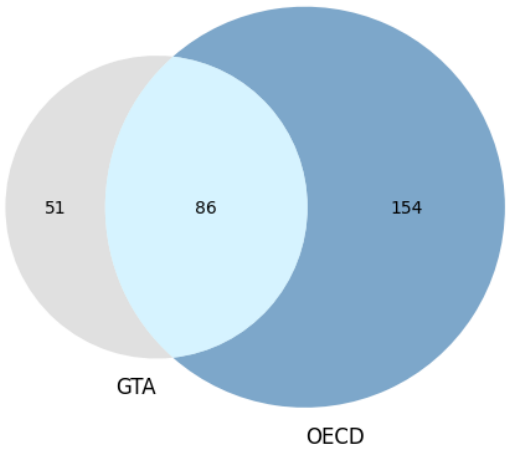



Figure D.1: Venn-Diagram of GTA and OECD Policies

*Notes:* Each unit in the Venn diagram is a unique piece of legislation or a policy document that is in the domain of the OECD and that is likely in the domain of the GTA. See Appendix D for more information.

The OECD's inventory lists export controls from 2009-2021 on 65 industrial primary commodities across metals, minerals, and wood. Policies are verified using official government sources (OECD, 2024). The OECD covers the 80 countries that are significant producers of any of these products. For each country, coverage is limited to the subset of products for which that country is a significant producer.

To understand the quality of the Global Trade Alert in terms of its ability to enumerate relevant policies, we hand-match policies across the two data sources. To do so, we need to transform and filter the data to render them comparable. This process involves a few steps. First, the OECD reports the stock of policies annually, whereas the GTA enumerates only new policies–i.e., it reports the flow. We thus transform the OECD data to also be defined in terms of annual flows.

8. We are grateful to our research assistant, Lottie Field, for her extensive work in constructing a comparable version of the OECD and GTA datasets and meticulously hand-matching them over many months.

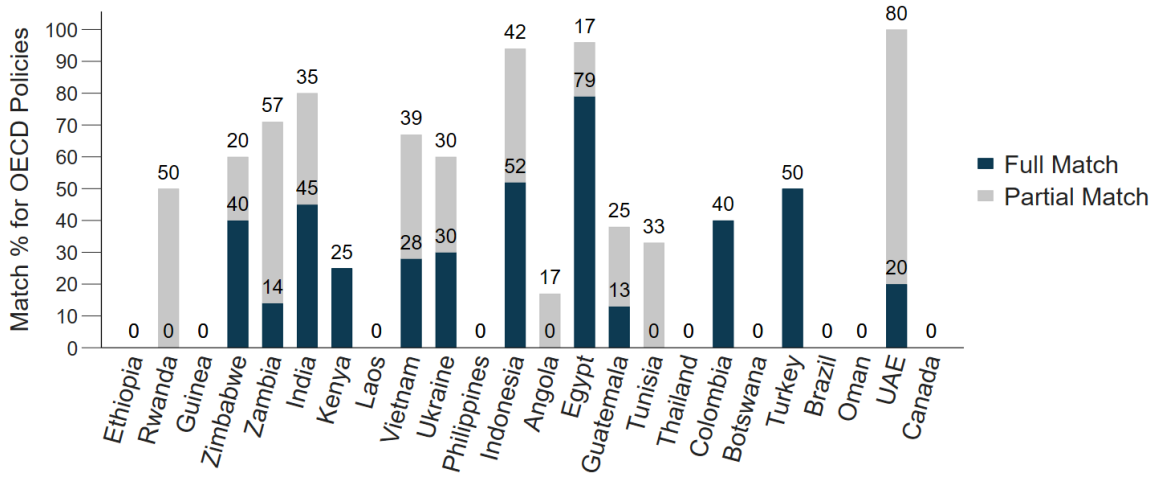



Figure D.2: Percentage of OECD Policies with GTA Matches by Country

*Notes:* Countries ordered from lowest (Ethiopia) to highest (Canada) 2010 GDP per capita. An OECD policy has a full match with a GTA policy if we can pinpoint the same policy document in the GTA. An OECD policy has a partial match with a GTA policy that uses the same 1) policy instrument, 2) affects the same industrial primary commodities, and 3) is announced within one year of the OECD policy being introduced. We provide the match rate out of all OECD policies in the GTA domain for each country. See Appendix D for more information.

Second, we need to filter both data sources for only those that fall under both organizations' remit. We count GTA policies as in the OECD domain if they 1) are in place at any point between 2009 and 2021, 2) have a listed "Measure Type" of "Export ban," "Export licensing requirement," "Export quota," or "Export tax," 3) affect products that the OECD covers for that country, and 4) are backed up by an official government source provided by the GTA. This provides a lower bound of GTA policies in the OECD domain as we do not include GTA measure types that may span both export controls (within the OECD remit) and other export policies (not in the OECD remit), such as "Export-related non-tariff measure."

We mark an OECD policy as in the GTA domain if it is 1) introduced after November 2008, 2) implemented unilaterally, and 3) there is some change, no matter how small, from the previous policy.[9] Our estimate of OECD policies in the GTA domain is an upper bound. The key reason is that we include policies with only minimal changes from previous policies. These policies are unlikely to meaningfully affect global trade

9. We also exclude OECD policies which we cannot identify using the information provided by the OECD. For example, for Mexico the only details on individual policies that the OECD provides are the date of introduction, type of policy instrument and affected sectors. This information was not sufficient to identify specific policies.

flows, so we likely include many OECD policies that do not meet the GTA's reporting thresholds.[10]

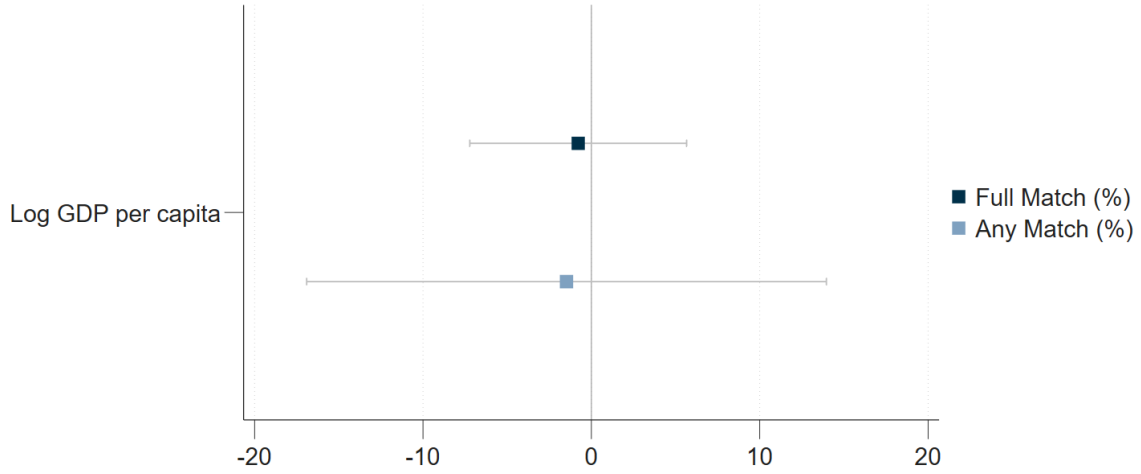



Figure D.3: Regression of Full and Any Match (%) on Log GDP per Capita

*Notes:* We regress the match rate (full or partial) on log GDP per capita and plot the coefficient of interest and its 95% confidence interval. The match rate refers to the share of policies in the OECD data that can be matched to a policy in the GTA. See Appendix D for more information.

We define two types of matched policies. An OECD policy has a full GTA match if we can pinpoint the same policy document for both entries (e.g., for Zambia, both the OECD and GTA list Statutory Instrument No. 40 of 2020 suspending the export duty on precious metals). A GTA policy partially matches an OECD policy if it 1) uses the same policy instrument, 2) affects the same industrial primary commodities, and 3) is announced within one year of the OECD policy being introduced. We think of GTA policies that partially match an OECD policy as being in the same "policy series." Generally, these are policies that precede, replace or amend the OECD policy. A partial match indicates that the GTA covers policies in the same area, even if it does not capture the exact policy.[11]

Given the large size of the data, we hand-matched the policies to a random subsample of countries stratified by income. The countries are as follows: Ethiopia, Rwanda, Guinea, Zimbabwe, Zambia, India, Kenya, Laos, Vietnam, Ukraine, Philippines, Indonesia,

10. For example, the OECD records when Brazil reduced its export quota on Lithium oxide from 50 to 10 metric tons in 2016. According to the Observatory of Economic Complexity (2025), Brazil is a negligible exporter in this category, so this small change in the export quota is unlikely to affect trade flows, and as such, will not be enumerated by the GTA.

11. For example, for Indonesia, the OECD lists a January 2009 Ministry of Trade regulation mandating domestic letters of credit to export certain goods. This partially matches with the GTA listing of the Ministry of Trade's March 2009 policy, which updates the earlier policy by, *inter alia*, specifying that the requirement only applies to exports worth over 1 million US dollars.

Angola, Egypt, Guatemala, Tunisia, Thailand, Colombia, Botswana, Turkey, Brazil, Oman, UAE, Canada.[12]

Appendix Figure D.1 shows that, in general, there is a fair amount of overlap across the two datasets. Of the policies identified in the OECD dataset, 36% have an exact match in the GTA, and 65% have a partial match. This is a conservative lower bound mainly for the reason (noted above) that many policies enumerated by the OECD are minor policy changes that do not satisfy the GTA's criteria of affecting global trade flows in a meaningful way. Similarly, 63% of the relevant policies in the GTA can be exactly matched to an OECD policy. That is, despite the much narrower focus of the OECD data collection effort, a meaningful share of relevant policies are not identified by the OECD, but *are* identified by the GTA.

Appendix Figure D.2 reports these match rates country by country. They vary considerably: several countries in the subsample have no matched policies at all, while for others nearly every OECD policy has at least a partial match in the GTA. This dispersion is not systematically related to income — regressing the match rate on log GDP per capita yields a coefficient indistinguishable from zero (Appendix Figure D.3) — so the GTA's coverage of this policy domain is no better for rich countries than for poor ones.

12. Our initial sample also included Mongolia, Peru, Mexico, Australia, the U.S. and Qatar. We were unable to perform hand-matching for these countries because none of the policies listed by the OECD were in the domain of the GTA.

## E. Agencies Implementing Industrial Policy

We extract data on "implementing agencies" — the organizations that administer and deploy industrial policy — from Global Trade Alert policy descriptions. We follow the entity-extraction method of Field (2024), as documented in Juhász and Lane (2024), applying it to the full GTA corpus.

We extract implementing agencies using OpenAI's GPT-3.5 (GPT-3.5-turbo-0613), fine-tuned for this task through OpenAI's API. The workflow is implemented in Python and was executed in November 2023. The workflow is as follows. We first draw training and test samples from the GTA data and construct a custom prompt specifying how agencies should be extracted and in what format; paired with their expected outputs, these examples form the labeled training set. We submit this set to OpenAI's fine-tuning endpoint, evaluate the resulting model on the held-out test sample, and then deploy it across the full corpus. The extracted agencies are then processed, cleaned, and validated by hand.

A proprietary model is appropriate for this step. Agency extraction is a named-entity task requiring accuracy rather than interpretability, and we validate its output manually. The construct-validity concerns that motivate our use of an inspectable model for the main classification task (Section 5.B) do not arise for this mechanical extraction.

We clean the extracted agency names in two steps. First, we standardize names to ensure consistency across spellings and forms. For example, the Italian development bank "Cassa Depositi e Prestiti" may appear under its full name, its abbreviation "CDP," or as "Italian National Development Bank." Second, we remove non-public entities, such as private companies — classified by shareholder composition or company history, following Field (2024) — along with regional and supranational organizations.

# F. Appendix Figures

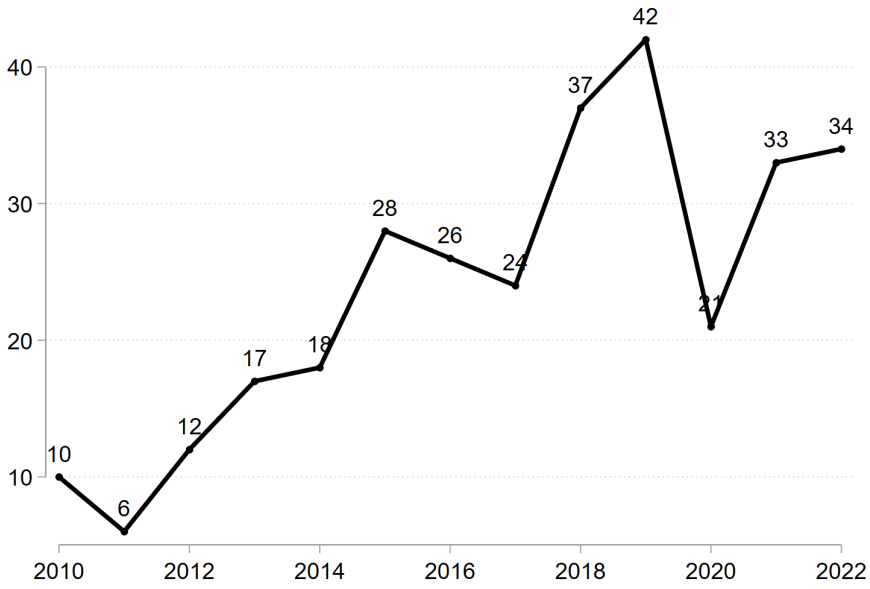



Figure F.1: Share of Industrial Policies (%) over Time

*Notes:* We calculate the share of industrial policies (%) out of all policies including those with and without identifiable goals. We follow GTA guidance and use only policies recorded by the GTA in the same year that they were announced for this exercise. This is due to the substantial back-filling of data in the GTA, which is a living dataset. By using only policies recorded in the same years as they were announced, we ensure the comparability of data across both more distant and recent years.

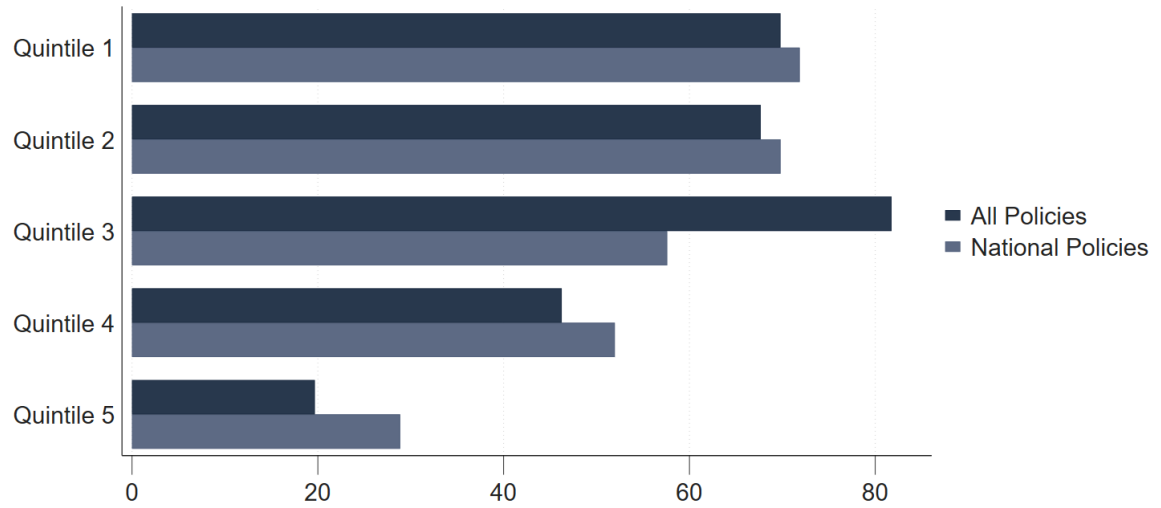



Figure F.2: Percentage of Policies Classified as "Not Enough Information" by BERT Three Class Model

*Notes:* This figure plots the share of policies classified as "Not Enough Information" by the BERT three-class model. "All Policies" is the baseline measure which enumerates all policies, "National policies" excludes those targeted at specific firms.

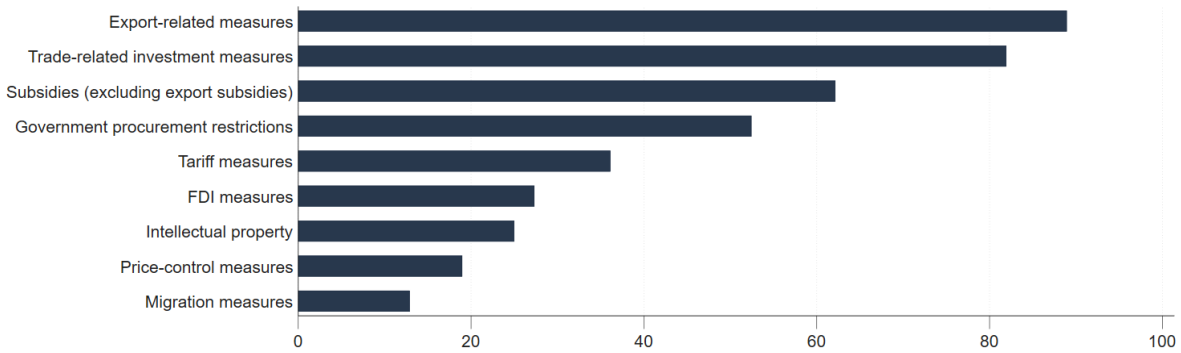



Figure F.3: Share of Policies Classified as Industrial Policy By Policy Instrument

*Notes:* This figure plots policies classified as industrial policy as a share of all policies with identifiable goals (i.e., excluding the not enough information class). Policy instruments defined based on UNCTAD's MAST chapter codes. We have added an additional category for import tariffs ("Tariff Measures").

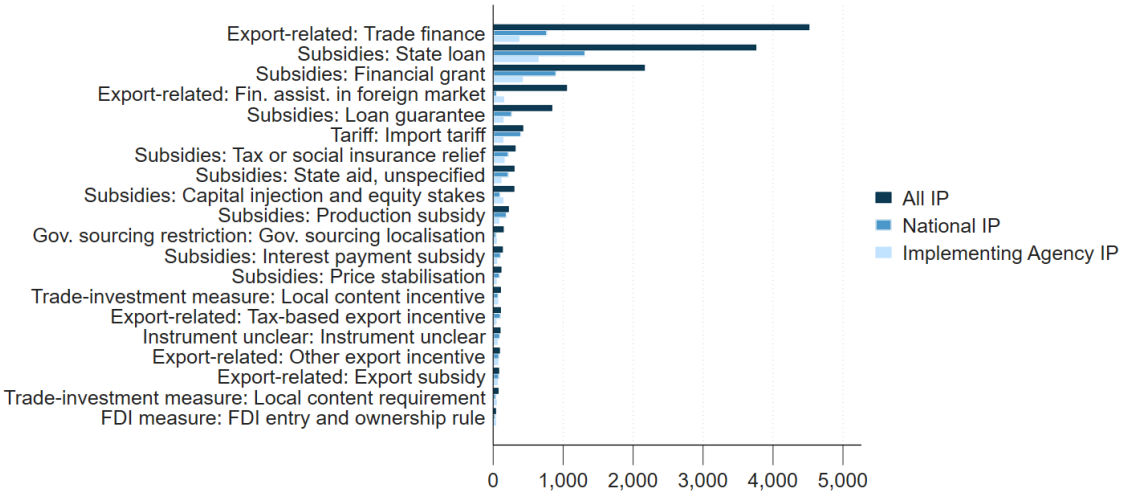



Figure F.4: The Instruments of Industrial Policy using GTA's In-House Classification System

*Notes:* We present the twenty policy instruments that have the highest average usage ranking across the three measures of IP activity. The policy instruments excluded from the graph are: export bans, FDI financial incentives, export taxes, in-kind grants, local value-added incentives, state aid (nes), import tariff quotas, internal taxation of imports, import bans, import-related non-tariff measures (nes), export licensing requirements, labour market access restrictions, export quotas, public procurement preference margins, local labour requirements, import licensing requirements, import incentives, export-related non-tariff measures (nes), FDI treatment and operations (nes), local operations requirements, controls on commercial transactions and investment instruments, public procurement access restrictions, public procurement measures (nes), import quotas, controls on credit operations, local supply requirements for exports, export tariff quotas, post-migration treatment policies, intellectual property protection measures, local operations incentives, local value-added requirements, localisation measures (nes), trade payment measures, anti-dumping measures, and repatriation and surrender requirements. For Implementing Agency IP, we calculate the number of agencies implementing at least one industrial policy via each policy instrument in each year, and then sum these counts across years.

# G. Appendix Tables

Table G.1: Implementation Level for Policies

| | |
|---|---|
| National | 33011 |
| National Financial Institution | 8764 |
| International Financial Institution | 3261 |
| Supranational | 2247 |
| Subnational | 1371 |
| Total | 48654 |

*Notes:* Distribution of the implementation level of the policies in the Global Trade Alert Database. We exclude the 1,371 subnational policies from our analysis.

Table G.2: Summary Statistics

| | BERT Three-Class Model Prediction | | | |
|---|---|---|---|---|
| | **Industrial Policy** | **Other Intention** | **Not Enough Info.** | **All** |
| **Panel (A): Implementation Level** | | | | |
| IFI | 2391 | 638 | 232 | 3261 |
| NFI | 7085 | 432 | 1247 | 8764 |
| national | 5502 | 8841 | 18668 | 33011 |
| supranational | 426 | 521 | 1300 | 2247 |
| **Panel (B): Firm-specific policies** | | | | |
| Not firm specific | 5484 | 7658 | 14022 | 27164 |
| Firm specific | 9920 | 2774 | 7425 | 20119 |
| **Panel (C): MAST Chapter Code** | | | | |
| Capital control measures | 11 | 250 | 169 | 430 |
| Contingent trade-protective measures | 1 | 2612 | 107 | 2720 |
| Export-related measures | 5925 | 737 | 2317 | 8979 |
| FDI measures | 88 | 234 | 704 | 1026 |
| Finance measures | 0 | 16 | 38 | 54 |
| Government procurement restrictions | 172 | 156 | 866 | 1194 |
| Instrument unclear | 115 | 121 | 290 | 526 |
| Intellectual property | 1 | 3 | 7 | 11 |
| Migration measures | 12 | 81 | 372 | 465 |
| Non-automatic licensing, quotas, prohibitions | 35 | 262 | 1748 | 2045 |
| Pre-shipment inspection and other formalities | 0 | 6 | 11 | 17 |
| Price-control measures | 26 | 111 | 378 | 515 |
| Subsidies (excluding export subsidies) | 8218 | 5004 | 7090 | 20312 |
| Tariff measures | 428 | 757 | 6267 | 7452 |
| Technical barriers to trade | 0 | 0 | 4 | 4 |
| Trade-related investment measures | 372 | 82 | 1079 | 1533 |
| Total | 15404 | 10432 | 21447 | 47283 |

*Notes:* This table presents the distribution of the 47,283 interventions at the core of our analysis. We report the implementation level, whether the intervention is targeted at specific firms or has broader scope, and the corresponding MAST Chapter Code for each intervention, categorized according to the labels generated by the BERT model. IFI and NFI denote international and national financial institutions (development banks and export credit agencies).

Table G.3: Examples of Agencies Implementing Industrial Policy

| **Implementing Agency name** | **Country** | **Policy Description (from GTA)** |
|---|---|---|
| Bangladesh Bank | Bangladesh | [...] *Bangladesh Bank*, while announcing the export subsidy for the year 2021-2022, added 4 categories to the list of goods that will be eligible for such a subsidy. |
| Banco Nacional de Desenvolvimento Econômico e Social | Brazil | [...] the *National Bank for Economic and Social Development* (BNDES in Portuguese) provided a BRL 77.83 million [...] loan to an undisclosed agricultural undertaking. The loan was issued under the PRONAF Investimento Faxia 2 scheme. |
| Government of Congo | Congo | [...] the *government of Congo* implemented an export ban on wood logs. According to news reports, this decision follows a governmental strategy to strengthen the wood processing industry within the country. |
| Société d'Enrichissement du Tricastin | France | [...] the European Investment Bank (EIB) and *Societe Denrichissement Du Tricastin (S.e.t)* signed an agreement worth EUR 400 million [...] for the project Areva Uranium Enrichment Facility from France. |
| Government of Ireland | Ireland | [. . . ] *Ireland* introduced a EUR 1176 (USD 1507) million price stabilisation scheme. The scheme will support undertakings producing electricity through renewable energy sources. |
| Secretariat of Agriculture and Rural Development | Mexico | [...] the *Secretary of Agriculture and Rural Development of Mexico* published an Agreement with the operating rules of the new Agriculture, Livestock, Fisheries and Agriculture Promotion Programme for the 2022 fiscal year. |
| Government of Puerto Rico | Puerto Rico | [...] the *government of Puerto Rico* approved Act No. 20 otherwise known as the "Act to Promote the Export of Services". |
| Ministry of Public Enterprises | South Africa | [...] the *South African Ministry of Public Enterprises* announced allocating R10.5 billion (USD 640.7 million) to finalize the business and rescue plan of South African Airways (SAA). |
| Türkiye Cumhuriyet Merkez Bankasi | Turkey | The Decree 2013/4 of *the Money-Credit and Coordination Council* replaces the Decree 2010/10 and encompasses an export subsidy programme for agricultural products. |
| Federal Crop Insurance Corporation | United States of America | [...] the *Department of Agriculture* approved a production subsidy worth USD 13 million to multiple US agricultural producers. [...] The program is administered by the *Federal Crop Insurance Corporation (FCIC* under [the] *Risk Management Agency (RMA) of* USDA. |

*Notes*: We provide a random example of a policy text associated with each implementing agency in a random sample. Policy descriptions (excerpts) from the Global Trade Alert. The implementing agency names were extracted from the policy text by us. There is not necessarily only one implementing agency identified per policy text. The text that refers to the names of the implementing agency/agencies associated with the policy text have been italicized by us.

Table G.4: Predictive Performance of Fine-Tuned Bert Model on Annotated Splits

| Data Split | Class/Metrics | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|---|
| Train | IP Goal | 0.99 | 1.00 | 0.99 | 448 |
| | No IP Goal | 0.99 | 1.00 | 1.00 | 329 |
| | Not Enough Information | 1.00 | 0.99 | 1.00 | 1128 |
| | Macro Avg | 0.99 | 1.00 | 0.99 | 1905 |
| | Weighted Avg | 1.00 | 1.00 | 1.00 | 1905 |
| Test | IP Goal | 0.89 | 0.88 | 0.88 | 104 |
| | No IP Goal | 0.90 | 0.93 | 0.92 | 76 |
| | Not Enough Information | 0.95 | 0.94 | 0.94 | 260 |
| | Macro Avg | 0.91 | 0.92 | 0.91 | 440 |
| | Weighted Avg | 0.93 | 0.93 | 0.93 | 440 |
| Validation | IP Goal | 0.99 | 0.99 | 0.99 | 138 |
| | No IP Goal | 1.00 | 0.99 | 0.99 | 102 |
| | Not Enough Information | 0.99 | 1.00 | 1.00 | 347 |
| | Macro Avg | 1.00 | 0.99 | 0.99 | 587 |
| | Weighted Avg | 0.99 | 0.99 | 0.99 | 587 |

*Notes:* This table presents the predictive performance of the BERT model across different data splits: train, test, and validation. For each split, it details standard classification metrics—Precision, Recall, and F1-Score—for the three class BERT Model. The classes are 'IP Goal', 'No IP Goal', and 'Not Enough Information'. Additionally, it provides the Macro Average and Weighted Average for these metrics across classes. The 'Support' column indicates the number of observations for each class (or average) by sample split.