Making Interpretable Discoveries from Unstructured Data: A High-Dimensional Multiple Hypothesis Testing Approach

Jacob Carlson
Harvard University*

First Draft: October 2025
This Draft: January 2026

Abstract

Social scientists are increasingly turning to unstructured datasets to unlock new empirical insights, e.g., estimating descriptive statistics of or causal effects on quantitative measures derived from text, audio, or video data. In many such settings, unsupervised analysis is of primary interest, in that the researcher does not want to (or cannot) pre-specify all important aspects of the unstructured data to measure; they are interested in "discovery." This paper proposes a general and flexible framework for pursuing discovery from unstructured data in a statistically principled way. The framework leverages recent methods from the literature on AI interpretability to map unstructured data points to high-dimensional, sparse, and interpretable "dictionaries" of concepts; computes statistics of dictionary entries for testing relevant concept-level hypotheses; performs selective inference on these hypotheses using algorithms validated by new results in high-dimensional central limit theory, producing a selected set ("discoveries"); and both generates and evaluates human-interpretable natural language descriptions of these discoveries. The proposed framework is robust to data snooping and other post-selection inference concerns, is fully replicable, and is cheap to implement—both in terms of financial cost and researcher time. Applications to recent descriptive and causal analyses of unstructured data in empirical economics are explored. An open source Jupyter notebook is provided for researchers to implement the framework in their own projects.

1 Introduction

Empowered by recent developments in machine learning and AI, researchers in the social sciences are increasingly leveraging sources of unstructured data—such as text, audio, images,

*Email: jacob_carlson@g.harvard.edu. I am grateful to Isaiah Andrews, Melissa Dell, Neil Shephard, Rahul Singh, Elie Tamer, Davide Viviano, and participants in the Harvard econometrics workshop for their feedback and suggestions. I thank Harvey Barnhard, Andrew Kao, and Karthik Tadepalli for early feedback and encouragement. All errors are my own.

and videos—in quantitative analyses. In economics, interest in unstructured data sources is especially widespread, ranging from using speech recordings from FOMC meetings to better understand monetary policy [Gorodnichenko et al., 2023]; to using videos of start-up pitches to study entrepreneurship and investment [Hu and Ma, 2025]; to using open-ended survey questions to probe the economic behavior and beliefs of individuals [Haaland et al., 2024]; to using mug shots to study judicial decisionmaking [Ludwig and Mullainathan, 2024]; to using visual art or written narratives to infer long-run living standards [Gorin et al., 2025, Lagakos et al., 2025]; to using qualitative interview transcripts to better understand the impacts of RCT treatments [Bergman et al., 2024, Krause et al., 2025].

In the social sciences, the promise of these unstructured data, fundamentally, is that they offer new opportunities to measure social phenomena that were previously unmeasurable. This includes both measures that a researcher is capable of pre-specifying before seeing the data as well as those that a researcher is not, and instead would like to "discover" through the data itself. When modern AI/ML methods are brought to bear on these goals, the former typically falls under the heading of "supervised learning," whereby the researcher makes predictions of their ex-ante known quantity of interest using the unstructured data as features, and the latter typically falls under the heading of "unsupervised learning," which is less amenable to a prediction-based framework, and demands that AI/ML methods uncover latent structure in the unstructured data with minimal researcher input. Though new statistical and econometric frameworks have recently become available for interpretable and statistically principled analyses of unstructured data in the supervised learning setting (e.g., Angelopoulos et al. [2023], Ludwig et al. [2024], Carlson and Dell [2025]), it is an open question as to how one best performs interpretable and statistically principled analyses of unstructured data in unsupervised learning settings that emphasize discovery. The framework proposed in this paper provides one possible answer to this open question.

Consider two motivating examples for the proposed framework, one focused on descriptive analysis, and another focused on causal inference. Both consider text as the unstructured data modality of interest, though the proposed framework applies more generally to other data modalities (see, e.g., the AI interpretability literature review in Section 2).

Example 1 (Information treatment RCT (c.f., Bursztyn et al. [2023])). A researcher is interested in running a randomized controlled trial (RCT) with an information treatment. A key outcome of the experiment is the participant's subjective reaction to the information treatment, and the researcher would like to discover any systematic differences in reactions across the treatment and control groups in an unsupervised way. As such, for all participants, the researcher collects text responses to open-ended questions eliciting reactions to the experimental intervention, and is interested in learning any relevant causal effects on the presence of concepts surfaced in these text responses.

Example 2 (Open-ended survey analysis (c.f., Stantcheva [2024])). A researcher is interested in understanding the attitudes of some population of interest towards some particular (economic) policy. To learn about population attitudes without priming respondents to discuss particular topics, the researcher runs a survey on a representative sample from the population of interest that asks many open-ended questions about attitudes and beliefs, to which survey respondents reply in text. The researcher is interested in describing any prominent concepts that are surfaced in this open-ended survey.

To address these examples—among many others—this paper proposes a general framework for conducting interpretable and rigorous statistical inference on unstructured data per the following four steps:

1. The researcher begins with the unstructured dataset they are interested in analyzing. Each unstructured data point is mapped to a "feature dictionary," which is a binary vector; each element of the vector indicates the presence (or lack thereof) of a single, human-interpretable ("monosemantic") feature in that unstructured data point.[1] The creation of these interpretable feature dictionaries is accomplished using *dictionary learning* models newly popular in the AI interpretability literature for deep neural networks (DNNs) such as large language models (LLMs). These feature dictionary vectors are *high-dimensional*, and catalog a vast number of concepts. Heuristically, the feature dictionary should be thought of as an inventory of all possible (monosemantic) concepts that, e.g., a LLM needed to learn for the purpose of next token prediction over a massive internet-scale corpus. By construction, these feature dictionary vectors are also *sparse*, i.e., only the few most "meaningful" concepts represented in the unstructured data point are cataloged as present by the dictionary.[2]

   Example 1 (Continued). Each experimental participant's post-treatment text response is passed through a LLM equipped with pretrained sparse autoencoder (SAE) [Bricken et al., 2023] (a popular implementation of a dictionary learning model); the SAE's feature activations are pooled and thresholded to serve as a (binary) feature dictionary. There is one feature dictionary per text response (and one text response per experimental participant), each of which is of dimension on the order $p \approx 10^5$, though there are only on the order of $n \approx 10^2$ participants across both treatment and control groups. Feature number 1 in the dictionary may indicate the presence of sentences ending with exclamation points, feature number 2 may indicate the discussion of politics, and so on.[3] △

2. The researcher, now with access to a new dataset composed of feature dictionaries, formulates p feature-specific hypothesis tests and associated test statistics—one for each concept cataloged in the feature dictionary. The researcher has great flexibility in designing relevant, feature-level hypotheses and t-stats at this stage of the framework, which may employ data sources beyond the feature dictionaries themselves.

   Example 1 (Continued). The researcher computes a difference in means across treatment and control groups for each feature indicator (binary entry in the feature dictionary) to test a total of p null hypotheses of zero average treatment effect (ATE) on the probability of a given feature appearing in a text response. △

3. All p feature-specific hypotheses are tested using statistical procedures with high-dimensional selective error guarantees validated by novel results in high-dimensional

[1] For more discussion of what "monosemanticity" means and its importance for LLM interpretability, see, e.g., the discussion in Bricken et al. [2023].

[2] Meaningfulness, in a technical sense, relates to the reconstruction loss associated with learning the dictionary feature under a sparsity penalty; for more on this, see, e.g., Bricken et al. [2023].

[3] For more examples of feature descriptions, see, e.g., `https://www.neuronpedia.org/gemma-scope`.

multiple hypothesis testing and central limit theory. Chiefly, these results prove that the procedures of Romano and Wolf [2007] for controlling the k familywise error rate (k-FWER)—the probability of making k or more false rejections—are valid in high-dimensions (under mild assumptions, using particular bootstrap methods and hyperparameter choices). k-FWER control is less conservative than simple FWER control,[4] leading to more powerful tests that permit the researcher to make more discoveries (rejections of true nulls). Moreover, the researcher may invert these tests to form generalized simultaneous confidence intervals, i.e., the probability that k or more parameters are not contained by the random intervals is less than some desired level. Because these new formal results make only very mild assumptions on the data generating process, they may also be of independent interest in other high-dimensional selective inference settings.

Example 1 (Continued). An algorithm based on the work of Romano and Wolf [2007] is deployed to produce a critical value (or sequence of critical values) for rejecting each of the p null hypotheses of no average treatment effect, guaranteeing k-FWER control in the selected set at level α (in large samples) for a researcher chosen value of k (that is sufficiently small), even though $p \gg n$. These bootstrapped critical values may also be used to form generalized confidence intervals for the difference-in-means point estimates of the average treatment effects. △

4. The output of the framework thus far is a selected set of rejected, feature-level null hypotheses (and corresponding generalized simultaneous confidence intervals and point estimates), a.k.a., "discoveries." By virtue of the technology used to create them, however, the concepts encoded by the feature dictionary are only indexed numerically, and do not come with natural language descriptions—we know, e.g., that feature 1 likely encodes a human-interpretable concept, and that we rejected the null hypothesis associated with it, but we do not yet have an English description for what that concept is. As such, in the final stage of the framework, the researcher generates natural language descriptions of the concepts in the feature dictionary using a variant of a popular *automatic interpretation* (or "autointerp") method that is adapted to the unstructured dataset being explored. In order to rigorously evaluate the quality of these autointerp descriptions, we propose a new statistical formalization of a workhorse autointerp evaluation method from Paulo et al. [2024], along with new estimators that permit statistically and conceptually principled autointerp quality evaluation.

   Example 1 (Continued). The researcher now has in hand some subset of the p null hypotheses that have been rejected: a set of indices associated with particular elements of the feature dictionary. In order to provide a natural language description for each of the selected features for which there is evidence of a nonzero ATE (which are otherwise just identified by their numeric index), a dedicated "explainer LLM" reads examples of texts in which a given feature is present (or presents with high SAE activations) and generates an English interpretation for what concept is encoded by the feature. The credibility of this LLM-derived description of the feature is then evaluated with a held-out evaluation sample using newly proposed estimators. △

[4] Note that FWER is equivalent to k-FWER when $k = 1$.

The final output of the proposed framework is therefore: a set of feature-level discoveries (rejections of nulls) that come with generalized familywise error rate guarantees; natural (e.g., English) language descriptions of those discoveries; and intuitive numeric scores measuring the quality of those descriptions.

Example 1 (Continued). The researcher discovers that the null of no average treatment effect is rejected for features indexed $\{4, 101, 5030, ...\}$, meaning that with high probability the information treatment induces nonzero causal effects on the probability of the appearance of these features in texts. Feature 4 is described by an explainer LLM as activating on "mentions of politics" and this description is evaluated to have a high "quality" score, i.e., the information treatment credibly causes increased discussion of politics; and so on. △

The proposed framework offers several distinct scientific benefits, including:

1. *Comprehensive discovery*: Meaningful scientific discovery rests on the ability to detect important and interpretable patterns in the data that the researcher would not have been able to predict the existence of ex ante. Towards this ideal, AI interpretability methods are repurposed to measure the presence of an enormous number of human-interpretable concepts in an unstructured dataset—every concept or feature that a modern LLM needed to "learn" to excel at pretraining on an internet-scale corpus—reducing the possibility that there are any truly important measures of concepts ignored by the researcher.

2. *Robustness to data snooping and other post-selection inference concerns*: In many settings featuring exploratory analysis of unstructured data, there is concern that researchers may cherry pick aspects of their data to measure in order to reverse engineer a specific conclusion. Even when data snooping is not motivated by some desired conclusion, defining and taking measurements based on some observed dataset realization leads to well known problems in *post-selection inference*.[5] By using dictionary learning methods to comprehensively measure almost any human-interpretable concept of interest to a researcher (without looking at the data first to define such concepts), the framework proposed by this paper can avoid data snooping and related post-selection inference concerns without the need for sample splitting at the discovery phase. Tailored automatic interpretation methods, especially when coupled with formal interpretation quality evaluation, make the process of describing concepts in natural language replicable and transparent, while further limiting a researcher's ability to guide analysis towards a preordained conclusion. Moreover, prior to analysis, a researcher can easily preregister a choice of dictionary learning model and explainer LLM, hampering the possibility that a researcher applies this framework with many different models to select on favorable results.[6]

3. *Cost and effort*: The proposed framework is inexpensive both in terms of financial cost and researcher burden; both empirical examples in this paper were computed on Google

[5] See, e.g., Kuchibhotla et al. [2022] for more on the problems of and solutions to post-selection inference.

[6] Even such a method for pursuing motivated data snooping is unlikely to succeed, given that, e.g., SAE features tend to exhibit some degree of universality across architectures (c.f., Wang et al. [2025]).

Colab notebooks using a single A100 GPU, access to which only costs tens of dollars per month (and is free for students).[7] This framework is also easy to implement, requiring no human-in-the-loop to label unstructured data points, no additional AI/ML model training, nor any additional data sources not already possessed by the researcher. The framework is nearly automatic in execution, and takes on the order of a single hour to run using the above-mentioned hardware (and with access to LLMs via API, it may only take minutes).

4. *Standardization and replicability*: The automaticity of the framework makes it easy to replicate discoveries and their descriptions and analyze the sensitivity of results with respect to those few choices a researcher must make (e.g., choice of LLM).

Thus, the primary contributions of this paper are:

1. The development of a flexible, general-purpose framework for making interpretable as well as statistically and scientifically principled discoveries from unstructured data, leveraging state-of-the-art intepretability methods for LLMs and other DNNs.

2. Novel, general-purpose results in high-dimensional central limit theory and high-dimensional multiple hypothesis testing that justify using the procedures of Romano and Wolf [2007] for controlling k-FWER in high-dimensional settings.

3. A new conceptual and statistical formalization of a popular framework for autointerp scoring from Paulo et al. [2024], including novel estimators for implementing it.

The remainder of this paper is structured as follows: Section 2 discusses related literature; Section 3 describes the framework in detail; Section 4 applies the framework to two recent papers making use of unstructured data for discovery [Bursztyn et al., 2023, Stantcheva, 2024]; and Section 5 concludes.

2 Related Literature

The framework proposed in this paper is related to recent works from literatures spanning economics, statistics, and computer science.

Econometric methods for unstructured data. Motivated by the capabilities of modern AI/ML methods for learning from unstructured economic data (see, e.g., Mullainathan and Spiess [2017], Gentzkow et al. [2019a], Ash and Hansen [2023], Dell [2025]), new econometric and statistical frameworks have been developed to facilitate principled statistical inference on low-dimensional features (predictions) learned from unstructured datasets under supervision, e.g., Angelopoulos et al. [2023], Ludwig et al. [2024], Carlson and Dell [2025], Rambachan et al. [2024]. This most recent wave of econometric literature emphasizes nonparametric frameworks compatible with black-box AI models coupled with debiasing methods, as opposed to model-based approaches, e.g., Gentzkow et al. [2019b], Battaglia et al. [2024]. Though the framework presented in this paper is concerned with principled inference on unstructured

[7] This pricing description is valid as of December 2025.

data without making parametric assumptions, it differs from this existing econometric literature in that its goal is primarily unsupervised discovery, as opposed to supervised detection, using AI/ML models.

The proposed framework is most similar to existing works such as Ludwig et al. [2017], Modarressi et al. [2025], both of which provide formal statistical frameworks for discovery from high-dimensional economic data. However, both of these works focus specifically on causal inference in RCTs and operate by reducing the effective dimensionality of the discovery problem, whereas the proposed framework is suited to settings beyond causal inference in RCTs, and develops high-dimensional selective inference methods that do not require dimensionality reduction.

High-dimensional selective inference. There is an extensive literature on multiple hypothesis testing (or "selective inference" more broadly) in statistics, biostatistics, and economics (see, e.g., Romano et al. [2010] for a review in econometrics). The methods in this literature span asymptotic and finite sample valid frameworks, and low- and high-dimensional settings. In particular, for the problem considered in this paper, the literature on asymptotically valid, high-dimensional selective inference is most relevant, as such methods permit making very few assumptions about the data (c.f., assuming independence or PRDS of p-values in the case of Benjamini and Hochberg [1995]) and readily incorporate resampling methods that improve power (relative to guarding against worst-case dependence, c.f., Bonferroni corrections or the method of Benjamini and Yekutieli [2001]). The new results for high-dimensional multiple hypothesis testing appearing in this paper are derived by extending recent work on high-dimensional central limit theory for order statistics [Chernozhuokov et al., 2022, Ding et al., 2025]. The formal results presented in this paper are the first stated of their kind that enable asymptotic high-dimensional control of the k-FWER for test statistics that may be arbitrarily correlated with one another, complementing existing methods for asymptotic high-dimensional control of the FWER [Belloni et al., 2018] and the FDR under weak dependence [Liu and Shao, 2014, Belloni et al., 2018].

Hypothesis generation. There is a growing literature in economics and computer science on "hypothesis generation" at the pre-scientific stage of empirical inquiry [Ludwig and Mullainathan, 2024]. From the perspective of this literature, the proposed framework may be viewed as leveraging dictionary learning as a high-dimensional hypothesis generator, on top of which discovery is performed at the "scientific" stage of inquiry using high-dimensional multiple hypothesis testing procedures. As such, the framework in this paper contributes to the literature on hypothesis generation by providing a statistically principled pathway from hypothesis generation to hypothesis evaluation given access to only one (potentially small) unstructured dataset—a pathway that involves little to no researcher discretion about what generated hypotheses should be tested. The recent framework of Movva et al. [2025] uses dictionary learned features (specifically, sparse autoencoder features) explicitly for the purposes of hypothesis generation, further motivating the use of dictionary learning methods as tools for hypothesis generation.

AI interpretability. There is a large literature in computer science on machine learning interpretability (see, e.g., Doshi-Velez and Kim [2017]). The machine learning interpretability methods leveraged by this framework originate from a nascent though highly active literature known as "mechanistic interpretability," which seeks to develop methods for interpreting LLM behavior via quantitative analyses of model internals (e.g., activations, weights). In particular, the dictionary learning methods for LLM interpretability implemented in this framework are called "sparse autoencoders" (SAEs), and were developed in Bricken et al. [2023], Templeton et al. [2024].[8] [9] Other dictionary learning methods that have been proposed as competitors to SAEs (e.g., transcoders [Paulo et al., 2025]) are similarly compatible with the framework proposed in this paper.

SAEs and other dictionary learning methods generate features, but not natural language (e.g., English) feature descriptions. Based on the popularity of these methods, a related literature on "automatic interpretability" (or "autointerp") methods has become active, which seeks to coherently use LLMs to describe the features discovered by dictionary learning methods at scale. Recent important papers in this literature include: Bills et al. [2023], Shaham et al. [2024], Paulo et al. [2024], Rajamanoharan et al. [2024]. Specifically, the autointerp framework proposed in this paper adapts and statistically formalizes insights from the "detection scoring" method of Paulo et al. [2024] in a way that emphasizes maintaining a high degree of relevance to the unstructured data generating process being considered.

Though the empirical examples considered in this paper handle one of the most popular unstructured data types—text—for which dictionary learning-based interpretability methods are the most mature, dictionary learning techniques have been successfully applied to many other modalities, including audio and images [Abdulaal et al., 2024, Bhalla et al., 2024, Fry, 2024, Daujotas, 2024, Pluth et al., 2025]. Future work would seek to apply the proposed framework to unstructured datasets in these other data modalities.

3 Framework

We now consider each stage of the proposed framework in detail, concluding with a pseudocode algorithm for a streamlined implementation of the framework that can be found in the Appendix.

[8] SAEs are very wide autoencoders inserted at various layers of an LLM that are trained to intercept and reconstruct activations from model internals (such as the residual stream, MLPs or the like) under a sparsity-inducing penalization on a reconstruction loss. Under the "linear representation hypothesis" and "superposition hypothesis" (see, e.g., Bricken et al. [2023] for more discussion), these sparse autoencoders are thought to act as an overcomplete basis of the space of concepts in text leveraged by a LLM to make next token predictions, encouraging learned autoencoder latents to have monosemantic interpretations.

[9] There has been much debate about the success of SAEs in pursuit of AI interpretability (e.g., Leask et al. [2025]), though most criticisms still support the notion that SAEs and other dictionary learning methods have a comparative advantage in discovering as opposed to detecting concepts of interest in text [Peng et al., 2025]—which is how they are applied in this framework.

3.1 Notation

In what follows, we define $[p] := \{1, \ldots, p\}$, and let $x \mapsto \log x$ be the natural logarithm. Further, for any vector $s \in \mathbb{R}^p$ and a subset of indices $K \subseteq [p]$, let $s_K \in \mathbb{R}^{|K|}$ denote the subvector indexed by K, such that $(s_K)_q = s_{j_q}$ for $j_q \in K$ and $j_1 < j_2 < \cdots < j_{|K|}$. We use $s \mapsto s_{[k]}$ as notation for the function that selects the k-th largest coordinate of the vector s.

3.2 Creating Feature Dictionaries

Consider that the researcher has access to a dataset of size n, $\{(W_i, Z_i)\}_{i=1}^n$, which is sampled i.i.d. from some (super-)population of interest represented by distribution P. The $Z_i \in \mathcal{Z}$ are unstructured data points and the $W_i \in \mathcal{W}$ are any other observed covariates of interest. The space $\mathcal{Z}$ is typically high-dimensional and semantically lean, e.g., if each Z_i was a 244×244 pixel image, $\mathcal{Z}$ might then be the space of all $3 \times 244 \times 244$ arrays of RGB values, or if Z_i was text that was truncated up to some maximum length, $\mathcal{Z}$ might be the space of all binary matrices of a certain dimensionality, which are concatenated one-hot encodings for each word, term, or token in the text with respect to a specific pre-defined vocabulary.

Example 1 (Continued). The Z_i is a text response elicited for experimental participant i; the $W_i \in \mathcal{W} = \{0, 1\}$ is a treatment status indicator for experimental participant i. $\triangle$

The first stage of the framework relies on the researcher having access to a function which creates semantically rich feature dictionaries from unstructured data points, which we may denote $\texttt{Dict} : \mathcal{Z} \to \{0, 1\}^p$. We suggest such a function can be computed by passing an unstructured data point to a DNN equipped with a dictionary learning model (e.g., a LLM equipped with pretrained SAE when Z_i are texts); recording the feature activations from the dictionary model; and then pooling and binarizing these dictionary feature activations to form a single indicator for each feature in the dictionary, indexed $j \in [p]$. There are many possible ways for a researcher to implement a `Dict` function, though an especially straightforward choice for the use case of text data is implementing a `Dict` function based on the SAE activations at a single layer of the DNN, for which the j-th entry of the output is equal to 1 if the j-th dictionary feature activated on any token of the input text, and 0 otherwise:

$$\texttt{Dict}(z)_j := \texttt{Dict}(z)_j^l := \begin{cases} 1 \text{ if SAE feature } j \text{ at layer } l \text{ activates on any token of } z, \\ 0 \text{ otherwise.} \end{cases}$$

Concrete implementation choices for the SAE architecture and layer l will be discussed in the empirical examples in Section 4.

We denote $Y_i := \texttt{Dict}(Z_i) \in \{0, 1\}^p$ as the output of the dictionary transformation of an unstructured data point: the "feature dictionary." Each "feature indicator" or "dictionary entry" $Y_{ij} \in \{0, 1\}$ has the natural interpretation that a particular dictionary feature j activated for the unstructured data point i.[10]

[10] Dictionary feature activations are typically positively valued scalars, where magnitude is thought to correspond to some notion of intensity of activation; other works in mechanistic interpretability consider using these raw activation values directly, perhaps with max or average pooling across tokens. However,

Example 1 (Continued). Each text response Z_i is passed through an open-source LLM equipped with pretrained SAE. The SAE activations for each token of Z_i at the l-th layer of the LLM are max-pooled and then thresholded at zero, yielding a $p \approx 10^5$-dimensional binary vector Y_i for each experimental participant i. Each $Y_{ij} \in \{0, 1\}$ is a binary indicator of a particular concept being expressed in experimental participant i's text response, e.g., Y_{i1} could indicate whether or not experimental participant i's text response contains phrases indicating uncertainty. △

The `Dict` function can be thought of as effectively producing a new, semantically rich and interpretable dataset of feature dictionaries and associated covariates $\{(Y_i, W_i)\}_{i=1}^n$ from our original unstructured dataset (with associated covariates), $\{(Z_i, W_i)\}_{i=1}^n$. The crux of the proposed framework is that inference on functionals of (the distribution of) each dictionary feature Y_{ij} (or each (Y_{ij}, W_i) pair) is desirable because:

1. Each Y_{ij} is monosemantic (indicates the presence of a single, human-intelligible concept) and therefore functionals for Y_{ij} for any given j (and associated hypotheses) are *interpretable*, e.g., $E_P[Y_{ij}]$ is the probability of human-interpretable feature j appearing in an unstructured data observation in population P.

2. Collectively, the p features in Y_i for any given i are exhaustive of some vast space of concepts of interest to the researcher. This means the researcher need not specify ex ante or ex post any particular concepts of interest to measure, reducing the conceptual and statistical complications associated with *data snooping* without limiting discovery.

Towards exploiting this wealth of interpretable information, we consider formulating formal hypothesis tests of parameters (estimands) of interest in the next stage of the framework.

3.3 Formulating Feature-Level Hypotheses

With an interpretable dataset of feature dictionaries and associated covariates $\{(Y_i, W_i)\}_{i=1}^n$ in hand, we may consider defining p "one-sided" concept-level hypotheses $\{\tilde{H}_{0,j}\}_{j \in [p]}$, where

$$\tilde{H}_{0,j} : \theta_j(P) := E_P[X_{ij}] \leq 0, \quad X_{ij} = h(W_i, Y_{ij}),$$

for some measurable function h, or, alternatively, a set of "two-sided" hypotheses of interest $\{H_{0,j}\}_{j \in [p]}$ as

$$H_{0,j} : \theta_j(P) := E_P[X_{ij}] = 0.$$

(As shorthand, we will denote $X := \{X_i\}_{i=1}^n$.)

Letting the feature-level parameter of interest $\theta_j(P)$ be an expectation of some flexible function h of dictionary feature indicators and covariates accommodates many natural estimands of interest in the social sciences. Inference on parameters that can still be "approximately" expressed as means (in an asymptotic sense) will be touched on in Section 3.4.4.

presently, magnitude of activation is not a well-understood or highly interpretable quantitative property of dictionary learning methods, and as such the default `Dict` function advocated for in the proposed framework does not incorporate this information. That said, leveraging feature activation intensity is empirically shown to be useful for deriving good autointerpretation generation strategies; see Section 3.5 for more on this.

Example 1 (Continued). To infer average treatment effects on dictionary features, for treatment indicator W_i, fixed probability of treatment π (bounded away from zero and one), and feature indicator potential outcomes $Y_{ij}(w)$ under SUTVA,

$$H_{0,j} : E[Y_{ij}(1) - Y_{ij}(0)] = E[X_{ij}] = E\left[\frac{W_i - \pi}{\pi(1-\pi)} Y_{ij}\right] = 0.$$

In other words, we set $X_{ij} = h(W_i, Y_{ij}) := \frac{W_i - \pi}{\pi(1-\pi)} Y_{ij}$, a Horvitz–Thompson transformation for computing a difference in means. △

We may form test statistics for such hypotheses as

$$T_{n,j} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{ij}, \quad T_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i,$$

and taking absolute values, as appropriate.

Though inference on all p hypotheses communicates rich and interpretable information about our unstructured dataset, inference in this setting is also challenging because:

1. The p is large, so inference on all $j \in [p]$ functionals is inherently a *multiple hypothesis testing* problem.

2. Not only is p large, but for most social science applications of interest $p \gg n$, so inference on Y_{ij} for each j is a *high-dimensional* multiple hypothesis testing problem.

3. Each estimator, test statistic, or p-value formed from the Y_{ij} for each $j \in [p]$ for the purposes of inference is plausibly statistically *dependent* on every other in a complicated way, ruling out multiple hypothesis testing approaches that assume independence or specific forms of dependence (e.g., PRDS).

4. Each Y_i is sparse, and the intent of the analysis is discovery, so the desired form of selective (familywise) error control *cannot be too conservative.*

What is needed, then, for principled inference on functionals of Y_{ij} is a high-dimensional selective inference procedure with control over a generalized selective error rate. To be as well-powered as possible, we want to focus on testing procedures that employ resampling methods for estimating the true covariance across test statistics, such that conservative protection against worst-case dependence is not required. To keep things as general as possible, we also want to allow for only asymptotically valid test statistics and p-values, as is common in much of econometric analysis. In the following section, theory and corresponding statistical procedures are developed to achieve exactly these aims.

3.4 High-Dimensional Multiple Hypothesis Testing

We now state and prove the validity of statistical procedures that provide high-dimensional k-FWER control asymptotically, i.e., control of the probability of making k or more rejections of true nulls in the large sample limit. Naturally, k-FWER control for $k = 1$ is control of the FWER, and larger choices of k make the permit more discoveries.

3.4.1 Background and Motivation

The fundamental challenge of high-dimensional statistical inference, characterized by settings where $p \gg n$, is that classic statistical theory no longer applies: classical asymptotic statistics treats p as fixed in the relevant asymptotic thought experiment as n grows large, whereas in high-dimensional settings the p is growing (much) faster than n in the relevant asymptotic thought experiment. As such, new statistical theory—built using new assumptions and theoretical tools—must be developed for high-dimensional settings, a literature that has grown popular in response to increasing interest in the statistical analysis of "big data."[11] Hypothesis testing (or selective inference, more broadly) in high-dimensions has received considerable attention, with methods now available that provide asymptotic high-dimensional control of the FWER under mild conditions [Belloni et al., 2018], and asymptotic high-dimensional control of the FDR under weak dependence across t-stats [Liu and Shao, 2014, Belloni et al., 2018]. Though valid in high-dimensions, these methods are typically quite conservative, or involve assumptions incompatible with our current setting of interest based on feature dictionary datasets.

These existing high-dimensional multiple hypothesis testing results are supported by advances in high-dimensional central limit theory. However, the central limit theory they draw upon is invalid for studying the behavior of workhorse methods for controlling the k-FWER in high-dimensions, such as the algorithms of Romano and Wolf [2007], Romano et al. [2008], which rely on the asymptotic validity of bootstrap approximations of the distribution of the k-th largest coordinate of a scaled sum of independent random vectors. Fortunately, very recent work in high-dimensional central limit theory (i.e., Ding et al. [2025]) has developed new theoretical tools suited to understanding the behavior of exactly these objects in high-dimensions. We apply such tools to study and ultimately validate the asymptotic control of the Romano and Wolf [2007] procedures in high-dimensions, and then extend these tools to develop new high-dimensional CLTs that support improved testing procedures.

The following subsections state formal assumptions and results in high-dimensional central limit theory and multiple hypothesis testing that validate the use of algorithms from Romano and Wolf [2007] to control the k-FWER in high-dimensions so long as, crucially, $p \gg k$ ("k is small"). Researchers not interested in theoretical details and formal statements may consider moving on to Section 3.5.

3.4.2 Assumptions

Assume $n \geq 3$ and $p \geq 3$. Let $b_1 > 0$ and $b_2 > 0$ be some constants such that $b_1 \leq b_2$, and let $B_n > 1$ be a sequence of constants, where it is possible that B_n diverges. Let us further define

$$S_{n,j} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_{ij} - E[X_{ij}]), \quad S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - E[X_i]).$$

We will denote $\Sigma := E\left[S_n S_n^{\mathsf{T}}\right] = n^{-1} \sum_{i=1}^{n} E[(X_i - E[X_i])(X_i - E[X_i])^{\mathsf{T}}]$, which simplifies to $\Sigma = E[(X_i - E[X_i])(X_i - E[X_i])^{\mathsf{T}}]$ under i.i.d. data.

Then we make the following assumptions.

[11] Concretely, challenges with developing statistical methods in high-dimensions include the lack of explicit limit distributions and complex dependence structures (see, e.g., Chernozhukov et al. [2023] for more).

Assumption M. For all $i \in [n]$, $j \in [p]$, assume: (i) $E\left[\exp\left(|X_{ij}|/B_n\right)\right] \leq 2$; (ii) $b_1^2 \leq \frac{1}{n}\sum_{i=1}^n E[X_{ij}^2]$; and (iii) $\frac{1}{n}\sum_{i=1}^n E[X_{ij}^4] \leq B_n^2 b_2^2$.

These are mild conditions on the tails and moments of the data X_{ij}, which are stated in a way that accommodates data that are independent but not identically distributed (i.n.i.d.). Part (i) of Assumption M simply requires that X_{ij} be sub-exponential, or, equivalently, have an Orlicz norm in Ψ_1 bounded by $B_n < \infty$. The B_n is indexed by n to accommodate growing tail thickness within the sub-exponential regime as $p = p_n$ grows in the asymptotic thought experiment with i.n.i.d. data. Part (ii) of Assumption M insists that the second moments of the data be bounded away from zero, appropriately stated for i.n.i.d. data. Part (iii) insists on bounded fourth moments in a similar fashion. As such, for i.i.d. data, the case considered in the applications of this framework, we may simply require for some fixed $B < \infty$ and fixed b_1, b_2 with $b_1 \leq b_2$: (i) $E\left[\exp\left(|X_{ij}|/B\right)\right] \leq 2$; (ii) $b_1^2 \leq E[X_{ij}^2]$; and (iii) $E[X_{ij}^4] \leq B^2 b_2^2$. Note that, in many use cases of interest for dictionary learned features, the X_{ij} will be both bounded (almost surely) and studentized, making these assumptions highly plausible.

Remark 1 (Degenerate features). Some fraction of dictionary features may be degenerate at zero for any given dataset (they are "dead" [Bricken et al., 2023]). Of course, these features are not of interest for the purposes of discovery. In practice, this infringement of Assumption M is well-tolerated by the framework because, in most realistic datasets of feature dictionaries, the upper quantiles of the k-th largest coordinate of S_n remain unchanged upon dropping all degenerate coordinates (in-sample or in-population).[12] As such, conceptually, one should think of j as ranging over the p non-degenerate features of the computed feature dictionaries. To perform studentization, these degenerate coordinates should be explicitly filtered out; studentization is discussed further in Section 3.4.4.[13] ◇

Assumption R. Assume that $B_n^2 \log^5(pn) = o(n)$.

This is the key rate condition needed for (sup-norm) Gaussian or bootstrap approximation error to go to zero asymptotically under the high-dimensional CLTs discussed in both the works of Ding et al. [2025] and Chernozhuokov et al. [2022]. Rewritten, it says that

$$\frac{B_n^2 \log^5(pn)}{n} = o(1)$$

which permits p growing very fast with n in the asymptotic thought experiment. In fact, p may be growing nearly exponentially in n, e.g., $p = e^{n^{1/6}}$ for fixed B_n. That this rate

[12] Consider, for example, the (with high probability) equivalence of the top quantiles of $S_{n,[k]}$ whether or not in-sample degenerate coordinates are dropped. As discussed in following sections, learning the quantiles of $S_{n,[k]}$ will be implemented with the multiplier bootstrap, which can be thought of as modeling the distribution of the k-th largest coordinate of normal random vector $N(0, \hat{\Sigma})$, where $\hat{\Sigma}$ is the empirical variance-covariance matrix (treated as fixed). Note then that the k-th largest coordinate in any bootstrap realization only changes when dropping in-sample degenerate features in the event $\{N(0, \hat{\Sigma})_{[k]} \leq 0\}$. As a benchmark, consider for $\hat{\Sigma}$ with zero off-diagonal entries (independent coordinates) that $P(N(0, \hat{\Sigma})_{[k]} \leq 0) \leq k\check{p}^{k-1}2^{-\check{p}}$ (where $\check{p}$ is the number of non-degenerate coordinates), which decays exponentially fast as $\check{p}$ grows large for small fixed k.

[13] A potential alternative to filtering out degenerate features is the strategy of injecting a small amount of continuously distributed, exogenous noise into dictionary feature indicators, e.g., define and analyze $\check{Y}_{ij} := Y_{ij} + \epsilon_{ij}$ with $E[\epsilon_{ij}] = 0$. As such, $E[\check{Y}_{ij}] = E[Y_{ij}]$, but $\check{Y}_{ij}$ is non-degenerate by construction. Such a strategy may be preferable in empirical settings exhibiting a large amount of degeneracy.

condition permits $p \gg n$ is an important positive result in high-dimensional central limit theory [Chernozhukov et al., 2017, 2023].

Remark 2 (State-of-the-art assumptions). Both Assumptions M and R are the state-of-the-art in the high-dimensional central theory from which this framework draws [Chernozhuokov et al., 2022, Ding et al., 2025]. Indeed, these conditions are the exact same as those used in both Chernozhuokov et al. [2022] and Ding et al. [2025]. Notably, these assumptions do not require non-degeneracy of Σ, another important positive result in this literature [Chernozhukov et al., 2023]. Under assumptions of non-degenerate Σ, improved rate conditions may be used in place of Assumption R. ◇

3.4.3 High-dimensional k-FWER Control

We will focus on the Gaussian multiplier bootstrap as our high-dimensional bootstrap method, which is both computationally light and theoretically suited to the setting at hand. We define the Gaussian multiplier bootstrap quantity

$$S_n^B := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \left(X_i - \bar{X}_n \right),$$

where $\xi_i \overset{iid}{\sim} N(0,1)$ and $\bar{X}_n := n^{-1} \sum_{i=1}^{n} X_i$. For the purposes of bootstrapping, we will denote the data-conditional probability measure $P^B(\cdot) := P(\cdot \mid X)$.

With these assumptions and notations in place, we now introduce statistical procedures for controlling the k-FWER for small k in large samples, based on the step-wise algorithms of Romano and Wolf [2007].

Theorem 1 (High-dimensional k-FWER control for small k, one-sided)**.** *Consider the method of Algorithm 2.1 or 2.2 in Romano and Wolf [2007] with test statistics T_n of hypotheses $\{\tilde{H}_{0,j}\}_{j \in [p]}$ and critical values $\hat{c}_{n,K}(1-\alpha, k)$ given by the $1-\alpha$ quantile of $S^B_{n,K,[k]}$ under P^B. Assume that k is fixed (i.e., not growing with n, p). Then under Assumptions M and R:*

(i) $\limsup_{n,p \to \infty} k\text{-}FWER_P \leq \alpha$.

(ii) If $\tilde{H}_{0,j}$ is false and $\theta_j(P) \gg B_n \sqrt{\log p / n}$, then the probability that the step-down method rejects $\tilde{H}_{0,j}$ tends to 1.

This result is the natural high-dimensional analog of Theorem 3.1 (as well as Theorem 3.3) of Romano and Wolf [2007]. It says that, so long as k is small (in the relevant asymptotic sense), Algorithm 2.1 or Algorithm 2.2 of Romano and Wolf [2007] may be used to asymptotically achieve control of the k-FWER at any desired level α in high-dimensional settings, and moreover that these algorithms are consistent against local alternatives that are not shrinking towards zero too quickly[14] (i.e., intuitively, in sufficiently large samples, we are able to appropriately reject a false null so long as $\theta_j(P) = E_P[X_{ij}]$ is not very close to zero). The proof of this result is available in Appendix Section 6.5, and, as previously alluded to, relies on the insight that the low-dimensional k-FWER controlling algorithms of Romano and

[14] Under Assumption R, note that $B_n \sqrt{\log p / n} = o(1)$.

Wolf [2007] can be adapted to and analyzed in the high-dimensional setting using new results in high-dimensional central limit theory for order statistics. In the following theorem, we state the sibling procedure for high-dimensional control of the k-FWER in two-sided testing settings; it is the high-dimensional analog of Theorem 3.2 of Romano and Wolf [2007].

Theorem 2 (High-dimensional k-FWER control for small k, two-sided)**.** *Consider the method of Algorithm 2.1 or 2.2 in Romano and Wolf [2007] with test statistics $|T_n|$ of hypotheses $\{H_{0,j}\}_{j\in[p]}$ and critical values $\hat{c}_{n,K}(1-\alpha, k)$ given by the $1-\alpha$ quantile of $|S^B_{n,K}|_{[k]}$ under P^B. Assume that k is fixed (i.e., not growing with n, p). Then under Assumptions M and R:*

(i) $\limsup_{n,p\to\infty} k\text{-}FWER_P \leq \alpha$.

(ii) If $H_{0,j}$ is false and $|\theta_j(P)| \gg B_n\sqrt{\log p/n}$, then the probability that the step-down method rejects $H_{0,j}$ tends to 1.

Remark 3 (Small k)*.* To comment explicitly on why the procedures of Theorems 1 and 2 are only valid for small k, note that the sup-norm bootstrap or Gaussian approximation error in the high-dimensional CLTs for the k-th largest coordinate introduced in Ding et al. [2025] only goes to zero if

$$\frac{k^8 B_n^2 \log^5(pn)}{n} = o(1),$$

meaning that k must be fixed or must grow incredibly slowly: for all practical purposes, k needs to be quite small. As such, the theory of Ding et al. [2025] is unsuitable for making progress on FDP exceedance control, which requires a k that may grow linearly ("quickly") with p. A productive goal for future research would be to establish valid high-dimensional FDP exceedance control using further extensions of high-dimensional central limit theory. However, as will be observed in the empirical applications of this paper, small $k > 1$ are often sufficient to power meaningful discoveries in relevant social science settings at conventional choices of α.[15] ◇

3.4.4 Studentization

Multiple hypothesis testing with studentized statistics is often important for the purposes of "balance": that all hypothesis tests considered are similarly powered, contributing to selective error control similarly. For example, when $X_{ij} = Y_{ij} \in \{0, 1\}$, more frequently occurring dictionary features are higher variance features, and we are typically interested in powering discoveries beyond just these frequently occurring concepts.

To facilitate studentization, we require a high-dimensional CLT that considers statistics with small estimation errors, as studentization relies on estimated variances. We therefore generalize the high-dimensional central limit theorems of Ding et al. [2025] to handle "approximate means" in the parlance of Belloni et al. [2018], stated in the following lemma.

[15] For applied researchers looking for guidance on choice of k, note that the given approximation error rate suggests a choice of k on the order of $\log(pn)$.

Lemma 1 (High-dimensional CLT for the small k-max coordinate of approximate means)**.** *Let $\hat{S}_n := S_n + R_n$, and assume that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$. Further assume that k is fixed (i.e., does not grow with n, p). If Assumptions M and R hold, then as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}} \left| P\left(\hat{S}_{n,[k]} \le t\right) - P\left(N(0,\Sigma)_{[k]} \le t\right) \right| \to 0.$$

This lemma is a natural generalization of Lemma A.6 of Ding et al. [2025], which proves the special case for $R_n = 0$. This lemma then indicates that, for appropriately small k, in the large sample limit, the distribution of the k-th largest coordinate of a scaled sum of independent random vectors is (uniformly) well-approximated by the distribution of the k-th largest coordinate of a Gaussian random vector with variance matrix Σ *whether or not* there is a small amount of noise R_n injected into the sum. The following lemma states that an analogous result is also true for the analogous (multiplier) bootstrap quantity under the relevant bootstrap law.

Lemma 2 (High-dimensional bootstrap for the small k-max coordinate of approximate means)**.** *Let $\hat{S}_n^B := S_n^B + R_n$, and assume that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$. Further assume that k is fixed (i.e., does not grow with n, p). If Assumptions M and R hold, then as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}} \left| P^B\left(\hat{S}_{n,[k]}^B \le t\right) - P\left(N(0,\Sigma)_{[k]} \le t\right) \right| \xrightarrow{P} 0.$$

Using these lemmas, we may now state two theorems that allow for studentization, recognizing that studentization with estimated variances can be cast as analyzing a particular S_n under estimation error. We define the diagonal matrix of asymptotic variances $\Lambda := \mathrm{diag}(\Sigma)$, as well as the corresponding correlation matrix $\Sigma_0 := \Lambda^{-1/2}\Sigma\Lambda^{-1/2}$. Let plug-in estimates be given by $\widehat{\Sigma}_{jj} := n^{-1}\sum_{i=1}^n \left(X_{ij} - \bar{X}_{n,j}\right)^2$ and $\widehat{\Lambda} := \mathrm{diag}\left\{\widehat{\Sigma}_{11}, \ldots, \widehat{\Sigma}_{pp}\right\}$.

Theorem 3 (High-dimensional CLT for the small k-max studentized coordinate)**.** *If Assumptions M and R hold, and $B_n = O(1)$ and k is fixed (i.e., does not grow with n, p), then, by application of Lemma 1, as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}} \left| P\left(\left(\hat{\Lambda}^{-1/2} S_n\right)_{[k]} \le t\right) - P\left(N(0,\Sigma_0)_{[k]} \le t\right) \right| \to 0.$$

This theorem validates that the relevant estimation error R_n in the studentization setting is indeed small enough in the sense required by Lemma 1 to lead to a standard high-dimensional CLT based on the relevant (asymptotic) correlation matrix. The next theorem proves the analogous result for the bootstrap case.

Theorem 4 (High-dimensional bootstrap for the small k-max studentized coordinate)**.** *If Assumptions M and R hold, and $B_n = O(1)$ and k is fixed (i.e., does not grow with n, p), then, by application of Lemma 2, as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}} \left| P^B\left(\left(\hat{\Lambda}^{-1/2} S_n^B\right)_{[k]} \le t\right) - P\left(N(0,\Sigma_0)_{[k]} \le t\right) \right| \xrightarrow{P} 0.$$

Combining these theorems, it is immediate that the bootstrap quantiles of studentized statistics well-approximate those of the studentized version of S_n. Further note that, in the setting where the data is not just i.n.i.d. but i.i.d., as is true for the use case of interest based on dictionary features extracted from an unstructured dataset obtained through i.i.d. sampling, $B_n = O(1)$ is assured.

3.5 Generating and Evaluating Automatic Interpretations

By applying the statistical procedures introduced in Theorems 1 and 2, the researcher obtains a selected set of rejected null hypotheses, or "discoveries." Concretely, these discoveries are recorded as set of coordinates $\hat{J} \subseteq [p]$ indexing the rejected nulls.

Though the machine learning theory that motivates sparse dictionary learning supposes that each hypothesis $\hat{\jmath} \in \hat{J}$ has single, human-interpretable meaning, the implementation of dictionary learning itself does not generate *natural language descriptions* of what each human-interpretable meaning may be. We now discuss a set of procedures for both automatically generating and evaluating the quality of such natural language feature descriptions from a statistical inference perspective.

Remark 4 (Why do we need automatic interpretation?). Human judgment alone could be deployed as a means of giving natural language descriptions to dictionary features, e.g., the researcher (or research assistants) could inspect many of the text examples that activate (highly) for a given dictionary feature in order to arrive at a plausible English language description for the concept signaled by that feature. However, such a process suffers from (1) problems related to motivated reasoning and data snooping (e.g., the researcher could simply come up with a feature description that is reverse engineered from a conclusion they wanted to find in the data); and (2) problems related to the scale of interpretation and associated interpretation effort/cost (e.g., there may be many hundreds of features discovered, all requiring interpretation; interpreting all p features would require even more labor). Automatic interpretation pipelines solve both of these problems by seeking to automate the process of creating and scoring natural language descriptions for any set of dictionary features by leveraging a dedicated "explainer LLM" which can reason about feature interpretations cheaply at scale when provided with sufficient information. ◇

3.5.1 Automatic Generation

We adopt the autointerp generation strategy pioneered in Bricken et al. [2023] for this framework: we pass text samples from the researcher's unstructured dataset that activate highly on a particular feature to an explainer LLM—annotating these exemplar texts with information about where and how strongly the feature activates—and ask for an English description of the possible meaning of the feature.

By drawing on text samples from the researcher's unstructured dataset, this generation strategy is tailored to the distribution P. Though off-the-shelf dictionary feature descriptions are sometimes available for pretrained dictionary learning models, these descriptions were necessarily learned on a different distribution of data (i.e., some large and opaque swath of the internet). As such, these descriptions are liable to be of lower quality than those based on

the researcher's own dataset. The specific prompts implementing this strategy are detailed in Appendix Section 6.3.

Example 1 (Continued). The null hypothesis of no average treatment effect is rejected for dictionary feature 4. Using the raw SAE feature activations (necessary to compute in the process of evaluating the `Dict` function), the researcher is able to rank text responses for experimental participants by largest activations for feature 4 on any token ("max pooling"). The researcher takes the top 10 largest activating text examples and annotates each of them with brackets delimiting the highest activating token in each response; these exemplar responses are then passed to an explainer LLM that reasons about and then outputs the concise English phrase "discussion of politics" to describe the concept feature 4 likely maps to (based on the activation patterns in the presented texts). △

3.5.2 Automatic Evaluation: A Statistical Inference Perspective

Autointerp pipelines ideally not only specify how to generate descriptions for dictionary features, but also *evaluate* the quality of these descriptions. We now consider a new statistical formalization of a mainstream autointerp evaluation technique called "detection scoring."

The fundamental principle of detection scoring is that "an interpretation should serve as a binary classifier distinguishing activating from nonactivating contexts" [Paulo et al., 2024]. In other words: description of a dictionary feature can be treated like a classification problem, one where the natural language description itself (incorporated in a LLM prompt) is the classifier, the input is a specific text example, and the label to be predicted is derived from the ground truth of the presence or absence of a feature's activation in that text example (which is always known to the researcher using their `Dict` function). Under this framing, binary classifier performance metrics like accuracy, precision, and recall characterize the quality of the description.

To formalize this autointerp strategy, consider a "held-out" unstructured dataset of size m drawn independently from the same distribution of interest P. This held-out dataset may be created, for example, by partitioning an original unstructured dataset drawn i.i.d. from P into two disjoint datasets via indices chosen at random, independently of all the data. Let us denote the set of indices of unstructured data points in the main sample as $\mathcal{I}^{\text{estim}}$, and those in the held-out dataset as $\mathcal{I}^{\text{eval}}$.

Using the data points with indices in $\mathcal{I}^{\text{estim}}$, we extract exemplar unstructured data points to be used to create an autointerp explanation of feature j, which we denote $\hat{\eta}_j$. For all $i \in \mathcal{I}^{\text{eval}}$, we define the random variable $\hat{Y}_{ij} \in \{0, 1\}$ as

$$\hat{Y}_{ij} := \texttt{CLS}(Z_i, \hat{\eta}_j).$$

The function $\texttt{CLS}(\cdot)$ is a non-stochastic classification prompt fed both an unstructured data point Z_i and the learned autointerp explanation $\hat{\eta}_j$. The $\hat{Y}_{ij}$ is the classifier output, a prediction of whether or not the unstructured data point Z_i possesses a feature with concept description $\hat{\eta}_j$.[16]

[16]Note that, even if a sample splitting implementation of this autointerp evaluation strategy is pursued, cross-fitting is inappropriate, because there is little reason to think, given the sample sizes available to researchers, that $\hat{\eta}_j$ would be stable across folds. As such, when sample splitting is employed, an uneven split is recommended, e.g., only 10% of the data is used for the held-out evaluation sample.

For the purposes of evaluating autointerp descriptions, we define the "accuracy score" (or "A-score") estimand as

$$\theta_j^{\text{acc}}(\hat{\eta}_j; P) := E_P\left[S_{ij} \mid \hat{\eta}_j\right] = P(\{Y_{ij} = \hat{Y}_{ij}\} \mid \hat{\eta}_j) = P\left(\{Y_{ij} = \texttt{CLS}(Z_i, \hat{\eta}_j)\} \mid \hat{\eta}_j\right)$$

where $S_{ij} := \mathbf{1}\{Y_{ij} = \hat{Y}_{ij}\}$. Conditioning on $\hat{\eta}_j$ in the estimand of interest should be thought of as evaluating a fixed autointerp description. Importantly, as the notation emphasizes, the $\theta_j^{\text{acc}}(\hat{\eta}_j; P)$ tells us exactly the quality of the description $\hat{\eta}_j$ for the population of interest P, unlike other autointerp evaluation methods for which the underlying distribution that implicitly defines the evaluation score is highly opaque (c.f., simulation scoring).

We can easily construct a plug-in estimator of the autointerp accuracy (A-score) estimand; its unbiasedness, consistency, and asymptotic normality under the $\hat{\eta}_j$-conditional law are established in the following proposition.

Proposition 1 (Conditional inference on A-score)**.** *Let $m := |\mathcal{I}^{eval}|$. Define the A-score estimator for dictionary feature j as*

$$\hat{\theta}_j^{acc} := \frac{1}{m}\sum_{i \in \mathcal{I}^{eval}} S_{ij} = \frac{1}{m}\sum_{i \in \mathcal{I}^{eval}} \mathbf{1}\{Y_{ij} = \hat{Y}_{ij}\} = \frac{1}{m}\sum_{i \in \mathcal{I}^{eval}} \mathbf{1}\{Y_{ij} = \mathit{CLS}(Z_i, \hat{\eta}_j)\}.$$

Assume that $P(S_{ij} = 1 \mid \hat{\eta}_j) \in [\delta, 1-\delta]$ for some δ bounded away from 0 and 1, almost surely. Then we have that $E_P\left[\hat{\theta}_j^{acc} \mid \hat{\eta}_j\right] = \theta_j^{acc}(\hat{\eta}_j; P)$ almost surely and as $m \to \infty$

$$\sup_{t \in \mathbb{R}} \left| P\left(\sqrt{m}(\hat{\theta}_j^{acc} - \theta_j^{acc}(\hat{\eta}_j; P)) \leq t \mid \hat{\eta}_j\right) - P(N(0, \text{Var}(S_{ij} \mid \hat{\eta}_j)) \leq t \mid \hat{\eta}_j) \right| \xrightarrow{a.s.} 0.$$

Example 1 (Continued). The null hypothesis of no average treatment effect is rejected for dictionary feature 4, and the autointerp generation strategy produced by the researcher describes feature 4 as activating on "discussion of politics." To evaluate the quality of this description, the researcher runs a LLM-based classifier over a held-out dataset of experimental participant text responses; the LLM classifier is given the prompt "Does the following text example indicate discussion of politics? Return 1 or 0 for yes or no:" followed by a text response from the held-out dataset. The researcher then uses the output of the LLM classifier to compute the A-score, which measures the accuracy of this autointerp description-based classifier at predicting the presence of feature 4 in the held-out dataset. △

Remark 5 (Oracle detection scores)*.* Towards better understanding the information communicated to the researcher by the A-score, it is useful to consider the function $\texttt{CLS}^*(\cdot)$, which we can define implicitly as

$$\bar{\theta}_j^{\text{acc}}(\hat{\eta}_j; P) := \sup_{f \in \mathcal{F}} P\left(\{Y_{ij} = f(Z_i, \hat{\eta}_j)\} \mid \hat{\eta}_j\right) = P\left(\{Y_{ij} = \texttt{CLS}^*(Z_i, \hat{\eta}_j)\} \mid \hat{\eta}_j\right),$$

where $\mathcal{F}$ is the space of all viable classification prompts and LLMs able to execute them. That is, $\texttt{CLS}^*(\cdot)$ is the "oracle" detection scoring model setup, which when evaluated produces the highest interpretative accuracy score possible under a given description $\hat{\eta}_j$ under distribution P. Arguably, $\bar{\theta}_j^{\text{acc}}(\hat{\eta}_j; P)$ is more informative to the researcher than $\theta_j^{\text{acc}}(\hat{\eta}_j; P)$ about the

quality of $\hat{\eta}_j$, as $\bar{\theta}_j^{\text{acc}}(\hat{\eta}_j; P)$ purely quantifies the quality of $\hat{\eta}_j$, without the wedge introduced by classification failures stemming from $\texttt{CLS}(\cdot)$ itself. Because any feasible choice of $\texttt{CLS}(\cdot)$ yields

$$P\left(\{Y_{ij} = \texttt{CLS}(Z_i, \hat{\eta}_j)\} \mid \hat{\eta}_j\right) \leq P\left(\{Y_{ij} = \texttt{CLS}^*(Z_i, \hat{\eta}_j)\} \mid \hat{\eta}_j\right)$$

one can reasonably interpret $\hat{\theta}_j^{\text{acc}}$ as a conservative (under)estimate of the more informative quantity $\bar{\theta}_j^{\text{acc}}(\hat{\eta}_j; P)$. ◇

We may also define estimands for autointerp precision and recall (the "P-score" and "R-score," respectively):

$$\theta_j^{\text{prec}}(\hat{\eta}_j) := P\left(Y_{ij} = 1 \mid \hat{Y}_{ij} = 1, \hat{\eta}_j\right) = \frac{E[\mathbf{1}\{Y_{ij} = 1\}\mathbf{1}\{\hat{Y}_{ij} = 1\} \mid \hat{\eta}_j]}{E[\mathbf{1}\{\hat{Y}_{ij} = 1\} \mid \hat{\eta}_j]},$$

$$\theta_j^{\text{rec}}(\hat{\eta}_j) := P\left(\hat{Y}_{ij} = 1 \mid Y_{ij} = 1, \hat{\eta}_j\right) = \frac{E[\mathbf{1}\{Y_{ij} = 1\}\mathbf{1}\{\hat{Y}_{ij} = 1\} \mid \hat{\eta}_j]}{E[\mathbf{1}\{Y_{ij} = 1\} \mid \hat{\eta}_j]}.$$

As above, these estimands may be consistently estimated with appropriate asymptotically valid confidence intervals under the $\hat{\eta}_j$ conditional law using the natural plug-in estimators and the delta method.

Remark 6 (Inference on the best interpretation). If a researcher is interested in inference on the best performing autointerp description across multiple generation prompts or strategies—say among a set $\{\hat{\eta}_{j,1}, \hat{\eta}_{j,2}, \dots\}$ for any j of interest—then the evaluation performance should be estimated with appropriate "inference on winners" corrections such as those discussed in Andrews et al. [2024]. ◇

Further, recall that we are not typically interested in autointerp evaluation for a fixed j, but instead for a particular data-driven choice $\hat{\jmath} \in \hat{J}$ suggested by a high-dimensional selective inference procedure.[17] However, so long as discovery is also performed using the main estimation sample indexed by $\mathcal{I}^{\text{estim}}$, the researcher may treat each $\hat{\jmath}$ as fixed for the purposes of evaluation inference. Formally, for all $j \in [p]$,

$$E_P\left[S_{i\hat{\jmath}} \mid \hat{\eta}_{\hat{\jmath}}, \hat{\jmath} = j\right] = E_P\left[S_{ij} \mid \hat{\eta}_j, \hat{\jmath} = j\right] = E_P\left[S_{ij} \mid \hat{\eta}_j\right]$$

where the last equality follows because, under the proposed independently held-out data (sample splitting) scheme,

$$[\mathbf{1}\{Y_{ij} = \texttt{CLS}(Z_i, \hat{\eta}_j)\} \perp\!\!\!\perp \hat{\jmath}] \mid \hat{\eta}_j.$$

Remark 7 (Evaluation of human-generated descriptions). It is worth noting that this autointerp evaluation strategy can also serve as a general-purpose description evaluation technology, e.g., for carefully crafted human descriptions of feature activations in settings where there is a small number of discoveries. The scientific benefit of evaluation in such a setting is also clear: even if there is concern that human-made descriptions are reverse-engineered to support a presupposed conclusion, any description (human or machine made, motivated or

[17] Autointerpretation typically incurs non-negligible compute costs, and, as such, researchers will often only be interested in forming meaningful autointerpretations for the subset of discovered concepts.

not) that is not supported by the data will be surfaced through the evaluation scores. In this way, description evaluation can be seen as even more critical when automatic methods are not employed by the researcher to generate descriptions. ◇

3.6 Pseudocode Algorithm

In Appendix Section 6.4, we state the entire framework for interpretable discovery from unstructured data as a single pseudocode algorithm. We focus on an implementation with text data, sample splitting, two-sided testing, one-step inference, and autointerp evaluation via accuracy scoring (A-scores).

4 Empirical Applications

As an illustration of the framework proposed in this paper, we now reanalyze two recent works in empirical economics that pursue discovery from unstructured data, and show how new, principled, and interpretable discoveries may be made using the same exact data at low cost.

4.1 Bursztyn et al. [2023]

Mirroring Example 1, Bursztyn et al. [2023] study how the provision of "social cover" affects willingness to publicly dissent to socially stigmatized causes, and the perception of this dissent. As a key application of their theory, Bursztyn et al. [2023] run an information treatment experiment (Experiment 2) in which participants are told they have been matched with another (fictional) respondent that chose to join a campaign to defund the police (a plausible expression of dissent in liberal American politics at the time the experiment was conducted), and then show the participant a tweet that the matched respondent is said to have agreed to publicly post. This tweet has been randomized to either include social cover or no social cover, namely whether or not the tweet indicated that the matched respondent joined the campaign before reading an evidence-based article in support of it (the no social cover condition) or after (the social cover condition).[18]

An important outcome collected by this experiment is the participant's open-ended text response to the question "Why do you think your matched respondent chose to join the campaign to oppose defunding the police?" This open-ended text response is meant to capture the causal effect of social cover on the perception of dissent. As such, this is a setting where discovery is of interest: ideally, to form a holistic understanding of how social cover affects perception of dissent, we do not want to pre-specify what aspects of perception we are interested in, and want to discover any interpretable systematic differences that exist in the text responses across treatment and control groups. In order to make progress on this analysis of unstructured data, Bursztyn et al. [2023] compute a Pearson's χ^2 statistic for all phrases of up to three words per Gentzkow and Shapiro [2010], which they use as an index to rank the phrases that are most differentially expressed in each condition's open-ended

[18]See Online Appendix Figure B.3 of Bursztyn et al. [2023] for a schematic of the experimental design.

responses.[19] The interpretability of the results from this analysis, is, naturally, hindered by the coarseness of the featurization of the text as three word phrases, as well as the fact that there are no obvious estimands or inferential guarantees. The only qualitative conclusion gleaned by Bursztyn et al. [2023] from this quantitative exercise is: "we find that respondents in the Cover condition are more likely to use phrases related to the article or the associated evidence—for example, 'article,' 'read,' 'convincing,' or 'increase in crime.'"[20]

Can we make more (and more interpretable) discoveries with the framework proposed in this paper? To investigate, we implement a `Dict` function based on the Gemma Scope autoencoders for Google's Gemma 2 2B LLM [Lieberum et al., 2024], and specifically the SAE trained on the residual stream of layer 12.[21] We implement an autointerpretation pipeline as discussed in Section 3.5 using as an explainer LLM the lightweight, open-source reasoning model GPT-OSS 20B [OpenAI, 2025]. A held-out evaluation dataset is created by sample splitting, using 10% of the original dataset in Bursztyn et al. [2023].

The hypotheses of interest are, for all $j \in [p]$,

$$H_{0,j} : E[Y_{ij}(1) - Y_{ij}(0)] = E\left[\frac{W_i - \pi}{\pi(1-\pi)} Y_{ij}\right] := E[X_{ij}] = 0,$$

for $\pi = 0.5$, and where the event $\{W_i = 1\}$ indicates the social cover condition. This yields a dataset $\{X_{ij}\}_{i\in[n],j\in[p]}$, where $n = 930$ and $p = 11901$ (non-degenerate features). We implement the simple one-step version of the statistical procedure in Theorem 2 as per Algorithm 1, which attains a lower bound on the power achieved by the step-wise procedure.

We may first consider whether or not, as anticipated, simply controlling the FWER would be too conservative for the purposes of discovery in this setting. To explore this, we set $k = 1$ and $\alpha = 0.05$ (equivalently, we implement the high-dimensional FWER controlling procedure of Belloni et al. [2018]), and we obtain only two significant discoveries: a 13.2 percentage point causal effect on the presence of dictionary feature 3518 and a 13.4 percentage point causal effect on the presence of dictionary feature 3426. The autointerprations of these features paint a picture that corroborates the original findings of Bursztyn et al. [2023]: feature 3426 is described as activating on mentions of "defunding the police associated with increased violent crime" and feature 3518 on "mention of an article as a source of evidence." However, we only discover two features out of approximately 12000, and not much new insight relative to the original analysis—even if the discovery and interpretation of these features was transparent, inexpensive, and automatic.[22]

In Figure 1, we make discoveries now setting $k = 5$, still with $\alpha = 0.05$. As can be seen, even with control of k-FWER at small $k = 5$, we obtain 9 rejections of the null, or 9

[19] Specifically, it is a Pearson's χ^2 statistic for a null hypothesis that the propensity to use a given phrase is equal across conditions, per Gentzkow and Shapiro [2010]. We interpret these statistics as simply forming an index, however, because the results of these hypothesis tests are not reported, and no multiple hypothesis testing corrections are implemented.

[20] The Online Appendix Table B.11 of Bursztyn et al. [2023] contains the top ten characteristic phrases in each condition based on the χ^2 index.

[21] It is a well-documented observation that SAEs trained on the middle layer of a LLM seem to capture "coarser" concepts and features (see dicussion in, e.g., Templeton et al. [2024]), which is plausibly more suitable for analyzing broad themes in text.

[22] The off-the-shelf Gemma Scope autointerpretations echo the same: feature 3426 is described as activating on "references to news articles and reports" and feature 3518 on "references to articles."

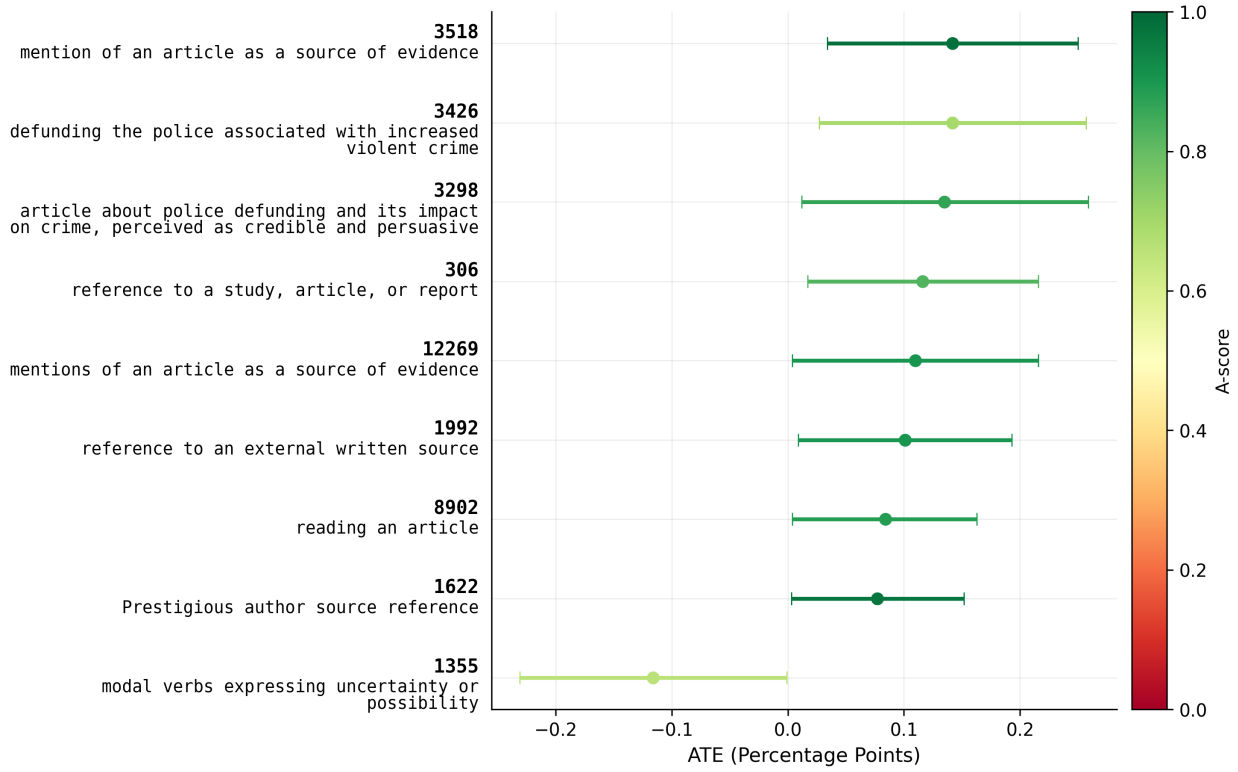

Figure 1: Discoveries for Experiment 2 of Bursztyn et al. (2023)

discoveries. By (large sample) guarantee of Theorem 2, the probability that 5 or more of these 9 discoveries is false is less than 5%. The previous features discovered with the FWER controlling procedure appear, as well as others that represent concepts both absent from the FWER controlled analysis and the original analysis in Bursztyn et al. [2023]. Though there are features that appear to activate on redundant concepts, in totality these discoveries contribute to a few entirely new characterizations of the causal effect of social cover on dissent. For example, feature 1622 responds to "Prestigious author source reference" and increases by nearly 8 percentage points in the treatment group, indicating that—exactly as Bursztyn et al. [2023] posit—the social cover mechanism plausibly relies on the perceived credibility of the article cited. Feature 1355 responds to "modal verbs expressing uncertainty or possibility," and its probability of activation is decreased by nearly 12 percentage points in the treatment group. This supports the idea that the causal effect of not having social cover is, to first order, generating uncertainty and speculation about dissent, as opposed to an alternative scenario where participants coalesce around a single narrative in the absence of cover. This insight is also congruous with the observation that few large negatively signed causal effects are discovered. The A-scores for all discoveries range from 0.98 to 0.66, indicating that some descriptions of discoveries deserve more skepticism than others, and that there is much scope for improving autointerp generation beyond small, open-source reasoning LLMs with the prompting strategy of Appendix Section 6.3.

Importantly, all of this analysis was performed automatically, in approximately one hour, on a single A100 GPU in Google Colab. Though this analysis was not preregistered, it is

possible to cheaply and automatically replicate these results, and assess their sensitivity to different autointerpretation strategies, dictionary learning methods, and LLMs. Notably, the space of possible concepts was not pruned in any way prior to this analysis, nor were they weighted in some way towards topics related to articles; the analysis was automatic and without human intervention, with no room for data snooping.

4.2 Stantcheva [2024]

Mirroring Example 2, Stantcheva [2024] conducts surveys on representative samples of the United States population investigating attitudes towards inflation. These surveys include open-ended text responses to a variety of questions, with the goal of discovering attitudes and opinions that the researcher should not (for concern about priming respondents) or could not (for lack of imagination) pre-specify.[23]

An important open-ended prompt that Stantcheva [2024] solicit an answer to is "High inflation is caused by..." In the original analysis, Stantcheva [2024] codes open-ended responses with a keyword-based method introduced in Ferrario and Stantcheva [2022] in which the researcher creates a list of topics and associated keywords in a discretionary way, which may range from "manual to semi-supervised or unsupervised" means. The result of applying this analysis to the open-ended responses to the above prompt yields Figure 3 in Stantcheva [2024], which finds, among other things, that mentions of "Biden and the administration," "Greed," "Monetary policy," "Fiscal policy," "War and foreign policy," "Demand vs supply," and "Supply-side mechanisms (other than input prices)" appear in more than 5% of all responses.

Do we make the same discoveries automatically if we apply the proposed framework instead? Do we make more and new interpretable discoveries? To find out, we use the same `Dict` function and explainer LLM as in Section 4.1, but reduce the space of features by half, filtering out all features that, in the corpus on which the Gemma Scope SAEs were trained, had greater than median empirical token activation frequency. This choice of dimensionality reduction is meant to screen out features that activate in many texts across domains, e.g., features related to grammatical aspects of text, which was handled implicitly in the previous causal analysis by virtue of differencing. (Ideally, such a dimensionality reduction choice would be preregistered, to prevent cherry-picking by filtering.) Again we split the sample randomly to form a held-out evaluation dataset consisting of 10% of the original dataset analyzed in Stantcheva [2024].

We are interested in testing hypotheses $H_{0,j} : E[X_{ij}] := E[Y_{ij}] = 0$, i.e., we test the probability of a given feature activating in the population of open-ended responses to the above prompt in the United States. This yields a dataset $\{X_{ij}\}_{i\in[n],j\in[p]}$ with $n = 453$ and $p = 3826$ non-degenerate features—still a very high-dimensional inference problem. We now apply 5-FWER control with $\alpha = 0.05$ based on the result in Theorem 2 per Algorithm 1, and we make 744 interpretable discoveries.

In Figure 2, we report point estimates and simultaneous confidence intervals for all discoveries with a generalized simultaneous lower confidence bound above 20%. Notably, we

[23]See, e.g., Haaland et al. [2024] for more discussion of the benefits of open-ended survey questions for understanding economic behavior.

discover, automatically, with no manual pruning or interpretation required, many sensible commonplace themes related to the topic at hand: feature 9804 responds to discussion of the economy; feature 13447 responds to mentions of inflation; and feature 13574 activates on broad discussion of negative economic conditions. However, we also recover many of the topics from the original analysis in Stantcheva [2024] without any human discretion or intervention: feature 4192 responds to discussion of monetary policy; feature 14747 responds to discussion of supply-side issues; feature 104 activates on discussion of excess and overabundance; feature 8316 responds to fiscal policy mentions; and feature 5719 responds to economic criticisms leveled at the current government. Undoubtedly, many other topics are recovered as we look deeper down the list of discoveries, beyond these 11 largest. However, even in this top 11, we learn that there is a great deal of pure uncertainty expressed in these open-ended responses based on the activation of feature 4794, and also that macroeconomic indicators get frequently discussed by virtue of feature 11036.

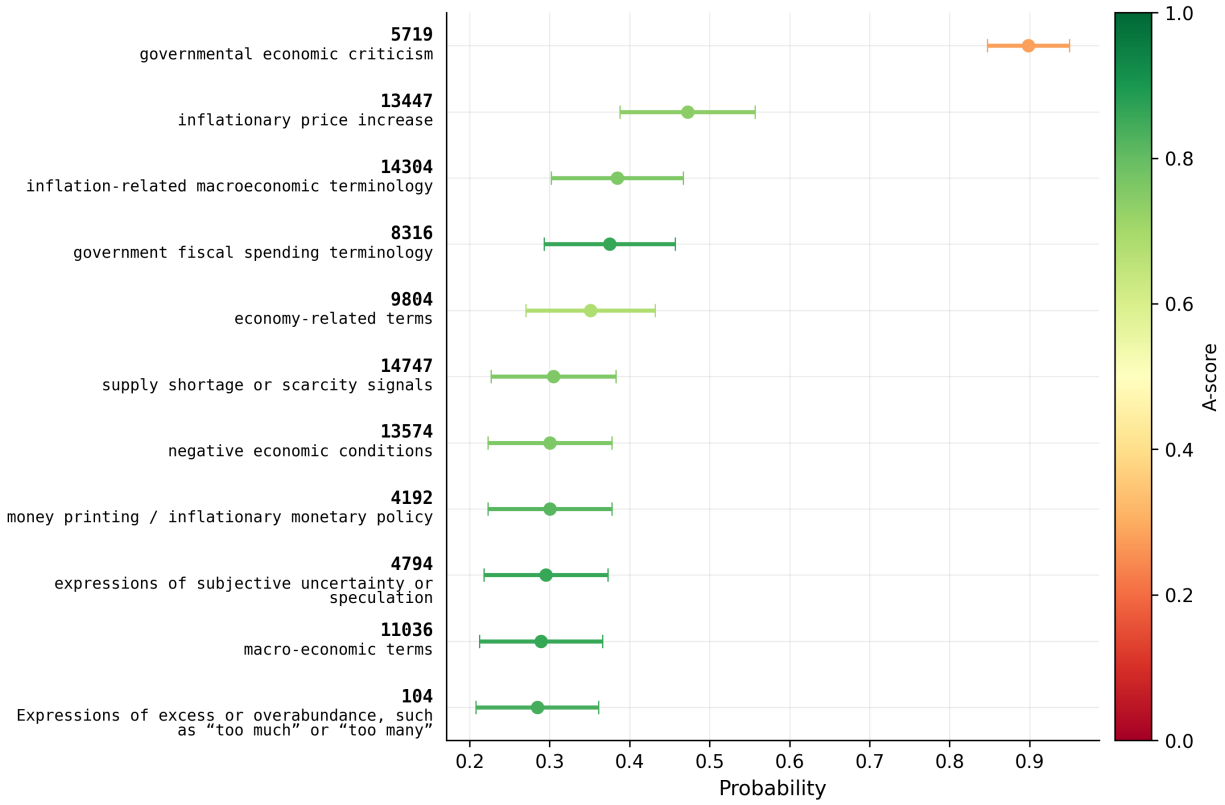

Figure 2: Ranked Discoveries for "High inflation is caused by..." in Stantcheva (2024)

Once again, this procedure was cheap and fast to implement, and could be quickly replicated and analyzed for sensitivity by any other researcher. Beyond filtering out greater than median activating dictionary features, no other choices were made to delimit the space of possible discoveries, or focus them in some way on the space of economic- or inflation-relevant concepts; the exact same explainer LLM and SAE that yielded the results of Section 4.1 yielded the results in this section.

5 Conclusion

Existing literature in empirical economics and econometrics has long suggested the importance of open-ended discovery from high-dimensional or unstructured data. The framework proposed in this paper shows how new statistical procedures for high-dimensional multiple hypothesis testing, when combined with the latest innovations in interpretability methods for machine learning models, can facilitate open-ended, interpretable discovery at scale with higher practicality and higher fidelity than previously possible.

The proposed framework also illustrates how treating (automatic) feature interpretation evaluation as a formal statistical inference problem can yield several scientific benefits in the process of making discoveries. A well-defined and intuitive quantitative method for evaluating the quality of feature descriptions is important in the context of the proposed framework, as it is these descriptions on which the scientific conclusions of discovery ultimately rest. By formulating autointerp evaluation as a statistical inference problem, the importance of evaluation with respect to the population of interest P becomes salient (as opposed to some other distribution implicitly defined through some other collection of unstructured data, common in most other autointerp evaluation methods); the need for sample splitting becomes clear; the role for confidence intervals becomes legible by appeal to sampling uncertainty (c.f., Miller [2024]); and we further unlock useful connections to the literature on post-selection inference, such that we may discipline settings evaluating a multiplicity of interpretations (c.f., Andrews et al. [2024]).

Importantly, the automaticity of the proposed framework makes it subject to very few researcher degrees of freedom, making it resilient to cherry-picking and data snooping without compromising the purpose of discovery in the first place. This is especially true if one couples this framework with minimal preregistration efforts, e.g., simply publicly declaring what dictionary learning model, LLM, and automatic interpretation pipeline (explainer LLM and prompts) are to be used for the analysis, and if any of the features are to be filtered on principled grounds.

The proposed framework is most naturally viewed as one tool of many in the empirical researcher's toolkit for making discoveries from unstructured data. Using this framework alongside others that researchers may already be implementing is entirely complementary, and would only serve to deepen insights into possible interpretations of inference on unstructured data.

References

Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. The Voice of Monetary Policy. *American Economic Review*, 113(2):548–584, February 2023. ISSN 0002-8282. doi: 10.1257/aer.20220129. URL https://pubs.aeaweb.org/doi/10.1257/aer.20220129.

Allen Hu and Song Ma. Persuading Investors: A Video-Based Study. *The Journal of Finance*, 80(5):2639–2688, October 2025. ISSN 0022-1082, 1540-6261. doi: 10.1111/jofi.13471. URL https://onlinelibrary.wiley.com/doi/10.1111/jofi.13471.

Ingar Haaland, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart. Understanding Economic Behavior Using Open-ended Survey Data. Technical Report w32421, National Bureau of Economic Research, Cambridge, MA, May 2024. URL http://www.nber.org/papers/w32421.pdf.

Jens Ludwig and Sendhil Mullainathan. Machine Learning as a Tool for Hypothesis Generation. *The Quarterly Journal of Economics*, 139(2):751–827, March 2024. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjad055. URL https://academic.oup.com/qje/article/139/2/751/7515309.

Clément Gorin, Stephan Heblich, and Yanos Zylberberg. State of the Art: Economic Development Through the Lens of Paintings. Technical Report w33976, National Bureau of Economic Research, Cambridge, MA, June 2025. URL http://www.nber.org/papers/w33976.pdf.

David Lagakos, Stelios Michalopoulos, and Hans-Joachim Voth. American Life Histories. Technical Report w33373, National Bureau of Economic Research, Cambridge, MA, January 2025. URL http://www.nber.org/papers/w33373.pdf.

Peter Bergman, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F. Katz, and Christopher Palmer. Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice. *American Economic Review*, 114(5):1281–1337, May 2024. ISSN 0002-8282. doi: 10.1257/aer.20200407. URL https://pubs.aeaweb.org/doi/10.1257/aer.20200407.

Patrick Krause, Elizabeth Rhodes, Sarah Miller, Alexander Bartik, David Broockman, and Eva Vivalt. The Impact of Unconditional Cash Transfers on Parenting and Children. Technical Report w34040, National Bureau of Economic Research, Cambridge, MA, July 2025. URL http://www.nber.org/papers/w34040.pdf.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, November 2023. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/10.1126/science.adi6000. Publisher: American Association for the Advancement of Science.

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Large Language Models: An Applied Econometric Framework, December 2024. URL http://arxiv.org/abs/2412.07031. arXiv:2412.07031 [econ].

Jacob Carlson and Melissa Dell. A Unifying Framework for Robust and Efficient Inference with Unstructured Data, 2025. URL https://arxiv.org/abs/2505.00282. Version Number: 2.

Leonardo Bursztyn, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. Justifying Dissent. *The Quarterly Journal of Economics*, 138(3):1403–1451, June 2023. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjad007. URL https://academic.oup.com/qje/article/138/3/1403/7000850.

Stefanie Stantcheva. Why Do We Dislike Inflation? Technical Report w32300, National Bureau of Economic Research, Cambridge, MA, April 2024. URL http://www.nber.org/papers/w32300.pdf.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Joseph P. Romano and Michael Wolf. Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4), August 2007. ISSN 0090-5364. doi: 10.1214/009053606000001622. URL https://projecteuclid.org/journals/annals-of-statistics/volume-35/issue-4/Control-of-generalized-error-rates-in-multiple-testing/10.1214/009053606000001622.full.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically Interpreting Millions of Features in Large Language Models, 2024. URL https://arxiv.org/abs/2410.13928. Version Number: 3.

Arun K. Kuchibhotla, John E. Kolassa, and Todd A. Kuffner. Post-Selection Inference. *Annual Review of Statistics and Its Application*, 9(1):505–527, March 2022. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-100421-044639. URL https://www.annualreviews.org/doi/10.1146/annurev-statistics-100421-044639.

Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Towards universality: Studying mechanistic similarity across language model architectures. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2J18i8T0oI.

Sendhil Mullainathan and Jann Spiess. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106, May 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.87. URL https://pubs.aeaweb.org/doi/10.1257/jep.31.2.87.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as Data. *Journal of Economic Literature*, 57(3):535–574, September 2019a. ISSN 0022-0515. doi: 10.1257/jel.20181020. URL https://pubs.aeaweb.org/doi/10.1257/jel.20181020.

Elliott Ash and Stephen Hansen. Text Algorithms in Economics. *Annual Review of Economics*, 15(1):659–688, September 2023. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev-economics-082222-074352. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-082222-074352.

Melissa Dell. Deep Learning for Economists. *Journal of Economic Literature*, 63(1):5–58, March 2025. ISSN 0022-0515, 2328-8175. doi: 10.1257/jel.20241733. URL https://pubs.aeaweb.org/doi/10.1257/jel.20241733.

Ashesh Rambachan, Rahul Singh, and Davide Viviano. Program Evaluation with Remotely Sensed Outcomes, 2024. URL https://arxiv.org/abs/2411.10959. Version Number: 2.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340, 2019b. ISSN 0012-9682. doi: 10.3982/ECTA16566. URL https://www.econometricsociety.org/doi/10.3982/ECTA16566.

Laura Battaglia, Timothy Christensen, Stephen Hansen, and Szymon Sacher. Inference for Regression with Variables Generated by AI or Machine Learning, 2024. URL https://arxiv.org/abs/2402.15585. Version Number: 5.

Jens Ludwig, Sendhil Mullainathan, and Jann Spiess. Machine-Learning Tests for Effects on Multiple Outcomes, 2017. URL https://arxiv.org/abs/1707.01473. Version Number: 2.

Iman Modarressi, Jann Spiess, and Amar Venugopal. Causal Inference on Outcomes Learned from Text, 2025. URL https://arxiv.org/abs/2503.00725. Version Number: 1.

Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. Hypothesis Testing in Econometrics. *Annual Review of Economics*, 2(1):75–104, September 2010. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev.economics.102308.124342. URL https://www.annualreviews.org/doi/10.1146/annurev.economics.102308.124342.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101. Publisher: [Royal Statistical Society, Oxford University Press].

Yoav Benjamini and Daniel Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2674075. Publisher: Institute of Mathematical Statistics.

Victor Chernozhuokov, Denis Chetverikov, Kengo Kato, and Yuta Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5), October 2022. ISSN 0090-5364. doi: 10.1214/22-AOS2193. URL https://projecteuclid.org/journals/annals-of-statistics/volume-50/issue-5/Improved-central-limit-theorem-and-bootstrap-approximations-in-high-dimensions/10.1214/22-AOS2193.full.

Yixi Ding, Qizhai Li, Yuke Shi, and Liuquan Sun. Gaussian Multiplier Bootstrap Procedure for the k-th Largest Coordinate of High-Dimensional Statistics, 2025. URL https://arxiv.org/abs/2508.14400. Version Number: 1.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato. High-Dimensional Econometrics and Regularized GMM, 2018. URL https://arxiv.org/abs/1806.01888. Version Number: 2.

Weidong Liu and Qi-Man Shao. Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics*, 42 (5), October 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1249. URL https://projecteuclid.org/journals/annals-of-statistics/volume-42/issue-5/Phase-transition-and-regularized-bootstrap-in-large-scale-t-tests/10.1214/14-AOS1249.full.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse Autoencoders for Hypothesis Generation, 2025. URL https://arxiv.org/abs/2502.04382. Version Number: 3.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, 2017. URL https://arxiv.org/abs/1702.08608. Version Number: 2.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse Autoencoders Do Not Find Canonical Units of Analysis, February 2025. URL http://arxiv.org/abs/2502.04878. arXiv:2502.04878 [cs].

Kenny Peng, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg. Use Sparse Autoencoders to Discover Unknown Concepts, Not to Act on Known Concepts, 2025. URL https://arxiv.org/abs/2506.23845. Version Number: 1.

Gonçalo Paulo, Stepan Shabalin, and Nora Belrose. Transcoders Beat Sparse Autoencoders for Interpretability, 2025. URL https://arxiv.org/abs/2501.18823. Version Number: 2.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A Multimodal Automated Interpretability Agent, 2024. URL https://arxiv.org/abs/2404.14394. Version Number: 2.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders, 2024. URL https://arxiv.org/abs/2407.14435. Version Number: 3.

Ahmed Abdulaal, Hugo Fry, Nina Montaña-Brown, Ayodeji Ijishakin, Jack Gao, Stephanie Hyland, Daniel C. Alexander, and Daniel C. Castro. An X-Ray Is Worth 15 Features: Sparse Autoencoders for Interpretable Radiology Report Generation, 2024. URL https://arxiv.org/abs/2410.03334. Version Number: 1.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE), 2024. URL https://arxiv.org/abs/2402.10376. Version Number: 2.

Hugo Fry. Towards multimodal interpretability: Learning sparse interpretable features in vision transformers, Apr 2024. URL https://www.lesswrong.com/posts/iYFuZo9BMvr6GgMs5/case-study-interpreting-manipulating-and-controlling-clip. Accessed: 2024-05-16.

Gytis Daujotas. Case study: Interpreting, manipulating, and controlling clip with sparse autoencoders, Aug 2024. URL https://www.lesswrong.com/posts/iYFuZo9BMvr6GgMs5/case-study-interpreting-manipulating-and-controlling-clip. Accessed: 2025-10-03.

Daniel Pluth, Yu Zhou, and Vijay K. Gurbani. Sparse Autoencoder Insights on Voice Embeddings, 2025. URL https://arxiv.org/abs/2502.00127. Version Number: 1.

Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike. High-Dimensional Data Bootstrap. *Annual Review of Statistics and Its Application*, 10(1):427–449, March 2023. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-040120-022239. URL https://www.annualreviews.org/doi/10.1146/annurev-statistics-040120-022239.

Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. FORMALIZED DATA SNOOPING BASED ON GENERALIZED ERROR RATES. *Econometric Theory*, 24 (02), April 2008. ISSN 0266-4666, 1469-4360. doi: 10.1017/S0266466608080171. URL http://www.journals.cambridge.org/abstract_S0266466608080171.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4), July 2017. ISSN 0091-1798. doi: 10.1214/16-AOP1113. URL https://projecteuclid.org/journals/annals-of-probability/volume-45/issue-4/Central-limit-theorems-and-bootstrap-in-high-dimensions/10.1214/16-AOP1113.full.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on Winners. *The Quarterly Journal of Economics*, 139(1):305–358, January 2024. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjad043. URL https://academic.oup.com/qje/article/139/1/305/7276491.

Matthew Gentzkow and Jesse M. Shapiro. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71, January 2010. ISSN 0012-9682. doi: 10.3982/ECTA7195. URL https://doi.org/10.3982/ECTA7195. Publisher: John Wiley & Sons, Ltd.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2, 2024. URL https://arxiv.org/abs/2408.05147. Version Number: 2.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Beatrice Ferrario and Stefanie Stantcheva. Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions. *AEA Papers and Proceedings*, 112:163–169, May 2022. ISSN 2574-0768, 2574-0776. doi: 10.1257/pandp.20221071. URL https://pubs.aeaweb.org/doi/10.1257/pandp.20221071.

Evan Miller. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations, 2024. URL https://arxiv.org/abs/2411.00640. Version Number: 1.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 1 edition, September 2018. ISBN 978-1-108-23159-6 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://www.cambridge.org/core/product/identifier/9781108231596/type/book.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, December 2022. ISSN 2049-8772. doi: 10.1093/imaiai/iaac012. URL https://academic.oup.com/imaiai/article/11/4/1389/6612958.

6 Appendix

6.1 Replication

The results from the empirical applications section of this paper can be replicated using the open source replication data from Bursztyn et al. [2023] and Stantcheva [2024] in addition to the Jupyter notebook hyperlinked here. This Jupyter notebook can readily be modified to make discoveries on other unstructured datasets.

6.2 Supplementary Tables

Table 1, shown below, supplements the information in Figure 1. Table 2 similarly supplements Figure 2.

Table 1: Discoveries for Experiment 2 of Bursztyn et al. (2023)

Feature	$\widehat{\text{ATE}}$	Description ("Feature activates highly on...")	t-stat	A-score
306	0.116	response to evidence that defunding police increases violent crime	4.518	0.709
1355	-0.116	modal verbs that express uncertainty or possibility	3.909	0.650
1622	0.077	references to a reputable source or author attribution	4.010	0.971
1992	0.101	reference to an article	4.244	0.932
3298	0.135	mentions of an article about police defunding and its impact on violent crime	4.240	0.883
3426	0.142	linking police defunding to increased violent crime	4.797	0.718
3518	0.142	article reference	5.072	0.515
8902	0.084	reading an article	4.080	0.864
12269	0.110	reference to an article used as evidence	4.004	0.893

Note: k-FWER control at $k = 5$ at level $\alpha = 5\%$.

Table 2: Ranked Discoveries for "High inflation is caused by..." in Stantcheva (2024)

Feature	Description ("Feature activates highly on...")	Sim. CI, Lower	Sim. CI, Upper	A-score
104	Expressions of excess or overabundance...	0.208	0.361	0.840
4192	money printing / inflationary monetary policy	0.223	0.378	0.820
4794	expressions of subjective uncertainty or speculation	0.218	0.373	0.860
5719	governmental economic criticism	0.847	0.950	0.280
8316	government fiscal spending terminology	0.293	0.457	0.860
9804	economy-related terms	0.270	0.432	0.680
11036	macro-economic terms	0.212	0.366	0.860
13447	inflationary price increase	0.388	0.557	0.740
13574	negative economic conditions	0.223	0.378	0.760
14304	inflation-related macroeconomic terminology	0.302	0.467	0.760
14747	supply shortage or scarcity signals	0.227	0.383	0.760

Note: k-FWER control at $k = 5$ at level $\alpha = 5\%$. The full description of feature 104 is "Expressions of excess or overabundance, such as 'too much' or 'too many.'" Autointerp descriptions are otherwise unabridged and unmodified.

6.3 Autointerpretation

Building autointerp pipelines for scalable description of features produced from dictionary learning methods is an active area of AI/ML research. Most best practices rely on leveraging LLMs themselves to interpret the features, per the pioneering work of Bills et al. [2023].

Heuristically, most of these pipelines operate by collecting text samples on which a given feature activates the most in a particular corpus (as well as perhaps sampling additional text samples across the empirical distribution of activation strengths) and weaving information about these activations—alongside the activating text—into prompts that LLMs are asked to interpret. Though autointerp descriptions are available for all features for the Gemma Scope models, they are learned based on the distribution of Gemma's pretraining data, and as such benefit from being refined through a tailored autointerpretation strategy based on text samples from the unstructured dataset distribution P.

For the purposes of generation, we adapt the methods of Paulo et al. [2024] using prompts of the form:

```
{"role": "system", "content": """
    You are a meticulous AI researcher conducting an important
    investigation into patterns found in language.
"""
},
{"role": "user", "content": f"""
    When a corpus of texts was passed through a LLM, a particular neuron
    most activated on the following examples, and specifically on the text
    delimited << like this >>. Provide a single phrase description of what
    the neuron likely responds to (in any corpus, not just this one), and
    delimit it as [[your concise description here]]. Do not mention the
    marker tokens ($<<$ $>>$) in your interpretation. The examples are:
    {texts}
"""
}
```

where `texts` is a concatenation of L texts—modified with delimiters as the prompt suggests—associated with the tokens that activated most highly on a given feature in the main sample of the researcher's unstructured dataset. We use $L = 20$ in the empirical examples of this paper, and the open-source reasoning model GPT-OSS 20B [OpenAI, 2025] to perform interpretations. GPT-OSS 20B's reasoning effort is set to "medium."

For the purposes of evaluation, we again adapt the detection scoring prompting strategy of Paulo et al. [2024], using prompts of the form:

```
{"role": "system", "content": """
    You are an intelligent and meticulous linguistics researcher.

    You will be provided a certain latent attribute of text, such as
    ``male pronouns" or ``text with negative sentiment".
```

```
        You will then be given a text example. Your task
        is to determine if the example possess the latent attribute.

        Return 1 if the text possess the latent attribute,
        and return 0 otherwise. Return only this number.
    """
    },
    {"role": "user", "content": f"""
        LATENT ATTRIBUTE: {description}
        TEXT EXAMPLE: {text}
    """
    }
```

where `description` is an autointerp description being evaluated and `text` is a text sample being analyzed for the presence of the relevant feature. We again use the open-source reasoning model GPT-OSS 20B [OpenAI, 2025] to perform evaluations; GPT-OSS 20B's reasoning effort is set to "low."

6.4 Pseudocode Algorithm

Algorithm 1 Interpretable Discovery from Text Data with k-FWER Control (Simplified)

Require: Data $\{(W_i, Z_i)\}_{i=1}^{n+m}$ drawn i.i.d. from distribution P; dictionary function $\texttt{Dict} : \mathcal{Z} \to \{0,1\}^{\tilde{p}}$; map $h : \mathcal{W} \times \{0,1\} \to \mathbb{R}$; level α; error control k; explainer LLM $\mathcal{E}$; eval dataset size m; bootstrap iterations β.

Require: $k \ll \tilde{p}$.

Split sample

Randomly split $[n+m]$ into $\mathcal{I}^{\text{estim}}$ and $\mathcal{I}^{\text{eval}}$ with $|\mathcal{I}^{\text{estim}}| = n$ and $|\mathcal{I}^{\text{eval}}| = m$.

Create feature dictionaries

for $i = 1, \ldots, n+m$ **do**

  $Y_i \leftarrow \texttt{Dict}(Z_i) \in \{0,1\}^{\tilde{p}}$.

end for

Drop degenerate features

Initialize $\mathcal{J} \leftarrow \emptyset$.

for $j = 1, \ldots, \tilde{p}$ **do**

  If $\exists\, i$ such that $Y_{ij} \neq 0$, $\mathcal{J} \leftarrow \mathcal{J} \cup \{j\}$.

end for

$p \leftarrow |\mathcal{J}|$.

Create feature-level estimates and t-stats

for $j \in \mathcal{J}$ **do**

  For $i \in \mathcal{I}^{\text{estim}} : X_{ij} \leftarrow h(W_i, Y_{ij})$.

  $\hat{\theta}_j \leftarrow \frac{1}{n} \sum_{i \in \mathcal{I}^{\text{estim}}} X_{ij}; \hat{\Sigma}_{jj} \leftarrow \frac{1}{n} \sum_{i \in \mathcal{I}^{\text{estim}}} (X_{ij} - \hat{\theta}_j)^2$.

end for

$T_{n,j} \leftarrow \sqrt{n}\, \hat{\theta}_j / \sqrt{\hat{\Sigma}_{jj}}$.

High-dimensional selective inference

Critical value $\hat{c}_n(1-\alpha, k) \leftarrow 1-\alpha$ quantile of $|S_n^B|_{[k]}$ under bootstrap law, approx. using β multiplier bootstrap iterations with $\{X_i\}_{i \in \mathcal{I}^{\text{estim}}}$.

Selected set $\hat{J} \leftarrow \{j : |T_{n,j}| > \hat{c}_n(1-\alpha, k)\} \subseteq [p]$.

For $j \in [p]$: gen. sim. CI $\mathcal{C}_j \leftarrow \left[\hat{\theta}_j - \hat{c}_n(1-\alpha, k)\sqrt{\frac{\hat{\Sigma}_{jj}}{n}}, \hat{\theta}_j + \hat{c}_n(1-\alpha, k)\sqrt{\frac{\hat{\Sigma}_{jj}}{n}}\right] \in \mathbb{R}^2$.

Autointerpretation

Initialize $\texttt{CLS}$ using $\mathcal{E}$ and fixed classification prompt.

for $\hat{\jmath} \in \hat{J}$ **do**

  Generate descriptions $\hat{\eta}_{\hat{\jmath}} \leftarrow \mathcal{E}(\{Z_i\}_{i \in \mathcal{I}^{\text{estim}}})$ (e.g., implemented per App. Sec. 6.3).

  for $i \in \mathcal{I}^{\text{eval}}$ **do**

    $\hat{Y}_{i\hat{\jmath}} \leftarrow \texttt{CLS}(Z_i, \hat{\eta}_{\hat{\jmath}}) \in \{0,1\}; S_{i\hat{\jmath}} \leftarrow \mathbf{1}\{Y_{i\hat{\jmath}} = \hat{Y}_{i\hat{\jmath}}\}$.

  end for

  A-score $\hat{\theta}^{\text{acc}}_{\hat{\jmath}} \leftarrow \frac{1}{m} \sum_{i \in \mathcal{I}^{\text{eval}}} S_{i\hat{\jmath}}$.

  Optionally: form CI for $\theta^{\text{acc}}_{\hat{\jmath}}(\hat{\eta}_{\hat{\jmath}})$ using plug-estimators under $(\hat{\eta}_{\hat{\jmath}}, \hat{\jmath})$-conditional law.

end for

Output: $\{(j, \hat{\theta}_j, \mathcal{C}_j) : j \in [p]\}$ and $\{(\hat{\jmath}, \hat{\eta}_{\hat{\jmath}}, \hat{\theta}^{\text{acc}}_{\hat{\jmath}}) : \hat{\jmath} \in \hat{J}\}$.

6.5 Proofs

Before stating proofs of the results in the main text, we state three additional useful lemmas.

Lemma 3. *For a sequence of random variables U_n and for a deterministic sequence r_n,*

$$U_n = o_P(r_n) \iff P^B(|U_n/r_n| \geq \varepsilon) = o_P(1) \text{ for any } \varepsilon > 0.$$

Proof. Note that, by definition, we have that $U_n = o_P(r_n)$ if for any $\varepsilon > 0$ that

$$\lim_{n\to\infty} P(|U_n/r_n| \geq \varepsilon) = \lim_{n\to\infty} E[Z_{n,\varepsilon}] = 0$$

where $Z_{n,\varepsilon} := P(|U_n/r_n| \geq \varepsilon \mid X)$, as by the law of total expectation $P(|U_n/r_n| \geq \varepsilon) = E[Z_{n,\varepsilon}]$. By Markov's inequality, because $Z_{n,\varepsilon}$ is always positive, for any $\delta > 0$,

$$P(Z_{n,\varepsilon} \geq \delta) \leq \frac{E[Z_{n,\varepsilon}]}{\delta}.$$

Thus we have that

$$0 \leq \lim_{n\to\infty} P(Z_{n,\varepsilon} \geq \delta) \leq \frac{\lim_{n\to\infty} E[Z_{n,\varepsilon}]}{\delta} = 0$$

and $\lim_{n\to\infty} P(|Z_{n,\varepsilon}| \geq \delta) = 0$ for any $\varepsilon, \delta > 0$. As such, $P(|U_n/r_n| \geq \varepsilon \mid X) = P^B(|U_n/r_n| \geq \varepsilon) = o_P(1)$ for any $\varepsilon > 0$ if $U_n = o_P(r_n)$.

For the other direction, note further that if $P(|U_n/r_n| \geq \varepsilon \mid X) = o_P(1)$ then the boundedness of $Z_{n,\varepsilon}$ permits using the bounded convergence theorem[24] to show for any $\varepsilon > 0$

$$P(|U_n/r_n| \geq \varepsilon) = o(1).$$

□

Lemma 4. *For $\{X_i\}_{i\in[n]}$ independent sub-exponential random vectors of dimension p, with each $\|X_{ij}\|_{\psi_1} \leq B_n$ for $i \in [n]$ and $j \in [p]$, then, under Assumption R,*

$$\|S_n\|_\infty = \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - E[X_i]) \right\|_\infty = O_P\left(B_n \sqrt{\log p}\right).$$

Proof. We start by noting Proposition 2.9.2 of Vershynin [2018], which implies here that

$$P\left\{ \left| \frac{1}{n} \sum_{i=1}^n (X_{ij} - E[X_{ij}]) \right| \geq t \right\} \leq 2\exp\left[-c \min\left(\frac{nt^2}{B_n^2}, \frac{nt}{B_n} \right) \right].$$

Thus by the union bound,

$$P\left\{ \max_{j\in[p]} \left| \frac{1}{n} \sum_{i=1}^n (X_{ij} - E[X_{ij}]) \right| \geq t \right\} \leq 2p\exp\left[-c \min\left(\frac{nt^2}{B_n^2}, \frac{nt}{B_n} \right) \right].$$

[24] See, e.g., Lemma 10.7 at this link.

Defining that

$$\varepsilon := 2p \exp\left[-c \min\left(\frac{nt^2}{B_n^2}, \frac{nt}{B_n}\right)\right]$$

then we have that for $\tilde{c} = 1/\sqrt{c}$

$$\tilde{c}\frac{\log(2p/\varepsilon)}{n} = \min\left\{t^2/B_n^2, t/B_n\right\}$$

and therefore

$$t = \tilde{c}B_n\frac{\log(2p/\varepsilon)}{n} \vee \tilde{c}B_n\sqrt{\frac{\log(2p/\varepsilon)}{n}},$$

meaning that

$$P\left\{\max_{j\in[p]}\left|\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - E[X_{ij}])\right| \geq \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{n} \vee \sqrt{\frac{\log(2p/\varepsilon)}{n}}\right)\right\} \leq \varepsilon.$$

As such, we have

$$P\left\{\max_{j\in[p]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_{ij} - E[X_{ij}])\right| \geq \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{\sqrt{n}} \vee \sqrt{\log(2p/\varepsilon)}\right)\right\} \leq \varepsilon.$$

However, under Assumption R, $\log(p)/\sqrt{n} = o(1)$, so the Gaussian tail term dominates for large n, and thus for large n we may set $M_\varepsilon := \tilde{c}\sqrt{\frac{\log(2/\varepsilon)}{\log 3} + 1}$ and observe that

$$P\left\{\frac{\max_{j\in[p]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_{ij} - E[X_{ij}])\right|}{B_n\sqrt{\log p}} \geq M_\varepsilon\right\} \leq \varepsilon.$$

We then conclude using the definition of stochastic boundedness that

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - E[X_i])\right\|_\infty = O_P\left(B_n\sqrt{\log p}\right)$$

as claimed. □

Lemma 5. *For sequences of random variables U_n and V_n, if, uniformly in t, for $\delta_n = o(1), \nu_n = o(1)$,*

$$P(|P(U_n > t \mid X) - P(V_n > t)| \geq \delta_n) \leq \nu_n$$

then, uniformly in t,

$$|P(U_n > t \mid X) - P(V_n > t)| = o_P(1).$$

Proof. Define $d_n(X) := |P(U_n > t \mid X) - P(V_n > t)|$. We want to show that, for any fixed $\delta > 0$,

$$\lim_{n\to\infty} P(d_n(X) \geq \delta) = 0$$

(the definition of convergence in probability). Note that, given the definition $\delta_n = o(1)$, for some n^*, for all $n \geq n^*$ we have that $\delta_n < \delta$. Thus, define the events:

$$A_n := \{d_n(X) \geq \delta\}, \quad B_n := \{d_n(X) \geq \delta_n\}$$

For all $n \geq n^*$, $A_n \subseteq B_n$. Thus, for all $n \geq n^*$

$$P\{d_n(X) \geq \delta\} \leq P\{d_n(X) \geq \delta_n\}.$$

Notice that then, using that $\nu_n = o(1)$,

$$0 \leq \lim_{n\to\infty} P\{d_n(X) \geq \delta\} \leq \lim_{n\to\infty} P\{d_n(X) \geq \delta_n\} = 0.$$

Because this holds for any choice of $\delta > 0$ and any t, we have proven the stated result. □

Lemma 6. *Under Assumptions M and R,*

$$\|S_n^B\|_\infty = O_P(B_n\sqrt{\log p}).$$

Proof. Using Assumptions M and R, the proof of Lemma 4 implies that

$$P\left\{\|S_n\|_\infty \geq \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{\sqrt{n}} \vee \sqrt{\log(2p/\varepsilon)}\right)\right\} \leq \varepsilon.$$

Using Lemma 4.6 of Chernozhuokov et al. [2022] and Lemma 5, notice that

$$b_n(X) := P\left(\|S_n^B\|_\infty > \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{\sqrt{n}} \vee \sqrt{\log(2p/\varepsilon)}\right) \mid X\right) \leq \varepsilon + o_P(1). \tag{1}$$

Define the event $A_n := \{b_n(X) > \varepsilon + \eta\}$ for any $\eta > 0$. Then the earlier statement implies that $P(A_n) = o(1)$. Now, by the law of total expectation and the law of total probability,

$$\begin{aligned}
&P\left(\|S_n^B\|_\infty > \tilde{c}B_n\left(\frac{\log(2p/\varepsilon)}{\sqrt{n}} \vee \sqrt{\log(2p/\varepsilon)}\right)\right) \\
&= E[b_n(X)] \\
&= E[b_n(X) \mid A_n]P(A_n) + E[b_n(X) \mid A_n^c]P(A_n^c) \\
&= E[b_n(X) \mid A_n^c] + o(1) && \text{(Set } \eta = o(1)\text{)} \\
&\leq \varepsilon + o(1).
\end{aligned}$$

It then follows, using the same logic as in Lemma 4, along with Assumption R, that

$$\|S_n^B\|_\infty = O_P(B_n\sqrt{\log p}).$$

□

We now prove the results from the main text. For any set of real numbers A with $|A| > k$, we use the notation k-max(A) to denote the k-the largest element of A, following Romano and Wolf [2007].

Theorem 1 (High-dimensional k-FWER control for small k, one-sided)**.** *Consider the method of Algorithm 2.1 or 2.2 in Romano and Wolf [2007] with test statistics T_n of hypotheses $\{\tilde{H}_{0,j}\}_{j\in[p]}$ and critical values $\hat{c}_{n,K}(1-\alpha,k)$ given by the $1-\alpha$ quantile of $S^B_{n,K,[k]}$ under P^B. Assume that k is fixed (i.e., not growing with n,p). Then under Assumptions M and R:*

(i) $\limsup_{n,p\to\infty} k\text{-}FWER_P \leq \alpha$.

(ii) If $\tilde{H}_{0,j}$ is false and $\theta_j(P) \gg B_n\sqrt{\log p/n}$, then the probability that the step-down method rejects $\tilde{H}_{0,j}$ tends to 1.

Proof. We consider Algorithm 2.1 from Romano and Wolf [2007], which at each stage takes as inputs critical values $\hat{c}_{n,K}(1-\alpha,k)$ and a vector of test statistics T_n; we denote this as `RW-2.1`$(T_n, \hat{c}_{n,K}(1-\alpha,k))$.

The choice of critical values we will use is based on a bootstrap of the $1-\alpha$ quantile of the k-max:

$$\hat{c}_{n,K}(1-\alpha,k) := 1-\alpha \text{ quantile of } S^B_{n,K,[k]} \mid X.$$

As in Romano and Wolf [2007], define $I(P)$ to be the set of indices of true null hypotheses under P. Note that, for any $K \supset I(P)$,

$$\hat{c}_{n,K}(1-\alpha,k) \geq \hat{c}_{n,I(P)}(1-\alpha,k)$$

because for any $K \supset I(P)$, and under any distribution, $S^B_{n,K,[k]} \geq S^B_{n,I(P),[k]}$ almost surely (i.e., the k-max statistic can only get larger if we include more test statistics without dropping the others). As such, Theorem 2.1 (i) holds, and we conclude that `RW-2.1`$(T_n, \hat{c}_{n,K}(1-\alpha,k))$ delivers:

$$k\text{-FWER}_P \leq P\left\{k\text{-max}\left(T_{n,j} : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\}.$$

Specifically, then, it is sufficient to show that

$$\limsup_{n,p} P\left\{k\text{-max}\left(T_{n,j} : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\} \leq \alpha.$$

To continue to use the notation of Romano and Wolf [2007], let $\hat{\theta}_{n,j} := \bar{X}_{n,j}$. Since $\theta_j(P) \leq 0$ for $j \in I(P)$, it follows that, almost surely,

$$\begin{aligned} k\text{-max}\left(T_{n,j} : j \in I(P)\right) &= k\text{-max}\left(\sqrt{n}\hat{\theta}_{n,j} : j \in I(P)\right) \\ &\leq k\text{-max}\left(\sqrt{n}\left[\hat{\theta}_{n,j} - \theta_j(P)\right] : j \in I(P)\right) \\ &= k\text{-max}\left(S_{n,j} : j \in I(P)\right) \end{aligned}$$

and therefore

$$\begin{aligned} &P\left\{k\text{-max}\left(T_{n,j} : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\} \\ \leq &P\left\{k\text{-max}\left(S_{n,j} : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\}. \end{aligned}$$

If we can show that the limit of the quantity on the right-hand side is no greater than α, the proof is complete. However, Theorem 2.2 of Ding et al. [2025] delivers exactly that, for any K,

$$P\left\{S_{n,K,[k]} > \hat{c}_{n,K}\left(1-\alpha,k\right)\right\} \leq \alpha + o(1)$$

under Assumptions M and R. So, chaining inequalities, we note that

$$k\text{-FWER}_P \leq P\left\{S_{n,I(P),[k]} > \hat{c}_{n,I(P)}\left(1-\alpha,k\right)\right\} \leq \alpha + o(1)$$

and so

$$\limsup_{n,p\to\infty} k\text{-FWER}_P \leq \alpha$$

which is precisely what we wanted to show.

To prove the second statement, consider the $\tilde{H}_{0,j}$ corresponding to all $\theta_j(P) > 0$. Note that, for the Gaussian multiplier bootstrap, using Assumptions M and R, by the tail inequality Equation 1 in the proof of Lemma 6,

$$\hat{c}_{n,[p]}\left(1-\alpha,k\right) \leq \hat{c}_{n,[p]}\left(1-\alpha,1\right) = O_P\left(B_n\sqrt{\log p}\right).$$

Furthermore, each $S_{n,j} = \sqrt{n}\left[\hat{\theta}_{n,j} - \theta_j(P)\right]$ has a limiting distribution, so

$$T_{n,j} = \sqrt{n}\hat{\theta}_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_j(P)) + \sqrt{n}\theta_j(P) \xrightarrow{P} \infty.$$

However, because $\sqrt{n}$ grows faster than $B_n\sqrt{\log p}$ by Assumption R, we have that also

$$\frac{T_{n,j}}{B_n\sqrt{\log p}} \xrightarrow{P} \infty.$$

Therefore, with probability tending to one, $T_{n,j} > \hat{c}_{n,[p]}\left(1-\alpha,k\right)$, resulting in the rejection of $\tilde{H}_{0,j}$ in the first step of Algorithm 2.1, so long as $\theta_j(P)$ is fixed or approaches zero slower than a rate of $B_n\sqrt{\log p/n}$.

The asymptotic validity of the streamlined Algorithm 2.2 follows immediately from having proved this, as it does in the Romano and Wolf [2007] proof of Theorem 3.3, given fixed or slow shrinking $\theta_j(P)$, using the same logic as for the proof of the second statement, i.e., $\min\left(T_{n,j} : j \notin I(P)\right)$ is diverging at rate $\sqrt{n}$, and if any $\theta_j(P) = 0$ then $\max\left(T_{n,j} : j \in I(P)\right) = O_P(B_n\sqrt{\log p})$. □

Theorem 2 (High-dimensional k-FWER control for small k, two-sided)**.** *Consider the method of Algorithm 2.1 or 2.2 in Romano and Wolf [2007] with test statistics $|T_n|$ of hypotheses $\{H_{0,j}\}_{j\in[p]}$ and critical values $\hat{c}_{n,K}(1-\alpha,k)$ given by the $1-\alpha$ quantile of $|S^B_{n,K}|_{[k]}$ under P^B. Assume that k is fixed (i.e., not growing with n,p). Then under Assumptions M and R:*

(i) $\limsup_{n,p\to\infty} k\text{-}FWER_P \leq \alpha$.

(ii) If $H_{0,j}$ is false and $|\theta_j(P)| \gg B_n\sqrt{\log p/n}$, then the probability that the step-down method rejects $H_{0,j}$ tends to 1.

Proof. We again start by considering Algorithm 2.1 from Romano and Wolf [2007], which at each stage takes as inputs critical values $\hat{c}_{n,K}(1-\alpha,k)$ and a vector of test statistics $|T_n|$; we denote this as `RW-2.1`$(|T_n|, \hat{c}_{n,K}(1-\alpha,k))$.

The choice of critical values we will use is based on a bootstrap of the $1-\alpha$ quantile of the k-max of absolute values:

$$\hat{c}_{n,K}(1-\alpha,k) := 1-\alpha \text{ quantile of } |S^B_{n,K}|_{[k]} \mid X.$$

As in Romano and Wolf [2007], define $I(P)$ to be the set of indices of true null hypotheses under P. Note that, for any $K \supset I(P)$,

$$\hat{c}_{n,K}(1-\alpha,k) \geq \hat{c}_{n,I(P)}(1-\alpha,k)$$

because for any $K \supset I(P)$, and under any distribution, $|S^B_{n,K}|_{[k]} \geq |S^B_{n,I(P)}|_{[k]}$ almost surely (i.e., the k-max statistic can only get larger if we include more test statistics without dropping the others). As such, Theorem 2.1 (i) holds, and we conclude that `RW-2.1`$(|T_n|, \hat{c}_{n,K}(1-\alpha,k))$ delivers:

$$k\text{-FWER}_P \leq P\left\{k\text{-max}\left(|T_{n,j}| : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\}.$$

Specifically, then, it is sufficient to show that

$$\limsup_{n,p} P\left\{k\text{-max}\left(|T_{n,j}| : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\} \leq \alpha.$$

To continue to use the notation of Romano and Wolf [2007], let $\hat{\theta}_{n,j} := \bar{X}_{n,j}$. Since $\theta_j(P) = 0$ for $j \in I(P)$, it follows that, almost surely,

$$\begin{aligned} k\text{-max}\left(|T_{n,j}| : j \in I(P)\right) &= k\text{-max}\left(\sqrt{n}|\hat{\theta}_{n,j}| : j \in I(P)\right) \\ &= k\text{-max}\left(\sqrt{n}\left|\hat{\theta}_{n,j} - \theta_j(P)\right| : j \in I(P)\right) \\ &= k\text{-max}\left(|S_{n,j}| : j \in I(P)\right) \end{aligned}$$

and therefore

$$\begin{aligned} &P\left\{k\text{-max}\left(|T_{n,j}| : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\} \\ =&P\left\{k\text{-max}\left(|S_{n,j}| : j \in I(P)\right) > \hat{c}_{n,I(P)}(1-\alpha,k)\right\}. \end{aligned}$$

If we can show that the limit of the quantity on the right-hand side is no greater than α, the proof is complete. However, Theorem 2.2 and Remark 2 of Ding et al. [2025] delivers exactly that, for any K,

$$P\left\{|S_{n,K}|_{[k]} > \hat{c}_{n,K}\left(1-\alpha,k\right)\right\} \leq \alpha + o(1)$$

under Assumptions M and R. So, chaining inequalities, we note that

$$k\text{-FWER}_P \leq P\left\{|S_{n,I(P)}|_{[k]} > \hat{c}_{n,I(P)}\left(1-\alpha,k\right)\right\} \leq \alpha + o(1)$$

and so

$$\limsup_{n,p\to\infty} k\text{-FWER}_P \leq \alpha$$

which is precisely what we wanted to show.

To prove the second statement, note that, for the Gaussian multiplier bootstrap, using Assumptions M and R, by the tail inequality Equation 1 in the proof of Lemma 6, and using the insight from Remark 2 of Ding et al. [2025] that quantiles for absolute values are identical to quantiles for a related $2p$ dimensional problem without absolute values,

$$\hat{c}_{n,[p]}\left(1-\alpha, k\right) \leq \hat{c}_{n,[p]}\left(1-\alpha, 1\right) = O_P\left(B_n\sqrt{\log p}\right).$$

Now notice that each $S_{n,j} = \sqrt{n}\left[\hat{\theta}_{n,j} - \theta_j(P)\right]$ has a limiting distribution, so consider that for the $H_{0,j}$ corresponding to all $\theta_j(P) > 0$ that

$$T_{n,j} = \sqrt{n}\hat{\theta}_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_j(P)) + \sqrt{n}\theta_j(P) \xrightarrow{P} \infty$$

and that for the $H_{0,j}$ corresponding to all $\theta_j(P) < 0$ that

$$T_{n,j} = \sqrt{n}\hat{\theta}_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_j(P)) + \sqrt{n}\theta_j(P) \xrightarrow{P} -\infty.$$

However, because $\sqrt{n}$ grows faster than $B_n\sqrt{\log p}$ by Assumption R, we have that also for all $H_{0,j}$ for which $\theta_j(P) \neq 0$

$$\frac{|T_{n,j}|}{B_n\sqrt{\log p}} \xrightarrow{P} \infty.$$

Therefore, with probability tending to one, $|T_{n,j}| > \hat{c}_{n,[p]}\left(1-\alpha, k\right)$, resulting in the rejection of $H_{0,j}$ in the first step of Algorithm 2.1, so long as $|\theta_j(P)|$ is fixed or approaches zero slower than a rate of $B_n\sqrt{\log p/n}$.

The asymptotic validity of the streamlined Algorithm 2.2 follows immediately from having proved this, as it does in the Romano and Wolf [2007] proof of Theorem 3.3, given fixed or slow shrinking $\theta_j(P)$, using the same logic as for the proof of the second statement, i.e., $\min\left(|T_{n,j}| : j \notin I(P)\right)$ is diverging at rate $\sqrt{n}$, and if any $\theta_j(P) = 0$ then $\max\left(|T_{n,j}| : j \in I(P)\right) = O_P(B_n\sqrt{\log p})$. □

Lemma 1 (High-dimensional CLT for the small k-max coordinate of approximate means)**.** *Let $\hat{S}_n := S_n + R_n$, and assume that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$. Further assume that k is fixed (i.e., does not grow with n, p). If Assumptions M and R hold, then as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}}\left|P\left(\hat{S}_{n,[k]} \leq t\right) - P\left(N(0,\Sigma)_{[k]} \leq t\right)\right| \to 0.$$

Proof. The proof of this lemma proceeds following the strategy of Chernozhukov et al. [2023].

First, recall that the function $t \mapsto t_{[k]}$ is 1-Lipschitz wrt to the sup-norm, meaning almost surely

$$\left|\hat{S}_{n,[k]} - S_{n,[k]}\right| = \left|(S_n + R_n)_{[k]} - S_{n,[k]}\right| \leq \|R_n\|_\infty.$$

Consider the event $\{\|R_n\|_\infty \leq \epsilon\}$, as well as the event $\{(S_n + R_n)_{[k]} \leq t\}$. Then observe that

$$\{\|R_n\|_\infty \leq \epsilon\} \cap \{(S_n + R_n)_{[k]} \leq t\} \subseteq \{S_{n,[k]} \leq t + \epsilon\}$$

As such notice that

$$\{(S_n+R_n)_{[k]} \le t\} = (\{(S_n+R_n)_{[k]} \le t\} \cap \{\|R_n\|_\infty \le \epsilon\}) \cup (\{(S_n+R_n)_{[k]} \le t\} \cap \{\|R_n\|_\infty > \epsilon\})$$

and so

$$P(\hat{S}_{n,[k]} \le t) = P((S_n + R_n)_{[k]} \le t) \le P(S_{n,[k]} \le t + \epsilon) + P(\|R_n\|_\infty > \epsilon).$$

We then have, using Assumptions M and R in addition to the high-dimensional CLT for the k largest coordinate of Lemma A.6 in Ding et al. [2025] (which requires Assumption M):

$$\begin{aligned} P\left(\hat{S}_{n,[k]} \le t\right) &\le P\left(S_{n,[k]} \le t + \epsilon\right) + P\left(\|R_n\|_\infty > \epsilon\right) \\ &= P\left(N(0,\Sigma)_{[k]} \le t + \epsilon\right) + o(1) + P\left(\|R_n\|_\infty > \epsilon\right). \end{aligned}$$

(Lemma A.6 + Assumption R)

We now need to apply an anti-concentration result, which can be found as Proposition A.1 of Ding et al. [2025]. Letting $G := N(0,\Sigma)$, it states, for any $\tilde{t}$,

$$P(\tilde{t} - \tilde{\epsilon} \le G_{[k]} \le \tilde{t} + \tilde{\epsilon}) \le Ck\tilde{\epsilon}\sqrt{1 \vee \ln(p/\tilde{\epsilon})}.$$

Thus, letting $t := \tilde{t} - \tilde{\epsilon}$ and $\epsilon := 2\tilde{\epsilon}$,

$$P(t \le G_{[k]} \le t + \epsilon) \le \frac{1}{2}Ck\epsilon\sqrt{1 \vee \ln(2p/\epsilon)}$$

and so because k is fixed,

$$P(G_{[k]} \le t + \epsilon) = P(G_{[k]} \le t) + O\left(\epsilon\sqrt{1 \vee \ln(2p/\epsilon)}\right)$$

meaning that

$$P\left(\hat{S}_{n,[k]} \le t\right) \le P\left(N(0,\Sigma)_{[k]} \le t\right) + O(\epsilon\sqrt{1 \vee \log(2p/\epsilon)}) + o(1) + P\left(\|R_n\|_\infty > \epsilon\right).$$

(Proposition A.1)

Then we may choose $\epsilon = \epsilon_n = o(1/\sqrt{\log p})$ in such a way that we get that

$$P\left(\hat{S}_{n,[k]} \le t\right) \le P\left(N(0,\Sigma)_{[k]} \le t\right) + o(1).$$

From the reverse direction, note that

$$\{\|R_n\|_\infty \le \epsilon\} \cap \{(S_n + R_n)_{[k]} \le t - \epsilon\} \subseteq \{S_{n,[k]} \le t\}$$

so it also holds that similarly, partitioning $\{(S_n + R_n)_{[k]} \le t - \epsilon\}$ using $\{\|R_n\|_\infty \le \epsilon\}$ and its complement, that

$$P\left(\hat{S}_{n,[k]} \le t\right) \ge P\left(S_{n,[k]} \le t - \epsilon\right) - P\left(\|R_n\|_\infty > \epsilon\right),$$

and so using identical arguments as above we conclude that, uniformly in t,

$$\left|P\left(\hat{S}_{n,[k]} \le t\right) - P\left(N(0,\Sigma)_{[k]} \le t\right)\right| = o(1),$$

proving the stated result. □

Lemma 2 (High-dimensional bootstrap for the small k-max coordinate of approximate means)**.** *Let $\hat{S}_n^B := S_n^B + R_n$, and assume that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$. Further assume that k is fixed (i.e., does not grow with n, p). If Assumptions M and R hold, then as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}} \left| P^B\left(\hat{S}_{n,[k]}^B \le t\right) - P\left(N(0,\Sigma)_{[k]} \le t\right)\right| \xrightarrow{P} 0.$$

Proof. The proof of this lemma proceeds following the strategy of Lemma 1. Using the same logic as the proof of Lemma 1, note that

$$P^B(\hat{S}_{n,[k]}^B \le t) = P^B((S_n^B + R_n)_{[k]} \le t) \le P^B(S_{n,[k]}^B \le t + \epsilon) + P^B(\|R_n\|_\infty > \epsilon).$$

Observe that, from Ding et al. [2025] Lemma A.8 and Lemma 5, using Assumptions M and R and the triangle inequality, that

$$\sup_{t\in\mathbb{R}} \left| P^B\left(S_{n,[k]}^B \le t\right) - P\left(N(0,\Sigma)_{[k]} \le t\right)\right| = o_P(1).$$

This equation then plays the role of Lemma A.6 of Ding et al. [2025] in the proof of Lemma 1; the stated result proceeds from this observation, continuing with identical logic as in Lemma 1, noting that if $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$ then $P^B(\sqrt{\log p}\|R_n\|_\infty > \epsilon) = o_P(1)$ for all $\epsilon > 0$ by Lemma 3. □

Theorem 3 (High-dimensional CLT for the small k-max studentized coordinate)**.** *If Assumptions M and R hold, and $B_n = O(1)$ and k is fixed (i.e., does not grow with n, p), then, by application of Lemma 1, as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}} \left| P\left(\left(\hat{\Lambda}^{-1/2} S_n\right)_{[k]} \le t\right) - P\left(N(0,\Sigma_0)_{[k]} \le t\right)\right| \to 0.$$

Proof. Let $\hat{S}_n := \hat{\Lambda}^{-1/2} S_n$. To show this theorem, note that

$$R_n = \hat{\Lambda}^{-1/2} S_n - \Lambda^{-1/2} S_n = (\hat{\Lambda}^{-1/2} - \Lambda^{-1/2}) S_n,$$

and thus we can use the machinery of Lemma 1 to prove the desired result so long as we can show that $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$.

Note that, by the sub-multiplicative induced matrix norm inequality,

$$\left\|\left(\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}\right) S_n\right\|_\infty \le \|\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}\|_\infty \, \|S_n\|_\infty = \|\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}\|_{\max} \, \|S_n\|_\infty$$

where the last equality follows from the fact that the max elementwise norm is equal to the induced operator ∞-norm for diagonal matrices.

To control the first term on the right-hand side, we may turn to Kuchibhotla and Chakrabortty [2022], Theorem 4.2, which shows that if X_{ij} are sub-Weibull with parameter $\alpha = 1$ (i.e., sub-exponential, granted by Assumption M), then Assumption R ensures that

$$\|\widehat{\Lambda} - \Lambda\|_{\max} = o_P(1/\log^2 p).$$

To see this, note that the condition discussed in Remark 4.1 of Kuchibhotla and Chakrabortty [2022],

$$(\log p)^{2/\alpha-1/2} = o\left(\sqrt{n}(\log n)^{-2/\alpha}\right)$$

is satisfied under Assumption R, meaning that $\|\hat{\Sigma}-\Sigma\|_{\max} = O_P\left(\sqrt{\log p/n}\right)$ if $B_n = O(1)$ (where $\|\cdot\|_{\max}$ is the maximum elementwise norm). As a consequence, Assumption R also delivers that $\|\widehat{\Lambda}-\Lambda\|_{\max} \leq \|\hat{\Sigma}-\Sigma\|_{\max} = o_P\left(1/\log^2 p\right)$, and thus $\|\widehat{\Lambda}^{-1/2}-\Lambda^{-1/2}\|_{\max} = o_P\left(1/\log^2 p\right)$ as well using a Taylor expansion argument and part (ii) of Assumption M.

For the second term, note that, using Lemma 4 via Assumptions M and R, that $\|S_n\|_\infty = O_P(B_n\sqrt{\log p}) = O_P(\sqrt{\log p})$, where the last equality follows because $B_n = O(1)$ by assumption. Putting everything together then, we conclude

$$\left\|\left(\widehat{\Lambda}^{-1/2}-\Lambda^{-1/2}\right)S_n\right\|_\infty \leq o_P\left(1/\log^2 p\right)O_P(\sqrt{\log p}) = o_P(1/\log^{3/2}(p)) = o_P(1/\sqrt{\log p}).$$

□

Theorem 4 (High-dimensional bootstrap for the small k-max studentized coordinate)**.** *If Assumptions M and R hold, and $B_n = O(1)$ and k is fixed (i.e., does not grow with n, p), then, by application of Lemma 2, as $n, p \to \infty$*

$$\sup_{t\in\mathbb{R}}\left|P^B\left(\left(\hat{\Lambda}^{-1/2}S_n^B\right)_{[k]} \leq t\right) - P\left(N(0,\Sigma_0)_{[k]} \leq t\right)\right| \xrightarrow{P} 0.$$

Proof. The proof of this theorem proceeds just as the proof of Theorem 3. Let $\hat{S}_n^B := \hat{\Lambda}^{-1/2}S_n^B$ and

$$R_n = \hat{\Lambda}^{-1/2}S_n^B - \Lambda^{-1/2}S_n^B = (\hat{\Lambda}^{-1/2}-\Lambda^{-1/2})S_n^B,$$

and thus we can use Lemma 2 to prove the desired result so long as we can show that, sufficiently, $\|R_n\|_\infty = o_P(1/\sqrt{\log p})$.

Note then that almost surely, as in Theorem 3,

$$\left\|\left(\widehat{\Lambda}^{-1/2}-\Lambda^{-1/2}\right)S_n^B\right\|_\infty \leq \|\widehat{\Lambda}^{-1/2}-\Lambda^{-1/2}\|_{\max}\left\|S_n^B\right\|_\infty.$$

Using the same arguments as in the proof of Theorem 3, we conclude that $\|\widehat{\Lambda}^{-1/2}-\Lambda^{-1/2}\|_\infty = o_P(1/\log^2 p)$. By Lemma 6 we have that $\|S_n^B\|_\infty = O_P(B_n\sqrt{\log p})$, and under $B_n = O(1)$ then $\|S_n^B\|_\infty = O_P(\sqrt{\log p})$. Thus, we complete the proof as in Theorem 3. □

Proposition 1 (Conditional inference on A-score)**.** *Let $m := |\mathcal{I}^{eval}|$. Define the A-score estimator for dictionary feature j as*

$$\hat{\theta}_j^{acc} := \frac{1}{m}\sum_{i\in\mathcal{I}^{eval}} S_{ij} = \frac{1}{m}\sum_{i\in\mathcal{I}^{eval}} \mathbf{1}\{Y_{ij} = \hat{Y}_{ij}\} = \frac{1}{m}\sum_{i\in\mathcal{I}^{eval}} \mathbf{1}\{Y_{ij} = \texttt{CLS}(Z_i, \hat{\eta}_j)\}.$$

Assume that $P(S_{ij} = 1 \mid \hat{\eta}_j) \in [\delta, 1-\delta]$ for some δ bounded away from 0 and 1, almost surely. Then we have that $E_P\left[\hat{\theta}_j^{acc} \mid \hat{\eta}_j\right] = \theta_j^{acc}(\hat{\eta}_j; P)$ almost surely and as $m \to \infty$

$$\sup_{t\in\mathbb{R}}\left|P\left(\sqrt{m}(\hat{\theta}_j^{acc} - \theta_j^{acc}(\hat{\eta}_j;P)) \leq t \mid \hat{\eta}_j\right) - P(N(0, \mathrm{Var}(S_{ij} \mid \hat{\eta}_j)) \leq t \mid \hat{\eta}_j)\right| \xrightarrow{a.s.} 0.$$

Proof. To show the conditional unbiasedness of the estimator, observe that

$$E_P\left[\hat{\theta}_j^{\text{acc}} \mid \hat{\eta}_j\right] = \frac{1}{m}\sum_{i\in\mathcal{I}^{\text{eval}}} E_P[S_{ij} \mid \hat{\eta}_j] = \theta_j^{\text{acc}}(\hat{\eta}_j; P)$$

because S_{ij} is drawn i.i.d. under the $\hat{\eta}_j$-conditional law.

To show $\sqrt{m}$-consistency and asymptotic normality, we make use of Lemma 1 from Modarressi et al. [2025], a clean statement of a conditional Berry-Essen theorem for i.n.i.d. data. Noticing that for all $i \in \mathcal{I}^{\text{eval}}$ that $\psi_i := S_{ij} - \theta_j^{\text{acc}}(\hat{\eta}_j; P)$ are mean zero and independent, and that their third moments are bounded under the conditional law almost surely (as $S_{ij} \in \{0,1\}$ and $\theta_j^{\text{acc}}(\hat{\eta}_j; P) \in [0,1]$), using the assumption that $P(S_{ij} = 1 \mid \hat{\eta}_j) \in [\delta, 1-\delta]$, this lemma provides that

$$\sup_{t\in\mathbb{R}}\left|P\left(m^{-1/2}\sum_{i\in\mathcal{I}^{\text{eval}}}\psi_i \le t \mid \hat{\eta}_j\right) - P(N(0, \text{Var}(S_{ij} \mid \hat{\eta}_j)) \le t \mid \hat{\eta}_j)\right| \le C_0 \frac{\sum_{i\in\mathcal{I}^{\text{eval}}} E\left[|\psi_i|^3 \mid \hat{\eta}_j\right]}{\left(\sum_{i\in\mathcal{I}^{\text{eval}}} E\left[\psi_i^2 \mid \hat{\eta}_j\right]\right)^{3/2}}$$

for some universal constant $C_0 < \infty$, almost surely. The left-hand side of this inequality is exactly the left-hand side of the convergence statement in the proposition being proved. As for the right-hand side, note that then, given our previous observations and assumptions, $E\left[|\psi_i|^3 \mid \hat{\eta}_j\right] \le 1$ almost surely, and $E\left[\psi_i^2 \mid \hat{\eta}_j\right] = \text{Var}(S_{ij} \mid \hat{\eta}_j) \ge \delta(1-\delta) = C_1 > 0$, almost surely. Thus we have that

$$C_0 \frac{\sum_{i\in\mathcal{I}^{\text{eval}}} E\left[|\psi_i|^3 \mid \hat{\eta}_j\right]}{\left(\sum_{i\in\mathcal{I}^{\text{eval}}} E\left[\psi_i^2 \mid \hat{\eta}_j\right]\right)^{3/2}} \le C_0 \frac{m}{(C_1 m)^{3/2}} = \frac{C_0}{C_1^{3/2}} m^{-1/2} \asymp m^{-1/2}.$$

Taking $m \to \infty$ then proves the stated claim.

□