Making Interpretable Discoveries from Unstructured Data

A High-Dimensional Multiple Hypothesis Testing Approach

Jacob (Jake) Carlson

Harvard University

August 2026

arXiv:2511.01680

Code: `github.com/jscarlson/hdmht-discovery`

Making Interpretable Discoveries from Unstructured Data

- **Unstructured data** (text, image, audio, video, + location data) increasingly popular.
  - → Why? Novel source of interpretable + econ. meaningful measurements.

Making Interpretable Discoveries from Unstructured Data

- **Unstructured data** (text, image, audio, video, + location data) increasingly popular.
- Growing number of use cases about **unsupervised discovery** of **rich** features.
  1. Surveys where open-ended responses key to understanding beliefs (Stantcheva, 2024).
  2. RCTs where important outcomes reported in text or audio interviews (Bergman et al., 2024).
  3. Obs. analysis of language change under quasi-experiment (Hansen et al., 2018; Ash et al., 2026).

Making Interpretable Discoveries from Unstructured Data

- **Unstructured data** (text, image, audio, video, + location data) increasingly popular.
- Growing number of use cases about **unsupervised discovery** of **rich** features.
- Usual "qual. coding" approaches are intuitive, but suffer from many sci. + stat. problems.

Making Interpretable Discoveries from Unstructured Data

- **Unstructured data** (text, image, audio, video, + location data) increasingly popular.
- Growing number of use cases about **unsupervised discovery** of **rich** features.
- Usual "qual. coding" approaches are intuitive, but suffer from many sci. + stat. problems.
- **Contributions**: (1) stat. + sci. principled framework for discovery; (2) extensions of 2 frontier econometric frameworks; (3) new formalization of AI interp. framework.

Running Empirical Example: Productivity RCT

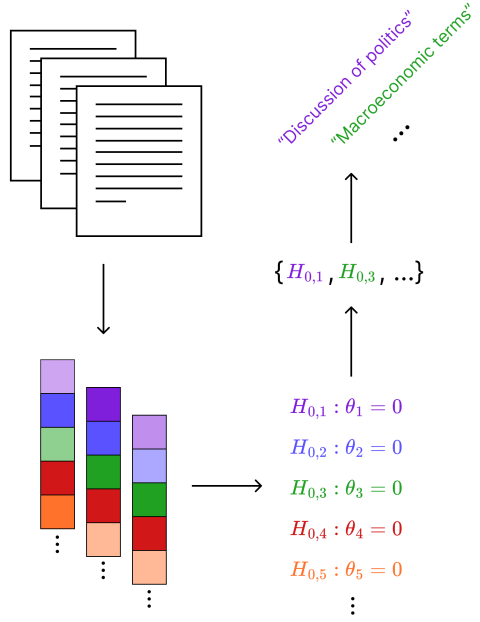
- Consider labor RCT on productivity effects of AI (Noy and Zhang, 2023):
  - → What is **causal effect** of LLM access on prof. tasks like writing HR email on sensitive topic?
  - → Assign writing tasks to $n = 453$ college-edu. professionals, randomize access to ChatGPT.
  - → Key outcomes: time spent on task ↓ 40%, human-assessed "quality grade" (1-7) ↑ 18%.
- *Why* higher quality? Can we **discover** systematic differences from **email text** itself?

Running Empirical Example: Productivity RCT

- Consider labor RCT on productivity effects of AI (Noy and Zhang, 2023):
- *Why* higher quality? Can we **discover** systematic differences from **email text** itself?
- How to proceed? Could manually explore data + def. measures based on salience, but...
  1. **Sci. concerns**: cherry-picking + fail to see important patterns ("streetlight effect").
     - Want: transparent, comprehensive accounting.
  2. **Stat. concerns**: how many things did you try to measure + what guarantees are meaningful? Would you have chosen to measure diff. things based on diff. sample?
     - Want: handle multiple testing + post-selection inference.
- Begs question: in light of difficulties + desiderata, what is a better way to proceed?

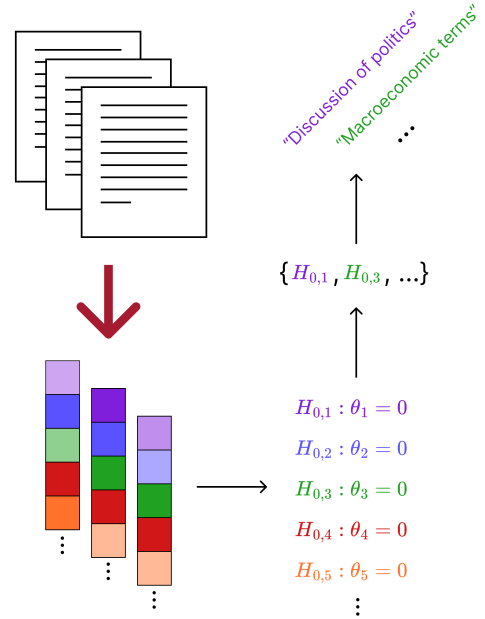
Framework

1. Measuring Concepts
2. Making Hypotheses about Concepts
3. Multiple Testing
4. Interpreting Discoveries



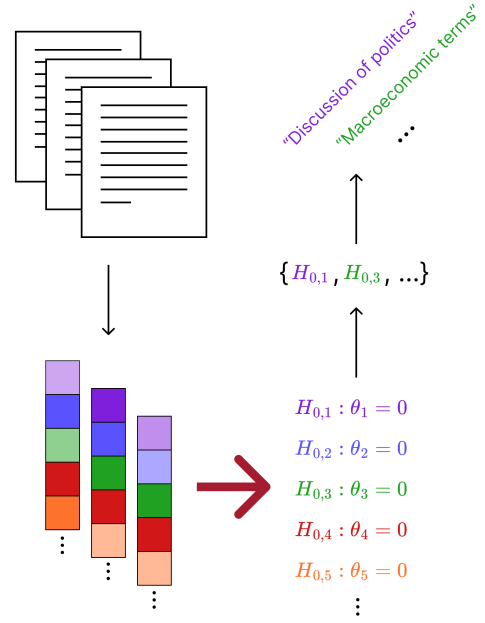
Framework

1. **Measuring Concepts**
2. Making Hypotheses about Concepts
3. Multiple Testing
4. Interpreting Discoveries



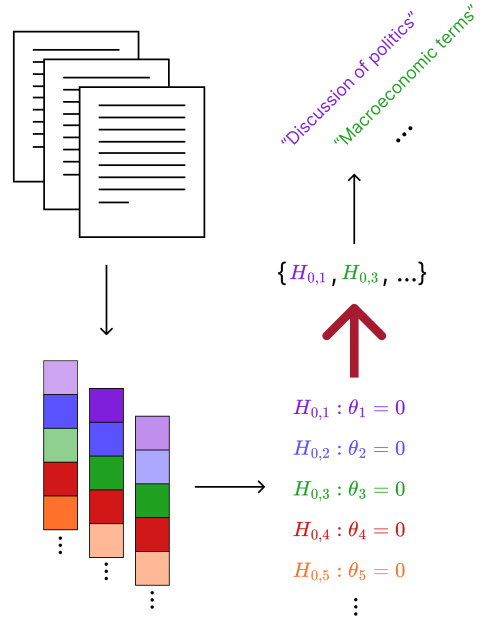
Framework

1. Measuring Concepts
2. **Making Hypotheses about Concepts**
3. Multiple Testing
4. Interpreting Discoveries



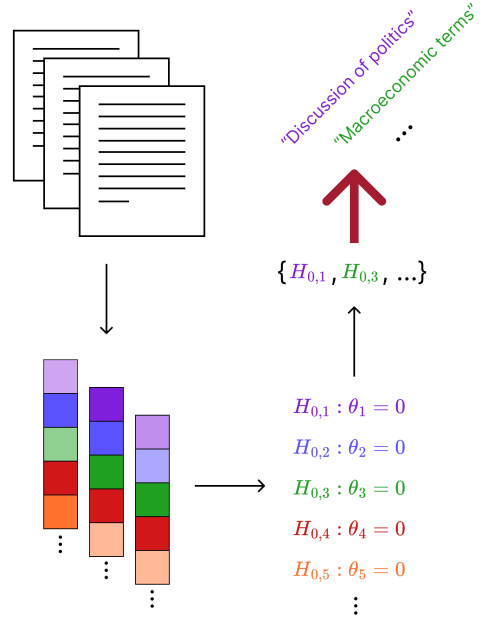
Framework

1. Measuring Concepts
2. Making Hypotheses about Concepts
3. **Multiple Testing**
4. Interpreting Discoveries



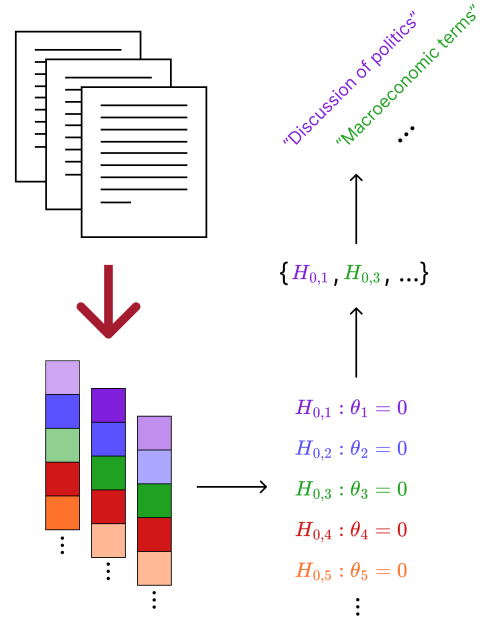
Framework

1. Measuring Concepts
2. Making Hypotheses about Concepts
3. Multiple Testing
4. **Interpreting Discoveries**



Framework

1. **Measuring Concepts**
2. Making Hypotheses about Concepts
3. Multiple Testing
4. Interpreting Discoveries



Concept Embeddings

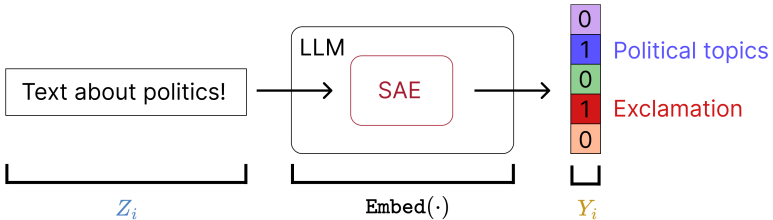
- Researcher observes unstructured data ($Z_i \in \mathcal{Z}$) and other covariates ($W_i \in \mathcal{W}$) sampled iid from P. iid?
  - $\rightarrow$ NZ: Z_i are emails, W_i are treatment indicators.

Concept Embeddings

- Researcher observes unstructured data ($Z_i \in \mathcal{Z}$) and other covariates ($W_i \in \mathcal{W}$) sampled iid from P. iid?
- Researcher has function $\texttt{Embed} : \mathcal{Z} \to \{0,1\}^p$ that measures presence of p interpretable features ("concepts") in a Z_i; call $Y_i := \texttt{Embed}(Z_i)$ a **concept embedding**.
  - $\to$ p should be *large*; rich index of concepts $\implies$ rich hypo. + discoveries. Dim. reduction?
  - $\to$ Each $Y_{ij} \in \{0,1\}$ has interpretation that concept $j \in \{1, \ldots, p\}$ appeared in Z_i.

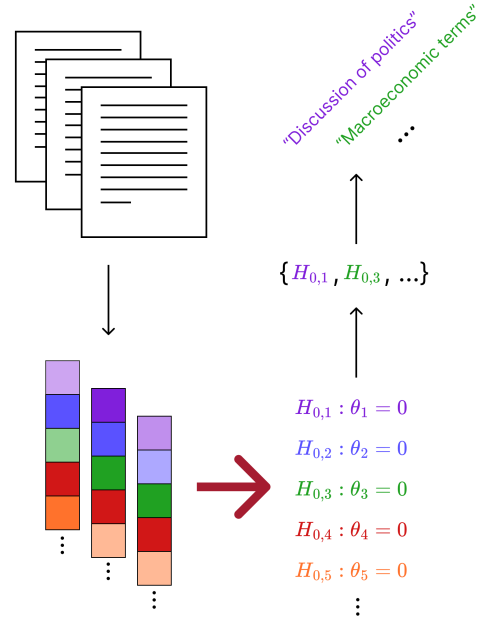
Concept Embeddings

- Researcher observes unstructured data ($Z_i \in \mathcal{Z}$) and other covariates ($W_i \in \mathcal{W}$) sampled iid from P. iid?
- Researcher has function $\texttt{Embed} : \mathcal{Z} \to \{0,1\}^p$ that measures presence of p interpretable features ("concepts") in a Z_i; call $Y_i := \texttt{Embed}(Z_i)$ a **concept embedding**.
- Promising default choice is **sparse autoencoder (SAE)** (Bricken et al., 2023). Others?
  - → SAE maps unstruct. data to sparse, binary vector of rich, interpretable concepts. How? Why?
  - → Join others saying SAEs useful for concept discovery (Peng et al., 2025; Movva et al., 2025).



Framework

1. Measuring Concepts
2. **Making Hypotheses about Concepts**
3. Multiple Testing
4. Interpreting Discoveries

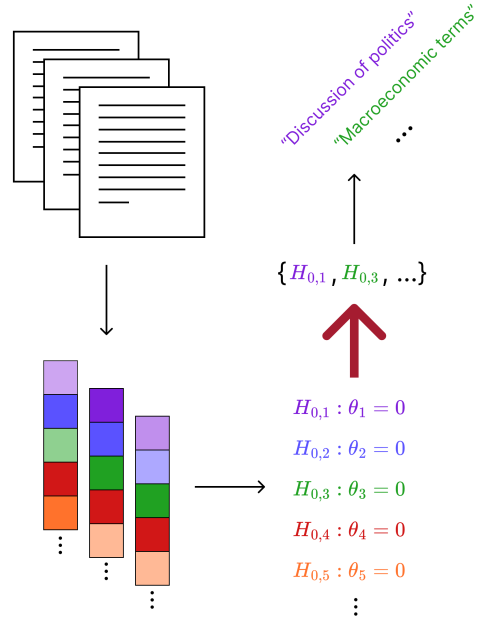


Concept-Level Parameters and Hypotheses

- The first step of the framework produced a **new dataset** of concept embeddings. (Y_i)
- Natural next step: explore rich concept variation wrt research question.
  - → NZ: What are causal effects of ChatGPT assistance on concepts in emails?
- Formalize by defining concept-level parameters $\{\theta_j\}_{j=1}^{p}$.
  - → NZ: θ_j is ATE on concept j ($\texttt{reg}\ Y_{ij} \sim W_i$) (happens to be causal—need not be).
- But we only have finite data, $\{(Y_i, W_i)\}_{i=1}^{n}$, so we must do **inference** on θ_js.
  - → NZ: $\hat{\theta}_j$ is estimate of ATE, $H_{0,j} : \theta_j = 0$.
  - → **Discoveries** are null hypotheses $H_{j,0}$ we reject w/ finite data.
- Inference, however, will be more challenging than usual...

Framework

1. Measuring Concepts
2. Making Hypotheses about Concepts
3. **Multiple Testing**
4. Interpreting Discoveries



High-Dimensional Multiple Hypothesis Testing

- Many tests $\implies$ want multiple hypo. testing (MHT) guarantees (FWER, FDX, FDR, ...).
  - $\rightarrow$ NZ: $p \approx 10^4$.

High-Dimensional Multiple Hypothesis Testing

- Many tests $\Longrightarrow$ want multiple hypo. testing (MHT) guarantees (FWER, FDX, FDR, ...).
- High-dim. ($p \gg n$), dep. test stats, not too conservative $\Longrightarrow$ need **new MHT tool**! Really?
  - $\rightarrow$ NZ: $n \approx 10^2$, $p \approx 10^4$.

High-Dimensional Multiple Hypothesis Testing

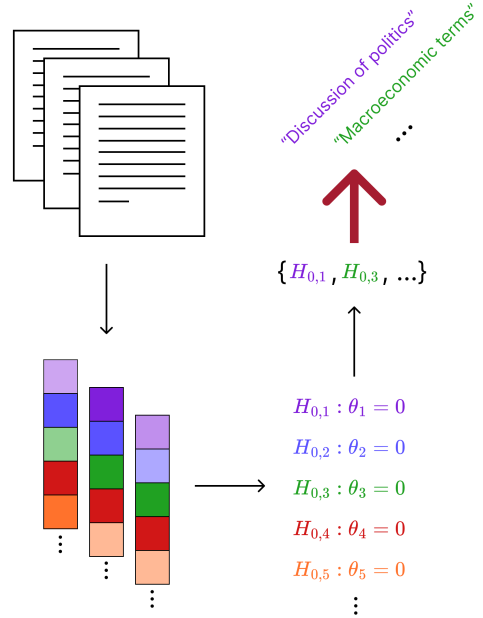
- Many tests $\Longrightarrow$ want multiple hypo. testing (MHT) guarantees (FWER, FDX, FDR, ...).
- High-dim. $(p \gg n)$, dep. test stats, not too conservative $\Longrightarrow$ need **new MHT tool**! Really?
- Under mild reg. cond., using new ext. of **high-dim. CLT**, can control generalized FWER under arbitrary dependence even in high-dimensions. More formality
  - $\rightarrow$ Many formal results in paper.

High-Dimensional Multiple Hypothesis Testing

- Many tests $\implies$ want multiple hypo. testing (MHT) guarantees (FWER, FDX, FDR, ...).
- High-dim. ($p \gg n$), dep. test stats, not too conservative $\implies$ need **new MHT tool**! Really?
- Under mild reg. cond., using new ext. of **high-dim. CLT**, can control generalized FWER under arbitrary dependence even in high-dimensions. More formality
- Mechanically, implement modified version of algo. from Romano and Wolf (2007) under guidance of new theory.

Framework

1. Measuring Concepts
2. Making Hypotheses about Concepts
3. Multiple Testing
4. **Interpreting Discoveries**



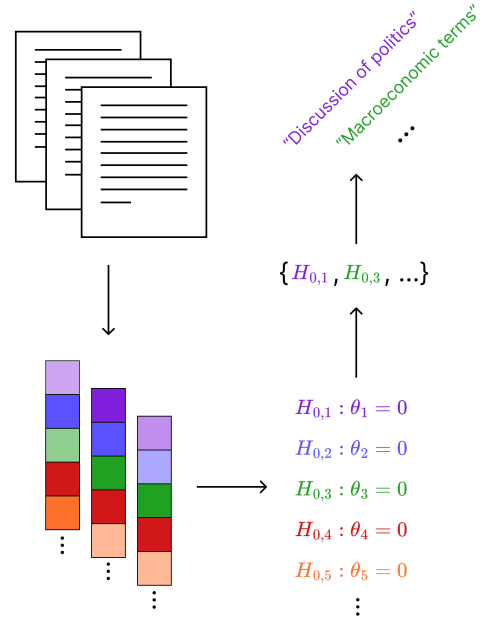
Local Automatic Interpretation

- SAEs tell us each of $Y_{i1}, Y_{i2}, \ldots$ *should* indicate a human-interpretable concept.
  - → *But* need to do more work to give them English descriptions (c.f., LDA topic models).
- Enter **automatic interpretation** ("autointerp") (Bills et al., 2023; Templeton et al., 2024).
  1. Get LLMs to generate concept descriptions. How?
  2. Use LLMs to evaluate description quality (compute eval. scores). How?
- Run autointerp "locally" (on our data) + do inference on eval. scores.
  - → Punchline: measure exactly how good concept desc. are wrt our data distr. of interest P.
  - → Stat. inf. view motivates sample splitting, CIs, "inf. on winners" corr. (Andrews et al., 2024).

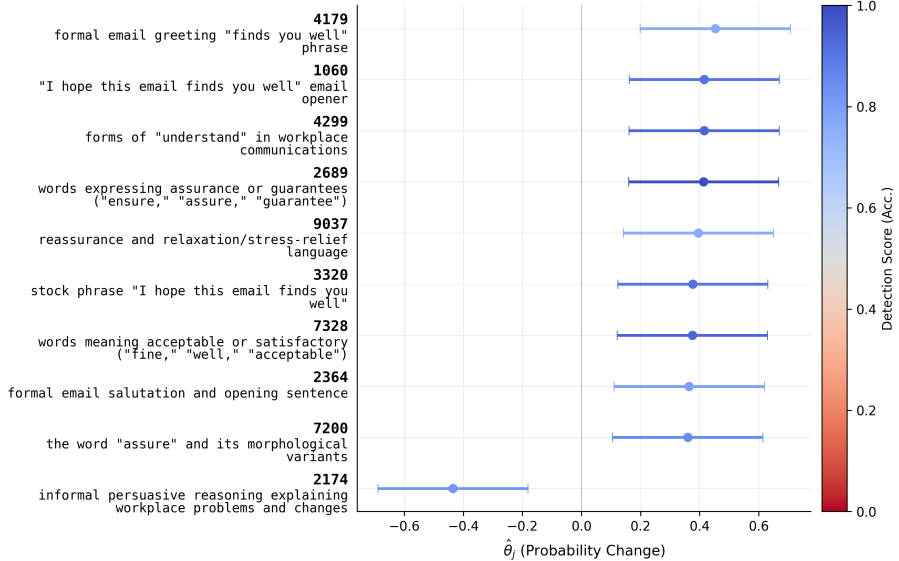
Framework

1. Measuring Concepts
2. Making Hypotheses about Concepts
3. Multiple Testing
4. Interpreting Discoveries

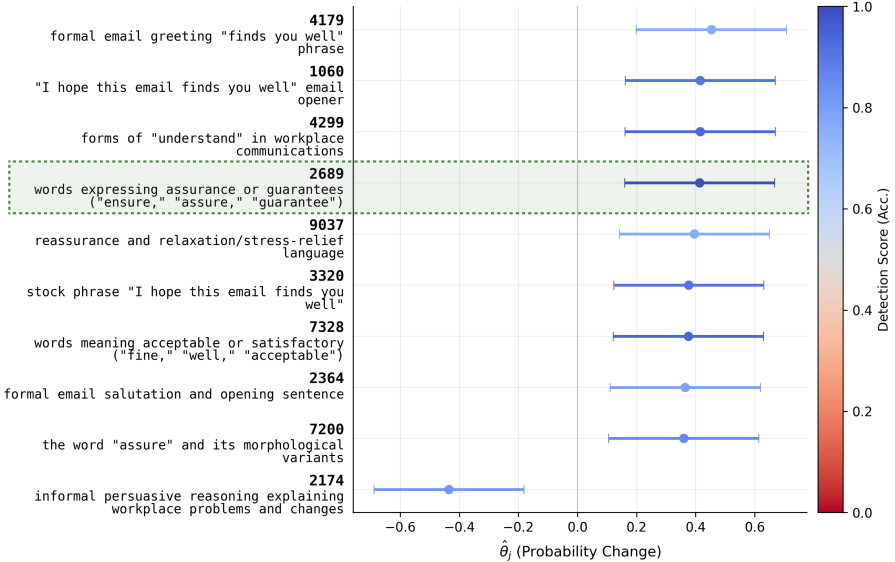
→ **Empirics**



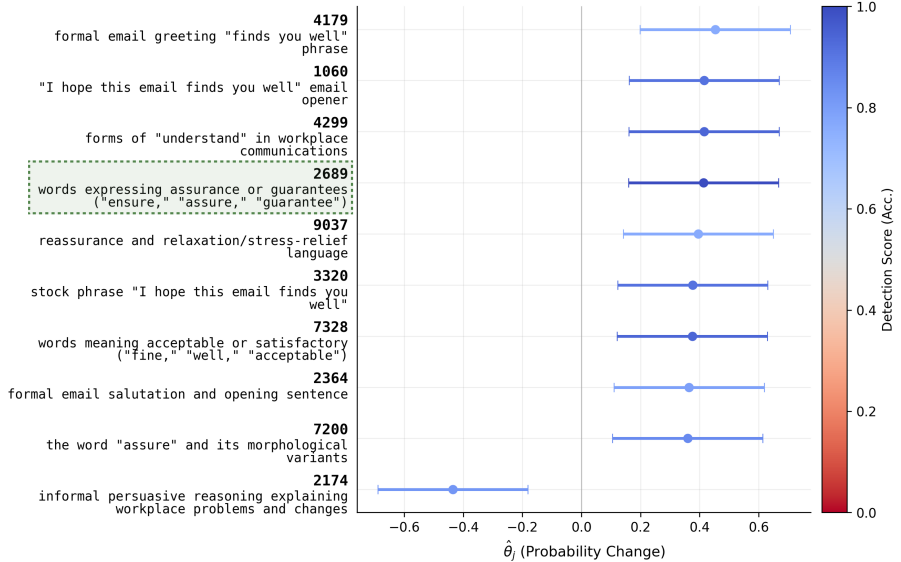
NZ: 128 Discoveries $\Longrightarrow$ ↑ Homogeneity, ↑ Positivity, ↓ Informality



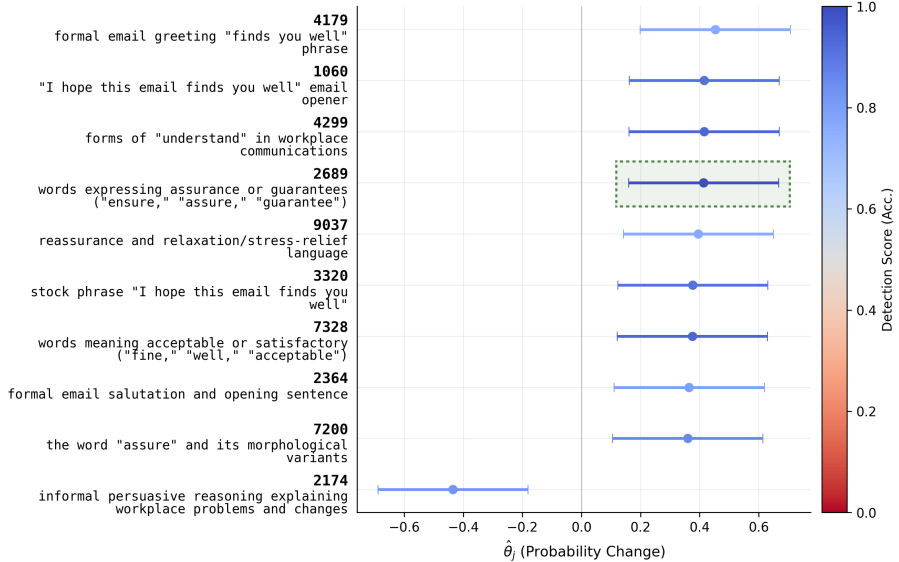
NZ: 128 Discoveries $\Longrightarrow$ ↑ Homogeneity, ↑ Positivity, ↓ Informality



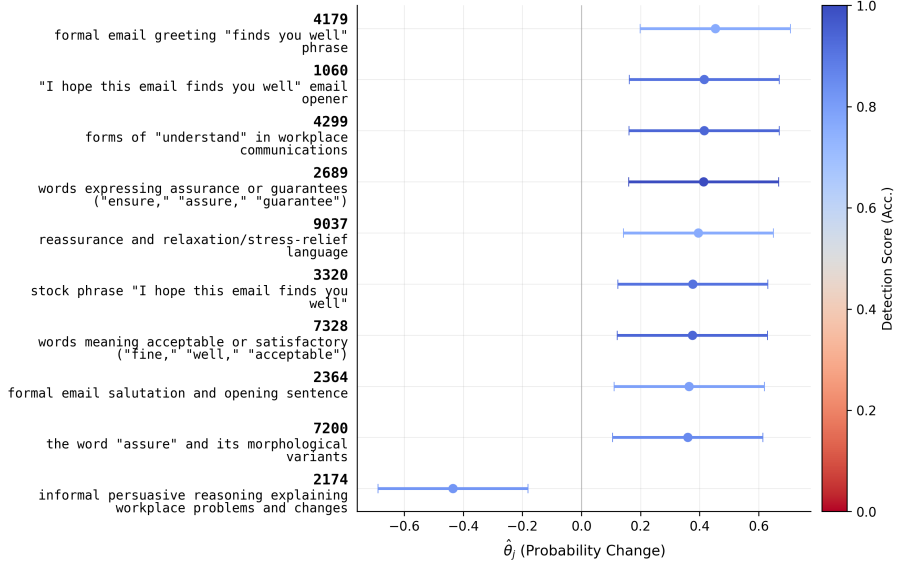
NZ: 128 Discoveries $\Longrightarrow$ ↑ Homogeneity, ↑ Positivity, ↓ Informality



NZ: 128 Discoveries $\Longrightarrow$ ↑ Homogeneity, ↑ Positivity, ↓ Informality



NZ: 128 Discoveries $\Longrightarrow$ ↑ Homogeneity, ↑ Positivity, ↓ Informality



Conclusion

- Human-in-the-loop not required $\Longrightarrow$ **automaticity**.
  - $\rightarrow$ Very few researcher DoF, highly replicable, easy sensitivity analysis.
- New results in high-dim. CLT and MHT useful in other settings.
- **Complementary** to existing discovery tools for unstructured data (Movva et al., 2025; Modarressi et al., 2025); think PAP and contingent analyses (Banerjee et al., 2020).

Thank you!

jacob_carlson@g.harvard.edu

Appendix

What is unstructured data? Back

- Unstructured data typically defined by example: text, image, audio, video, location.
- Towards a **unifying definition**, all these modalities have 2 things in common:
  1. Serve as source of idiosyncratic, meaningful "structured" features.
  2. Features are quant. illegible from raw form of data $\implies$ structure/meaning must be learned/extracted.
- Today: deep neural networks (DNNs)—LLMs, etc.—serve as meaningful representation learners and/or structured feature extractors. (Historically: "feature engineering.")
  - $\rightarrow$ Modern "AI" trained to do precisely this; substrate of all "AI" *is* unstructured data.
  - $\rightarrow$ Interpretable prediction *is* structured feature extraction.
  - $\rightarrow$ DNNs simultaneously scalable and high fidelity.

What is qualitative coding? What is *rich* discovery? Back

- Natural reference class of discovery methods is qualitative coding.
  - → Broadly popular in soc. sci, increasingly so in econ. (Haaland et al., 2024).
  - → Family of semi-systematic approaches for rich, interpretable classif. of unstructured data.
    - E.g., read through interviews, "code up" salient themes, map themes to keywords, and label passages w/ keyword-theme map (Ferrario and Stantcheva, 2022).
  - → Approaches are sensible, but run into earlier discussed scientific and stat. shortcomings.
- Most popular alternatives are classic NLP methods (Gentzkow et al., 2019).
  - → More scientifically principled (though not always statistically).
  - → Not very rich in detail or nuance.
    - Analysis typically uses coarser featurization based on, e.g., n-grams.
- This framework: retain richness, maintain scientific + statistical rigor.

What if the data is not iid? Back

- Various extensions to clustered sampling are arms reach, e.g., under single-cluster sampling (data iid across clusters, but may be dependent within clusters).
  - → Intuition: still apply the iid LLN and CLT by working w/ rep. of data in terms of clusters.
- For example, can write means as

$$\hat{\theta}_j := \frac{\frac{1}{C}\sum_c \sum_{i \in \mathcal{I}(c)} Y_{ij}}{\frac{1}{C}\sum_c N_c},$$

  ratio of two cluster-level averages, where $\mathcal{I}(c)$ is set of cluster c indices, C is cluster count, N_c is number of obs. in cluster c.
- B/c $\hat{\theta}_j$ is asy. lin. estimator of $\theta_j := E[Y_{ij}] = \frac{E\left[\sum_{i \in \mathcal{I}(c)} Y_{ij}\right]}{E[N_c]}$, new theory goes through.

Is this dimensionality reduction? If not, why not do that? Back

- Not always case that $\dim(\mathcal{Z}) \gg p \implies$ spirit of `Embed` is not dimensionality reduction.
  - → Role of `Embed` is to go from semantically thin to semantically rich representation.
- Framework aims to embrace inherent high-dimensionality of unstructured data.
  - → Text, images are inherently high-dim., even in the space of interpretable concepts.
  - → Tackle this directly with high-dim. inference methods.
- Alt. approaches and rationales exist to do dim. reduction (Modarressi et al., 2025; Ludwig et al., 2017), but that is explicitly *not* the goal of this project.
  - → Such approaches can be seen as complementary; have different profiles of scientific costs and benefits.

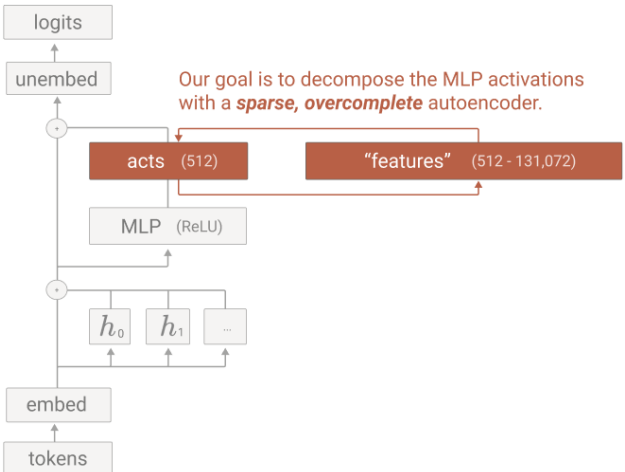
Alternative Concept Embeddings Back

- SAEs + other sparse dictionary learning methods useful default choice of concept embedding, but can readily consider others.
  - → Nothing in framework requires SAE $\implies$ just need data-indep. curation.
  - → Other choices may not require step 4 of the framework, as possible benefit.
- Two other examples could be:
  1. n-gram dictionary: each index catalogs of presence or absence of particular n-gram.
     - Comp. is cheap, richness/interp. is worse.
  2. Wiki categories + GABRIEL: each index indicates absence or presence of Wikipedia category (`https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories`); implement via zero-shot LLM classification (Asirvatham et al., 2026).
     - Comp. is expensive, richness/interp. is better.

Sparse Dictionary Learning + SAEs Back

- Sparse dictionary learning popular tool in "mechanistic interpretability" + AI safety.
  - → Most popular implementation are SAEs, others exist (Paulo et al., 2025).
  - → Not just for text models (Abdulaal et al., 2024; Pluth et al., 2025; Fel et al., 2026).
- "Read minds" of AI models by mapping internal activations to interpretable concepts.
  - → Implement: autoencoders attached to residual stream of LLM, with dim(SAE hidden layer) $\gg$ dim(residual stream); trained with sparsity-inducing penalization on reconstruction loss.
  - → Idea: "Linear rep. hypothesis" + "superposition hypothesis" $\implies$ SAEs $\approx$act as overcomplete linear basis of activations $\implies$ "monosemantic" features. (Bricken et al., 2023)
  - → Insight: mind of LLM says a lot about what it's reading! (Echoed in Jiang et al. (2025).)
- Lots of debate about the success of SAEs (Leask et al., 2025; Bhalla et al., 2026a); most criticisms still consistent w/ SAEs having role in discovering novel concepts (Peng et al., 2025).

SAE Visualized (Bricken et al., 2023) Back



Examples of Off-the-shelf SAE Concepts Back

- SAE concepts vary by LLM, hyperparam. (sparsity penalty, dim.), + training distr.
- 20 random examples from the T-SAE (Bhalla et al., 2026b): numbers and dates (#62), I2C related terms (#409), substrings with underscores and angle brackets (#577), phrases with on the (#635), general concepts (#667), historical dates and years (#781), geographic locations (#1011), import statements (#1149), understanding and explanation concepts (#1180), health issues (#1196), relationships and romance (#1325), nouns related to concepts and objects (#1377), North America and United States references (#1549), words related to change in shape (#1590), effort and time investment (#1676), license terms (#2003), proper nouns and names (#2259), words related to volunteering (#2341), words in angle brackets (#2610), units per something (#2826).
- Local autointerp can overwrite these descriptions *or* use as prior.

Do we really need a new MHT procedure? Back

- Recall that inference is challenging b/c:
  1. p is large, so inference on all θ_js is multiple hypothesis testing (MHT) problem.
  2. $p \gg n$, so inference is high-dimensional MHT problem.
  3. Each estimator, test stat., or p-value formed plausibly stat. dependent on every other.
  4. For discovery, selective (familywise) error control ideally not too conservative.
- Need dep.-agnostic, high-dim. MHT procedure w/ control over generalized error rate.
  - → Also want: (1) well-powered (uses resampling); (2) allow for asy. valid test stat.'s.
- Bonferroni? Benjamini and Yekutieli (2001)? Too conservative.
  - → Benjamini and Hochberg (1995)? Fithian and Lei (2022)? Need indep., or known dep. (e.g., PRDS).
  - → Barber and E. J. Candès (2015), E. Candès et al. (2018), and Barber and E. J. Candès (2019)? Low-dim., operational mismatches, and/or known distr.
  - → Romano and Wolf (2007)? Low-dim. $\Longrightarrow$ Ding et al. (2025) or Feng (2026) provide way forward!

Generalized Selective Error Control Back

Under mild reg. cond., using new ext. of **high-dim. CLT**, can show version of Algorithm 2.1 (2.2) of Romano and Wolf (2007) achieves "k-FWER control" in high-dim.:

$$\limsup_{n,p\to\infty} P(\geq k \text{ false rejections}) \leq \alpha$$

for "small" (fixed) k, researcher chosen α, p growing nearly expo. in n.

Algorithm 2.1 (Romano and Wolf, 2007) Back

1. Let $A_1 = \{1, \ldots, p\}$. If $\max\left(T_{n,j} : j \in A_1\right) \leq \hat{c}_{n,A_1}(1-\alpha, k)$, then accept all hypotheses and stop; otherwise, reject any $H_{0,j}$ for which $T_{n,j} > \hat{c}_{n,A_1}(1-\alpha, k)$ and continue.
2. Let R_2 be indices j of $H_{0,j}$ prev. rejected, let A_2 be indices of remaining hypotheses. If $|R_2| < k$, then stop. Otherwise, let

$$\hat{d}_{n,A_2}(1-\alpha, k) = \max_{I \subset R_2,\, |I| = k-1} \left\{\hat{c}_{n,K}(1-\alpha, k) : K = A_2 \cup I\right\}.$$

   Then, reject any $H_{0,j}$ with $j \in A_2$ satisfying $T_{n,j} > \hat{d}_{n,A_2}(1-\alpha, k)$. If no further rejections, stop.

$\vdots$

ℓ. Let R_ℓ be indices j of $H_{0,j}$ prev. rejected, A_ℓ be indices of remaining hypotheses. Let

$$\hat{d}_{n,A_\ell}(1-\alpha, k) = \max_{I \subset R_\ell,\, |I| = k-1} \left\{\hat{c}_{n,K}(1-\alpha, k) : K = A_\ell \cup I\right\}.$$

   Then, reject any $H_{0,j}$ with $j \in A_\ell$ satisfying $T_{n,j} > \hat{d}_{n,A_\ell}(1-\alpha, k)$. If no further rejections, stop.

Local Automatic Generation Back

- Goal: get LLMs to generate English descriptions of the concepts $j \in \{1, \ldots, p\}$.
- Canonical procedure from Bricken et al. (2023) and Templeton et al. (2024), integrating insights from Bills et al. (2023); for $j \in \{1, \ldots, p\}$:
  1. Rank texts in dataset by max token activation for SAE feature j.
  2. Collect $L \approx 10$ top ranked texts to be used as "exemplars."
  3. Annotate high activation tokens w/ activation info (brackets, quant. summaries, etc.).
  4. Prompt LLM with context + L annotated texts, asking for English desc. of plausible concept that produced activation pattern.
  5. Extract description and use as label for concept j.
- Locality: collect exemplars drawn from P (*not* from some opaque sample of web data).

Local Automatic Evaluation

Back

- Goal: get LLMs to "score" quality of English descriptions of concepts $j \in \{1, \ldots, p\}$.
- Popular eval. scoring choice from Paulo et al. (2024) called "detection scoring."
- Idea: good desc. is good classifier of SAE feature activation in a given text.
  - $\rightarrow$ Classifier *is* desc. (oper. via LLM), produces pred. label $\hat{Y}_{ij} \in \{0, 1\}$
  - $\rightarrow$ Label is true feature presence $Y_{ij} \in \{0, 1\}$.
  - $\rightarrow$ Binary classifier metrics, e.g., accuracy, recall, *become* quality metrics for desc.
- Procedurally, using indep. hold-out sample, for concepts $j \in \{1, \ldots, p\}$:
  1. Show LLM each text in sample and ask "does the concept {`concept desc.` j} appear?"
  2. Compare each answer $\hat{Y}_{ij}$ to ground truth activation Y_{ij} + compute, e.g., accuracy.
- Locality: use indep. hold-out sample from P for estimating det. scores and CIs.

References I

Abdulaal, Ahmed et al. (2024). *An X-Ray Is Worth 15 Features: Sparse Autoencoders for Interpretable Radiology Report Generation*. Version Number: 1. DOI: 10.48550/ARXIV.2410.03334. URL: https://arxiv.org/abs/2410.03334 (visited on 10/03/2025).

Andrews, Isaiah, Toru Kitagawa, and Adam McCloskey (Jan. 2024). "Inference on Winners". en. *The Quarterly Journal of Economics* 139.1, pp. 305–358. ISSN: 0033-5533, 1531-4650. DOI: 10.1093/qje/qjad043. URL: https://academic.oup.com/qje/article/139/1/305/7276491 (visited on 11/20/2025).

Ash, Elliott, Daniel L Chen, and Suresh Naidu (Jan. 2026). "Ideas Have Consequences: The Impact of Law and Economics on American Justice". en. *The Quarterly Journal of Economics* 141.1, pp. 845–887. ISSN: 0033-5533, 1531-4650. DOI: 10.1093/qje/qjaf042. URL: https://academic.oup.com/qje/article/141/1/845/8241352 (visited on 06/07/2026).

References II

Asirvatham, Hemanth, Elliott Mokski, and Andrei Shleifer (Feb. 2026). *GPT as a Measurement Tool*. en. Tech. rep. w34834. Cambridge, MA: National Bureau of Economic Research, w34834. DOI: 10.3386/w34834. URL: http://www.nber.org/papers/w34834.pdf (visited on 06/08/2026).

Banerjee, Abhijit et al. (Apr. 2020). *In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics*. en. Tech. rep. w26993. Cambridge, MA: National Bureau of Economic Research, w26993. DOI: 10.3386/w26993. URL: http://www.nber.org/papers/w26993.pdf (visited on 04/03/2026).

Barber, Rina Foygel and Emmanuel J. Candès (Oct. 2015). "Controlling the false discovery rate via knockoffs". *The Annals of Statistics* 43.5. ISSN: 0090-5364. DOI: 10.1214/15-AOS1337. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-43/issue-5/Controlling-the-false-discovery-rate-via-knockoffs/10.1214/15-AOS1337.full (visited on 08/04/2026).

References III

Barber, Rina Foygel and Emmanuel J. Candès (Oct. 2019). "A knockoff filter for high-dimensional selective inference". *The Annals of Statistics* 47.5. ISSN: 0090-5364. DOI: 10.1214/18-AOS1755. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-47/issue-5/A-knockoff-filter-for-high-dimensional-selective-inference/10.1214/18-AOS1755.full (visited on 04/22/2025).

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1. Publisher: [Royal Statistical Society, Oxford University Press], pp. 289–300. ISSN: 00359246. URL: http://www.jstor.org/stable/2346101 (visited on 10/06/2025).

References IV

Benjamini, Yoav and Daniel Yekutieli (2001). “The Control of the False Discovery Rate in Multiple Testing under Dependency”. *The Annals of Statistics* 29.4. Publisher: Institute of Mathematical Statistics, pp. 1165–1188. ISSN: 00905364, 21688966. URL: http://www.jstor.org/stable/2674075 (visited on 10/06/2025).

Bergman, Peter et al. (May 2024). “Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice”. en. *American Economic Review* 114.5, pp. 1281–1337. ISSN: 0002-8282. DOI: 10.1257/aer.20200407. URL: https://pubs.aeaweb.org/doi/10.1257/aer.20200407 (visited on 10/02/2025).

Bhalla, Usha et al. (2026a). *Do Sparse Autoencoders Capture Concept Manifolds?* Version Number: 1. DOI: 10.48550/ARXIV.2604.28119. URL: https://arxiv.org/abs/2604.28119 (visited on 06/27/2026).

References V

Bhalla, Usha et al. (2026b). "Temporal Sparse Autoencoders: Leveraging the Sequential Nature of Language for Interpretability". *The Fourteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=bojVI4l9Kn.

Bills, Steven et al. (2023). *Language models can explain neurons in language models*. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Bricken, Trenton et al. (2023). "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Candès, Emmanuel et al. (June 2018). "Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection". en. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3, pp. 551–577. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/rssb.12265. URL: https://academic.oup.com/jrsssb/article/80/3/551/7048447 (visited on 04/15/2025).

References VI

Ding, Yixi et al. (2025). *Gaussian Multiplier Bootstrap Procedure for the k-th Largest Coordinate of High-Dimensional Statistics*. Version Number: 1. DOI: 10.48550/ARXIV.2508.14400. URL: https://arxiv.org/abs/2508.14400v1 (visited on 08/31/2025).

Fel, Thomas et al. (2026). *Structuring Sparsity: Block-Sparse Featurizers Capture Visual Concept Manifolds*. Version Number: 1. DOI: 10.48550/ARXIV.2606.25234. URL: https://arxiv.org/abs/2606.25234 (visited on 07/30/2026).

Feng, Long (Apr. 2026). *High Dimensional Bootstrap and Asymptotic Expansion for the k-th Largest Coordinate*. arXiv:2604.04785 [math.ST]. DOI: 10.48550/arXiv.2604.04785. URL: http://arxiv.org/abs/2604.04785 (visited on 07/07/2026).

References VII

Ferrario, Beatrice and Stefanie Stantcheva (May 2022). "Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions". en. *AEA Papers and Proceedings* 112, pp. 163–169. ISSN: 2574-0768, 2574-0776. DOI: 10.1257/pandp.20221071. URL: https://pubs.aeaweb.org/doi/10.1257/pandp.20221071 (visited on 10/09/2025).

Fithian, William and Lihua Lei (Dec. 2022). "Conditional calibration for false discovery rate control under dependence". *The Annals of Statistics* 50.6. ISSN: 0090-5364. DOI: 10.1214/21-AOS2137. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-50/issue-6/Conditional-calibration-for-false-discovery-rate-control-under-dependence/10.1214/21-AOS2137.full (visited on 03/31/2025).

References VIII

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy (2019). "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech". en. *Econometrica* 87.4, pp. 1307–1340. ISSN: 0012-9682. DOI: 10.3982/ECTA16566. URL: https://www.econometricsociety.org/doi/10.3982/ECTA16566 (visited on 10/04/2025).

Haaland, Ingar et al. (May 2024). *Understanding Economic Behavior Using Open-ended Survey Data*. en. Tech. rep. w32421. Cambridge, MA: National Bureau of Economic Research, w32421. DOI: 10.3386/w32421. URL: http://www.nber.org/papers/w32421.pdf (visited on 09/12/2025).

Hansen, Stephen, Michael McMahon, and Andrea Prat (May 2018). "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*". en. *The Quarterly Journal of Economics* 133.2, pp. 801–870. ISSN: 0033-5533, 1531-4650. DOI: 10.1093/qje/qjx045. URL: https://academic.oup.com/qje/article/133/2/801/4582916 (visited on 06/07/2026).

References IX

Jiang, Nick et al. (2025). *Interpretable Embeddings with Sparse Autoencoders: A Data Analysis Toolkit*. Version Number: 1. DOI: 10.48550/ARXIV.2512.10092. URL: https://arxiv.org/abs/2512.10092 (visited on 03/04/2026).

Leask, Patrick et al. (Feb. 2025). *Sparse Autoencoders Do Not Find Canonical Units of Analysis*. arXiv:2502.04878 [cs]. DOI: 10.48550/arXiv.2502.04878. URL: http://arxiv.org/abs/2502.04878 (visited on 04/01/2025).

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess (2017). *Machine-Learning Tests for Effects on Multiple Outcomes*. Version Number: 2. DOI: 10.48550/ARXIV.1707.01473. URL: https://arxiv.org/abs/1707.01473 (visited on 10/04/2025).

Modarressi, Iman, Jann Spiess, and Amar Venugopal (2025). *Causal Inference on Outcomes Learned from Text*. Version Number: 1. DOI: 10.48550/ARXIV.2503.00725. URL: https://arxiv.org/abs/2503.00725 (visited on 05/26/2025).

References X

Movva, Rajiv et al. (2025). *Sparse Autoencoders for Hypothesis Generation*. Version Number: 3. DOI: 10.48550/ARXIV.2502.04382. URL: https://arxiv.org/abs/2502.04382 (visited on 08/12/2025).

Noy, Shakked and Whitney Zhang (July 2023). "Experimental evidence on the productivity effects of generative artificial intelligence". en. *Science* 381.6654, pp. 187–192. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.adh2586. URL: https://www.science.org/doi/10.1126/science.adh2586 (visited on 06/07/2026).

Paulo, Gonçalo, Stepan Shabalin, and Nora Belrose (2025). *Transcoders Beat Sparse Autoencoders for Interpretability*. Version Number: 2. DOI: 10.48550/ARXIV.2501.18823. URL: https://arxiv.org/abs/2501.18823 (visited on 10/06/2025).

Paulo, Gonçalo et al. (2024). *Automatically Interpreting Millions of Features in Large Language Models*. Version Number: 3. DOI: 10.48550/ARXIV.2410.13928. URL: https://arxiv.org/abs/2410.13928 (visited on 10/06/2025).

References XI

Peng, Kenny et al. (2025). *Use Sparse Autoencoders to Discover Unknown Concepts, Not to Act on Known Concepts*. Version Number: 1. DOI: 10.48550/ARXIV.2506.23845. URL: https://arxiv.org/abs/2506.23845 (visited on 10/06/2025).

Pluth, Daniel, Yu Zhou, and Vijay K. Gurbani (2025). *Sparse Autoencoder Insights on Voice Embeddings*. Version Number: 1. DOI: 10.48550/ARXIV.2502.00127. URL: https://arxiv.org/abs/2502.00127 (visited on 10/03/2025).

Romano, Joseph P. and Michael Wolf (Aug. 2007). "Control of generalized error rates in multiple testing". *The Annals of Statistics* 35.4. 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 (visited on 05/12/2025).

References XII

Stantcheva, Stefanie (Apr. 2024). *Why Do We Dislike Inflation?* en. Tech. rep. w32300. Cambridge, MA: National Bureau of Economic Research, w32300. DOI: 10.3386/w32300. URL: http://www.nber.org/papers/w32300.pdf (visited on 10/03/2025).

Templeton, Adly et al. (2024). "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet". *Transformer Circuits Thread*. URL: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.