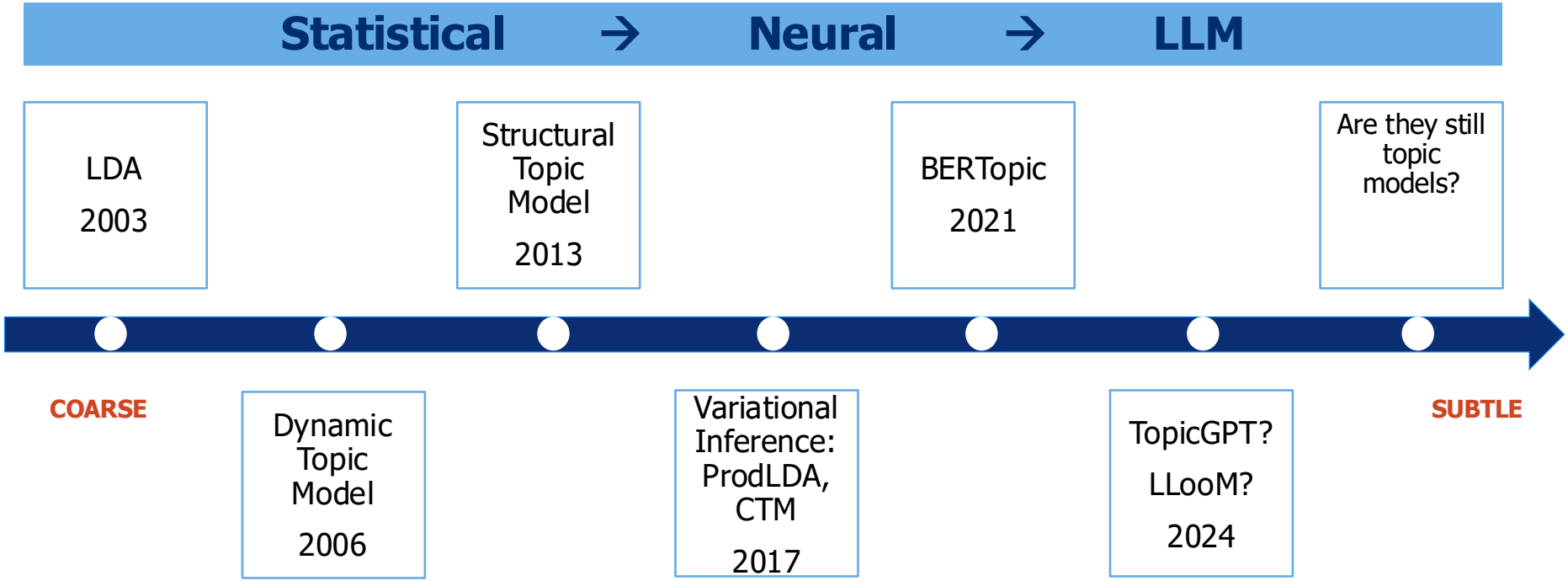
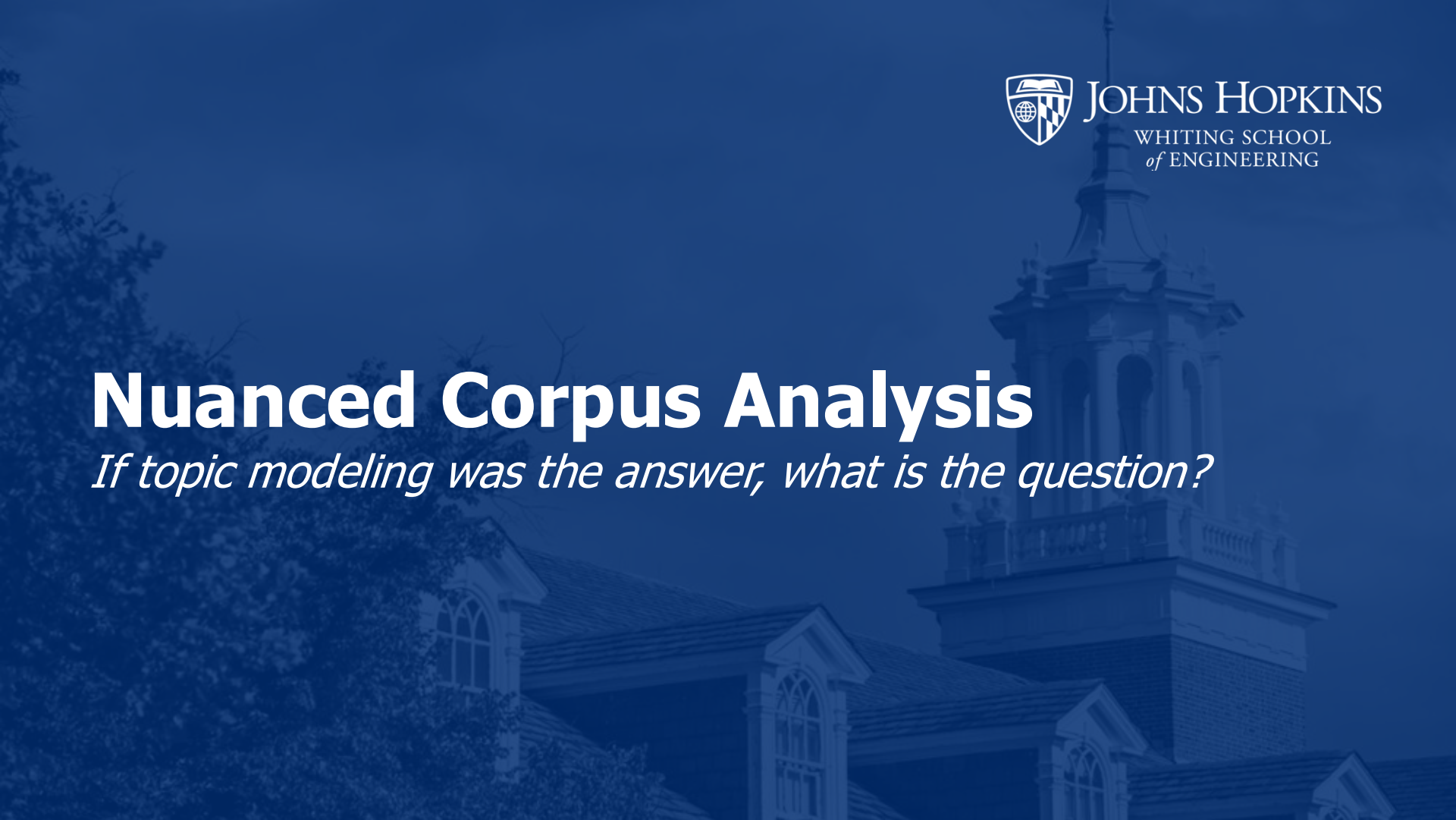
Our Favorite Topic Models





Nuanced Corpus Analysis

If topic modeling was the answer, what is the question?

Little Innovation for Nuanced Analysis on Large Corpora

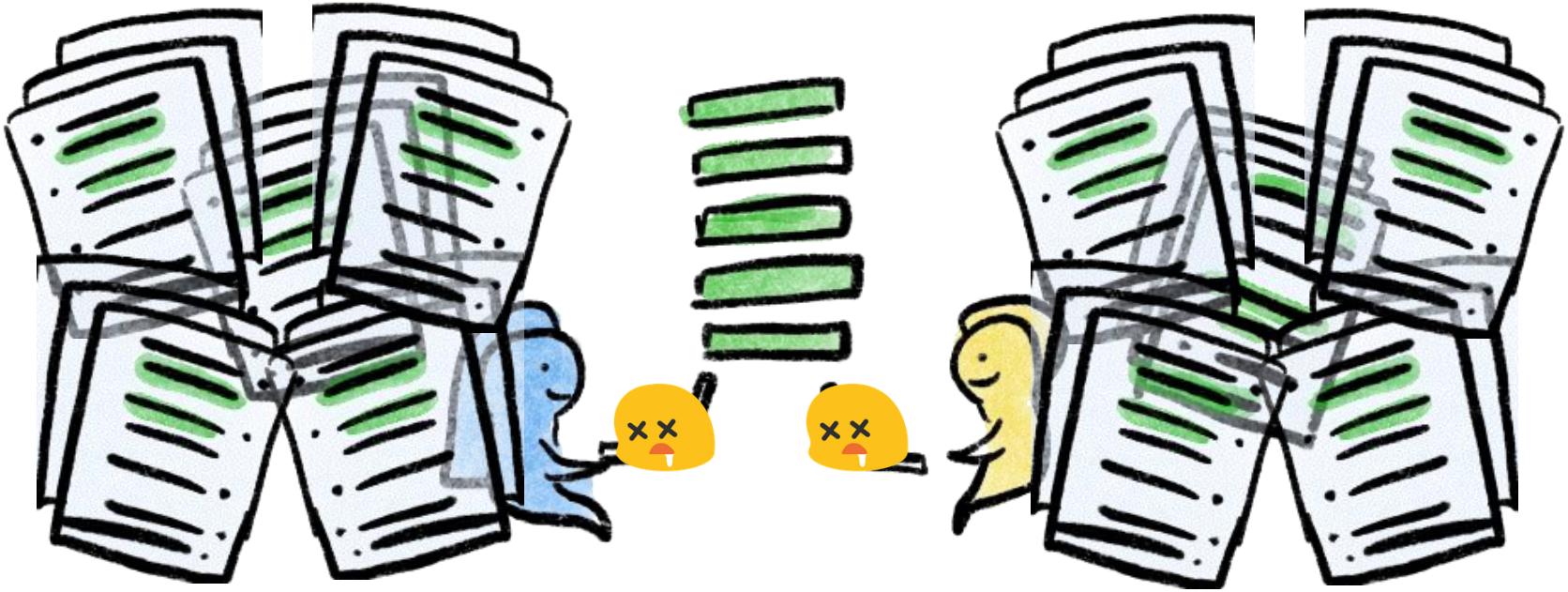
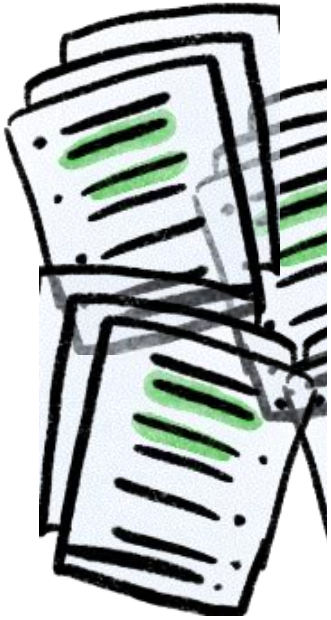


Image credit: https://delvetool.com/blog/deductiveinductive

Little Innovation for Nuanced Analysis on Large Corpora

Inside "Operation Change Agent": Mallinckrodt's Plan for Capturing the Opioid Market



Daniel Eisenkraft Klein
University of Toronto

Ross MacKenzie
University of New South Wales

Ben Hawkins
Cambridge University

Adam D. Koon
Johns Hopkins University

Little Innovation for Nuanced Analysis on Large Corpora

The Values Encoded in Machine Learning Research

Abeba Birhane*
abeba@mozillafoundation.org
Mozilla Foundation & School of Computer Science, University College Dublin
Dublin, Ireland

Pratyusha Kalluri*
pkalluri@stanford.edu
Computer Science Department, Stanford University
Palo Alto, USA

Dallas Card*
dalc@umich.edu
School of Information, University of Michigan
Ann Arbor, USA

William Agnew*
wagnew3@cs.washington.edu
Paul G. Allen School of Computer Science and Engineering, University of Washington
Seattle, USA

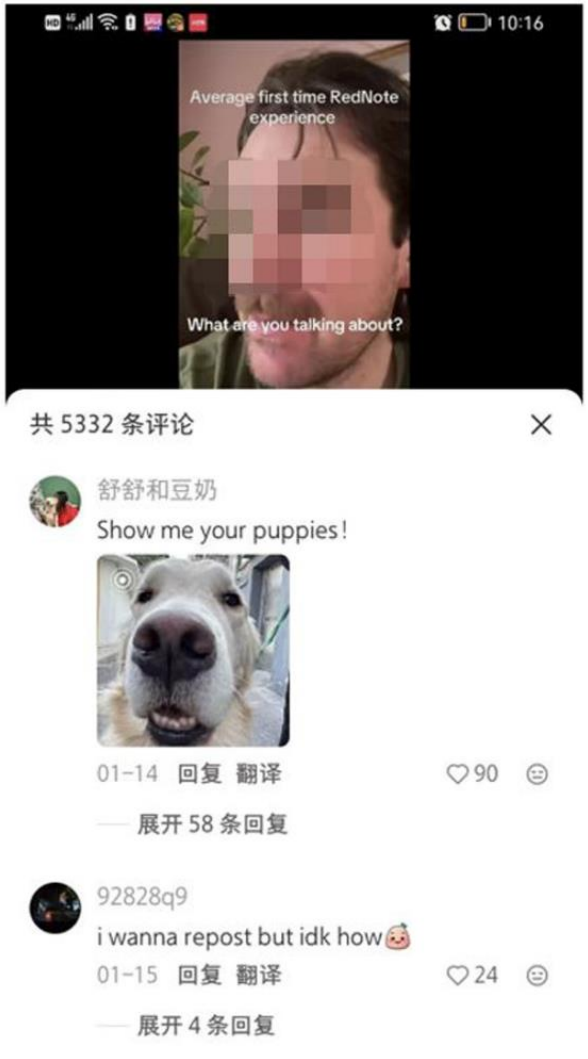
Ravit Dotan*
ravit.dotan@berkeley.edu
Center for Philosophy of Science, University of Pittsburgh
Pittsburgh, USA

Michelle Bao*
baom@stanford.edu
Computer Science Department, Stanford University
Palo Alto, USA

Inductive Coding at Scale

Breaking News January 2025

DeepSeek R1



TikTok "Refugee"?

- ~ 63,000 posts
- ~700,000 new users

https://www.bbc.com/news/articles/c2475l7zpqyo

https://www.npr.org/2025/01/15/nx-s1-5260742/tiktok-china-rednote-xiaohongshu-app

Table A1. Open Coding (178 Initial Codes)

Open coding was conducted line by line across all first-post texts. Discrete meaning units were identified and coded descriptively. A total of 178 open codes were generated and refined through constant comparison and analytic memo writing.

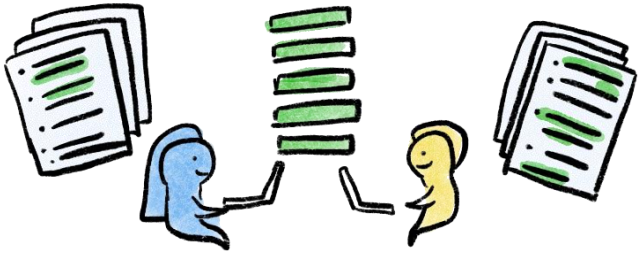
Illustrative Open Codes	Empirical Focus
"I'm a TikTok refugee"	Identity declaration
References to TikTok ban or migration	Digital displacement
"Thank you for welcoming me"	Gratitude / emotional stance
Apologies for language ability	Emotional modesty
"你好 (Ni Hao) new Xiao Hongshu friends"	Bilingual shift
Requests to learn Mandarin	Linguistic learning
Praise of Chinese users or culture	Cultural admiration
Comparisons with Western platforms	Comparative evaluation
"Let's follow each other"	Interactional ritual
Self-deprecating humor	Authenticity signaling

Inductive coding in the wild: From exile to visibility

- Analysis Question:
  - What kinds of **identity roles** do TikTok Refugees enact in their first posts on Xiaohongshu?
- Data: ~ 200 first posts
- Line by line
- -> 5 roles (Refugee, Witness, Connector, Mediator, Collaborator)

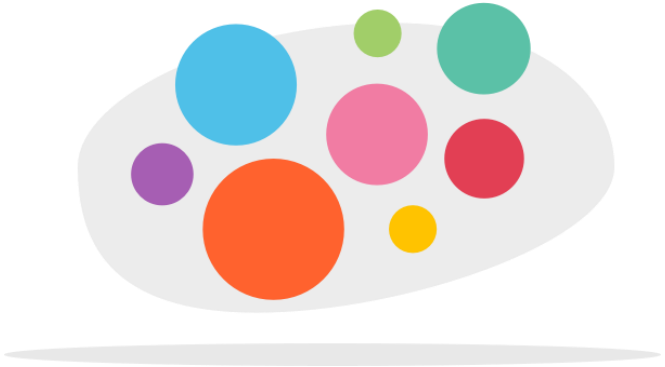
Current approaches

Downsample data



Not scalable

Topic modeling



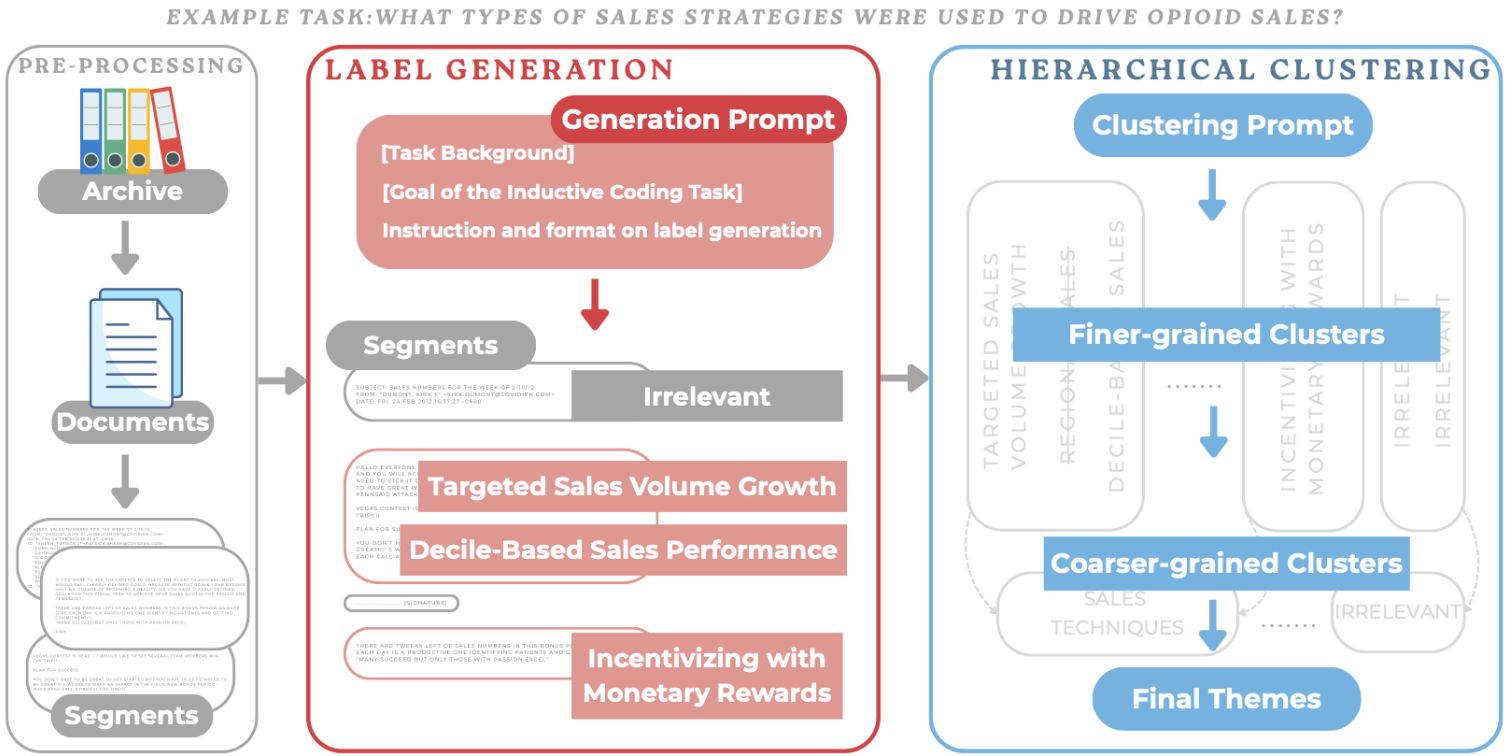
Bad Control

Hierarchical v.s. Incremental

- Pass all documents to generate codes
- Hierarchically merge codes into groups by iterations

- Start with a batch of documents
- Generate codes: preliminary set
- Add, recode/merge, drop codes
- Repeat

HICode: Pipeline Overview



Zhong et al., 2025

Prompting that works

```
{Background Information}
{Goal of Inductive Coding}

Instruction:
- Label the input only when it is HIGHLY
RELEVANT and USEFUL for {Goal of Inductive
Coding}.
- Then, define the phrase of the
label. The label description should be
observational, concise and clear.
- ONLY output the label and DO NOT output
any explanation.

Format:
- Define the label using the format
\"LABEL: [The phrase of the label]\".
- If there are multiple labels, each label
is a new line.
- If the input is irrelevant, use \"LABEL:
[Irrelevant]\".
- The label MUST NOT exceed 5 words.
```

- Nuanced literature review
- Analysis Question: What are values encoded in machine learning research?
- Data: Abstracts and introductions of ML papers

Zhong et al., 2025

Prompting that works

```
{Background Information}
{Goal of Inductive Coding}

Instruction:
- Label the input only when it is HIGHLY
RELEVANT and USEFUL for {Goal of Inductive
Coding}.
- Then, define the phrase of the
label. The label description should be
observational, concise and clear.
- ONLY output the label and DO NOT output
any explanation.

Format:
- Define the label using the format
\"LABEL: [The phrase of the label]\".
- If there are multiple labels, each label
is a new line.
- If the input is irrelevant, use \"LABEL:
[Irrelevant]\".
- The label MUST NOT exceed 5 words.
```

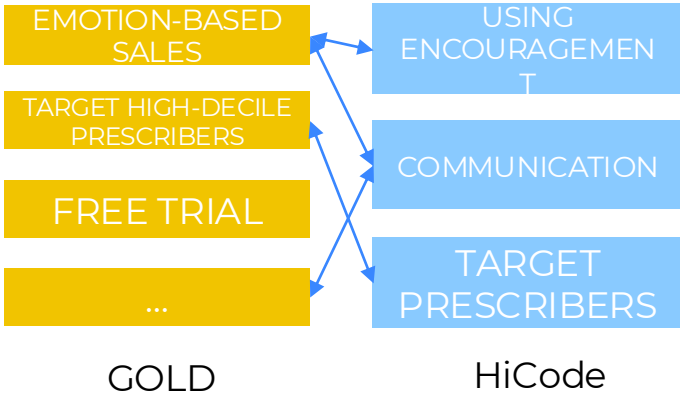
- Background information: clarify definitions and describe data
  - A value of an entity is a property that is considered desirable for that kind of entity. Highly cited papers are collected to identify and analyze emergent values of the machine learning field. We are using scientific papers from machine learning to conduct INDUCTIVE LABELING.
- Goal of Inductive Coding
  - what specific values regarded and justify as desirable attributes for machine learning research

Evaluation is hard

Goal: comprehensive + minimal. ➔ Modified precision and recall

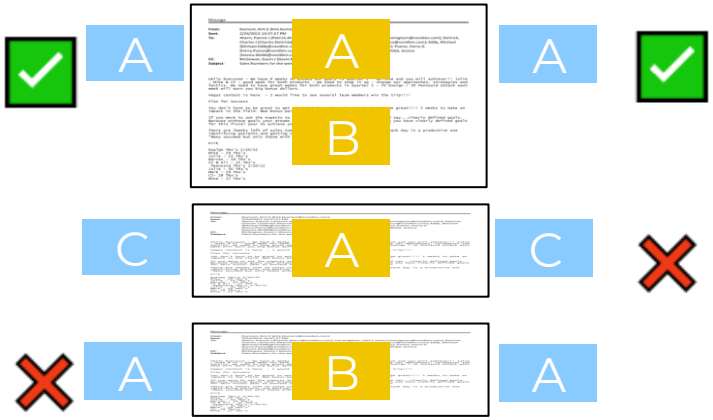
Theme-level

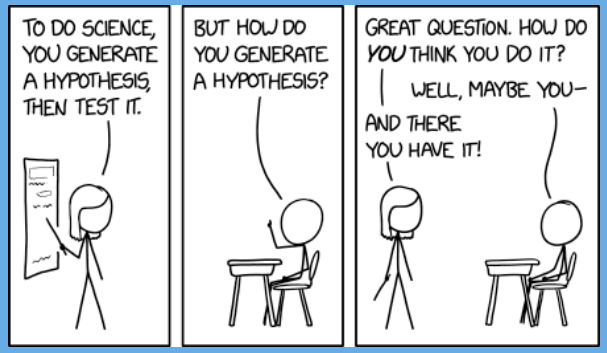
- How much does model capture the themes annotated by human?



Segment-level

- Given the correct final theme, does it label the text correctly?





Hypothesis Generation with Sparse Autoencoders

Image Credit: https://xkcd.com/2569/

Autoencoder

- Goal: Output $\hat{\mathbf{x}}$ = Input $\mathbf{x}$
- $\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = ||\mathbf{x} - \hat{\mathbf{x}}||^2$
- Idea: Approximate an identity matrix with constraints
- Encoder: Compresses into low-dimensional latent representation $\mathbf{h}$
  .

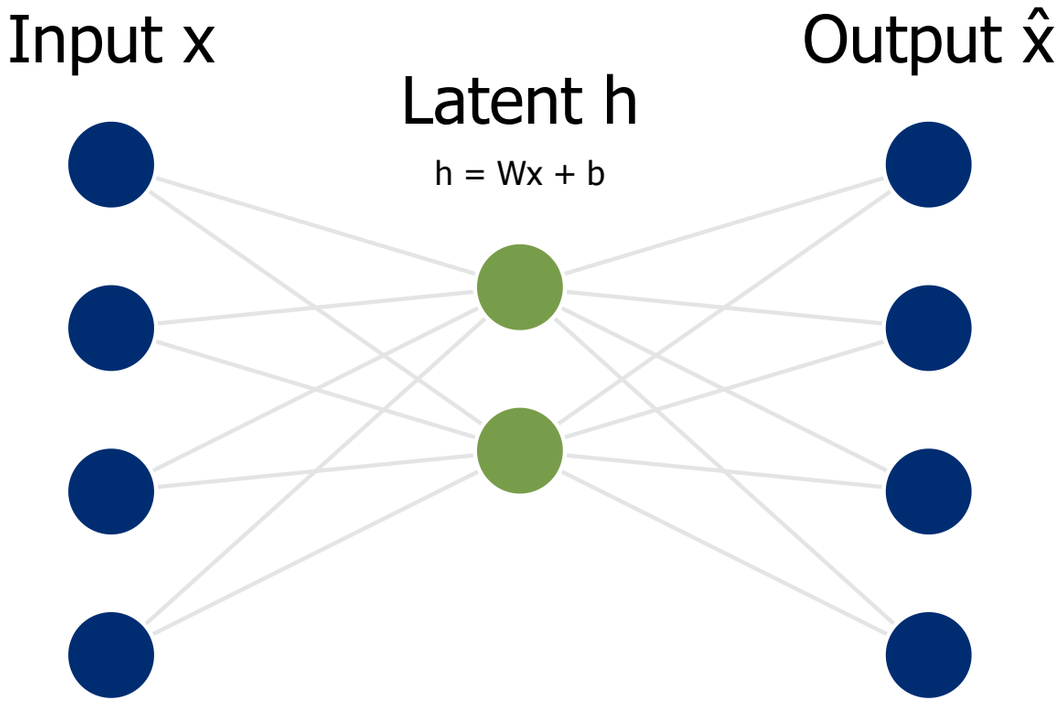
$$\mathbf{h} = \mathbf{W}\mathbf{x} + \mathbf{b}$$

- Decoder: Reconstructs $\hat{\mathbf{x}}$ from $\mathbf{h}$.

$$\hat{\mathbf{x}} = \mathbf{W}'\mathbf{h} + \mathbf{b}'$$

$$\hat{\mathbf{x}} = \mathbf{W}'\mathbf{W}\mathbf{x} + bias$$

Looks familiar? A linear autoencoder can be equivalent to PCA.



Going to non-linear

Activation functions to rescue

Latent vector:

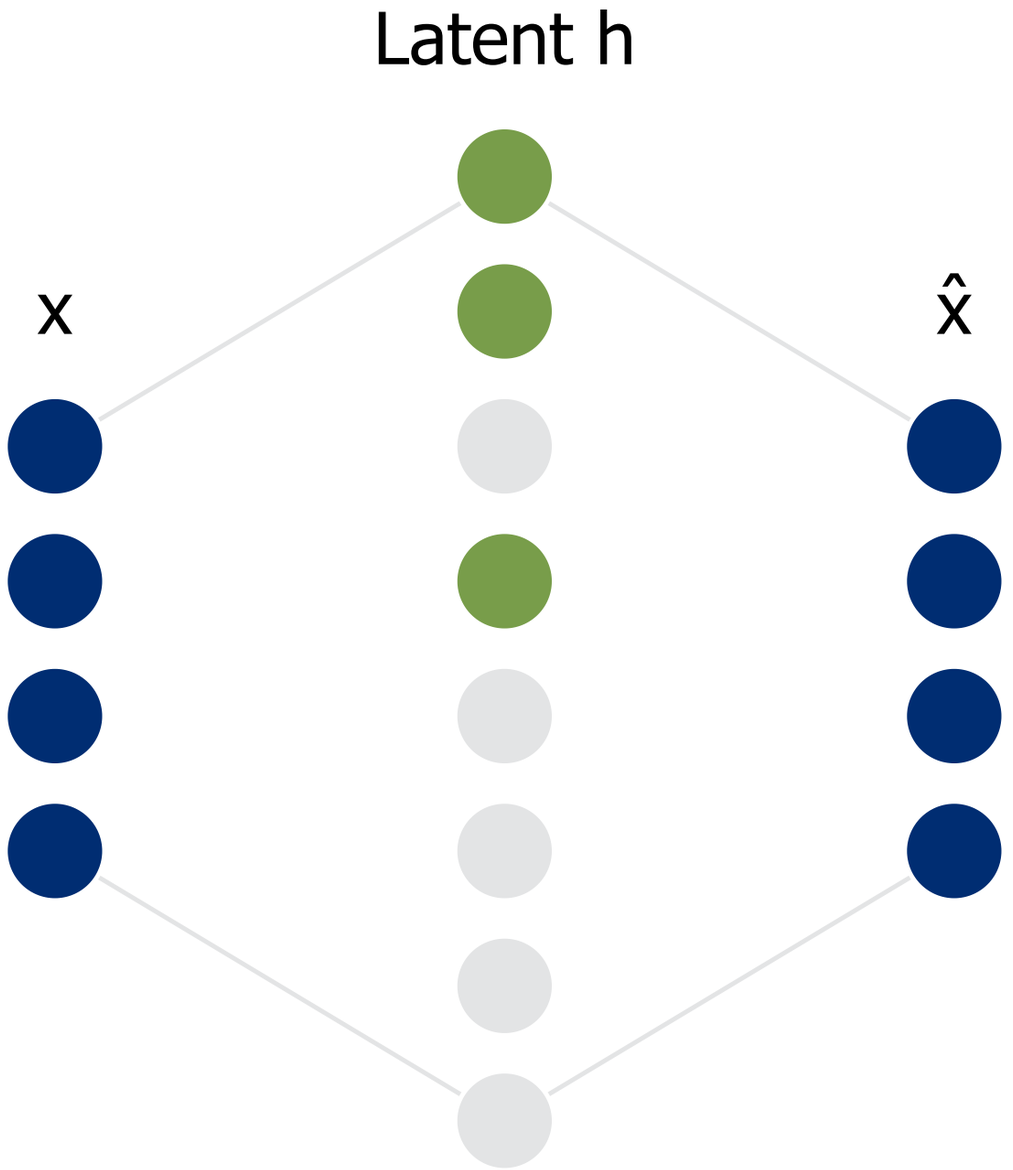
$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$$

Sigmoid activation:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Enforce Sparsity

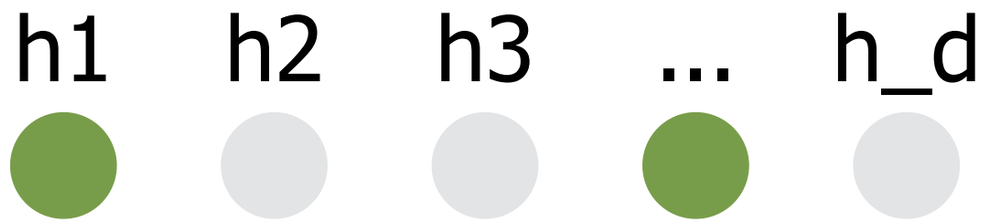
- Motivation: sparse coding hypothesis from neuroscience
- Some neurons are useful -> active (push it to 1), others are not -> inactive (push it to 0)
- Often larger latent dimension than input dimension -> overcomplete



Sparse Autoencoder: KL divergence penalty

Add a sparsity penalty term to the loss function.

If a vector is sparse, the average activation should be small.



$\hat{h} := \frac{1}{d}(h_1 + h_2 + ... + h_d) \to 0$

Sparse Autoencoder: KL divergence penalty

Formally, for each $h_i = [h^1, ..., h^d]$, enforce $\hat{h_i} = \rho$, e.g., $\rho = 0.05$

$$KL(\rho||h_i) = \rho \log \frac{\rho}{\hat{h}_i} + (1-\rho) \log \frac{1-\rho}{1-\hat{h}_i}$$

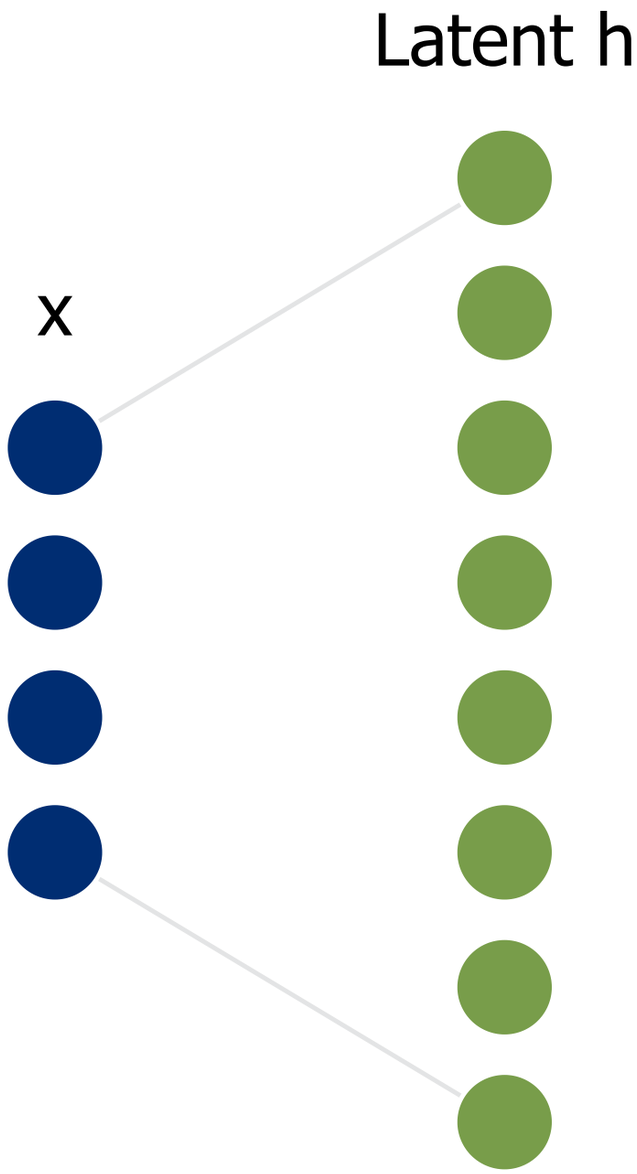
Add this penalty to the original loss function:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) + \beta \sum KL(\rho||\hat{h}_i)$$

k-Sparse Autoencoder

- Makhzani and Frey, 2014
- Constraint: keep exactly the top k highest activations, and zero out the rest.

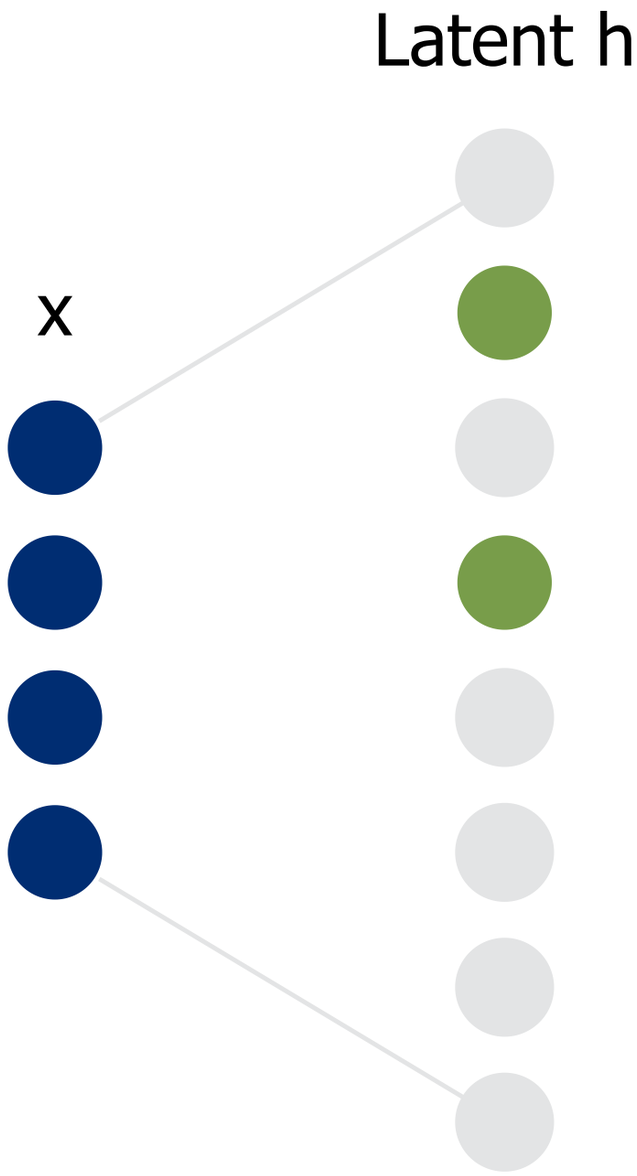
1. Linear feedword pass $h = Wx + b$



k-Sparse Autoencoder

- Makhzani and Frey, 2014
- Constraint: keep exactly the top k highest activations, and zero out the rest.

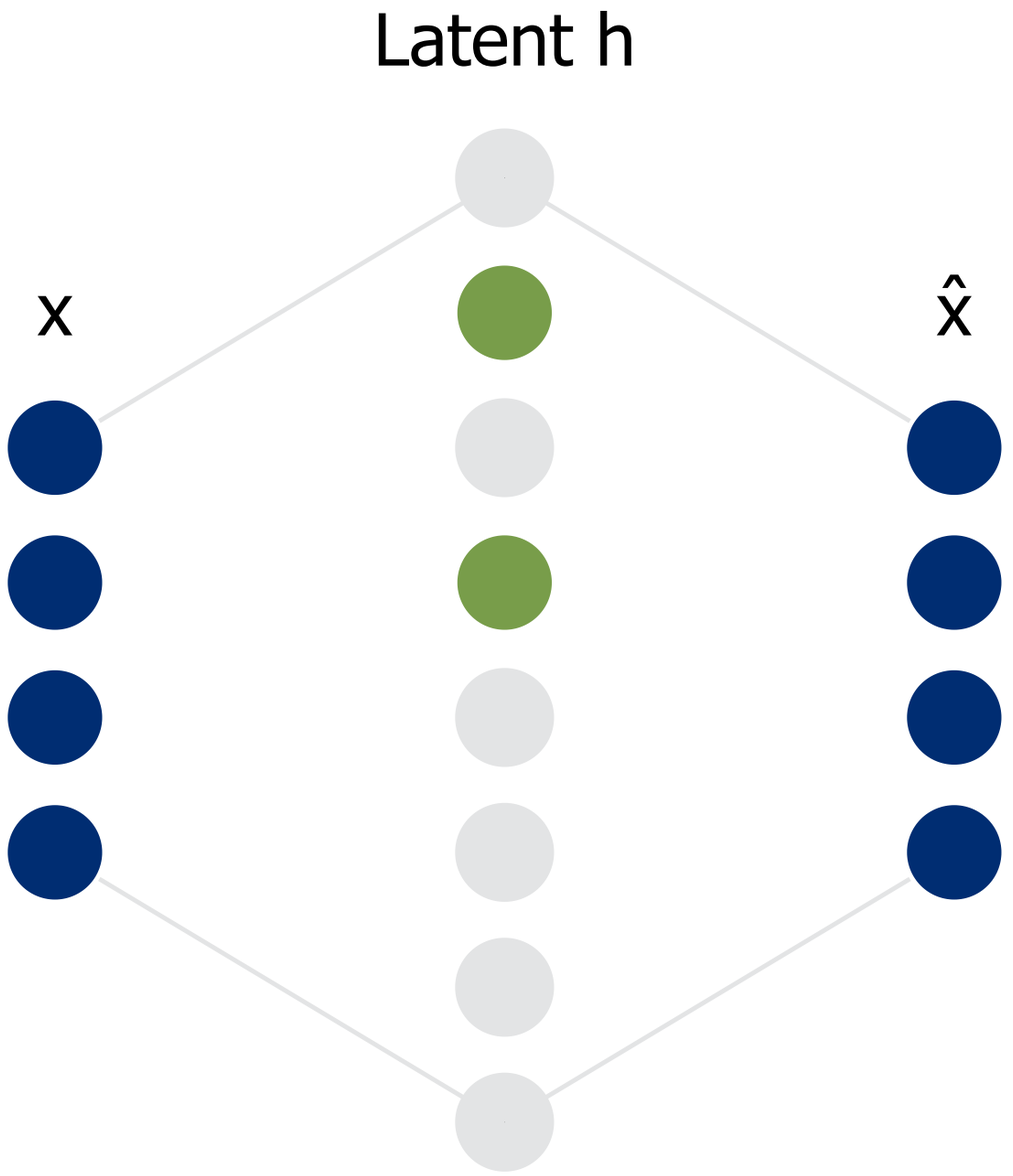
1. Linear feedword pass $h = Wx + b$
2. Find k largest activations and zero out the rest



k-Sparse Autoencoder

- Makhzani and Frey, 2014
- Constraint: keep exactly the top k highest activations, and zero out the rest.

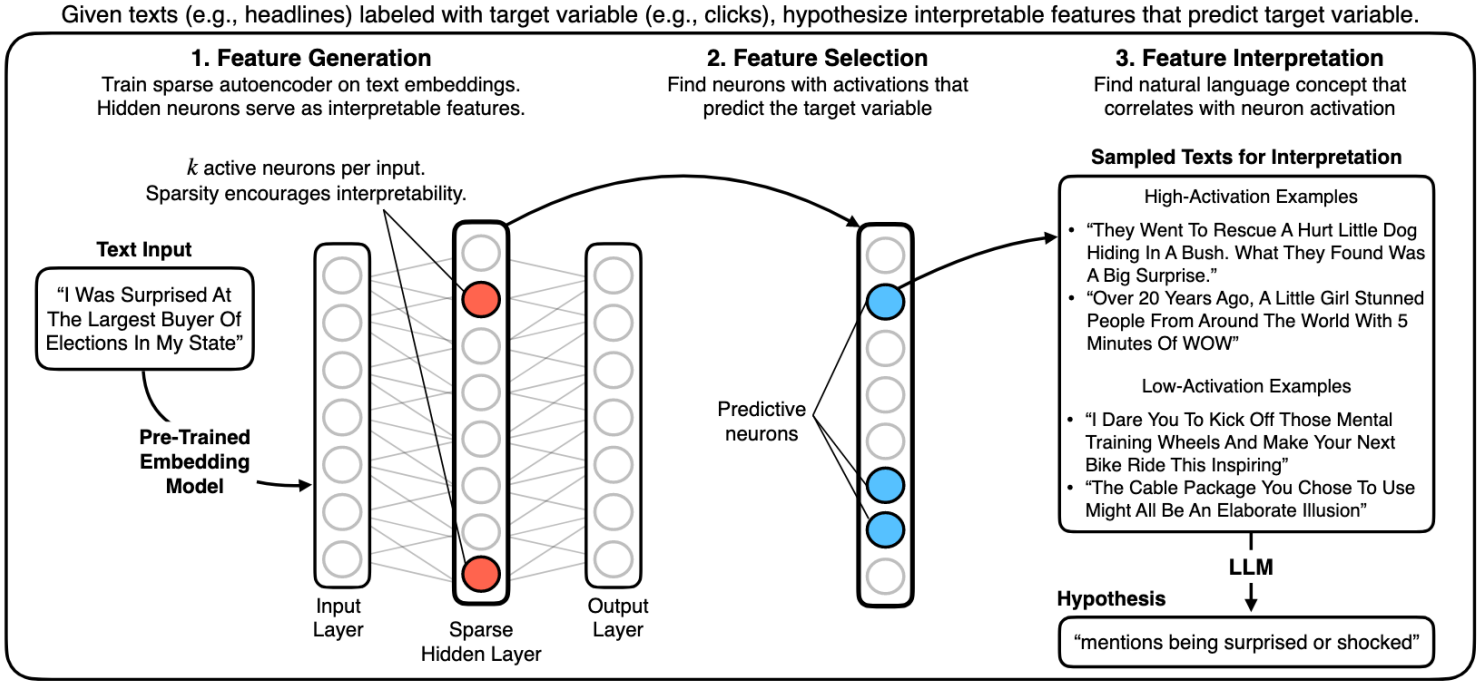
1. Linear feedword pass $h = Wx + b$
2. Find k largest activations and zero out the rest
3. Reconstruct $\hat{x} = W^T h + b'$ and backpropagate



Hypothesis Generation

- Discover themes from corpus
  - You have an analysis question to study, e.g. political framing
    - -> target variable "Party affiliation (D/R)"
  - Generate (interpretable) predictive features
    - Fightin' words
    - Topics from your favorite topic models
    - Inductive codes from HiCode
- Hypothesis: an interpretable feature realized by natural language description that can be further tested for generalization

HypotheSAEs: Pipeline



Movva et al., 2025

HypotheSAEs: example result

Significant Hypotheses: Red: ↑ Republican ; Blue: ↑ Democrat	Sep.	AUC
contains the phrase 'I ask unanimous consent'	0.30	0.53
mentions hearings conducted by Senate Committees or Subcommittees	0.18	0.52
mentions scheduling or details about Senate votes or amendments	0.17	0.55
discusses illegal immigration and its associated implications	0.15	0.51
mentions victories or successes in military or political contexts	0.15	0.51
mentions freedom or liberty	0.06	0.50
discusses economic growth or growth rates	0.01	0.50
discusses government spending or budgetary issues	-0.09	0.54
mentions the need for reform or proposes reforms	-0.17	0.58
mentions key figures related to the civil rights movement	-0.22	0.51
criticizes government policies or actions	-0.29	0.63
mentions the U.S. national debt or raising the debt limit	-0.40	0.51
criticizes tax breaks or advantages for the wealthy	-0.44	0.52
criticizes Republican leadership or policies	-0.45	0.60
criticizes the administration's handling of the Iraq war	-0.45	0.52

HypotheSAEs: comparing with baselines

- Less cost
- More significant hypotheses
- Qualitative evaluation
  - Helpfulness
  - Interpretability

Source	Overall AUC	# Sig	Significant Hypotheses: Red: ↑ Republican ; Blue: ↑ Democrat	Sep.	AUC
HypotheSAEs	0.702	15	*contains the phrase 'I ask unanimous consent'*	0.30	0.53
			mentions hearings conducted by Senate Committees or Subcommittees	0.18	0.52
			mentions scheduling or details about Senate votes or amendments	0.17	0.55
			discusses illegal immigration and its associated implications	0.15	0.51
			mentions victories or successes in military or political contexts	0.15	0.51
			mentions freedom or liberty	0.06	0.50
			discusses economic growth or growth rates	0.01	0.50
			discusses government spending or budgetary issues	-0.09	0.54
			mentions the need for reform or proposes reforms	-0.17	0.58
			mentions key figures related to the civil rights movement	-0.22	0.51
			criticizes government policies or actions	-0.29	0.63
			mentions the U.S. national debt or raising the debt limit	-0.40	0.51
			criticizes tax breaks or advantages for the wealthy	-0.44	0.52
			criticizes Republican leadership or policies	-0.45	0.60
			criticizes the administration's handling of the Iraq war	-0.45	0.52
BERTopic	0.636	10	*Senate procedural requests and adjournment schedules*	0.39	0.53
			Requests for unanimous consent to authorize committee meetings	0.32	0.53
			procedural discussions and scheduling of votes	0.20	0.56
			discusses immigration and border security	0.13	0.51
			mentions individuals with professional titles or roles	0.08	0.53
			debates on earmarks and federal spending processes	0.05	0.50
			discusses the Darfur conflict and international intervention	-0.30	0.50
			Congressional oversight of defense contracting in Iraq	-0.33	0.50
			discusses Asian Pacific American communities	-0.36	0.50
			Critiques Republican majority leadership	-0.46	0.55
NLParam	0.650	8	*expresses strong patriotism*	0.09	0.51
			contains specific legislative references	0.08	0.53
			mentions specific individuals or organizations	-0.03	0.52
			includes numerical data	-0.08	0.53
			addresses a critical national issue	-0.14	0.56
			expresses strong support or opposition	-0.15	0.58
			advocates for specific groups or communities	-0.17	0.55
			employs emotional or dramatic language	-0.23	0.58
HypoGeniC	0.675	5	*focuses on national security and immigration enforcement*	0.18	0.51
			discusses social issues such as healthcare and civil rights	-0.22	0.58
			mentions of government inefficiencies or calls for reform	-0.27	0.62
			advocates for civil liberties and critiques government overreach	-0.29	0.57
			emphasizes the need for government accountability	-0.37	0.62

Movva et al., 2025

Last Slide Today!

- HiCode: Hierarchical Inductive coding at Scale
  - Consider to use it in your projects!
- Hypothesis Generation via Sparse Autoencoder
  - Potential caveats? Feature splitting/absorbtion
- Going forward…
  - Multimodal Corpus Analysis?
  - Better evaluation that captures human judgement?

References

- Yang, J., & Li, P. (2026). From exile to visibility: Performative translation and identity construction of TikTok refugees on Xiaohongshu. *Discourse, Context & Media*, *71*, 101006. https://doi.org/10.1016/j.dcm.2026.101006
- Zhong, M., Wang, P., & Field, A. (2025). HICode: Hierarchical Inductive Coding with LLMs. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 31060–31078). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1580
- CS 294A lecture notes: https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf
- Adam Karvonen's blog post on SAE: https://adamkarvonen.github.io/machine_learning/2024/06/11/sae-intuitions.html
- Makhzani, A., & Frey, B. (2014). *K-Sparse Autoencoders* (arXiv:1312.5663). arXiv. https://doi.org/10.48550/arXiv.1312.5663
- Movva, R., Peng, K., Garg, N., Kleinberg, J., & Pierson, E. (2025). *Sparse Autoencoders for Hypothesis Generation* (arXiv:2502.04382). arXiv. https://doi.org/10.48550/arXiv.2502.04382