From Legacy Autocoders to Large Language Models: New Approaches to Coding Industry and Occupation Data

National Academies AI Day, April 30, 2026

Collaborators: Julia Beckhusen[1], Lynda Laughlin[1], Ana J. Montalvo[1], Madison Danton[1], Jackson Chen[2], Yezzi Angi Lee[2], and Jiahui Xu[2]

[1]U.S. Census Bureau, [2]Reveal Global

United States® Census Bureau

Industry and Occupation, or the Most Common Icebreaker in Washington, D.C.: What's Your Job?

- Produce workforce and economic statistics.
- Support federal funding and workforce policies.
- Produce reports, public-use datasets, and detailed tables.
- Maintain and inform industry and occupational classification systems:
  - Standard Occupational Classification (SOC) system.
  - North American Industrial Classification System (NAICS).

2025 American Community Survey (ACS) Questionnaire

Industry Items:

- Industry data describe the kind of business conducted by a person's employing organization.
- 3 industry questions:
  - 2 write-ins.
  - 1 checkbox.
- Industry questions:

1. What is the name of your employer, business, agency, or branch of the Armed Forces?

   Example response: United States Census Bureau.

2. What kind of business or industry is this?

   Example Response: Federal statistical agency.

Occupation Items:

- Occupation describes the kind of work a person does on the job.
- 2 occupation items:
  - Main occupation.
  - Occupation duties.
- Occupation questions:

1. What is your main occupation?

   Example Response: Data scientist.

2. Describe the most important activities or duties of your occupation.

   Example Response: Building machine-learning models.

2019 ACS Example of Write-ins*

Column A: 2018 Public-Use Microdata Sample (PUMS) occupation (OCC) code.

Column B: Occupation write-in: *What was this person's main occupation?*

Column C: Occupation write-in: *Describe this person's most important activities or duties.*

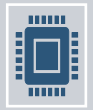
Column D: 2018 Public-Use Microdata Sample (PUMS) industry (IND) code.

Column E: Industry write-in: *What kind of business or industry was this?*

* Public-use sample of Industry and Occupation (I&O) write-ins from the 2019 ACS questionnaire.

2018 PUMS OCC Code	Occupation Write-in (OCW1)	Occupation Write-in (OCW2)	2017 PUMS IND Code	Industry Write-in (INW3)
0010	CHIEF FINANCIAL OFFICER	HANDLE THE MONEY	7071	REAL ESTATE
0020	GENERAL MANAGER	PROFIT & LOSS EMPLOYEE RETENTION OVERALL OPERATIONS	8680	RESTAURANT
0205	FARMER	FARM OPERATER	0170	FARMING
0220	GENERAL CONTRACTOR	REMODELING	0770	CONSTRUCTION
1021	SOFTWARE ENGINEER	MAINTAIN GOVERNMENT SATELLITE PROGRAMS	7290	GOVERNMENT CONTRACTOR
1800	ECONOMIST	DATA ANALYSIS	9570	GOVERNMENT
2300	PRESCHOOL TEACHER	INSTRUCT AND EVALUATE STUDENTS IN ALL AREAS OF CURRICULUM AND CREATE LESSON PLANS	7860	SCHOOL DISTRICT
3090	PHYSICIAN	PRACTICE MEDICINE	8191	HEALTHCARE
3602	CARING FOR SENIORS	CLEANING	8170	HOMECARE
3870	STATE TROOPER	LAW ENFORCEMENT	9470	LAW ENFORCEMENT
4150	HOSTESS	SEATING PEOPLE AND TAKINMG ORDERS	8680	RESTARAUNT
4760	RETAIL ASSOCIATE	CUSTOMER SERVICE	4971	GROCERY STORE
4840	SELLS AND DELIVERY	SELLS AND DELIVERY	1990	PRINTING SHOP
5740	OFFIVE MGR	ACCOUNTING	9180	LABOR UNION
7700	SUPERVISOR	OVERSEE EMPLOYEES SHIP PRODUCTS	1870	PAPER COMPANY
7810	MEAT CUTTER	BUTCHER	4971	GROCERY
8030	MACHINEST	MAKING PARTS FOR PLANT EQUIPMENT	1090	FROZEN FRUIT PROCESSING
8320	SEWER	SEW MARINE PRODUCTS FOR WHOLESALE DISTRIBUTORS	3291	COMMERCIAL SEWING
8740	QUALITY INSPECTOR	MEASURING AND VISUAL INSPECTION	3390	MANUFACTURING
8810	PAINTER	PAINTS AUTOMOBILES	8770	AUTO BODY
9121	BUS DRIVER	DRIVES SCHOOL BUS	7860	EDUCATION
9130	TRUCK DRIVER	DRIVE TO STORES THEN UNLOAD	6170	TRUCKING COMPANY
9620	PRODUCE CLERK	UNLOAD AND BREAK PALLETS OF PRODUCT INTO BACK ROOM FILL PRODUCT OUT ON SALES FLOOR	4971	GROCERY STORE
9620	LABORER	BUILDING FORMS MOVING MATERIALS	7580	TEMP AGENCY
9645	STOCKER	UNLOAD TRUCKS STOCK SHELVES	5275	RETAIL
9720	SANITATION WORKER	GARBAGE REMOVAL	7790	GARBAGE REMOVABLE

Challenges and Goals



Coding Industry and Occupation (I&O) data for the ACS is a massive operation.

Every year over 2 million industry and occupation write-ins are assigned Census Bureau I&O codes.

Until ACS data year 2012, 100% of cases were reviewed by an I&O coder at the National Processing Center (NPC) and assigned an industry and an occupation code.

To lower costs, an automated coding system was designed and implemented to assign industry and occupation codes to a percentage of these write-ins. Since 2012, around 70% of ACS cases are sent to NPC for clerical coding.

Legacy Autocoding Process

01

Match responses to a "data dictionary".

- Data dictionary is built from past clerical-coded responses and contains word bits (single/multi-word phrases) linked to industry and occupation codes.

02

Use logistic regression.

- Predicts the best I&O code when multiple matches exist and outputs a probability score (PHAT) for each possible code.
- PHAT threshold is set at 90%.

03

Apply hard codes (if applicable).

- Rules manually added to correct common misclassification.

04

Results

- High PHAT (above 88%) → autocoded.
- Low PHAT (below 88%) → sent for clerical review.
  - Should either industry or occupation code have a PHAT score below 88%, then both codes are sent for clerical review.

Need for Improved I&O Autocoder

Challenges in I&O coding:

- Evolving language and occupation titles.
- Complexity of responses.
- Ensuring consistency between autocoder and clerical codes.

ACS autocoder is effective but requires continuous quality checks to ensure:

- High accuracy and reliability.
- Early detection of systematic errors and biases.
- Continuous refinement and improvement.

New Opportunities for Coding Text Data on Federal Surveys

- Traditional approaches relied heavily on rules-based systems such as word bit dictionaries.
  - Heavy dependence on manually maintaining dictionaries.
  - Poor performance on rare, emerging, or new/unseen responses.
  - Limited scalability.

- Large Language Models have significantly expanded what is possible in coding text data.
  - Better capture context and meaning in free-text responses.
  - Learn relationships across related classification systems or coding rules.
  - Process text with far greater flexibility.

Approach 1: Semantic Search

LLM semantic search: Experimented with various embedding models, context retrieval windows, and reranking strategies to build a robust, LLM-driven search engine.

- Pros:
  - Grounded outputs: Matches only to predefined documents or code lists.
  - Flexible updates: Easily incorporate new or emerging terms, no retraining required.
  - Robust to messy input: Handles misspellings, abbreviations and unstructured text.
  - Easier to update and doesn't require a large amount of training data.
- Cons:
  - Limited interpretability: Can struggle with fine distinctions.
  - Potential mismatches: It may retrieve plausible but wrong matches.
  - Decision layer needed: May need fine-tuning or an additional decision layer to assign final codes.

Approach 2: LLM Autocoder

Multi-task transfer learning: Used a two-head, multi-task architecture to predict industry and occupation codes simultaneously, ensuring high accuracy while eliminating model hallucinations.

- Pros:
  - Highly consistent with structured coding like the Census Bureau's Occ/Ind code lists.
  - Can code both industry and occupation and use to inform one another.
  - Efficient at scale and doesn't necessarily need a lot of computing power.
  - Prevents hallucinations because the model is restricted to set code lists.
- Cons:
  - Needs lots of high-quality coded training data.
  - Can struggle with new or emerging occupation or industry terms.
  - Updating the model can take a lot of time.

Using Large Language Models

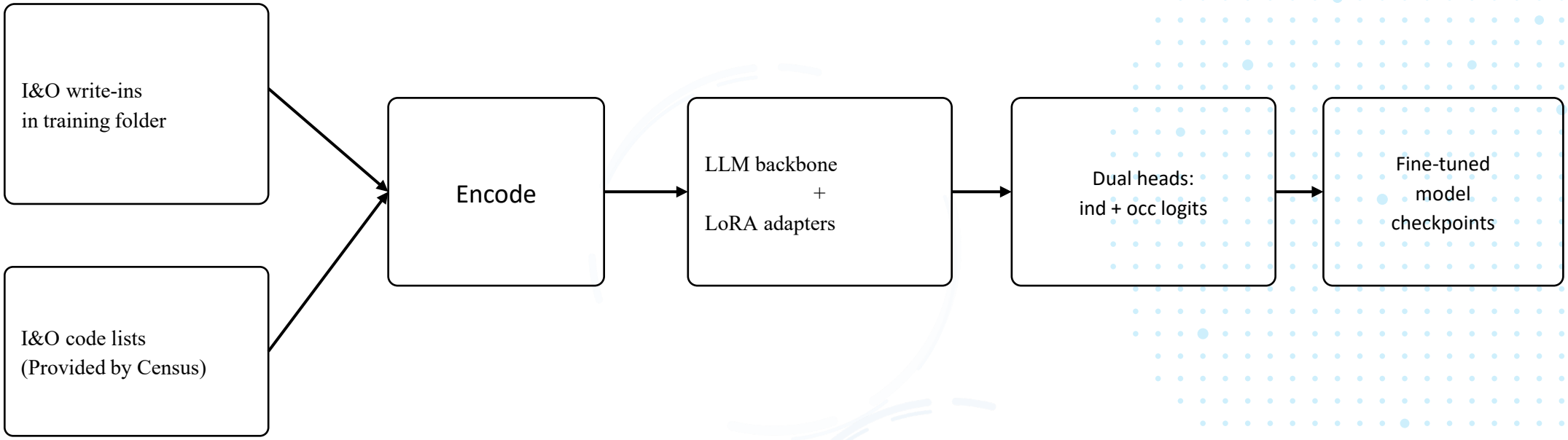
Choosing the best LLM:

- Large language models (LLMs) represent sentences as vectors, capturing semantic information that helps them understand the meaning of sentences.
- With numerous models available, optimization is focused on both size and performance to maximize efficiency.
- Contractors tested a variety of models and ultimately chose **Approach 2, the multi-task transfer learning** model for the LLM autocoder for production coding purposes.

Training and running model:

- Model trained with 2024 clerically coded write-in data. No other inputs required.
- Choose probability threshold: 0.82 leads to error rate of about 6%.
    - Clerical coders are required to have an error rate of 6% or less.
- After initial review and before going into production:
    - Further fine-tuning.
    - Additional training with select cases from 2023.
    - Additional training with “problematic” cases.

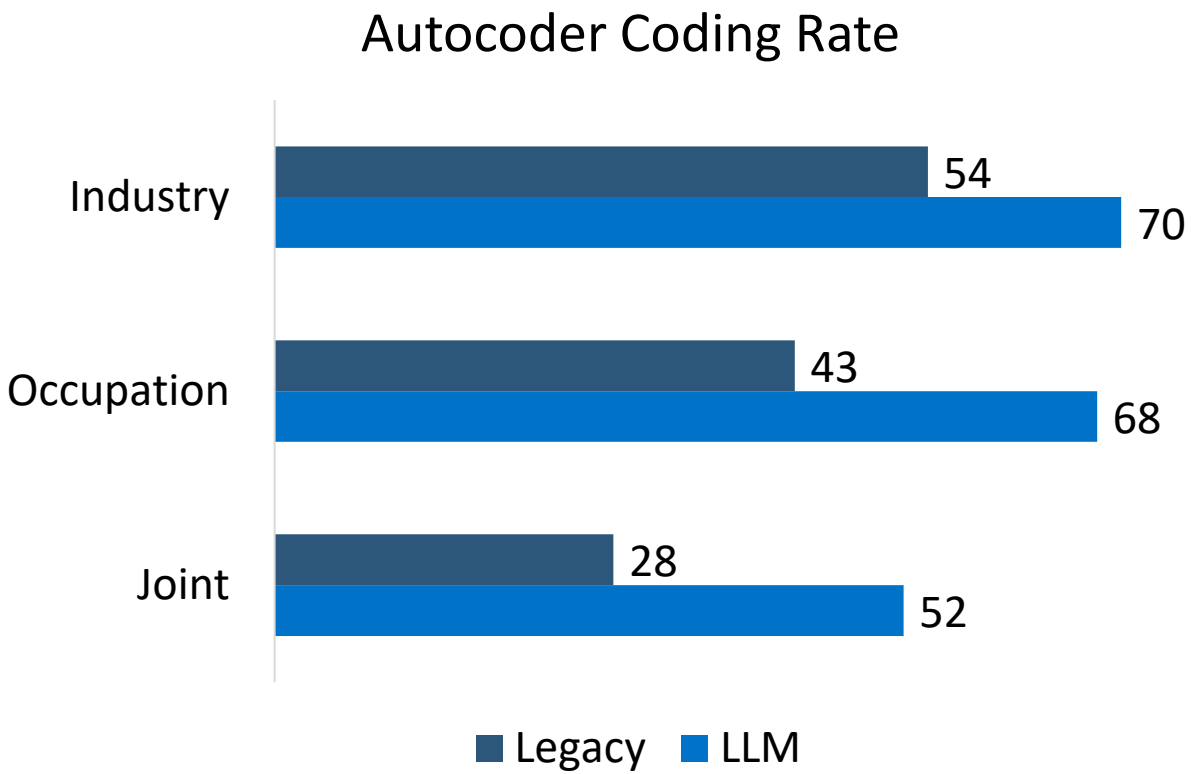
LLM Autocoder: Fine-Tuning (Training) Pipeline



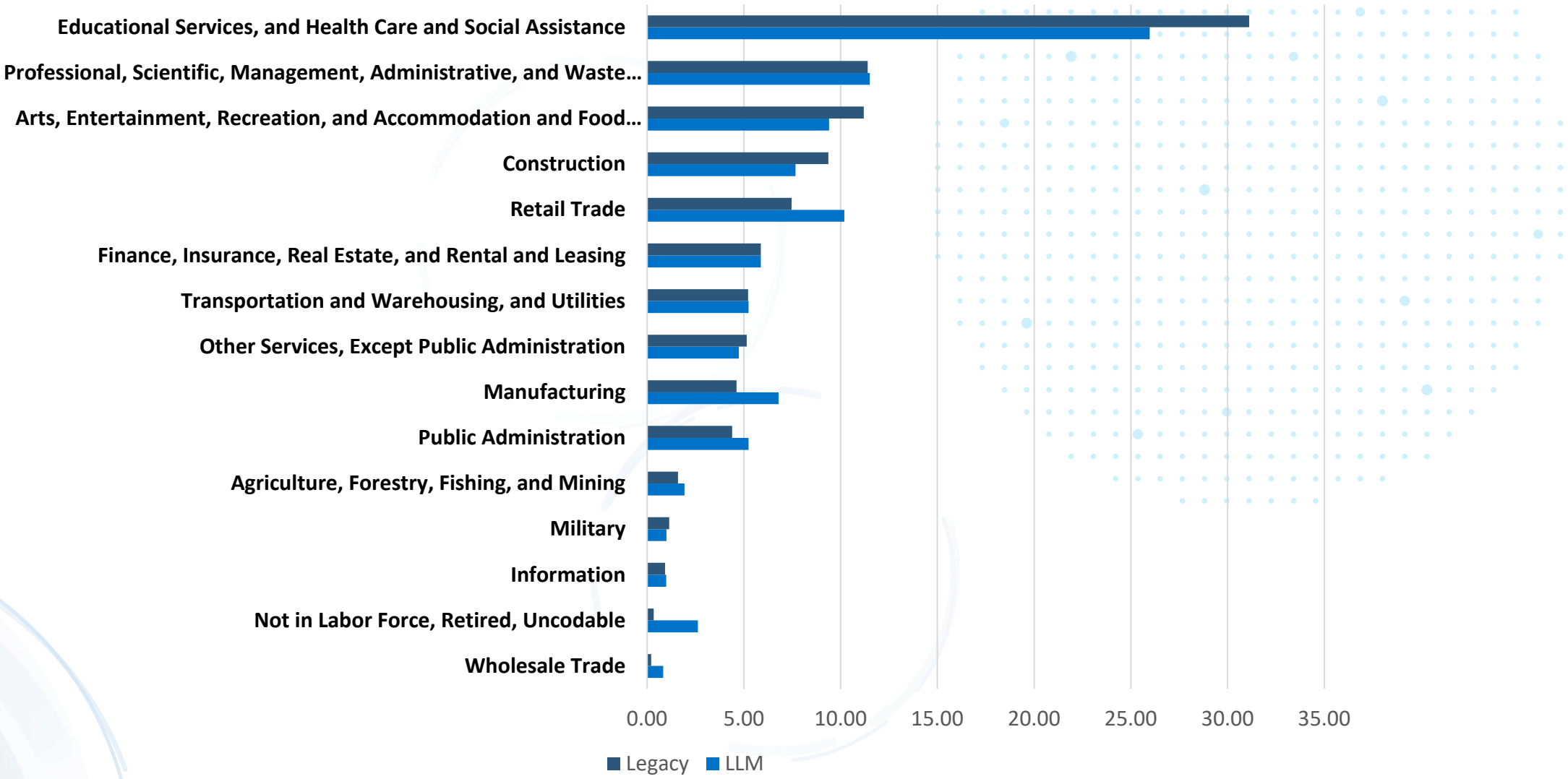
Our core model or "backbone" is Llama – an open-source decoder-only transformer by Meta.

- One shared language model reads all text inputs.
- Uses clues from employer name to improve occupation coding.
  - Example: *"Hospital" → likely healthcare jobs*.
- Uses clues from occupation title/duties to improve industry coding.
  - Example: *"Registered nurse" → likely healthcare industry*.
- Industry + occupation are predicted together.

Single and Joint Coding Rates

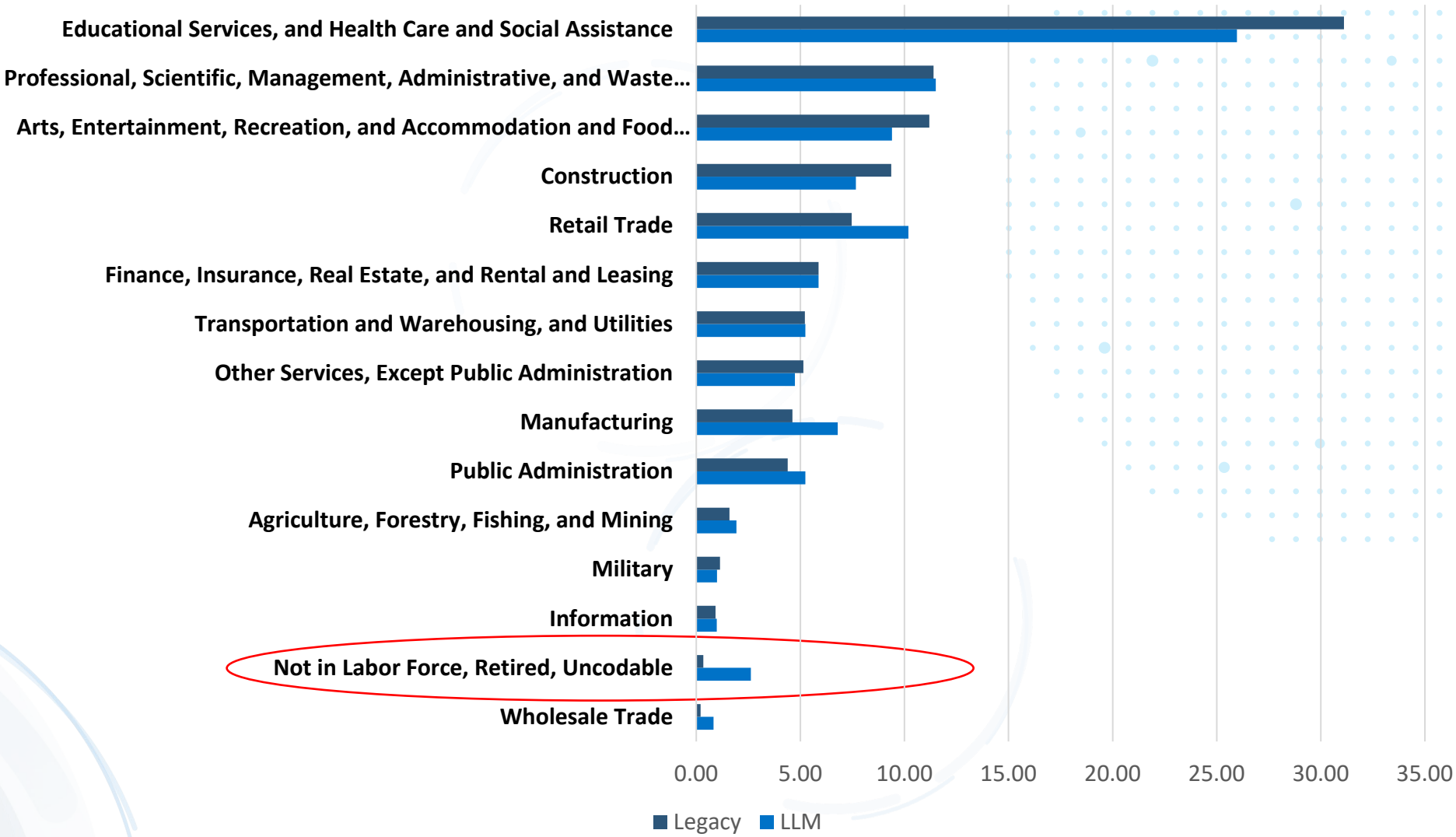


Distribution of Industry Assigned by LLM vs Legacy Autocoder



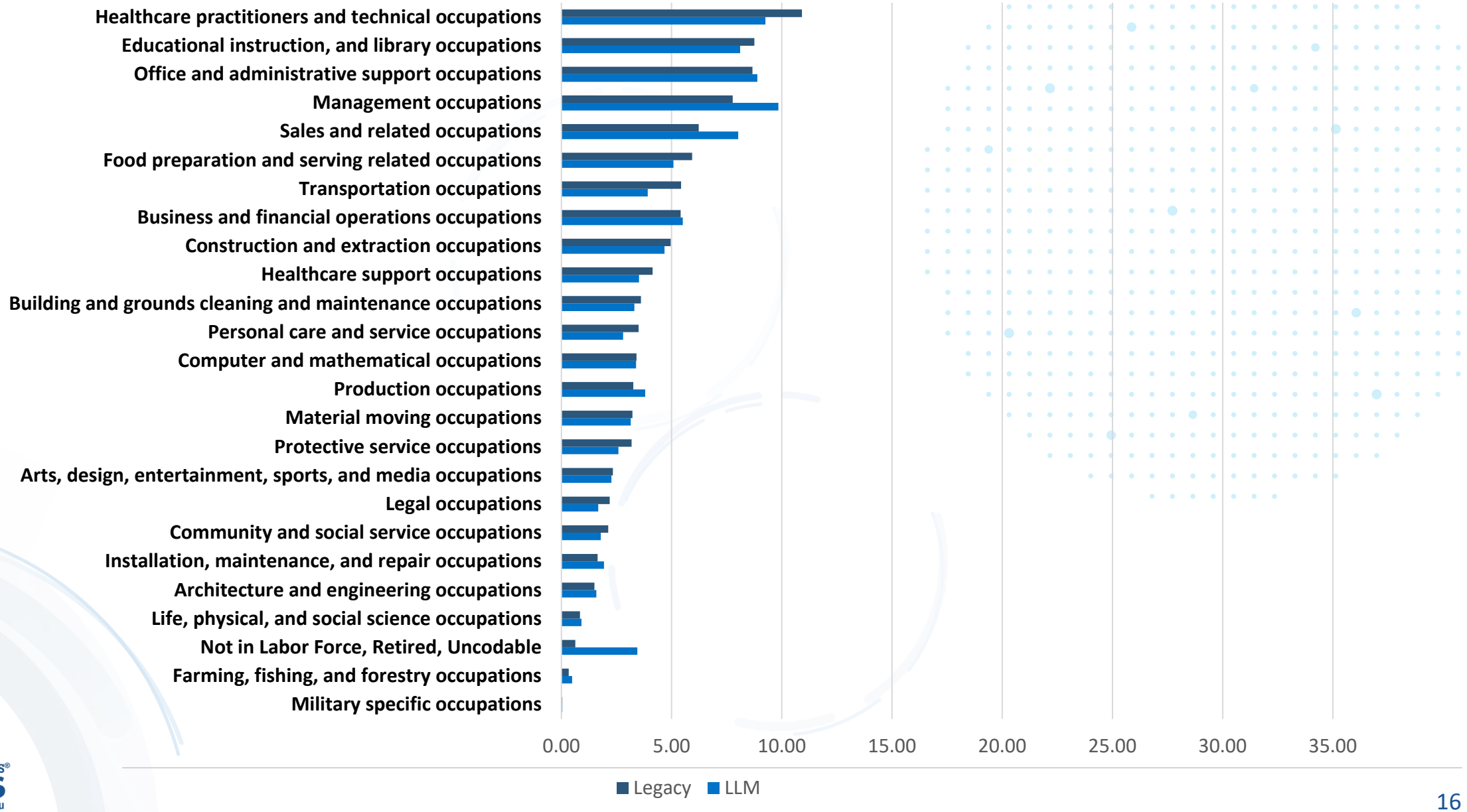
United States® Census Bureau

Distribution of Industry Assigned by LLM vs Legacy Autocoder



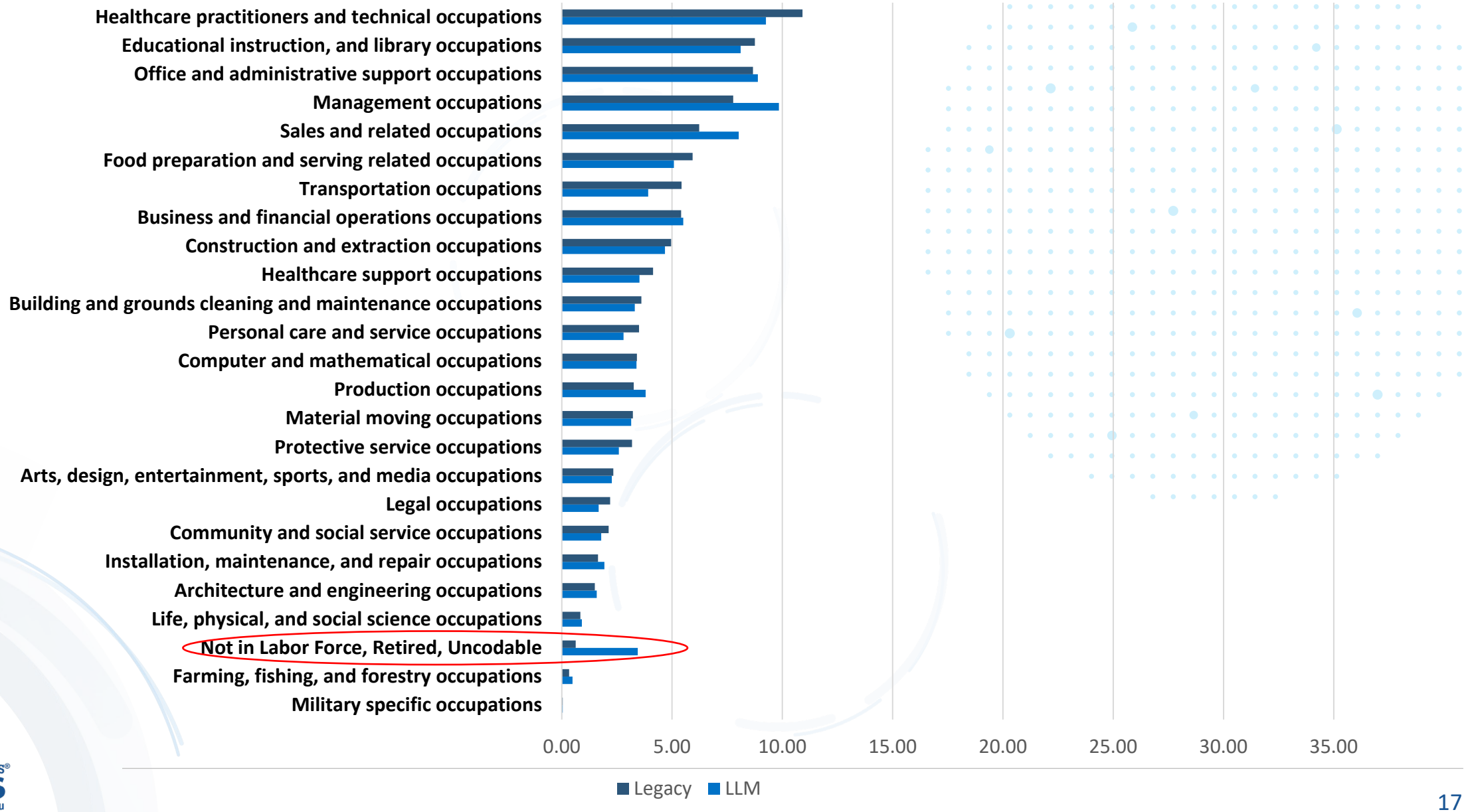
United States® Census Bureau

Distribution of Occupation Assigned by LLM vs Legacy Autocoder



United States® Census Bureau

Distribution of Occupation Assigned by LLM vs Legacy Autocoder



United States® Census Bureau

Where We Started and Where We Are Today

- Resources and tools we developed along the way:
  - 2019 public-use write-in file: A great source of training data.
  - Working papers and conference presentations.
  - Forthcoming methodology paper on how the LLM coder was built and steps on how to maintain and retrain.

Key Takeaways

- Joint coding rate increased from 28 percent to 52 percent.
  - 32 percent fewer cases sent to NPC each month or 500,000 fewer cases per year.
- Access to a cloud environment.
  - AWS or cloud-based computing environment is vital for our LLM coder.
  - LLM autocoder takes around 6 hours to run; legacy coder took 20 mins.
- Pay close attention to the referral rate:
  - Does increase in referral rate undo cost savings?
  - Identify model improvements to increase coding rate of most-referred codes.
- Legacy coder still provides value.
  - Automate updating the word bit dictionaries.
- Not all text coding tasks require LLMs.
  - Some write-ins are better handled with fuzzy matching or deterministic rules.
  - Computational costs are lower and easier to maintain.

Thank You!

Contact information:

Lynda Laughlin

lynda.l.laughlin@census.gov

United States® Census Bureau