

# Using Large Language Models for Other-Specify Coding

---

**FedCASIC 2024**

April 17, 2024

Weihuang Wong



# National Survey of Early Care and Education

## About the NSECE

- This study uses data from the 2019 NSECE
- Nationally representative survey of ECE supply and demand
- Sponsored by the Office of Planning, Research, and Evaluation in the Administration for Children and Families
- Previously conducted in 2012 and 2019; data collection is ongoing for the 2024 NSECE



NATIONAL SURVEY OF EARLY CARE & EDUCATION

# Agenda

01 Method and Findings

---

02 Why do respondents choose Other?

---

03 Making LLM-onade from LLM-ons

---



---

# Method & Findings

## OS-coding

A7.


In what kind of building is your program located? Please choose one only for each building your program occupies.

1. Religious building
2. Public School
3. Private School
4. University or College
5. Work Place
6. Community Center or Municipal Building
7. Commercial Structure
8. Independent Structure (i.e., ORGANIZATION is the sole occupant)
9. Home, apartment, or other residential structure
10. Other, specify: \_\_\_\_\_

## OS-coding

A7.

In what kind of building is your program located? Please choose one only for each building your program occupies.

1. Religious building
  2. Public School
  3. Private School
  4. University or College
  5. Work Place
  6. Community Center or Municipal Building
  7. Commercial Structure
  8. Independent Structure (i.e., ORGANIZATION is the sole occupant)
  9. Home, apartment, or other residential structure
  10. Other, specify: Ground floor commercial space of condo
- 

# Objective

## **OS-coding is necessary but potentially burdensome**

- Conventionally: 2 human coders + 1 expert adjudicator
- Major advances in large language models (LLMs) in the last 1-2 years
- Can we reduce labor burden? E.g. 1 AI + 1 human coder + 1 expert adjudicator?

## Objective 2

### **Using open-source LLMs deployed on consumer-grade hardware**

- Commercial LLMs may not be an option due to confidentiality policies
- GPU-accelerated hardware may not be an option in some cases
- If experiments are successful using open-source LLMs deployed on consumer-grade hardware, then using LLMs in OS-coding can be a feasible option across a wide range of projects



# Setup

A7.

In what kind of building is your program located? Please choose one only for each building your program occupies.

1. Religious building
2. Public School
3. Private School
4. University or College
5. Work Place
6. Community Center or Municipal Building
7. Commercial Structure
8. Independent Structure (i.e., ORGANIZATION is the sole occupant)
9. Home, apartment, or other residential structure
10. Other, specify: \_\_\_\_\_

# Setup

A7.

In what kind of building is your program located? Please choose one only for each building your program occupies.

1. Religious building
2. Public School
3. Private School
4. University or College
5. Work Place
6. Community Center or Municipal Building
7. Commercial Structure
8. Independent Structure (i.e., ORGANIZATION is the sole occupant)
9. Home, apartment, or other residential structure
10. Other, specify: \_\_\_\_\_

**V01**

You are a helpful survey assistant. Your task is to categorize open-ended text into one of the following categories:

The categories are:

(1) Religious building

(2) Public School

[snip]

Tell me only the category number of the category that best fits the open-ended response. Then, provide your reasoning.

Please categorize the following text: {open text}

Answer: (

# Setup

A7.

In what kind of building is your program located? Please choose one only for each building your program occupies.

1. Religious building
2. Public School
3. Private School
4. University or College
5. Work Place
6. Community Center or Municipal Building
7. Commercial Structure
8. Independent Structure (i.e., ORGANIZATION is the sole occupant)
9. Home, apartment, or other residential structure
10. Other, specify: \_\_\_\_\_

V02

You are a helpful survey assistant. Your task is to categorize open-ended text into one of the following categories:

The categories are:

(1) **My program is located in a Religious building**

(2) **My program is located in a Public School**

[snip]

Tell me only the category number of the category that best fits the open-ended response. Then, provide your reasoning.

Please categorize the following text: **My program is located in the following type of building:** {open text}

Answer: (

# Setup

A7.

In what kind of building is your program located? Please choose one only for each building your program occupies.

1. Religious building
2. Public School
3. Private School
4. University or College
5. Work Place
6. Community Center or Municipal Building
7. Commercial Structure
8. Independent Structure (i.e., ORGANIZATION is the sole occupant)
9. Home, apartment, or other residential structure
10. Other, specify: \_\_\_\_\_

V03

You are a helpful survey assistant. Your task is to categorize open-ended text into one of the following categories:

Please categorize the following text: My program is located in the following type of building: {open text}

The categories are:

(1) My program is located in a Religious building

(2) My program is located in a Public School

[snip]

Tell me only the category number of the category that best fits the open-ended response. Then, provide your reasoning.

Answer: (

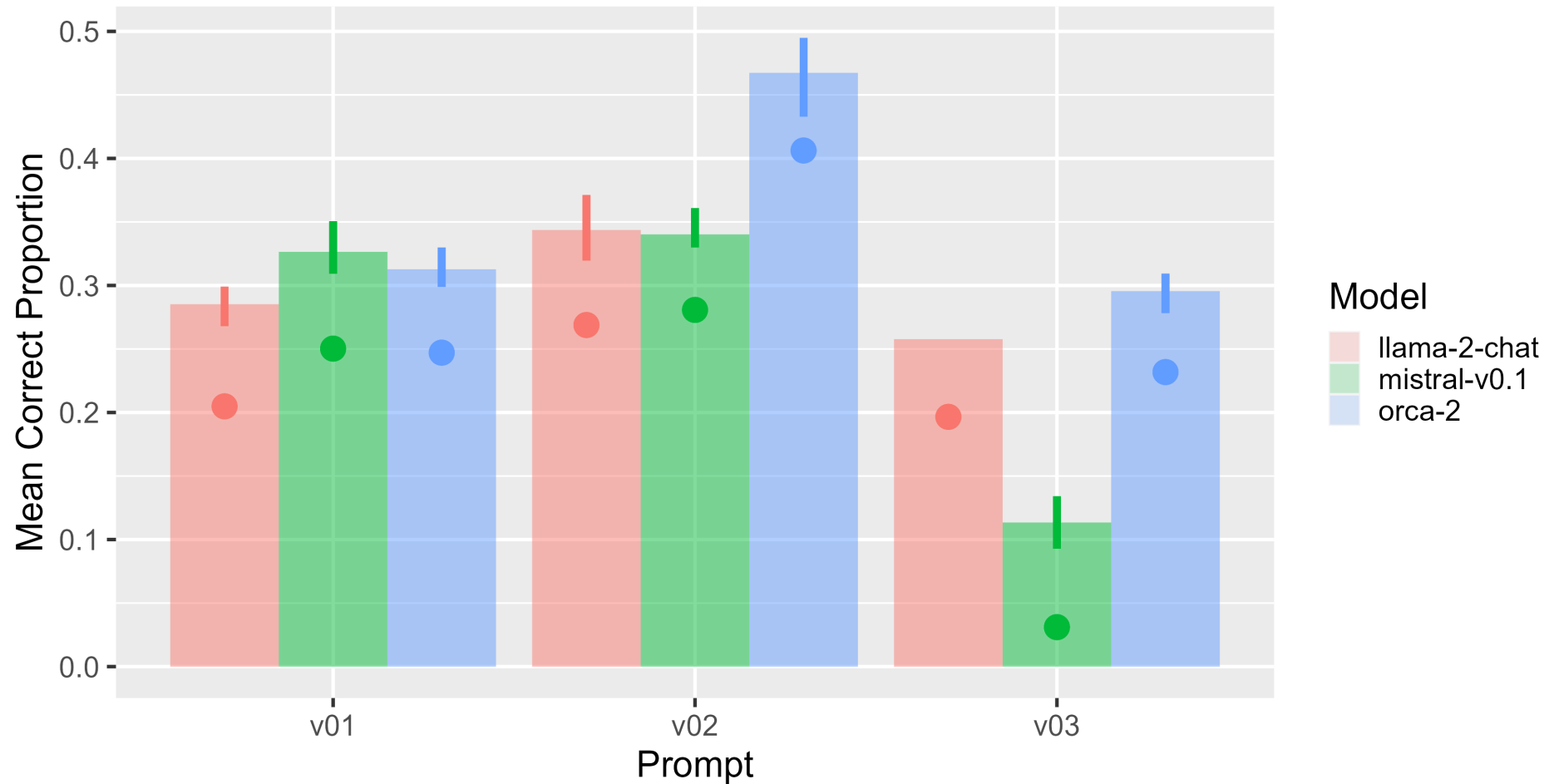
## Setup, continued

### **Design of experiments**

- 100 randomly selected verbatims (open-ended response)
- Compare LLM output against actual codes from 2019 NSECE
- 3 open-source models: Llama-2-13B, Orca-2-13B, Mistral v0.1
- Fully cross prompts and models = 9 prompt-model pairs
- 3 trials for each pair = 27 trials

# Findings

Proportion of LLM backcodes matching manual coding  
Lines show range across 3 trials for each model-prompt combination  
Dots show average Cohen's  $\kappa$  across trials



---

Why do respondents choose Other?

## Unsure about question wording

A7.

In what kind of building is your program located? Please choose one only for each building your program occupies.

1. Religious building
2. Public School
3. Private School
4. University or College
5. Work Place
6. Community Center or Municipal Building
7. Commercial Structure
8. Independent Structure (i.e., ORGANIZATION is the sole occupant)
9. Home, apartment, or other residential structure
10. Other, specify: Commercial space in condo building



?



## None of the above?

A7.

In what kind of building is your program located? Please choose one only for each building your program occupies.

1. Religious building
2. Public School
3. Private School
4. University or College
5. Work Place
6. Community Center or Municipal Building
7. Commercial Structure
8. Independent Structure (i.e., ORGANIZATION is the sole occupant)
9. Home, apartment, or other residential structure
10. Other, specify: Military base

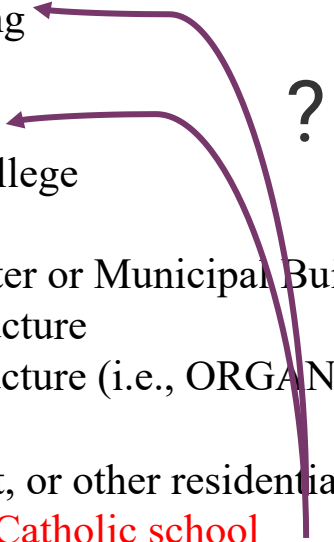
?



## More than one?

A7.

In what kind of building is your program located? Please choose one only for each building your program occupies.

1. Religious building
  2. Public School
  3. Private School
  4. University or College
  5. Work Place
  6. Community Center or Municipal Building
  7. Commercial Structure
  8. Independent Structure (i.e., ORGANIZATION is the sole occupant)
  9. Home, apartment, or other residential structure
  10. Other, specify: Catholic school
- 

# OS-coding is not just any classification task

## Potentially difficult task for LLMs

- **We are asking the LLM to do something that a (non-expert) human couldn't do!**
- LLM does not have the substantive and policy context of our study (what is the category that "best" fits the verbatim?)
  - May be technically correct but substantively wrong

---

# Making LLM-onade from LLM-ons

# What if LLMs are more useful when they are wrong?

## **Alternative approach**

- Find the most common human-LLM deviations (i.e. most common LLM “mistakes”) and use it to develop guidance for human coders

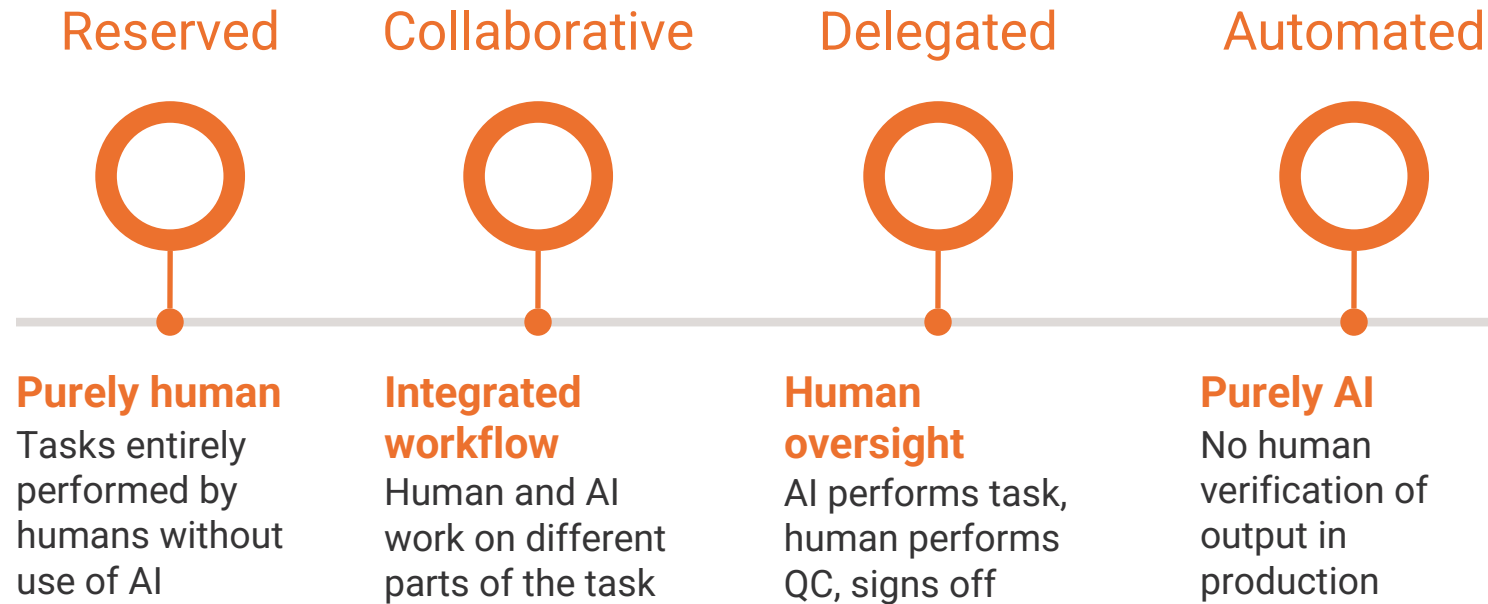
## **Main idea: LLM has “entry-level” tacit knowledge**

- Exploit this basic level of tacit knowledge to tease out what coders may not know (but we would like them to know)
- Rather than spend time to improve models, use models to improve quality of human coding

We asked an LLM to provide feedback on why “human” category was preferred

Open-end	LLM	Human	Feedback
Catholic school	Private school	Religious building	Category (1): "My program is located in a Religious building" is more appropriate than category (3): "My program is located in a Private School" because the former more accurately reflects the religious nature of the location, and it is more specific than the latter, which implies a private educational institution.
Leased facility from housing authority	Independent structure	Home, apartment, or other residential structure	The category (9) "My program is located in a Home, apartment, or other residential structure" is more appropriate because it specifically mentions "housing authority," which is a government-run program that provides housing assistance, and such programs are typically located in residential structures. Category (8) "My program is located in an Independent Structure where we are the sole occupant" is more general and does not capture the specific context of a housing authority.

# Experiment with different categories when considering whether/how to use AI



# Conclusion

## **Delegated/Automated tasks are growing categories as LLMs become more capable, but consider alternatives**

- OS-coding is a challenging task for LLMs, almost by construction
- This does not mean LLMs are not useful. Think instead about how LLMs can be used differently to complement human expertise
- In particular, how they can help to level-up less experienced or non-expert staff



# Thank you.



**Weihuang Wong**  
wong-weihuang@norc.org

---

 Research You Can Trust™

---

 **NORC** at the  
University of  
Chicago