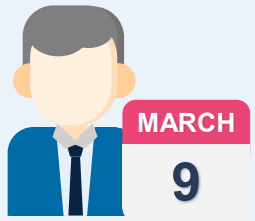
Coding classification with AI: Auto-Coding for the 2026 Population Census

9.3.2026

Michael Ming-yeung CHOI

Statistician

Census and Statistics Department, Hong Kong, China



Agenda



01 Background

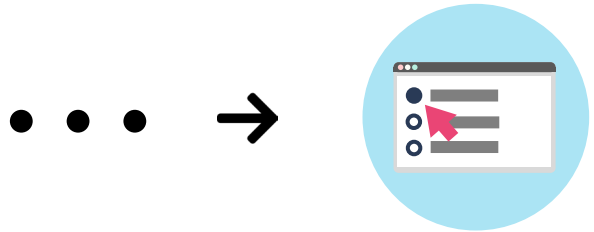
02 The AI Classification Models

03 Our Experience – The Application & Implementation

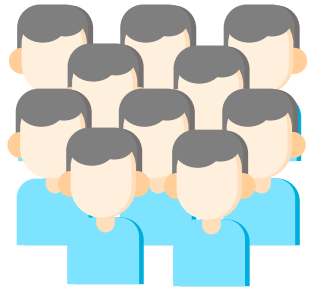
04 Conclusion and Q&A

Background

The coding and editing process in previous Censuses and General Household Survey (GHS)



Manual coding on Industry and Occupation





Free text description on open-ended occupation and industry questions

Industry	Occupation
560	1412
470	1211
550	5120

Our work – The Auto-coding

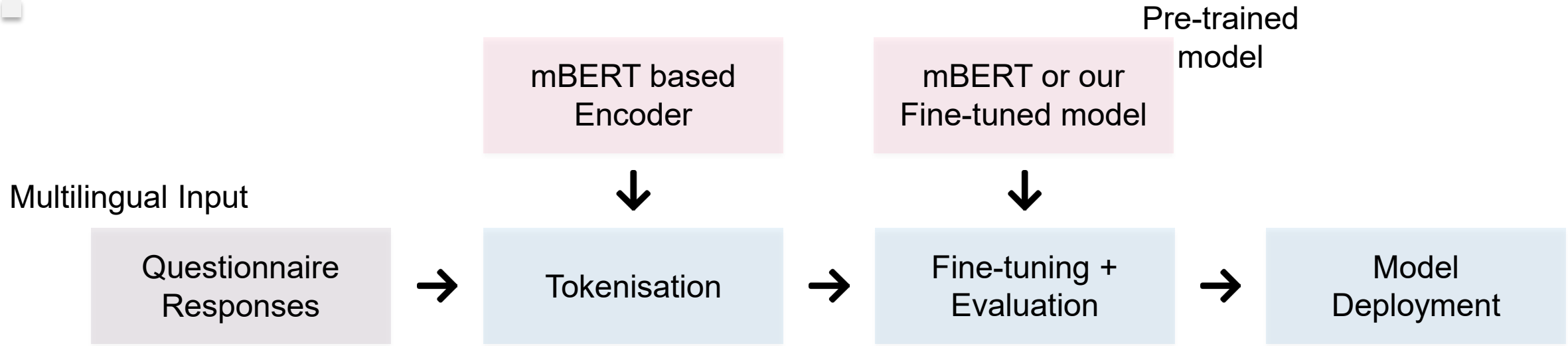
- The AI Classification Models for Occupation and Industry
- A risk-based mechanism with human-in-the-loop application
- Proof-of-concept and data analysis on later survey rounds and pilot runs
- The extendable auto-coding framework integrated with our data collection platform

Classification code	Number of classes
2-digit occupation code (adopted in GHS)	38
4-digit occupation code (adopted in Census)	281
3-digit industry code (adopted in Census and GHS)	113



The AI Classification Model – Training

- Pre-trained multilingual Large Language Model (LLM) for Natural Language Processing (NLP) with Transformer Architecture
- Fine-tune model parameters through transfer learning



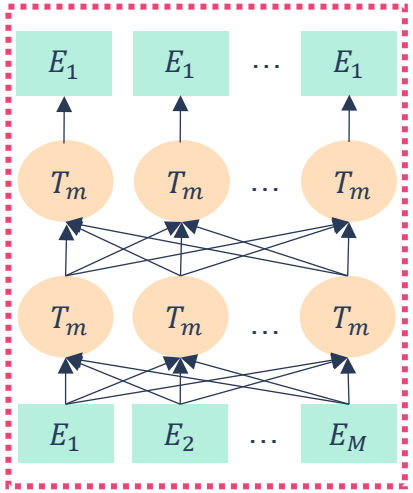
The AI Classification Model – Modelling

1. Pre-training mBERT

Data

Wikipedia data from over 100 languages

Transformer

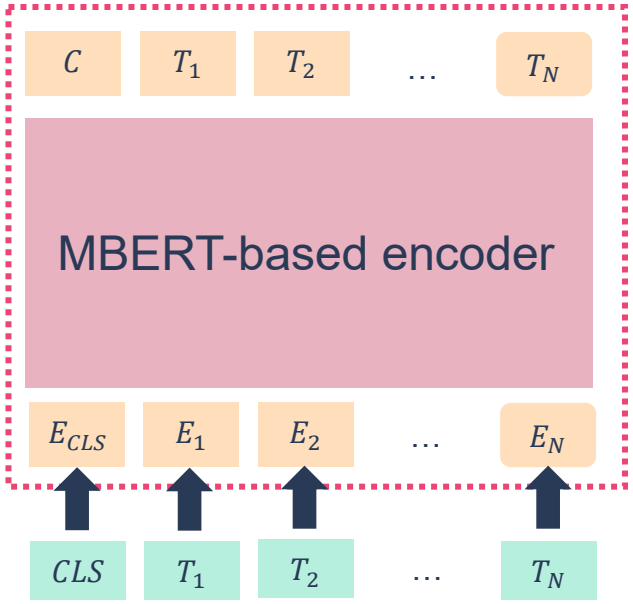


2. Fine-tuning for multi-label classification

Train data

2021 Census + GHS data

Fine-tuning



3. Prediction

Data

Testing data

Prediction

Softmax

- Class 1
- Class 2
- Class 3
- ⋮
- Class Y

Coding Classification Results (Models)

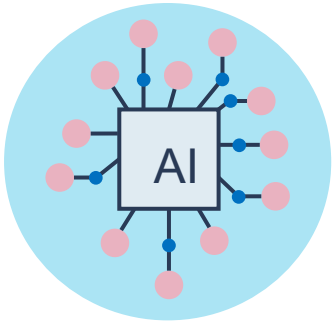
Data size	2021 Census	GHS	Overall
Training dataset	~142k	~186k	~328k
Validation dataset	~30k	~41k	~71k
Testing dataset	~31k	~42k	~73k
Total no. of records	~203k	~269k	~472k

Classification code	Number of classes	Accuracy	F1-score
2-digit occupation code (adopted in GHS)	38	85.8%	0.8793
4-digit occupation code (adopted in Census)	281	89.8%	0.8927
3-digit industry code (adopted in Census and GHS)	113	90.8%	0.9072

The AI Classification Model – Application



Multilingual Natural Language Input



Fine-tuned LLMs

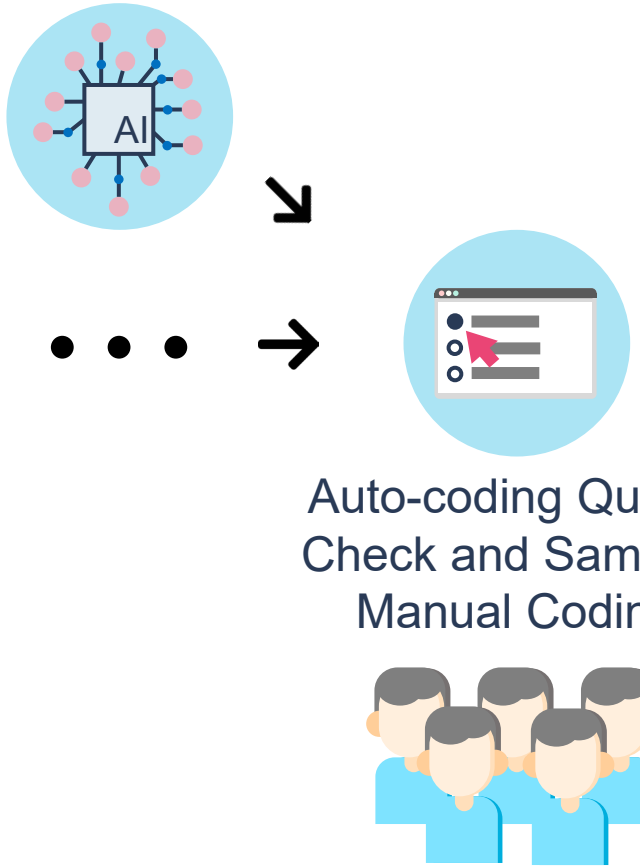
Predicted Occupation & Industry Code

+

Probability of the Predicted Results

From Proof-of-Concept to Application

Fine-tuned LLMs



To improve the accuracy, a risk-based mechanism with human-in-the-loop is applied in the auto-coding framework.

01 Predicted results with low probability are sampled for Manual Coding

02 The predicted results in classes with low historical accuracy are sampled for Manual Coding

03 Sampled some model trusted results for quality check of the models

04 Manual coding quality check

Risk-based Mechanism and Human-in-the-loop

The overall accuracy under the mechanism and human-in-the-loop

Classification code	Accuracy
Occupation Code	97.0%
Industry Code	95.1%

Also, the application reduced the manual coding efforts

To improve the accuracy, a risk-based mechanism with human-in-the-loop is applied in the auto-coding framework.

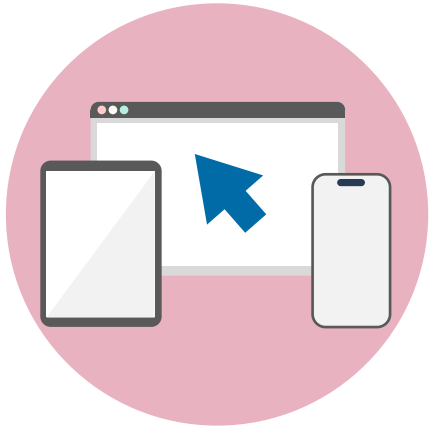
01 Predicted results with low probability are sampled for Manual Coding

02 The predicted results in classes with low historical accuracy are sampled for Manual Coding

03 Sampled some model trusted results for quality check of the models

04 Manual coding quality check

The Extendibility of Auto-coding Framework



Departmental Data Collection Platform

01 Support the fieldwork management, data collection and coding

02 Generic for different surveys

03 Configurable data fields

04 Support multimodal data collection methods

The Extendibility of the Auto-coding Framework in the system

01 Other surveys can use the same architecture of application with the API approach

02 Applicable for other data fields on fine-tuned models easily

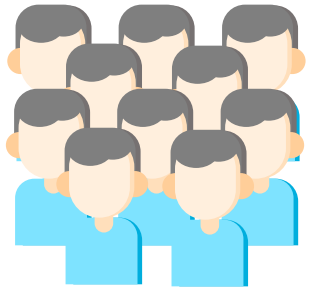
03 Adoptable on retrained models or modern architecture models in configuration

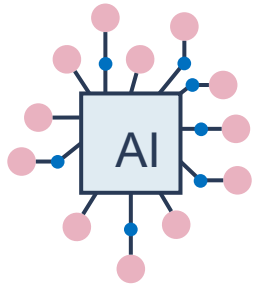
04 Configurable sampling rules and risk-based mechanism

Conclusion



Fully Manual Code



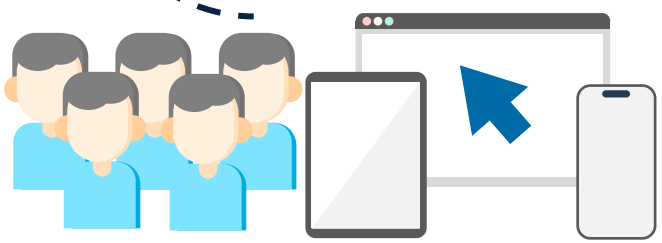


AI Classification Models Development with LLMs

Risk-based Mechanism with Human-in-the-loop



Integration with our Departmental Data Collection Platform



Auto-coding and Sampled Manual Coding

For a Smarter Future

Re-train the Models
Extend the Application
Adopt latest Architecture
Improve the Performance

Q&A

