The Use of AI to Aid Survey Text Classification

Colin Beavan-Seymour
Head of Statistical Design
Social Survey Transformation
UK Office for National Statistics

The Challenge

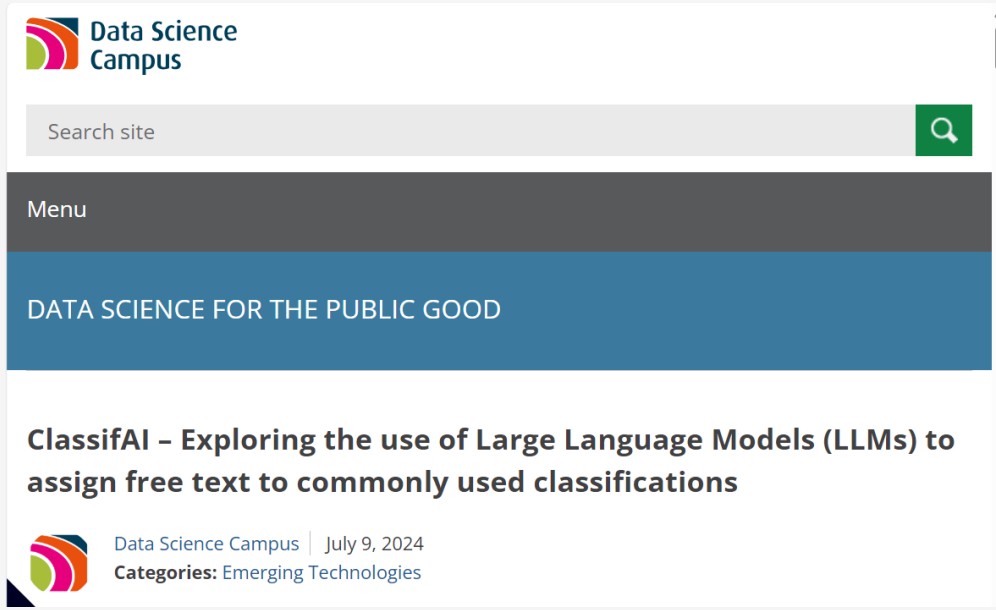
- Transition to an online Transformed Labour Force Survey (TLFS)
- TLFS Standard Industrial Classification (SIC) and Standard Occupational Classification (SOC) data quality is not at level required for linked statistical outputs.
- Online = no interviewer interaction to gather more detail on responses
- We are developing a novel generative AI solution, **Survey Assist**, to improve data quality at the point of collection.
- **Survey Assist** aims to replicate an interviewer's interaction with the respondent.

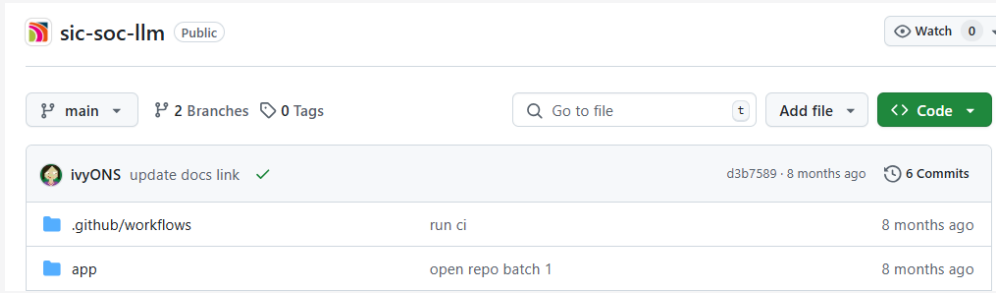
Survey Assist

Generative AI survey collection tool

ClassifAI SIC LLM: An enabler for Survey Assist

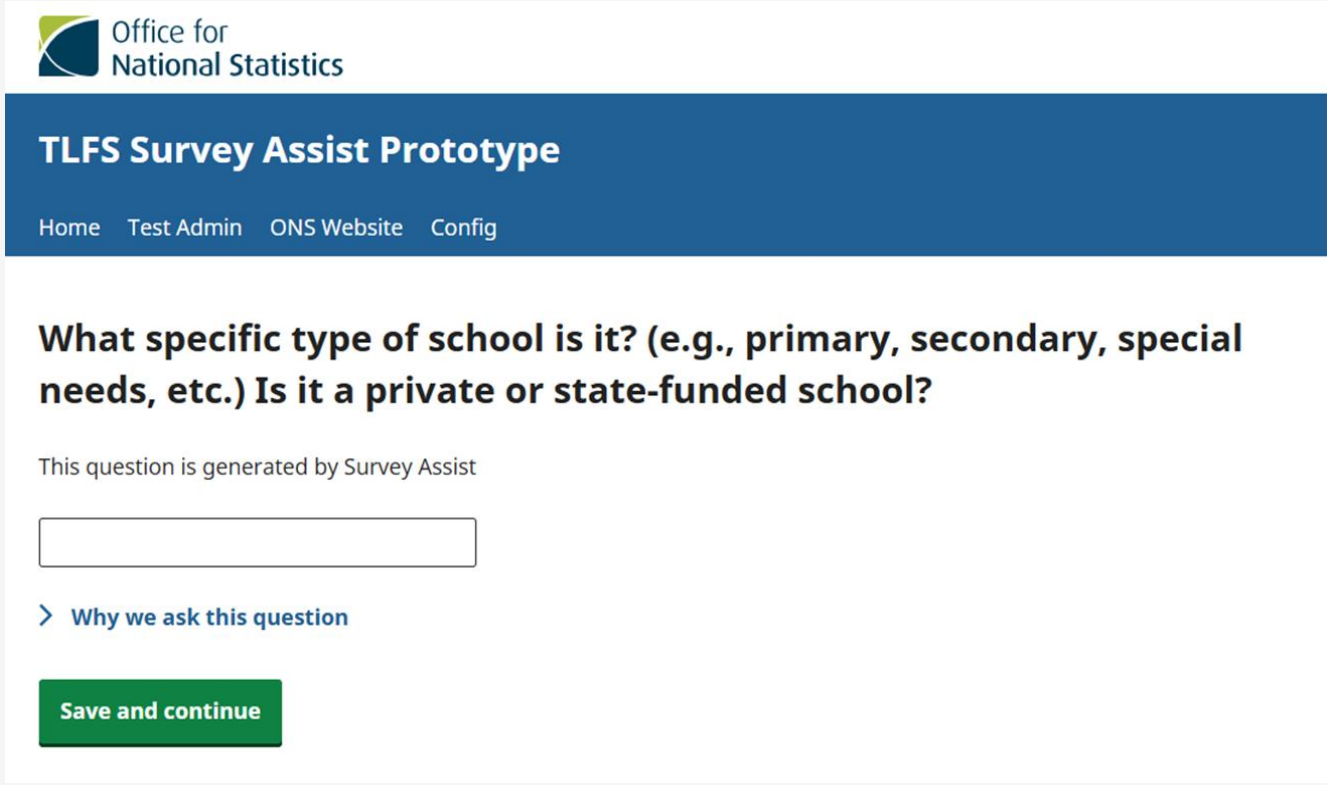
- Assigning free text to categories is a common activity for National Statistical Institutes (NSIs), where a combination of manual, rules-based, and machine learning techniques are currently employed.
- Data Science Campus explored a large language model (LLM) Retrieval Augmented Generation (RAG) approach to classify anonymised free text from labour market surveys to SIC industry codes in early 2024.
- Initial results were promising, demonstrating a 5 and 11 percentage point improvement on existing approaches when classifying data to the 2-digit and 5-digit SIC levels. See blog and github repo for more details.
- Developed as **SIC ClassifAI** after the initial research.





Survey Assist: adaptive questioning with genAI

- Prototype aims to improve survey data quality through targeted follow-up questions.
- The respondent's free text answers about their job and employer are routed to our classification service.
- First stage look-up checks for an exact match.
- If the response cannot be coded automatically, follow-up questions (open and closed) are generated by AI.
- Responses are coded to 5-digit SIC



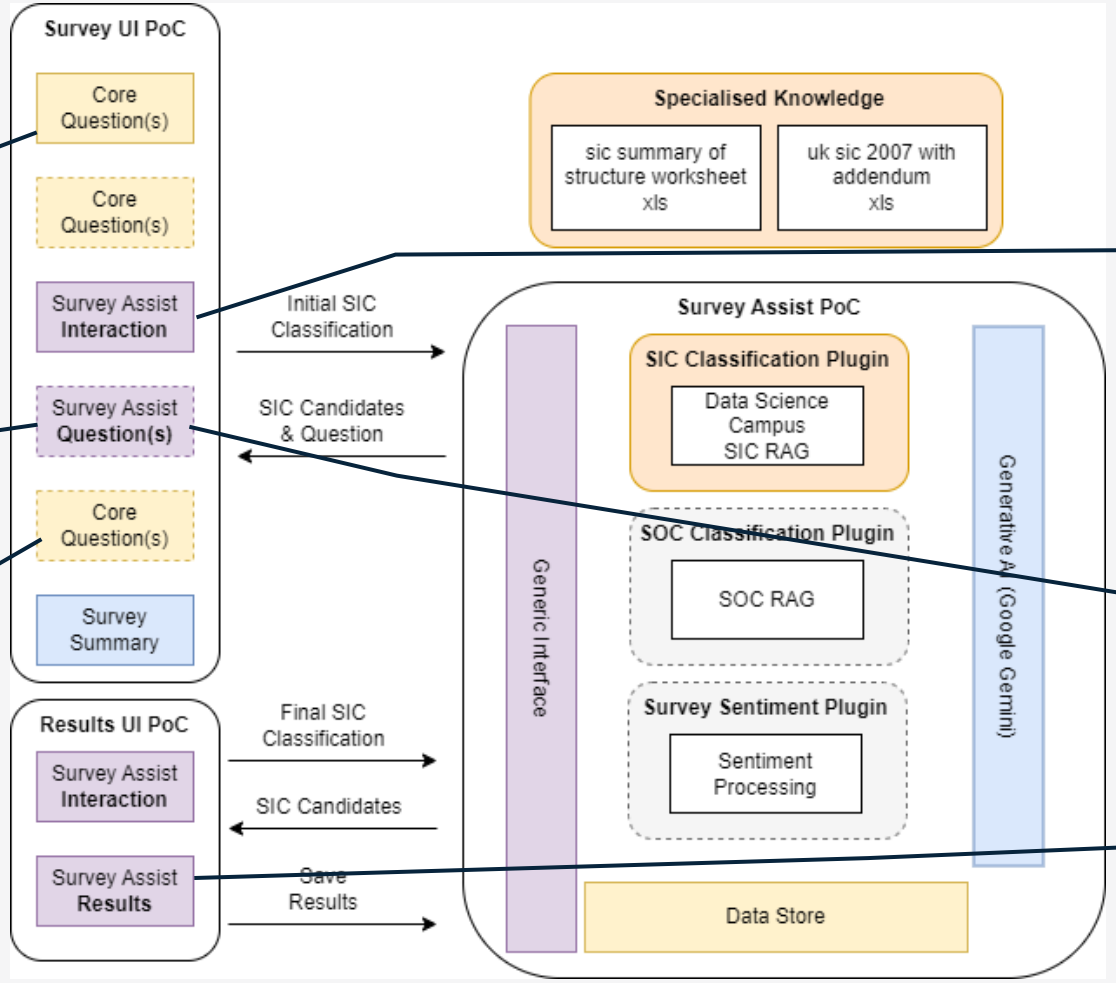
Survey Assist (prototype version)

Respondent

STEP 1
User completes the first three TLFS questions (job, job activity, industry)

STEP 4
User responds to customised AI generated question(s)

STEP 5
User completes the final TLFS industry question



Survey Assist

STEP 2
2a. If "industry" answer matches a known SIC, Survey Assist does not ask custom questions and routes the user to STEP 5.
OR
2b. Survey Assist makes an initial SIC classification.

STEP 3
Survey Assist uses a Large Language Model (LLM) to create customised questions.
The prototype allows 1 open question + 1 closed question (derived from possible SIC matches)

STEP 6
Survey Assist makes a final SIC classification based on initial answers and answers to AI generated questions

Key:
UI = user interface
PoC = proof of concept
RAG = Retrieval Augmented Generation

Design Considerations

AI-generated questions	User Acceptance
Questions change for each respondent and are not standardised or validated Quality review and design iteration Respondent-centred frame to support meaningful and accessible questions	Respondent attitudes to AI Transparency about AI use Consent Design preferences
Respondent Experience	**Evaluation and assurance**
Focus on survey length and drop-off (missingness) How and when to deploy Survey Assist Feedback on AI questions	Evaluation against “truth” dataset Assurance against existing solutions Periodic evaluation for LLM ‘model drift’ Periodic scrutiny of AI questions

Survey Assist User Research (PoC version)

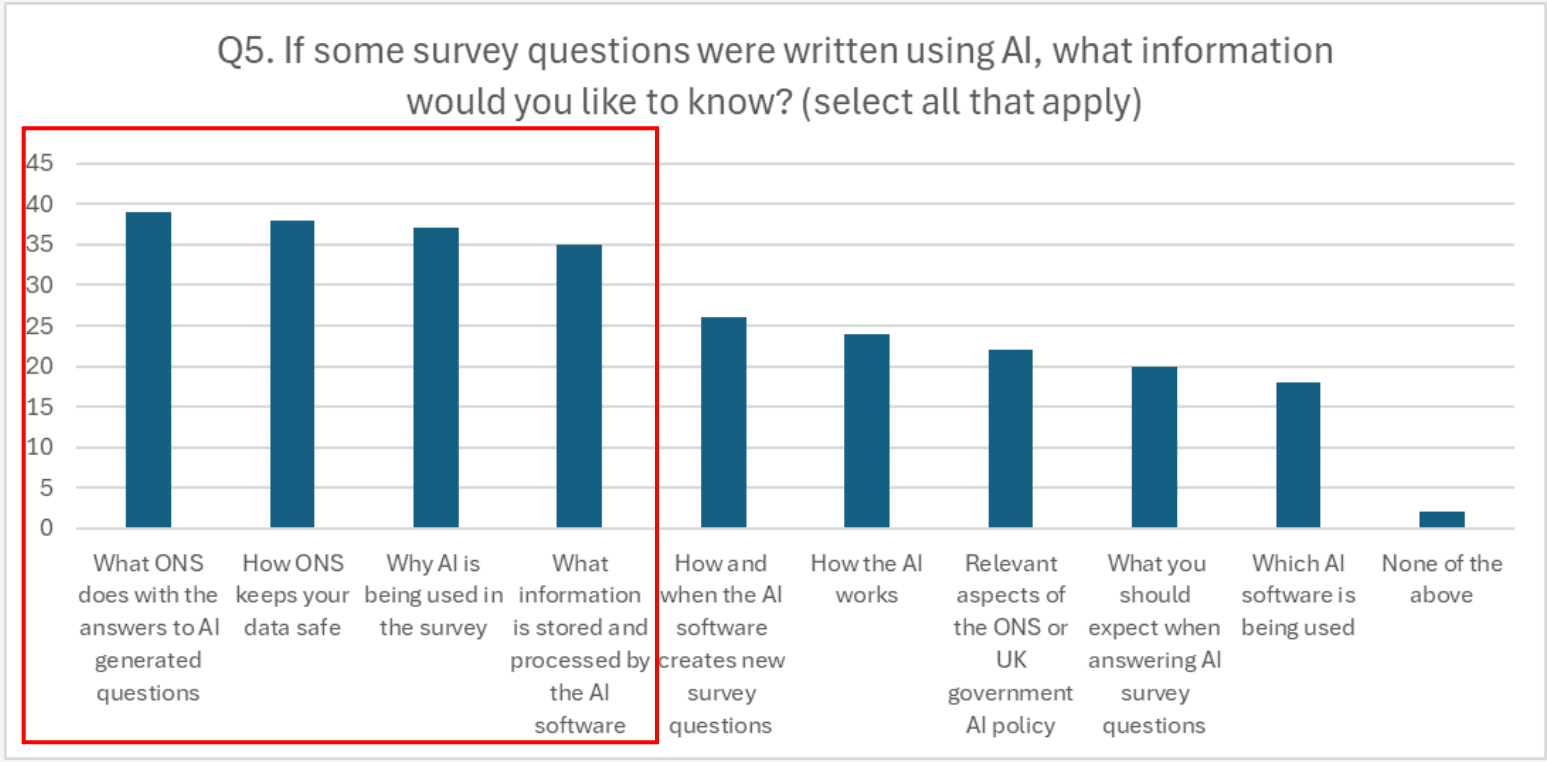
Key Findings

84% felt positive or neutral about answering survey with some **AI generated questions**

76% agree or strongly agree *'I **trust the ONS** to use AI in its surveys appropriately'*

70% agree or strongly agree that *'it's **important to know if a survey question has been written by AI**'*

37% concerned about **privacy and data security** when AI is used in a survey



Resources & Contacts

ClassifAI further reading:

Blog: https://datasciencecampus.ons.gov.uk/classifai-exploring-the-use-of-large-language-models-llms-to-assign-free-text-to-commonly-used-classifications/

Github repo: https://github.com/datasciencecampus/sic-soc-llm

Contacts:

SIC/SOC Project: Labour.Market.Survey.Project@ons.gov.uk

Survey Assist & ClassifAI LLM: digital.innovation@ons.gov.uk

Special thanks:

Lizanne Duckworth & Debbie Curtis at ONS