Examination of Estimates of Expenditure 2026-27 Reply Serial No.

FSTB(FS)142

CONTROLLING OFFICER'S REPLY

(Question Serial No. 0059)

Head: (26) Census and Statistics Department

Subhead (No. & title): (-) Not Specified

Programme: (4) General Statistical Services

Controlling Officer: Commissioner for Census and Statistics (Leo YU)

Director of Bureau: Secretary for Financial Services and the Treasury

Question:

In paragraph 48 of the Budget Speech, it is mentioned: "The C&SD will launch a new online platform for interactive data dissemination service next month (i.e. in March). Through integrating different types of statistical data, this platform enables enterprises and the public to conduct cross-subject analysis. A natural language data query feature will be introduced into the platform in the third quarter." Please inform the Council:

(1) Are these data limited only to the original data of the Census and Statistics Department? Or will it be connected with the existing "Common Spatial Data Infrastructure" or databases of other bureaux?

(2) In advancing the "Northern Metropolis" and various large-scale urban planning projects, enterprises and the academic sector have a strong need to combine demographic statistics, economic activities, and geospatial data for analysis. When building this platform, has the Government formulated relevant open data standards and an integration timetable specifically targeting such macro urban planning needs?

(3) Regarding the "natural language data query feature" to be introduced in the third quarter, given that large language models generally suffer from the "AI hallucination" problem at present, what specific technical mechanisms or review processes will the authorities adopt to ensure that the official statistical data provided by the system is absolutely accurate and authoritative, so as to avoid misleading the public or business decision-making?

(4) Is the foundation AI model for this feature trained internally by the Government, or is it procured from a third-party cloud service provider? When handling specific business queries from the public and enterprises, how will sensitive commercial query intentions and personal data be prevented from leaking, in order to safeguard local information security?

Asked by: Hon LAM Siu-lo, Andrew (LegCo internal reference no.: 27)

Reply:

(1) and (2)

The online platform for interactive data dissemination service covers the aggregate socio-economic statistics of Hong Kong released by the Census and Statistics Department (C&SD). These data are currently disseminated in the form of statistical tables and reports of different themes on C&SD's Website. The new platform provides an additional channel that further integrates cross-theme statistical data to enable users to conduct analyses more flexibly and efficiently. The platform is not connected to the systems of other bureaux or departments.

Besides, C&SD has been sharing and updating statistics having geospatial features through the Common Spatial Data Infrastructure (CSDI) portal (portal.csdi.gov.hk) of the Spatial Data Office under the Development Bureau. The CSDI portal uses locational information as the common infrastructure, enabling government departments and organisations to spatialise and standardise their data, and then freely share it through this one-stop portal with the community, government departments, and both public and private organisations. It also provides application programming interface (API) services to support various spatial data analytics, policy planning and smart city applications.

(3) The "natural language data query feature" uses a large language model (LLM) to analyse unstructured textual question input by a user, and creates the required statistical tables using the platform's functions. Since the query feature can only find answers from the aggregate statistics in the platform's database and display the results in the form of the statistical tables set by the platform, there is no risk of fabricating statistics in the process.

(4) The core workflow of the "natural language data query feature" is designed by staff of C&SD, while the underlying LLM is provided by a third-party cloud service. During the LLM operation, only the aggregate statistics in the platform's database can be accessed, and the platform does not store any raw data of individual persons or organisations, or any other confidential information. Therefore, the entire operation poses no risk of leakage of confidential information. The design of the platform complies with government information security requirements, and the platform has already passed an independent third-party security risk assessment and audit. In addition, before launching the new feature, C&SD will engage an independent third party to conduct the security risk assessment and audit, and privacy impact assessment, to ensure that the feature complies with government information security requirements.

- End -