

AI needs cultural policies, not just regulation

The future of Artificial Intelligence (AI) will not be secured by regulation alone. To ensure safe and trustworthy AI for all, we must balance regulation with policies which promote high-quality data as a public good. This approach is crucial for fostering transparency, creating a level playing field, and building public trust. Only by giving fair and wide access to data can we realise AI's full potential and distribute its benefits equitably.

Data are the lifeblood of AI. In this regard, the laws of neural scaling are simple: the more, the better. The more volume and diversity of human-generated text is available for unsupervised learning, for example, the better the performance of Large Language Models (LLMs) will be. Alongside computing power and algorithmic innovations, data arguably are the most important driver of progress in the field.

A data race at the expense of ethics

But there is a problem. Humans do not produce enough digital content to feed these ever-growing beasts. Current training datasets are already huge: Meta's LLaMa 3, for example, is trained on 15 trillion tokens, equivalent to over 10 times the British Library's book collection. According to a recent study, the demand for pristine text is such that we might reach something akin to 'peak data' before 2030. Other papers caution against the dangers of public data contamination by LLMs themselves, causing feedback loops that amplify biases and deplete diversity.

Fears of an 'AI winter' reflect the relentless race for data in which researchers and industry players are engaged, sometimes at the expense of quality and ethics. A prime example is 'Books3', a trove of pirated books widely believed to feed leading LLMs. Whether such practice falls under fair-use policy is a debate for lawyers. What is more disturbing is that these books are being hoarded without any clear guiding principle.



Clément Godbarge

Lecturer in Digital Humanities, School of Modern Languages, The University of St Andrews, U.K.

Only by giving fair and wide access to data can AI's full potential be realised and its benefits distributed equitably

Even if progress is being made, notably thanks to regulation, LLMs are still largely trained on an inscrutable morass of licensed content, 'publicly available data', and 'social media interactions'. However, studies show that these data reflect, and sometimes even exacerbate, the current distortions of our cyberspace: an overwhelmingly anglophone and presentist world.

The absence of primary sources

The notion that LLMs are trained on a universal compendium of human knowledge is a fanciful delusion. Current LLMs are far from the universal library envisioned by the likes of Leibniz and Borges. While stashes of stolen scriptures like 'Books3' may include some scholarly works, these are largely secondary sources written in English: commentaries that merely skim the surface of human culture. Conspicuously absent are the primary sources and their myriad tongues: the archival documents, oral traditions, forgotten tomes in public depositories, inscriptions etched in stone – the very raw materials of our cultural heritage.

These documents represent an untapped reservoir of linguistic data. Consider Italy. The State Archives of this nation alone harbour no less than 1,500 kilometres of shelved documents (in terms of linear measurement) – excluding the vast holdings of the Vatican. Estimating the total volume of tokens that could be derived from this heritage is difficult. However, if we include the hundreds of archives spreading across our five continents, it is reasonable to believe that they could reach, if not surpass, the magnitude of data currently used to train LLMs.

If harnessed, these data would not only enrich AI's understanding of humanity's cultural wealth but also make it more accessible to the world. They could revolutionise our understanding of history, while safeguarding the world's cultural heritage from negligence, war, and climate

change. They also promise significant economic benefits. As well as helping neural networks scale up, their release into the public domain would mean that smaller companies, startups, and the open-source AI community could use those large pools of free and transparent data to develop their own applications, levelling the playing field against Big Tech while fostering innovation on a global scale.

Examples from Italy and Canada

Advances in the digital humanities, notably thanks to AI, have drastically reduced the cost of digitisation, enabling us to extract text from printed and manuscript documents with unprecedented accuracy and speed. Italy recognised this potential, earmarking €500 million of its 'Next Generation EU' package for the 'Digital Library' project. Unfortunately, this ambitious initiative, aimed at making Italy's rich heritage accessible as open data, has since been deprioritised and restructured. Short-sightedness prevailed.

Canada's Official Languages Act offers an instructive lesson in this regard. Long derided as wasteful, this policy requiring bilingual institutions eventually yielded one of the most valuable datasets for training translation software.

However, recent debates about adopting regional languages in the Spanish Cortes and European Union institutions have overlooked this key point. Even advocates have failed to recognise the cultural, economic, and technological benefits of promoting the digitisation of low-resource languages as complementary.

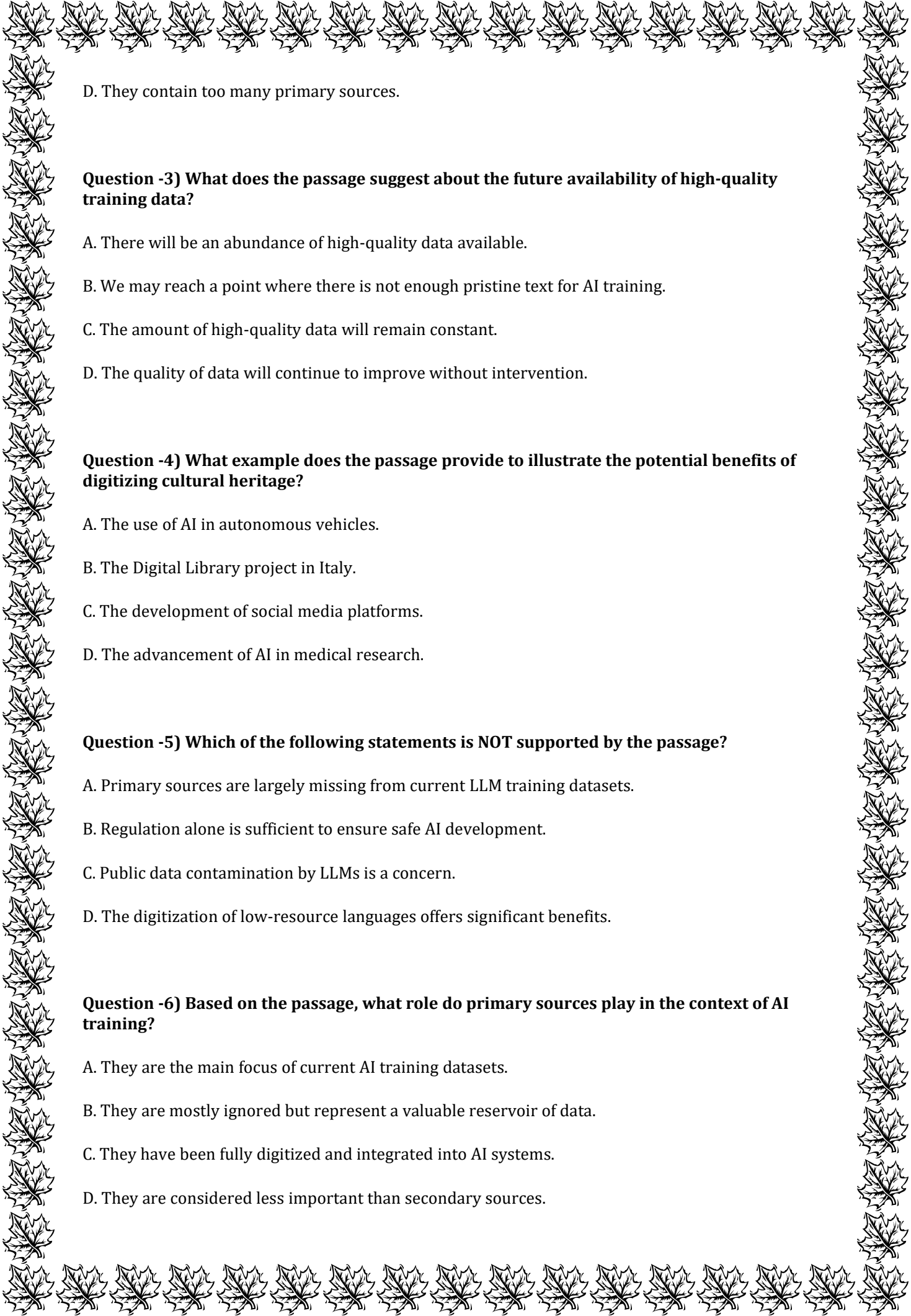
As we accelerate the digital transition, we must not overlook the immense potential of our world's cultural heritage. Its digitisation is key to preserving history, democratising knowledge, and unleashing truly inclusive AI innovation.

Question -1) Which of the following best captures the main argument of the passage?

- A. AI progress should be regulated to ensure ethical use of data.
- B. AI development relies solely on algorithmic innovations and computing power.
- C. The digitization of cultural heritage can enhance AI and promote equitable access to data.
- D. Current LLMs are trained on a comprehensive database of global knowledge.

Question -2) According to the passage, what is the primary issue with current AI training datasets?

- A. They are too small to be effective.
- B. They rely too much on human intervention.
- C. They lack diversity and can amplify biases.



D. They contain too many primary sources.

Question -3) What does the passage suggest about the future availability of high-quality training data?

- A. There will be an abundance of high-quality data available.
- B. We may reach a point where there is not enough pristine text for AI training.
- C. The amount of high-quality data will remain constant.
- D. The quality of data will continue to improve without intervention.

Question -4) What example does the passage provide to illustrate the potential benefits of digitizing cultural heritage?

- A. The use of AI in autonomous vehicles.
- B. The Digital Library project in Italy.
- C. The development of social media platforms.
- D. The advancement of AI in medical research.

Question -5) Which of the following statements is NOT supported by the passage?

- A. Primary sources are largely missing from current LLM training datasets.
- B. Regulation alone is sufficient to ensure safe AI development.
- C. Public data contamination by LLMs is a concern.
- D. The digitization of low-resource languages offers significant benefits.

Question -6) Based on the passage, what role do primary sources play in the context of AI training?

- A. They are the main focus of current AI training datasets.
- B. They are mostly ignored but represent a valuable reservoir of data.
- C. They have been fully digitized and integrated into AI systems.
- D. They are considered less important than secondary sources.