The Schumacher Institute

How can LLMs transform the delivery of the functions of government?

By Jack Bobin and Barnaby Frith

Introduction

The technology of the 21st century is evolving rapidly, and although the private sector innovates and adapts quickly to maximise their productivity by utilising the latest technology, government systems often lag. Inefficiencies and limited functions result. In this report, we will examine how the AI revolution and the development of Large Language Models (LLMs) can improve the functions of government. We will look at the ways in which the implementation of LLM technology can improve the delivery of government functions, and the effects that will have on our systems in a broader context.

At a time when budgets are stretched, our public service infrastructure has its capacity stretched, and vulnerable citizens are frequently let down, we believe that the strengths of Large Language Model technology plays to the weaknesses of our present systems. Technology like this must be a central part of our strategy to deal with the problems threatening our systems.

Natural Language Processing

To understand what problems LLMs are well placed to solve, we must first understand what their unique strengths are and have a basic understanding of how they work. Natural Language Processing (NLP) is the domain of Artificial Intelligence focused on the interaction between humans and computers through written language. This technology encompasses LLMs.

The models ‘understand’ language via a process known as vector mapping. Words and phrases are ‘tokenised’ to convert the meaning carried into a vector, a stack of numbers. These are complex arrangements to convey the meaning, context and connotations carried by a certain word. (For example, the word ‘it’ would be mapped differently in different contexts, depending on what it refers to - part of the encoding would be different to convey the connection between

that word and the noun it stands in for in the context). The complex, multi-dimensional vectors enable computers to understand vast complex connections between words in a text to be able to convey the information carried. Text generation works by predicting the most likely words and phrases to come in a text that could be described by the prompt given.

This is important ground knowledge for understanding how we can deploy LLMs because we must understand the strengths and limitations of chatbots and software built on NLP. It is adapted for pattern recognition, not creative manoeuvring of uncharted territory, so we must be careful in where the technology is used and oversight must be maintained.

However, the ability to map relationships between ideas in a text or across thousands of texts, can be almost infinitely more powerful than the capabilities of a single human mind, making LLMs extremely useful for tasks involving spotting connections or patterns from language.

This ability to measure relationships between ideas also makes Natural Language Processing excellent at picking out key information and avoiding redundancies in very large volumes of text, a laborious task for humans. Therefore, by understanding the capabilities of the technology, we can identify the areas to target to deal with weaknesses in the current systems and human limitations.

Implementation

The simplest applications of LLM technology are to enable interaction between government systems and citizens with reduced reliance on employed specialists. This produces direct fiscal savings, but more importantly it can enable services to be delivered at higher quality, greater consistency, and significantly improved convenience for citizens. In the UK public sector, where paying the direct salaries of government employees accounts for over 20%*(Institute for Fiscal studies) of operating expenditure, even modest substitution or augmentation effects can generate large efficiency gains.

The key applications that we have identified as high-impact areas for Large Language Model application are:

- Health
- Compliance
- Citizens advice
- Interviews and eligibility assessments
- Consultancy
- Policy making

Health

The shortage of GPs and the maximum capacity of centres make it very difficult to get appointments for all the patients needing one, creating long waiting lists. As of 2024–2025, England has roughly 0.46 fully qualified GPs per 1,000 patients. In 2015, during austerity, this was 0.52 BMA). Patients wait on average over 2 weeks for a routine appointment, with significant regional variation. Limitations create inequality and reduced quality of care for all.

NHS 111, as an alternative, is often ineffective: it handles over 20 million calls per year, yet frequently defaults to risk-averse escalation (frequently directing patients to A&E), increasing system strain. It also has high operational costs due to staffing and triage inefficiencies.

It takes a long time and is expensive to train GPs (typically 10–15 years of education and training, costing the state well over £200,000 per doctor). GPs must be experts both in dealing with people and in a wide breadth of medical subject matter. Doctor's visits also create inconvenience and, for many patients, anxiety or avoidance behaviour, particularly among vulnerable groups. Clearly, the solution cannot simply be to increase the number of GPs - even if we could afford to do so, increasing budgets without delivering more strategic solutions will not solve our problems. However, Large Language Models could be adapted to help in several ways.

AI triage / diagnostics

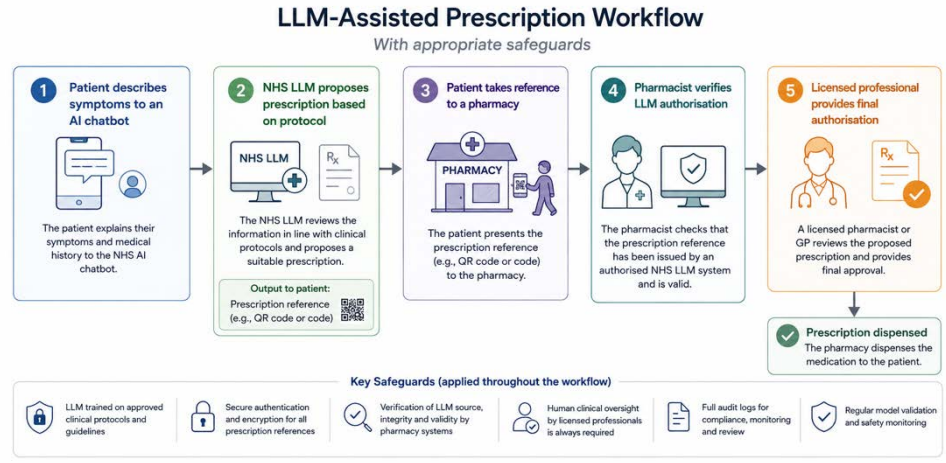
I. At home

LLMs trained on medical textbooks, clinical guidelines (such as NICE), and verified NHS content can act as a first port of call. This replaces both unnecessary GP appointments and unreliable self-diagnosis via unverified internet sources.

Approximately 20–30% of GP consultations involve self-limiting conditions (minor infections, mild dermatological issues etc). For a large proportion of these cases, an LLM system could recommend over-the-counter treatments and suggest monitoring thresholds for when to escalate. This can reassure and care for patients without taking time out of a busy doctor's day. It would also reduce pressure on the NHS system, especially from the large number of low-level cases that take time away from vulnerable or complex patients who require human medical attention.

Technology picking up the burden of work for low level cases would reduce demand on primary care without reducing quality. Early trials of AI triage systems in 2024 have shown diagnostic accuracy comparable to or above the level of junior doctors in structured scenarios, which improves when constrained to guideline-based outputs. In 2026, a Harvard study found OpenAI's o1 model had significant higher success rate at triaging real ER cases compared to the doctors on duty at the time.

With appropriate safeguards, such systems could also support prescription workflows. For example:



This model, keeping humans in the loop, reduces risk while still achieving large efficiency gains, in particular saving time for GPs and patients (as no GP appointment is required).

II. Adapted GP appointments

Specialist medical LLMs, designed as decision-support systems rather than replacements, can reduce the breadth of expertise required from each practitioner.

A new kind of appointment has been envisioned by Richard and Daniel Susskind in their book 'The Future of the Professions' where NLP models take on some expertise requirements for the human operator. This looks like nurses or 'physician associates' inputting structured patient data, the LLM suggesting differential diagnoses, tests, and next steps. This leaves the human clinician to focus on patient interaction, judgement, and communication

This specialisation in the human side of appointments may also improve patient experience. For vulnerable patients (children, elderly patients, or those for whom the social aspect of appointments causes difficulty), reducing cognitive burden on clinicians allows more time for communication and reassurance, which is often the binding constraint in care quality.

Treatment plans

The volume of medical literature is vast and growing exponentially. PubMed indexes over 1.3 million new biomedical articles per year, making it impossible for any individual doctor to remain fully up to date even within a narrow specialty.

This creates a structural inefficiency: treatments improve in theory faster than they diffuse into practice. This is where NLP technology can unlock much more of the potential gains brought by accumulating research and expertise.

Providing summaries of long texts is a well-known strength of this technology. LLMs are specialised to identify relationships between ideas in text. They can summarise individual

papers or entire bodies of literature and explain the differences between competing studies. This reduces reading time from hours to minutes while preserving key findings.

Simply identifying relevant literature also optimises the efficiency of human doctors. Rather than searching manually, clinicians can ask questions to a system that retrieves all relevant studies for a specific condition or patient profile. This reduces search costs and improves coverage. By embedding historical patient data (appropriately anonymised), LLM systems can compare new cases to prior cases and outcomes. This allows the identification of treatment patterns; probabilistic outcome forecasting; and more personalised solutions.

Even a 5–10% improvement in treatment optimisation across major conditions would translate into large gains in QALYs (quality-adjusted life years) and cost savings at the system level.

Citizens advice

Another area where vulnerable members of the public can benefit from interactions with government trained LLMs is by providing advice to resolve legal, financial and social problems, in an anonymous and frictionless way. This important social work has previously been undertaken largely by charities, but this creates regional inequality and insecurity over resourcing.

Much of the work of the Citizens Advice Bureau could be performed by LLM systems, reducing reliance on volunteers and public funding. The organisation handles over 2.5 million clients annually in England and Wales, with common issues including debt, housing, and benefits.

Once trained, a ‘citizens advice’ LLM could provide all these services with 24 hour availability, at no additional marginal cost per user, with greater consistency and reduced chance of human error.

The anonymity of interacting with a chatbot may also increase usage among individuals who are uncomfortable seeking help in person, particularly in cases involving debt, domestic issues, or legal uncertainty.

While government web pages theoretically make answers accessible, they are often fragmented and difficult to navigate. Digital exclusion remains significant: around 10% of UK adults lack basic digital skills. An LLM interface consolidates information into a single, conversational access point. Furthermore, the interactivity of an AI chatbot makes it more familiar and personalised to assist people in need.

In addition, LLMs can power benefits calculators, eligibility checks and step-by-step guidance through application processes to reduce both user error and administrative burden.

Interviews and Eligibility Assessment

The use of human interviewers in determining eligibility for benefits, housing programmes, or asylum decisions introduces significant inefficiencies and inconsistencies. Systems that rely on individual assessors inevitably inherit variation in judgement and interpretation, even where formal guidelines exist. This is visible in outcomes: in the UK, around 60–70% of Personal Independence Payment (PIP) decisions that reach tribunal are overturned, indicating substantial inconsistency at the initial assessment stage. Similar variation has been observed in asylum decisions, where outcomes can differ across caseworkers and regions despite broadly similar cases.

Attempts to standardise these processes through training, guidance, and oversight add further cost and complexity. Assessors must be trained not only in policy, but also in interviewing and documentation, while appeals systems are required to correct errors. In effect, the system compensates for human variability by adding additional layers of administration. Moreover, when policy directives conflict with the personal values or intuitions of frontline staff, implementation becomes less predictable, as discretion and bias—conscious or not—shape outcomes.

LLM-based systems offer a way to reduce these inefficiencies by introducing consistency at the point of assessment. A system could conduct structured interviews, ensuring all relevant questions are asked and responses are interpreted against policy in a uniform way. Unlike human interviewers, such systems do not vary due to fatigue or individual bias, and early

evidence from administrative automation suggests potential reductions in processing time of 20–40%.

In practice, this would likely take a hybrid form: the LLM conducts the initial interview and produces a structured case assessment, while human caseworkers review complex or ambiguous cases. This concentrates human judgement where it is most valuable, improves consistency in routine decisions, and reduces the volume of appeals. Given the scale of administrative backlogs and tribunal costs, even modest improvements in first-instance decision accuracy could generate significant efficiency gains while improving fairness and accessibility.

Compliance

Tax regulation in the UK is extremely complex. HMRC's primary tax manuals (the 'Yellow and Orange books') together are over 23,000 pages, without including case law and supplementary guidance.

This complexity creates economic restrictions and sources of inequality because of high costs to make businesses compliant, (estimated at £15–20 billion annually across UK businesses). Complexity has created barriers to entry for small firms and sole traders and given disproportionate advantage for large firms with in-house legal and accounting teams.

It also requires significant time expenditure simply to understand obligations, rather than to create economic value. Complexity further introduces uncertainty and loopholes, making government revenue less predictable.

LLMs simplify

LLMs could be used to simplify interaction with legislation, even where the underlying body of law remains highly complex.

Vector mapping of lengthy documents enables LLM technology to identify relationships between different sections of text. This understanding of extremely complex connections between different rules, around tax for example, make a government-trained LLM well placed

to identify loopholes or complexities caused by rules made redundant by a constantly evolving rule book.

Even before any legislative reform takes place, the technology can be used directly by citizens and firms to interpret legal documents without needing professional intermediaries. The vector embeddings process to identify relationships between different ideas in a text, or between two texts, can quickly identify all the relevant rules or legal precedents in a given scenario described by the user. For example, a small business owner could query their tax obligations in plain English, and the system would return a structured, source-linked explanation tailored to their specific situation. In this way, compliance becomes faster, cheaper, and more accurate, particularly for individuals and smaller firms that lack access to in-house accounting or legal expertise.

Evidence from early deployments of legal AI tools suggests time savings of 30–70% on routine legal research tasks, implying substantial productivity gains if such systems were applied at scale across the economy.

Policy making

The potential for optimal policy has improved exponentially in recent decades. Vast quantities of data are collected and produced continuously across the economy, from administrative records and tax data to real-time labour market and health statistics. In the UK alone, the public sector generates and stores data on tens of millions of individuals and transactions each year, while the wider digital economy produces far more again. At the same time, policy proposals and reports are published daily by a dense network of institutions. There are roughly 150–160 think tanks in the UK, employing only a few thousand researchers, yet collectively producing on the order of 20,000–60,000 reports and policy papers annually.

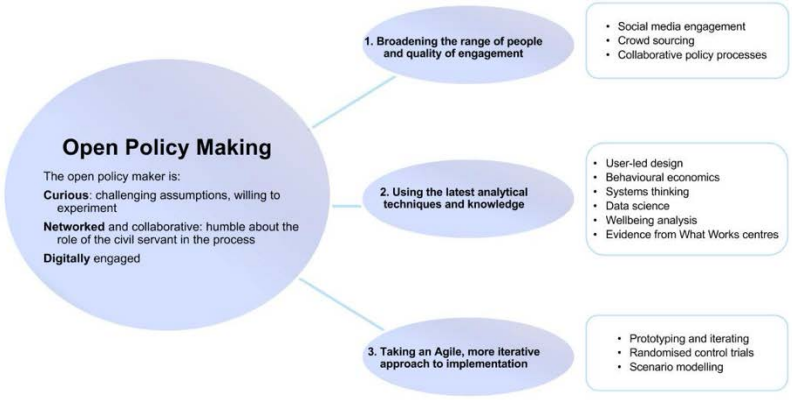
This creates a paradox. The raw inputs for high-quality policymaking, data, analysis, and competing proposals, are more abundant than ever before, yet the capacity of government to process and synthesise this information has not kept pace. Civil servants operate under constraints of time, expertise, and institutional silos. Even within a single department, it is difficult to systematically compare competing proposals, track the evidence base behind them,

or evaluate how similar policies have performed historically. As a result, policymaking often relies on a relatively narrow subset of available information, shaped by what is most visible, recent, or politically salient, rather than what is most comprehensive or robust.

LLMs offer a way to address this bottleneck. By ingesting large corpora of policy documents, academic research, and administrative data, they can identify patterns, compare proposals, and surface relevant evidence in a way that would be infeasible for human analysts alone. This does not replace human judgement, but it expands its effective scope. Instead of reading dozens of reports, policymakers can interrogate thousands, with the system highlighting areas of consensus, disagreement, and empirical support.

In this sense, the constraint on better policy is no longer primarily the availability of ideas, but the ability to organise and evaluate them. LLMs directly target this constraint, turning an overwhelming volume of information into something that can be systematically used.

The civil service has previously launched initiatives to democratise policy making and take advantage of a wider range of expertise, with the ‘Open Policy Making’ programs. This is a diagram they have produced to show the methods of the program:



This program could be heavily augmented by using LLM technology:

Broadening range

Social media engagement has huge potential to expand the range of inputs, however is very difficult at scale. LLMs enable civil servants to make use of much broader public responses. Where previously, to collect valuable data might have required closed question surveys, open ended social media prompts and posts on topics as a whole can be utilised.

LLMs can cluster media on certain themes for easier interpretation by civil servants. They can also identify recurring ideas or concerns, or highlight minority concerns or new ideas that are further from the mainstream or could be sidelined otherwise. This shows how AI can democratise policy making, by making use of a far broader range of opinions than we can traditionally account for.

The latest analytical techniques

Data science techniques are lauded as modern methods to optimise policy, however traditional policy information frequently comes in unstructured forms that are difficult to analyse in this way: reports, articles and qualitative evidence. However, LLMs can convert these into structured data that can be analysed systematically. Extracting variables, identifying arguments and recurring ideas that can be counted or compared and turning text into vector-mapped data for a computer to analyse hugely extends the reach of data analysis, enabling optimised, modern policy making.

Iterative approach to implementation

AI modelling is a rapidly growing sector. Retrieval Augmented Generation (RAG) of live databases enable the iterative approach that is needed (yet hard to scale with human analysts) by updating assumptions and variables as the data changes to adjust predicted outcomes and test policy ideas quickly, in a cheap way.

Consultancy

Government consultancy contracts are one of the most attacked areas of government expenditure. Recent policy has committed to £1.2 Billion on consultancy, however most of this

has focused on the procurement process as opposed to seeking a fundamentally more efficient service.

Consultants are primarily used not for uniquely human insight, but for tasks such as policy research, report writing, programme evaluation, and digital transformation support. These are largely forms of structured analytical work, tasks like synthesising documents, identifying options, and presenting recommendations. These tasks are well suited to the strengths of AI language processing. The National Audit Office points out that consultants are most often used where skills are temporarily unavailable internally, not because of a unique strength of consultancy firms. However, localised LLMs could guarantee the provision of skills within a department and can adapt to fill spaces previously plugged simply by hiring external consultants, at significant cost to the taxpayer.

This category of work is the one in which LLMs are most effective. Much of the value produced by consultants consists of processing large volumes of information and converting it into structured outputs. LLMs can perform these tasks at negligible marginal cost and at far greater scale with greater reliability, for example by analysing thousands of policy documents simultaneously or generating draft reports in seconds. Evidence from comparable domains suggests 30–70% reductions in time spent on routine analytical tasks. This seems an area that LLMs could easily save a large chunk of government spending.

In public facing and internal tasks, the strengths of NLP models applied to modern day problems and tight budgets can revolutionise the work of government. To further analyse the effects and particularly the second order impacts of these changes, we take a systems thinking approach to consider the holistic impact of systems level changes.

Effects

The most significant effects of the large-scale usage of LLMs in public services are unlikely to be the direct efficiencies described above, but rather the second order consequences that emerge once these systems are embedded into the UK’s complex social, economic, and institutional networks. Often, I find that these effects are harder to predict, often delayed, and frequently non-linear. Understanding them fully requires a systems thinking approach: seeing

healthcare, legal compliance, and citizen support not as isolated services, but as interconnected systems shaped by incentives, information flows, and human behaviour.

One immediate second-order impact is a shift in demand dynamics. By lowering the cost, both financial and time, of accessing services, LLMs will significantly increase usage. In healthcare an AI GP assessment system available at home removes friction that currently discourages people from seeking help. While this is beneficial in catching previously neglected conditions earlier, it also risks overwhelming downstream services. Paradoxically, making access easier may increase pressure on the system overall, because more people who would previously have delayed treatment now enter the system earlier. The NHS waiting list as of January 2026 was 7.25 million, and the NHS, already capacity-constrained by the COVID-19 pandemic and austerity, could see GP and specialist referrals rise rather than fall, unless assessments are calibrated with strong thresholds and feedback loops that regulate flow. As the NHS is free at the point of consumption, the lack of a market mechanism to help manage fluctuating demand further increases the number of patients that the NHS must manage and care for.

Lower barriers to entry increase demand, which then puts pressure on d services like GPs and specialists. In the short term, this often can make the system feel worse rather than better. However, there is a counterpoint. Earlier intervention often reduces the severity and cost of medical problems later on. So while demand might spike initially, over time the system could become more efficient as fewer cases reach critical stages. The difficulty is that these benefits are delayed, while the costs are immediate, which makes policymaking increasingly complex

A second major impact concerns the labour market and professional roles. Rather than simply replacing jobs, LLMs are likely to reconfigure them. In healthcare, the “adapted GP appointment” model suggests a decoupling of technical expertise from interpersonal care. Nurses or less specialised practitioners, augmented by LLMs, could handle a larger share of diagnostic processes. This creates a more modular system, where human labour is reallocated toward empathy, communication, and trust-building. These are areas where machines remain weaker. Yet this reconfiguration introduces risks. Over-reliance on LLMs may lead to skill atrophy among professionals. If clinicians increasingly depend on AI-generated suggestions, their independent diagnostic ability may degrade over time. This creates a fragile system: one

that performs well under normal conditions but is vulnerable to failure when the AI is wrong or unavailable.

Furthermore, using AI might seem the best option for a employer in governmental organisations to replace low skilled beginner jobs, but an unintended consequence is that the lack of beginner jobs creates a lack of skills needed to thrive at a more senior level. For example, using LLMs in place of junior doctors to assess patients may lead to a lack of doctors with the skills to diagnoses rarer or more complicated conditions. These information gaps must be addressed in public services to ensure that AI workers don't have unforeseen consequences.

A useful case study can be drawn from early deployments of clinical decision support systems in the United States, where doctors began to follow algorithmic recommendations even when they conflicted with their own judgement. While outcomes initially improved, later analysis showed that in edge cases, rare conditions or atypical presentations, over-reliance on LLMs led to worse decisions. Translating this to an NHS context suggests that LLM integration must preserve human override capacity and actively train professionals to question, rather than defer to, machine outputs.

In the domain of compliance and taxation, second-order effects are likely to be even more profound. Currently, the UK tax code is 20000 pages long. If LLMs make it significantly easier for individuals and small businesses to understand and optimise their tax obligations, this could reduce the informational advantage currently held by large corporations. In theory, this levels the playing field and encourages economic participation. Lower barriers to entry could increase entrepreneurship, particularly among individuals who are currently deterred by administrative complexity. However, this same capability could also accelerate the discovery of tax avoidance strategies.

LLMs trained on legislation may identify loopholes far more efficiently than humans, enabling widespread optimisation behaviour that erodes the tax base. This creates an adversarial dynamic: as governments use AI to simplify and enforce rules, citizens and firms use similar tools to exploit them. The result is a co-evolutionary system, where regulation and avoidance continuously adapt to each other. This dynamic has precedent in financial markets, where

algorithmic trading systems rapidly evolved in response to one another, increasing both efficiency and systemic risk. In a tax context, the equivalent risk is instability in government revenue, making fiscal planning more difficult. One potential mitigation is the use of AI by governments not just for interpretation, but for simulation: stress-testing legislation against millions of hypothetical behaviours to identify vulnerabilities before they are exploited.

The introduction of LLMs into citizen-facing advisory services also raises important questions about trust and legitimacy. While chatbots may provide more consistent and accessible information than human advisors, they fundamentally alter the relationship between citizens and the state. Trust in institutions is partly built through human interaction; replacing this with automated systems risks creating a perception of distance or impersonality. At the same time, anonymity and lack of judgement may encourage greater engagement from vulnerable groups. For example, individuals dealing with debt, immigration issues, or sensitive legal matters may be more willing to seek help from an AI system than from a human advisor. This suggests a dual effect on trust: increased accessibility for some, but potential erosion of institutional legitimacy for others.

As seen in Estonia's extensive use of digital government services, automation has increased efficiency and user satisfaction overall, but it has also required significant investment in transparency and accountability mechanisms to maintain public trust. Citizens must understand how decisions are made, particularly when those decisions affect entitlements or legal status. For LLMs, this is challenging due to their inherently opaque reasoning processes.

This leads to the broader issue of accountability and governance. When an LLM provides incorrect advice, whether medical, legal, or financial, responsibility becomes diffuse. Is the fault with the model developer, the institution deploying it, or the regulatory framework that allowed its use? Without clear lines of accountability, there is a risk of responsibility gaps, where harms occur but are difficult to attribute or remedy.

From a systems perspective, this weakens feedback mechanisms that are essential for improvement. In traditional systems, errors lead to identifiable consequences and adjustments. In AI-mediated systems, errors may be harder to detect, trace, and correct, particularly if they

occur at scale. This underscores the importance of robust monitoring and audit systems, as well as maintaining human oversight in critical decision points.

Another significant second-order effect is the impact on inequality. While LLMs have the potential to democratise access to information, their benefits may not be evenly distributed. Individuals with higher digital literacy, better access to technology, and greater ability to critically evaluate AI outputs are more likely to benefit. Conversely, vulnerable populations may be more susceptible to misinformation or misuse.

This creates a risk of a new form of inequality: not just access to information, but ability to effectively use and interpret it. In education, similar patterns have been observed with the introduction of digital learning tools, where outcomes improved overall but gaps between high- and low-performing students widened. Applying this to public services suggests that LLM deployment must be accompanied by efforts to improve digital literacy and provide alternative access routes.

There is also a geopolitical dimension to consider. The development and deployment of advanced LLMs are currently dominated by a small number of technology companies, most of which are based in the US, like Chat-GPT, Claude and Anthropic. Reliance on external providers introduces issues of sovereignty and strategic dependency. If critical public services depend on foreign owned models, governments may have limited control over their operation, cost, and evolution.

One response is the development of public-sector or open-source models, trained on curated datasets and aligned with national priorities. This approach, however, requires significant investment and technical expertise. The trade-off is between control and efficiency: proprietary systems may be more advanced and cost-effective in the short term, but public systems offer greater long-term resilience and alignment.

Finally, the cumulative effect of LLM integration across multiple domains may lead to a broader transformation in the nature of the state itself. As information processing becomes cheaper and more scalable, governments can, in principle, provide more personalised and

responsive services. Policies could be dynamically adjusted based on real-time data, and citizens could receive tailored advice and support.

This points toward a more adaptive state but also raises concerns about surveillance and autonomy. The same systems that enable personalised services could also enable detailed tracking of individual behaviour. Balancing these capabilities will be a central challenge, requiring clear ethical frameworks and democratic oversight.

In conclusion, while the first-order benefits of LLMs in public services such as efficiency, accessibility, and cost reduction are significant, the true impact lies in the second-order effects that reshape systems over time. These include shifts in demand, labour market reconfiguration, adversarial dynamics in compliance, changes in trust and legitimacy, and new forms of inequality and dependency. Many of these effects involve trade-offs rather than clear gains or losses, and their outcomes will depend heavily on how systems are designed, regulated, and integrated.

A systems thinking approach highlights the importance of feedback loops, incentives, and resilience. From examples across the world, I have seen that Successful implementation will require not just technological capability, but careful attention to governance, accountability, and remembering that in an increasingly AI world, it is humans who come first. If these challenges are addressed, I believe LLMs have the potential to significantly enhance the functioning of public institutions. If not, they risk introducing new complexities and vulnerabilities into already strained public systems.