OpenUK—written evidence (LLM0115)

House of Lords Communications and Digital Select Committee inquiry: Large language models

OpenUK is a globally unique organisation representing the business of Open Technology in the UK and this spans open source software, open hardware, open data and open standards as well as increasingly open AI. It sits at the intersection of software engineering, business, law and policy and is a world leader in Open Technology and focuses on the people in the UK who work in the business of open source and the UK companies creating and using open source. It collaborates globally with open organisations, including the open source foundations which are the custodians of open source software. It is a member of many such organisations and projects run by them. OpenUK is recognised within those as an important part of the global open source leadership[1] and creates a cohesive voice for UK open source.

OpenUK is uniquely placed within the UK to offer comment and clarification on Open Source and to bring together the UK's deep expertise in open source software, open data and AI to support UK Government, regulators and the public sector in building their understanding.

It has provided an initial group of software engineers, data scientists, lawyers and policy experts at a round table for the Office of AI's White Paper Consultation in July 2023 and is working to support various departments in this way. Its second annual conference[2] in February 2024 will offer a consultation room in which public sector departments can consult through round tables, workshops and the like via direct engagement with both the local and global Open Source communities.

UK Open Source Software Market:

The UK's open source software engineering community is number one in Europe by number of developers and lines of code contributed, and number five (generally) on a global basis. In 2022 27% of GVA contributed by the UK tech sector was based upon open source software businesses and individuals working in this space.[3]

In 2023 96% of all software was found to have open source software "dependencies" requiring open source software to run or including open source software. This was the case across open source and proprietary software. And 76% of the software stacks were open source software.[4]

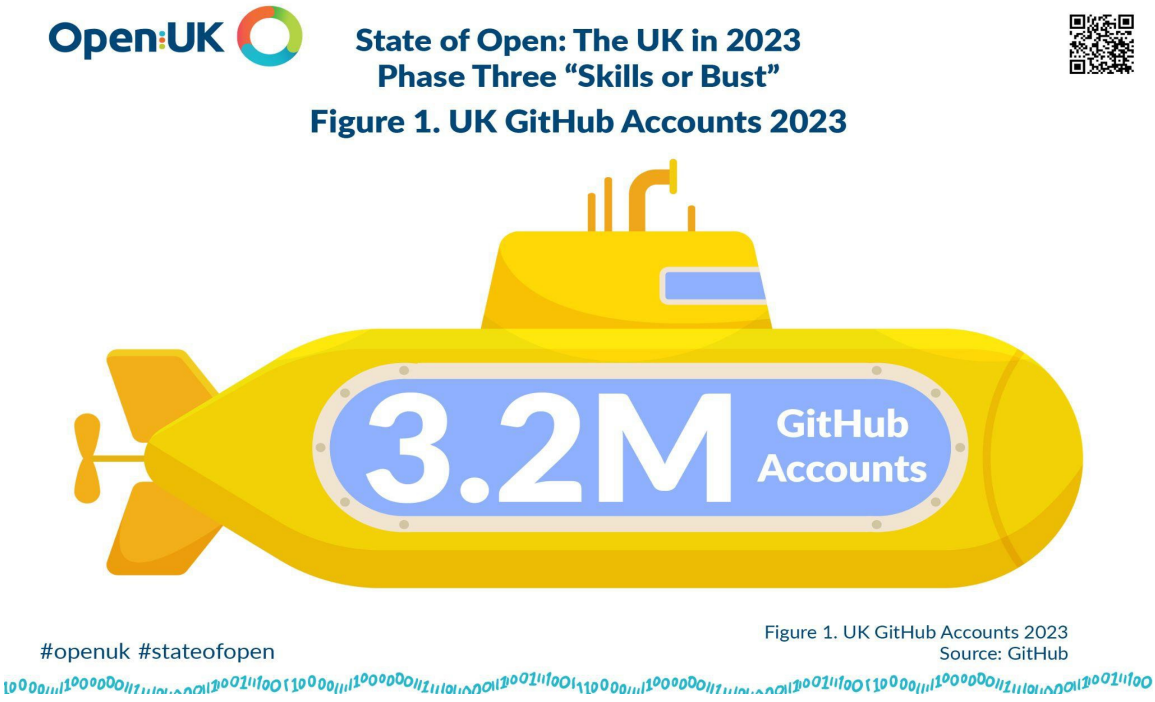
Primarily home-working this community collaborates on global technology projects creating software, providing governance and community building and developer relations as well as commercialisation skills. Employed by global

---

1 https://openuk.uk/participants/our-memberships/
2 https://stateofopencon.com/
3 chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://openuk.uk/wp-content/uploads/2023/07/FINAL-State-of-Open-The-UK-in-2023-Phase-Two-Part-1.pdf
4 https://www.synopsys.com/blogs/software-security/open-source-trends-ossra-report.html

companies including the Big or High Tech companies these individuals are often not well known within the UK yet have "rock star" status in the global tech sector. Open source software may be considered the submarine under the UK digital economy.

With 3.2m GitHub accounts, the standard method of measuring open source software developers, this is 4.5% of the UK population and more per capita than any country in the world.



The UK was the first country in the world to have an open source software first public sector policy and Government Digital Services and GCloud were built on this.[5]

Security vulnerabilities like 2021's Shell4J demonstrated the risk of proprietary software which did not disclose the use of open source software failing to fix such vulnerabilities and understanding the code that is being used is critical to trust.

In considering any requirements or regulations around open source not only must the term be better understood and defined but also the impact of decisions on open source.

The UK's open source expertise and AI leadership enable it to be uniquely placed to succeed in "open AI" as what that means become clear.

---

5 https://www.gov.uk/guidance/be-open-and-use-open-source

Questions:

Capabilities and trends

1. How will large language models develop over the next three years?

a) Given the inherent uncertainty of forecasts in this area, what can be done to improve understanding of and confidence in future trajectories?

Answer:

Opening up AI will democratise technology, build trust, improve innovation, break lock-in and allow competition.

1.1 Access to data

Development of large language models (LLMs) today and in future necessitates training models on data. It is understood that no LLMs are being trained in the UK due to confusion around the ability to use publicly available data. To enable understanding of and confidence in future trajectories the ability to train LLMs on UK data must be clarified and confirmed.

The Text and Data Mining exception believed to enable LLMs to be trained on UK publicly available data is currently shrouded in unnecessary uncertainty. Following the Vallance Report a Code of Conduct confirming this ability was expected but concerns about content use led to this being retracted. There are now grave concerns about the creation of sage AI in the UK and that innovation in the UK will be further stifled by a new Code of Conduct restricting this legitimate use.

Whilst the need for our creative industries to be financially supported is absolutely recognised and supported, the revenue streams to do so must be forward-thinking and progressive. They cannot inhibit innovation or, irrespective of UK actions, they will fail on the global arena. The ability to innovate safely in AI on an equal footing with competitive nations is critical. Options to protect this sector must be explored but not at the detriment of innovation. Possibilities like bulk payments or taxes may be more suited than the outdated royalty model.

It is understood that the Intellectual Property Office will shortly produce a code of conduct and this must clarify the UK's current position to allow use of publicly available data if the UK is to succeed and be pro innovation.[6]

Many nations have specifically enacted exceptions to copyright law to allow for this training whilst the US has a fair use provision, all allowing LLMs to be trained.

---

[6] https://www.ipfederation.com/news/text-data-mining-tdm-uk/

In order to have trust in the data upon which an LLM is trained there must be transparency, which in turn allows for safe AI and control. Additionally this may support alleviating concerns around bias.

Opening up data will enable a more competitive marketplace and support the removal of lock-out, a current blocker to innovation currently impacting the potential for competition.

1.2 Opening up code/software:

There are various aspects of AI that must be considered - weights, models and algorithms primarily as well as documentation and research and the data upon which it was trained. Each and all of those individual aspects may be "open" and shared. The US approach may focus on the weights being open to define what makes an LLM open.

For all countries beyond the US and possibly China the opening up of AI and LLMs will be critical to its ability to control its technology future.

Opening each component and the whole must be considered to determine what will amount to open source AI.

In opening up the software elements we see both "open source software" and "open innovation". These are fundamentally different and have different characteristics and do not both offer the same advantages. In turn this impacts the risk profiles of each.

Risk is of course not only exponential but sits also at a commercial and societal level with respect to AI. Such risks include the lack of access to LLMs and ensuing inability to innovate and that technology may sit in the hands of only a few large companies with the ability to preclude others. History will judge today's decision makers on the choices made with respect to AI and in particular to its being opened up. Their learning from our recent tech history. Learning from history and making informed decisions based on that knowledge is entirely reasonable.

The LLM landscape faces the risk of market dominance and foreclosing competition and innovation if a few players are the exclusive holders of critical technology that others must pay to use and may not inspect or modify.

During the next three years LLMs will inevitably and increasingly open up creating transparency and trust. At the same time some will understandably remain closed and there is room for each. The opening up of AI will involve multiple "shades of open" which must be explored, understood and appropriately accommodated in regulation, codes of conduct etc.

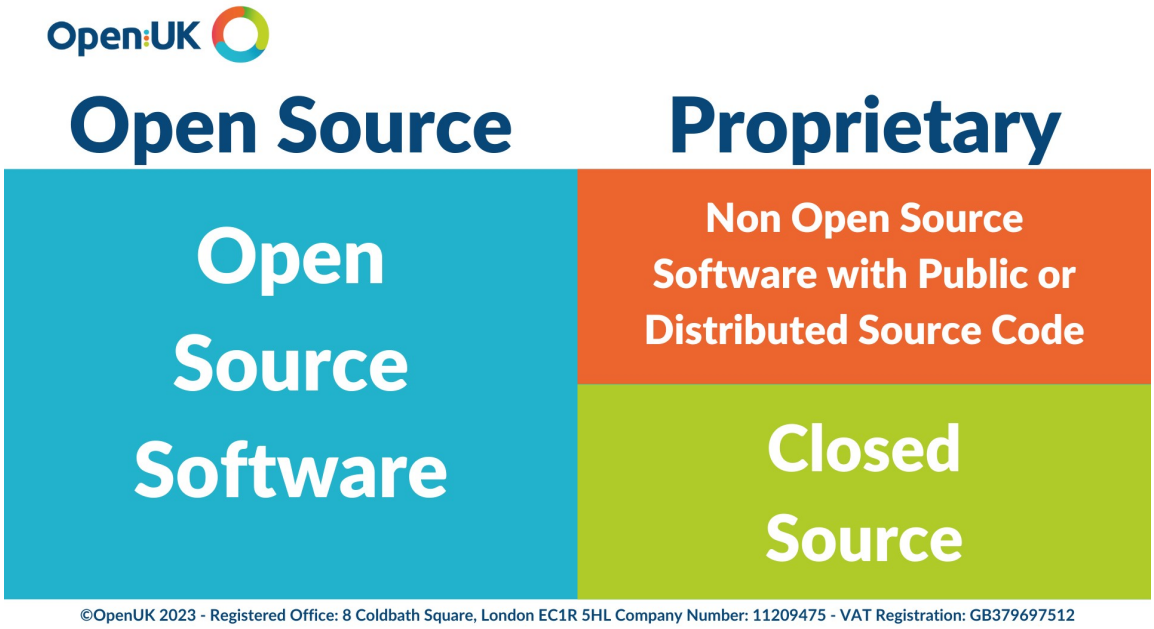
Unfortunately at the present time the varying shades of open do not all have labels or definitions that are universal even amongst those with understanding.

Regulation must recognise this evolution of the levels of openness and different benefits and impacts.

Differing Understandings and shades of openness in software licensing

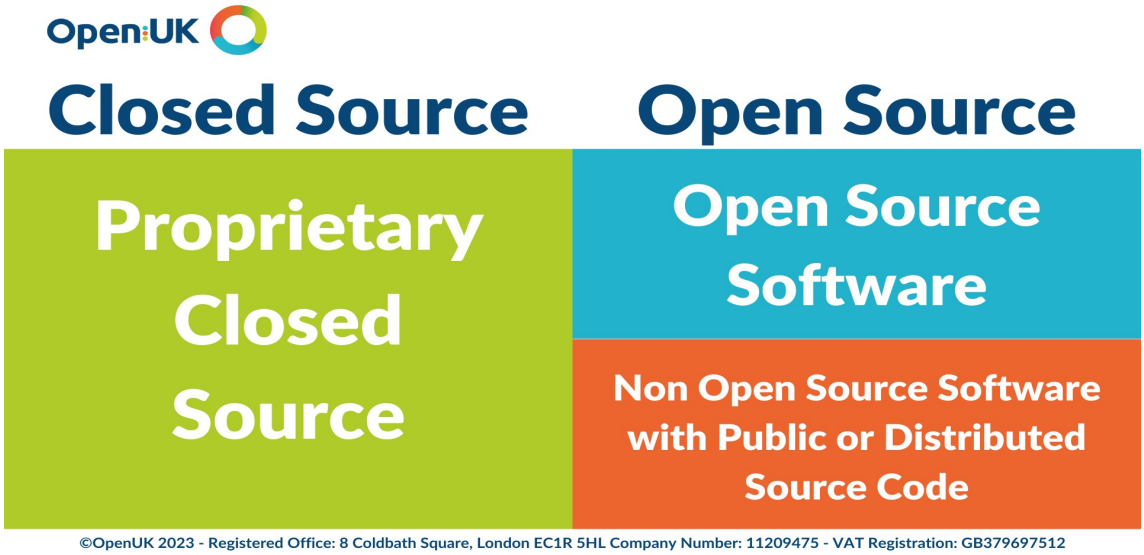
The open source software community categorises software as either open source or proprietary whilst regulatory approaches often categorise it as open or closed. There is a disconnect in understanding.

Open innovation - code that does not meet the Open Source Definition and which may also be labelled distributed source, public source or shared source - due to its licensing is deemed to be proprietary. Proprietary code can be open or closed. This means that code like LLama 2 with its commercial restrictions would be deemed to be proprietary.



The proposed EU AI Act will offer an exception for "free and open source software" which will likely utilise the recognised definition and this term may be purposefully used to avoid the confusion being created by LLMs like Llama 2 being described as "open source" when it is not.

However regulators have generally failed to grasp this nuance and generally contrast open and closed placing the code which has openly available source but does not meet the Open Source Definition under the heading open source software.



These diagrams demonstrate the confusion that currently exists around categorisation of software as open source and the shades of open created by licensing software with the source available but with commercial restrictions.

Attempting to regulate all AI software currently labelled "open source" may be the equivalent to saying that all vehicles will be regulated with identical law, as vehicles rather than understanding the differences between a lorry, a car and a bicycle. Clearly understanding what each is and its impact means that they would never have the same benefits or

The OECD created a new definition of AI being utilised by the EU in its AI Act but we do not have the same regulatory clarity for "open source". This is unfortunate and extremely problematic.

The varying levels of openness might also mean that there ought to be differing levels of benefits correlated to the differing levels of openness and liability.

The shades of openness, their impact and different regulations will be essential to improve understanding of and confidence in future trajectories.

Opening up of Llama 2 and Falcon and shades of openness

The initial leak of Llama 1 LLM in the Spring was not on an open source licence and not open source (despite many wrongly describing it as such) but rather it gave access to an LLM for the open communities. The LLM, a hugely expensive piece of the AI jigsaw, had been missing and access to it enabled faster innovation than had been previously seen across AI through the work of the open communities. This was recognised in the now infamous "Google We Have no Moat Memo."[7] In short this memo recognised that with the level and pace of innovation from the open source communities large company AI despite its finesse was not protected by enough of an IP Moat to allow it to sell AI when

[7] https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

close equivalents created by the open source communities were readily available. What the big techs could provide over and above what the open source communities were able to freely deliver through their collaborative innovation would not be enough of a differentiator.

The memo was of course simply the opinion of one employee but speaks to the value of opening up this technology had on the pace of innovation through collaboration and community contribution.

The initial provision of Llama was a leak and not open source software. It was not provided with a licence that allowed its use let alone an open source one. For the sake of clarity, open source software is licensed on standard approved licences approved by the Open Source Initiative as complying with the Open Source Definition. Use of open source software is based upon a copyright-respecting approved licence. Licences for open source comply with the Open Source Definition which celebrated its 30th birthday in 2023. Definitions 5 and 6 allow anyone to use the code for any purpose subject to complying with the licence.

In a commercial context distributors of open source software enable their competitors with their own innovation. Not a choice lightly made.

This has led to further and legitimised opening up of LLMs with the "Open Innovation" of Llama 2 in July, giving not only a formal if not open source software licence and documentation. OpenUK partnered with Meta on the release of Llama 2 and was the only "open source" organisation to do so. Unlike the pure open source organisations OpenUK focuses on the gambit of opens and was able to partner on the basis that it viewed the shade of open offered by the Llama Community Licence which is not open source as a positive step in the right direction. Meta's web site clearly states that Llama 2 is open innovation and does not claim that it is open source.[8]

The Falcon LLM also released in 2023, by the UAE, is distributed on an open source software licence, the Apache 2.0 licence which meets the Open Source Definition and does not allow for any commercial restrictions. Falcon is open source software.

Disclaiming liability

Almost all open source software licences disclaim developer liability "to the fullest extent permitted by law" and require attribution of the creator of the code. Note - The open source community recognises that laws trump licences and commercial providers of open source (providers of services or enterprise/curated editions of open source not the base open source itself which is freely licensed) as a matter of course comply with laws in their businesses. There appear to be many who have misconceptions around this.

Open source is not about law breaking, it is about building the best software to fix a challenge.

---

8 https://ai.meta.com/llama/

Code of Conduct to manage Risk

Development of AUP/ Code of Conduct requirements will provide guidelines for this innovation encouraging responsible practices from the "Open AI Communities".

This will be particularly important to innovators in AI outside of the US and China giving access to LLMs which had been prohibitively expensive in terms of resource, compute and finances, meaning that these had prior to their being opened up been locked into companies in the US and China.

2. What are the greatest opportunities and risks over the next three years?

a) How should we think about risk in this context?

Risk is not something that is always bad, but something that must be understood.

Once understood a risk appetite must be applied to the facts to allow an informed decision. The UK has a reputation for being risk averse which has historically hindered its ability to innovate.

From understanding will come the ability to balance the need to protect citizens against encouraging innovation. Without the latter there will be no future for those citizens. Work must be done to understand better as the technology evolves.

The approach to risk must be a modern tech friendly one, agile by nature as opposed to the prescriptive waterfall approaches taken in the past. This will allow flexibility and the ability to adapt as the technology emerges.

Following a light touch principles based approach will allow the regulation of LLMs and AI not to fall into the trap we saw with the internet regulation of 30 years ago which has been so painfully replaced when long since redundant and certainly not fit for purpose in the unimaginable future it was not designed for.

Much of the regulation needed relates to use of technology and is already in place. Exciting as LLMs are, they are another form of technology which is subject to the law of the use case and users and distributors must be responsible and exercise discernment. Clearly this is stricter in a regulated sector like finance or health care, as opposed to general commerce.

A limited amount of AI specific regulation would be adequate to manage risk and this could well be managed appropriately through a code of conduct. Certainly this would also potentially offer the UK a leadership position in tech regulation.

The approach to risk must also be collaborative and span geo-politics, engaging with as many countries as possible.

3. How adequately does the AI White Paper (alongside other Government policy) deal with large language models? Is a tailored regulatory approach needed?

a) What are the implications of open-source models proliferating?

Proliferation of open source models - the first risk is the lack of understanding and all shades of open being treated the same and the second is whether the software, models and weights are open but

1. Dealing first with the understanding of the meaning of open source, is a proper name and ought not to be hyphenated. See an explanation of the shades of open and risk of treating a model with any level of commercial restriction in its licensing the same way as a truly open source model.

1.1 The failure to understand the meaning of open source across the discussions we have seen to date and whilst a few companies offering open source products have been included in discussions the representatives of the open source community have not been included. Likely as a consequence of the failure to understand what open source is and how it works.

It is extremely nuanced and this will be a critical failure in understanding and the risk that bad law will be made.

1.2 There are many loud voices shouting and clambering to be part of a conversation to be relevant which it is clear have no understanding. There needs to be an understanding of both nuances of open source development and the wider open source benefits and value to society and economic benefit to the UK.

Greater recognition and engagement of the open source communities - its foundations and representative bodies - is a critical next step if this is to work and risk is to be managed.

1.3 There must also be recognition that blurring the lines of definition and understanding the impacts of the shades of open is in the commercial interests of the commercial parties concerned. If an LLM has commercial restrictions in its licensing the party releasing that may have long term control of a commercial ecosystem. The commercial terms are unknown and the long term impact is unclear.

1.4 Creates a need for regulators and Governments to understand that the open communities respect laws in the same way as society as a whole does and are generally people with a collaboration value set driven by fixing challenges and improving systems.

2. Open source brings many benefits:

- Ability to build on the technology opened up - "to stand on the shoulders of giants" and not repeat unnecessary and costly

innovation, access to LLMs and other technologies

- Better innovation through collaboration

- Community contribution allowing ongoing participation and input

- Better response to security vulnerabilities through a collective response - "many eyes make bugs shallow"

- Democratisation of technology allowing skills development in the UK through access to otherwise restricted technology

- Allowing individuals to gain experience in key technologies opened up which may access international and local jobs

- Removing Lock-in to large vendors of critical technology which may be abused over time

- Access to critical innovation allowing new market entrants and enabling competition

- Allowing local autonomy

- Understanding of the data used to train assuming an appropriate open data licence is also used

3. The risks - the risks in opening up LLMs are largely the same as for closed systems

- Opening up AI may allow bad actors to access innovation, however bad actors are generally able to access innovation and the leak of Llama is a key example of this

- Undermine dominant positions

4. Do the UK's regulators have sufficient expertise and resources to respond to large language models?[5] If not, what should be done to address this?

No they have not taken adequate stock of the voice of the open source software community or its 30 year history and must now engage.

5. What are the non-regulatory and regulatory options to address risks and capitalise on opportunities?

a) How would such options work in practice and what are the barriers to implementing them?

b) At what stage of the AI life cycle will interventions be most effective?

c) How can the risk of unintended consequences be addressed?

International context

6. How does the UK's approach compare with that of other jurisdictions, notably the EU, US and China?

a) To what extent does wider strategic international competition affect the way large language models should be regulated?

b) What is the likelihood of regulatory divergence? What would be its consequences?

The UK would be well advised to look to the US approach on opening weights and to learn from the EU that being first mover is not always best. Overly complex and overly prescriptive legislation will not only fail it will create regulatory capture leaving only a few companies able to comply and to take the contractual risk and liability required to supply LLM based products.

13 November 2023