Hugging Face Comments on the UK Parliament Call for Evidence

Hugging Face commends the UK Parliament Communications and Digital Committee on its ongoing work to examine the opportunities and risks of large language models (LLMs). The following comments are informed by our experiences as an open platform for state-of-the-art (SotA) AI systems, working to make AI accessible and broadly available to researchers for responsible development. Comments are organized by questions listed in Call for Evidence. If a section is not highlighted, we do not have specific, actionable feedback.

About Hugging Face

Hugging Face is a community-oriented company working to democratize good Machine Learning (ML), and has become the most widely used platform for sharing and collaborating on ML systems. We are an open-source and open-science platform hosting machine learning models and datasets within an infrastructure that supports easily processing and analyzing them; conducting novel AI research; and providing educational resources, courses, and tooling to lower the barrier for all backgrounds to contribute to AI. Hugging Face is based in the U.S., with an office in London and a global developer community.

Capabilities and trends

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

Beyond the opportunities of LLMs shared in corporate communications, opportunities with LLMs that are not currently common in practice include:

- Providing the opportunity for creators of data to consent to, and/or be compensated for, their work
- Opening development discussions to people from different backgrounds, a process called "participatory design". For example, non-verbal individuals can share how they would benefit from LLMs that generate multiple possible utterances for them to select from.
- Helping with English language writing, personalized to what the needs are of the user. For example, children learning English may benefit from working with an LLM to create English stories.
- Creating on-the-fly games and entertainment, for example, "Balderdash" for a single player.

The risk landscape, and corresponding harms, is continually evolving and regulatory action should address both present-day harms and foreseeable risks, especially those that affect marginalized communities. For LLMs, this includes representational harms and risks such as discriminatory stereotypes, which can prove catastrophic in high-stakes applications. **The type of risk and prioritization of each risk is contentious. Research to taxonomize existing risks include foundational work on dangers, examining harms to people, and scoping sociotechnical harms.**

Measurable social risks from LLMs include but are not limited to: bias, stereotypes, and representational risks; cultural values and sensitive content; disparate performance; privacy and data protection; financial costs; environmental costs; and data and content moderation labor costs. Risks to society include but are not limited to: trustworthiness and autonomy; inequality, marginalization, and violence; concentration of authority; labor and creativity; and ecosystem and environment.

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

Risks in this context specifically refer to risks of *harm*, where harm includes problematic outcomes for different populations. Risks can arise along the system development and deployment process, meaning that all components and processes – from training datasets to intended application – can embed a certain level of risk. **Efforts to evaluate social risks and conduct comprehensive risk assessments are crucial.** The U.S. National Institute of Standards and Technology's AI Risk Management Framework is one of the leading tools for managing risk, and will soon profile generative AI. Auditing frameworks also give insight to risk management.

Accounting for harms and mitigating risks requires the ability to evaluate them and their severity. Evaluations for large language models, especially for complex social impacts such as biases and environmental costs, are not standardized and have large gaps across risk areas. For example, evaluating biases in large language models often skews to quantification and more evaluations exist for certain protected classes, such as gender, than others, such as age, religion or disability. Specific types of language-based systems, such as code generation, benefit from assessing safety in context. **Better understanding risks requires more resourcing and central fora for testing, which will require better researcher infrastructure, system access, and transparency and deployment disclosure as risk is best assessed in context.**

Domestic regulation

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

We support the proposed pro-innovation approach's methods such as transparency measures and feedback mechanisms. We provided comments on the White Paper in the call for evidence.

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

What constitutes an open-source model does not currently have unanimous agreement. We find it useful to take a step back from specifically open-"*source*" to open models more broadly. Large language models are composed of many components throughout their training and development process that should all be considered in a release. Available options for releasing foundation models vary across a spectrum from fully closed to fully open, each option with its own challenges and tradeoffs. Openness and increased access creates many opportunities for broader community research, including empowering researchers to create safeguards by being able to test on an accessible model. Ethical openness, as exemplified by Hugging Face's approach, requires implementing many types of safeguards. **Increased access to artifacts such as models and datasets enables researchers and external parties to better understand systems, conduct audits, mitigate risks, and find high value applications.**

Specific to models, risks and harms arise from a model regardless of how accessible it may be. A fully closed model risks not having external expertise to guide alignment or risk mitigation. A hosted model can still be used to generate harmful disinformation but can gather user feedback. A fully open model can foster broader research but risks misuse. **The complexity of risk tradeoffs along release methods is why safeguard research parallel to model development is necessary.**

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?

Expertises throughout sectors must complement each other; each sector and organization within provides insights that may not be represented in another. UK regulators can leverage the many available expertises by prioritizing regulatory mechanisms that use transparency to enable stakeholders to meaningfully engage with AI systems; investing in standardization mechanisms designed to work at the component level and are easily implementable; protecting open source and open science when naturally aligned with the requirements of more accountable and democratic technology; and consulting academic stakeholders and open source developers when designing exemption regimes. Government-provided resources for researchers to access infrastructure such as computing power can increase expertise and research on risks. This can take lessons from the U.S. National AI Research Resource.

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

There is no one panacea against all risks from AI; instead, safeguards should span regulatory, policy, legal, and technical tools and levers. Different pieces of legislation can address different aspects of AI systems, such as privacy legislation and IP law's impact on training data being separate but often complementary. Options can address many risks at once. Transparency requirements can be one vector of influence. Requiring model documentation, such as model cards, can address transparency and research reproducibility concerns.

Concretely, non-regulatory options we recommend are: transparency guidance, researcher access and protections, public infrastructure for AI. Regulatory options include proportional requirements for systems by use (sector, risk, popularity) and use case.

b) At what stage of the AI life cycle will interventions be most effective?
As stated in question 2.a. of this response, since risk may arise along the life cycle, research and risk management processes should be applied across many system components. More research and evaluation tools are needed for examining training data, such as being able to examine attributes of large datasets. As we shared in our response to the U.S. Department of Commerce's National Telecommunications and Information Administration's Request for Comments, accountability mechanisms such as audits should focus on all stages of the development process by requiring transparency via good documentation and external access processes and inviting broad contribution across affected stakeholders

Conclusion

The LLM regulatory landscape requires many expertises. We thank the UK Parliament for the opportunity to provide our insights and look forward to supporting ongoing and future efforts.

Respectfully,

Irene Solaiman
Policy Director
Hugging Face

Margaret Mitchell
Chief Ethics Scientist
Hugging Face

Submitted: 1 September 2023