# OpenAI—written evidence (LLM0113)

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

OpenAI welcomes the opportunity to provide written evidence to the House of Lords Communications and Digital Select Committee's inquiry into Large Language Models. This inquiry and forthcoming report will contribute significantly to UK and international governance efforts in better understanding the trajectory of large language models and considering how best the UK can respond to opportunities and risks raised by this technology.

We are grateful for the opportunity to share our experiences developing and deploying cutting-edge AI technology, as well as studying and advancing AI safety. This work is consistent with our long-standing commitment to working collaboratively with governments to ensure the development and widespread availability of safe and beneficial AI tools.

### 1. Future Trajectories: Please could you describe how the next generation of large language models is likely to develop over the next 3 years, and what are the main drivers of future capability leaps beyond that?

Large language models are becoming important to society in many ways. Future models, including near-future models, will be significantly more capable than their current iterations. We see both capability and safety continuing to improve in tandem, as the technology delivers immense, tangible benefits to society.

Future generations of these models will be more capable, more flexible, and more accessible. We recently launched GPTs, these are custom versions of ChatGPT that you can create for a specific purpose. These are based on today's model but give a sense of one way in which we expect future models to develop, toward an experience that is increasingly tailored to what the individual user needs. We also expect that models will become increasingly multi-modal, as our own have recently done. You can now interact with ChatGPT by voice, and the system can understand and generate images.

We know that making models more capable is only one part of the picture – we are also working constantly to make them safer. We are investing in preparedness to address potential catastrophic risks, and in safety systems that use our own technology to help make today's products safer. For instance, we're leveraging the immense capabilities of GPT-4 to innovate on safety, trimming the time it takes to undertake some safety processes down from months to hours. We also continue to invest in monitoring systems to detect and address misuse of our systems.

Our Application Programming Interface (API) allows third-party developers to integrate the capabilities and benefits of our models into their own applications. Thousands of organizations are building new features, applications, and businesses using our API. We're also working with a number of major UK brands in food delivery, logistics, and financial services, including challenger banks and asset management.

**2. Frontier risk and preparedness: How credible is the likelihood of catastrophic risk which involves threat to life or financial instability? Do we have credible warning indicators for the next generation of models?**

As part of our mission to ensure that artificial general intelligence benefits all of humanity, we take seriously the full spectrum of safety risks related to AI, from the systems we have today to the furthest reaches of superintelligence.

We work to ensure safety is built into our systems at all levels. For example, prior to releasing any new system we conduct rigorous testing, engage external experts for feedback, work to improve the model's behavior with techniques like reinforcement learning with human feedback, and build safety and monitoring systems.

In July 2023, alongside other leading AI labs, we made a set of voluntary commitments to reinforce the safety, security and trustworthiness of AI technology and our services. We published an update on our voluntary commitments, including our progress on frontier AI safety, as part of our preparation for the UK AI Safety Summit.

These commitments have three parts: safety, security, and trust. They include a range of best practices such as internal and external red-teaming, publicly reporting model or system capabilities and limitations, and working with other labs to further develop best practices via the Frontier Model Forum (FMF). FMF is an industry body to promote the safe and responsible development of frontier AI systems. FMF recently hired an Executive Director, Chris Meserole, and announced a $10 million AI Safety Fund, which is designed to support independent researchers to accelerate new academic research into AI safety.

We also recently announced that we are creating a team called Preparedness. This team will develop a Preparedness Framework, which will detail our approach to developing rigorous frontier model capability evaluations and monitoring and establishing a structure for operations and governance across that development process. The framework is meant to complement and extend our existing risk mitigation work, which contributes to the safety and alignment of new, highly capable systems, both before and after deployment.

As the capabilities of these models get more advanced, so does the scale and severity of the risks they may pose, particularly if under direction from a malicious

actor or if the model is not properly aligned with human values. Rigorously measuring advances in potentially dangerous capabilities is essential for effectively assessing and managing risk. We are addressing this by exploring and building evaluations for potentially dangerous capabilities that range from simple, scalable, and automated tools to bespoke, intensive evaluations performed by human experts.

We are also committed to progressing science-based measurements of frontier system risks, which are still nascent. To support this work, we are collaborating with academic and industry experts, and aim to contribute to the development of a diverse suite of evaluations that can contribute to the formation of best practices for assessing emerging risks in highly capable foundation models. We believe dangerous capability evaluations are an increasingly important building block for accountability and governance in frontier AI development. To that end, we're supportive of the UK AI Safety Institute, which is poised to make an important contribution to filling this crucial gap.

### 3. Regulation: Do you believe that further regulatory intervention is needed, or would that stifle innovation? What should that look like?

We believe it is essential to develop regulations that incentivize AI safety while ensuring that people are able to access the technology's many benefits.

Given the emergence of increasingly powerful AI systems, the stakes for global cooperation have never been higher. While we know different countries will make different choices on some aspects of regulating AI, we think it's important that these efforts are as coordinated as possible so that we can fully realize the benefits of AI.

In addition to supporting a greater degree of international coordination overall, we are supportive of policy efforts in two key areas:

- **Progressing the science of measuring and mitigating risks posed by the most advanced systems:** Defining evaluations that measure frontier capabilities is an active area of research and more investment is needed to strengthen evidence-based methods and usher this field into maturity. By improving capability evaluations, policymakers and other experts will be better able to evaluate risks and determine which systems should be subject to heightened safety obligations.

- **Independent expert red-teaming of system safety:** We believe it is important that our safety approaches are externally validated by independent experts, and that our decisions are informed at least in part by independent safety and risk assessments, so we can develop appropriate mitigations prior to deployment.

We're aso supportive of policies the UK government have put forward to further develop the AI R&D ecosystem, including the UK AI Research Resource's new supercomputer facility at the University of Bristol, the £118M skills package, the AI Standards Hub, and the UK AI Safety Institute.

**4. Copyright: Please could you explain your position on the use of copyrighted works in LLM training data? What are the options for building models without using copyrighted data?**

We respect the rights of content creators and owners, and look forward to continuing to work with them to expand their creative opportunities. Creative professionals around the world use ChatGPT as a part of their creative process, and we have actively sought their feedback on our tools from day one. By democratizing the capacity to create, AI tools will expand the quantity, diversity, and quality of creative works, in both the commercial and noncommercial spheres. This will invigorate all creators, including those employed by the existing copyright industries, as these tools increase worker productivity, lower the costs of production, and stimulate creativity by making it easier to brainstorm, prototype, iterate, and share ideas.

We believe that AI tools are at their best when they incorporate and represent the full diversity and breadth of human intelligence and experience. In order to do this, today's AI technologies require a large amount of training data and computation, as models review, analyze, and learn patterns and concepts that emerge from trillions of words and images. OpenAI's large language models, including the models that power ChatGPT, are developed using three primary sources of training data: (1) information that is publicly available on the internet, (2) information that we license from third parties, and (3) information that our users or our human trainers provide. Because copyright today covers virtually every sort of human expression–including blog posts, photographs, forum posts, scraps of software code, and government documents–it would be impossible to train today's leading AI models without using copyrighted materials. Limiting training data to public domain books and drawings created more than a century ago might yield an interesting experiment, but would not provide AI systems that meet the needs of today's citizens.

In training our models, OpenAI complies with the requirements of all applicable laws, including copyright laws. Nevertheless, although we believe that legally copyright law does not forbid training, we also recognize that there is still work to be done to support and empower creators. We have been industry leaders in allowing creators to express their preferences with respect to the use of their works for AI training. For example, we've led on websites' ability to exclude their content and we provide an easy way to disallow for our "GPTBot" web crawler to access a site, as well as an opt-out process for creators who want to exclude their images from future DALL·E training datasets. While we look forward to continuing to develop additional mechanisms to empower rightsholders to opt-out of training,

we are actively engaged with them to find mutually beneficial arrangements to gain access to materials that are otherwise inaccessible, and also to display content in ways that go beyond what copyright law otherwise allows. We have already struck partnership deals with publishers like the Associated Press, and we expect our ongoing negotiations with others to yield additional partnerships soon.

*5 December 2023*