GitHub

88 Colin P Kelly Jr Street
San Francisco, CA 94107
United States of America

11 March, 2024

RE: European Commission call for contributions:
Competition in generative AI

GitHub is dedicated to providing a home for open source development and accelerating building proprietary and open source software alike with AI-powered developer tools, including our GitHub Copilot product. Over 15.5 million developers in the EU build and share software on our platform.[1] A 2019 European Commission-sponsored study found that open source software development—as measured by GitHub activity—contributes at least €65-95 billion per year to European GDP.[2] As an independent subsidiary of Microsoft, we welcome the opportunity to share our expertise and perspective on competition in generative AI.

I. Open source competition in the AI stack

Competition in generative AI applications for consumers or business is shaped by the extensive and nascent market that supports the development, distribution, and use of AI applications. The technical stack that supports AI development and deployment as well as its value chain is complex. We offer GitHub's perspective on this value chain as both a collaboration platform hosting software development relevant at various layers of the stack and as a downstream application developer that uses AI models in our products. For example, while consumers may label some software as open source and others as proprietary, an analysis of widely used tech stacks reveals that on average 76% of a tech stack is composed of open source software code, regardless of whether the final product is marketed as closed or open.[3] Looking beyond the completed AI product will prove similarly informative.

At a high-level, the development and deployment of AI systems are a function of compute, data, and expertise. In practice, expertise manifests in innovative architectures and software that manage each stage of pre-training, application building, and serving. The availability of open source software has spread this expertise, widely distributing the knowledge and capability to build AI. Below we present a software-focused AI stack that illustrates the software components that are used to create and serve AI products, expanding on our

[1] https://innovationgraph.github.com/economies/eu#developers
[2] https://digital-strategy.ec.europa.eu/en/news/commission-publishes-study-impact-open-source-european-economy
[3] https://www.synopsys.com/content/dam/synopsys/sig-assets/reports/rep-ossra-2023.pdf#page=7

previous stack shared in an AI Act position paper.[4] Featured examples are open source software projects openly available on the GitHub platform. In many cases, start-ups will freely share open source software that may also be integrated into paid product offerings; individual developers will often openly share components that they have developed and maintain as personal projects.[5] This offers one point-in-time view of a rapidly evolving value chain.[6]

The AI Software Stack		
Layer	*Component*	*Example*
Application		
	Interface	Gradio supports the development of graphical user interface-based applications for AI inference.
	Monitoring and safety	Helicone documents and provides limits on model requests and inference costs. Rebuff and Guardrails protect against known prompt-injection attacks and provide specification validation for output quality guarantees, respectively.
	Orchestration and non-model enhancements	LangChain supports chains of model outputs and prompts, retrieval-augmented generation (RAG) using databases, and more. JARVIS, AutoGen, and AutoGPT are additional open source frameworks for orchestrating models and outputs from one model to other model instances.
	Model serving	Many open source frameworks can be used to deploy models for inference, including Transformers, Truss, LoRAX, and web-LLM.
Model		
	Pre-trained model	ONNX Model Zoo has a variety of AI and generative AI models available directly on GitHub. Other popular repositories of openly

[4] https://github.blog/wp-content/uploads/2023/02/GitHub_Position_Paper-AI_Act.pdf#page=3
[5] For additional context on open source software development organizations and business models, please see https://openfuture.eu/blog/open-source-is-unavoidable-open-source-policies-and-digital-sovereignty/; https://sifted.eu/articles/open-source-ai-make-money.
[6] Depictions of the AI stack may vary depending on the audience and objective. Venture capital firm Andreessen Horowitz has published an LLM application stack with services commonly used by providers of AI systems; Menlo Ventures has more recently published a simplified stack focused on enterprise AI. The UK Competition and Markets Authority and Microsoft, among others, have offered stylized stacks that contextualize software-based AI systems in the compute infrastructure required to train them.

		available pre-trained models for download include Ollama and Transformers.
	Evaluations	Model evaluations vary widely in purpose, from question answering performance to safety. Popular evaluation frameworks include Evals and HELM; popular benchmarks include MMLU and ARC. Numerous more purpose specific evaluations exist, including Fairlearn for bias and Purple Llama for cybersecurity.
	Model enhancement	Models are commonly modified to improve performance following pre-training. Methods include Reinforcement Learning with Human Feedback (RHLF, example open source implementation available here), Low-Rank Adaptation (LoRA), quantization (popular example), and fine-tuning on additional data (popular example).
Training		
	Monitoring	Libraries including Weights and Biases and Sacred organize and document AI training experiments for reproducibility.
	Architecture selection	At the outset of AI model training, the architecture must be specified. This is commonly done in custom software code that makes calls to frameworks like TensorFlow, PyTorch, and JAX. Example: GPT-NeoX
	Infrastructure management	In order to perform training, especially for large datasets, resource virtualization is required to distribute the computational load across multiple processors, whether locally or in the cloud. Example: Ray. At a lower level, GPU optimization tools support AI training and inference. Example: ZLUDA.
	Data management	Prior to training, a number of data processing steps are required to clean, format, and serve the data. Example: Data-Juicer.
Infrastructure		
	Hardware and data	Although outside the scope of this software-focused stack, both inputs are essential to train and run AI applications. Each input has its own value chain that should be assessed for competition.

I a. *Open source software supports competition within each layer of the stack*

As can be seen above, there are examples of open source software at nearly every layer of the stack, software that is often developed and shared on GitHub. Open source software may be developed by companies as part of their business model, shared by academic institutions, created by hobbyists, and/or maintained by governments. Regardless of its origin, open source licenses ensure that enterprises, new or old, big or small, can use the code in

developing their products and services. Openly available models, in particular, give businesses building AI-powered applications many options along the build-buy spectrum, at every layer of the stack. One business may elect to purchase API-hosted services for a model, while another may fine-tune or otherwise modify a model and host directly, while still others may train an established architecture from scratch or pursue novel research.

I b. *Open source software has positive effects across layers of the stack*

Competition in one layer in the stack supports competition in other layers. Open source AI components and software in layers of the AI stack can thus prove complements to competition in other layers, including in closed applications. For example, the Transformers library written by Hugging Face, licensed under Apache-2.0, and hosted on GitHub is used in open source and proprietary AI applications alike, including OpenAI's ChatGPT.[7] Similarly, Apple has reportedly adapted Google's JAX training framework for the development of its proprietary models.[8] Ultimately, the availability of open source components at nearly every layer of the stack lowers barriers to entry, as a new entrant need not write their entire software stack in order to build competing solutions.[9]

II. The market for AI-powered developer tools is competitive

Focusing on AI applications, in contrast to other layers of the stack, and to the market for AI applications that support software development, GitHub Copilot is a ground-breaking AI developer tool. It leverages large language models to suggest code and entire functions in real-time, as a developer writes code. Users experience GitHub Copilot as an extension for their integrated development environment (IDE). IDEs supported today include open source offerings Vim and Neovim, Microsoft offerings Visual Studio Code, Visual Studio, and Azure Data Studio, in addition to a suite of IDEs offered by Czech software firm JetBrains. Copilot Chat expands on this functionality to answer coding-related questions directly in the IDE, supporting cases ranging from generating test cases, to explaining highlighted code and proposing fixes.

There is considerable and growing competition in the AI-powered developer tool market. Meta, AWS, Google, Replit, ServiceNow, and others have developed large language models specifically designed for coding tasks.

[7] https://www.theregister.com/2023/03/24/column/
[8] https://analyticsindiamag.com/apple-springs-a-surprise-embraces-open-source-training-method/
[9] Research from Harvard and the University of Toronto estimates that corporate users of open source software would need to spend $USD 8.8 trillion to rewrite the code from scratch if it were not available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4693148.

Meta's CodeLlama and the StarCoder model are two such prominent models to be released openly, for further modification and/or integration into products. AWS, Replit, ServiceNow, Snyk, and others today offer commercial services for AI-powered code generation. Systems and applications using general purpose AI models, like ChatGPT, Claude, and others, also offer code suggestions in response to developers' prompts. Startups including Magic AI have raised significant funding to bring AI tooling to software development.

Amid competition, the GitHub Copilot product emphasizes a focus on developer experience and enterprise trust. The product is designed for developers, to work in their favorite IDEs while abstracting away challenges in ensuring low latency and otherwise strong AI model performance. The product uses multiple filter systems to ensure that suggestions from models are scrutinized for privacy, security, and copyright risks prior to being surfaced to a user. These filters and our approach to responsible AI development give GitHub confidence to be transparent and provide legal certainty to customers on risks that may otherwise prevent adoption.[10]

GitHub customers use Copilot to accelerate their software development, bringing its users competitive advantage in numerous markets. GitHub features stories from our customers on our public webpage. Customers from a range of sizes, industries, and geographies have found value from GitHub Copilot, including leading European firms like Mercedes-Benz, Philips, Carlsberg, Indra, Amplifon, and Doctolib.

Mercedes-Benz uses GitHub to accelerate its over 23,000 software developers. On GitHub, they build and maintain 90% of Mercedes's software source code, stored across 65,000 projects and 4,100 organizations. Mercedes began testing GitHub Copilot in January 2023, and within 6 months elected to make the AI-powered developer tool available to all of its developers. Mercedes developers using Copilot accept 30% of code suggested made in their IDE and report saving 30 minutes per day.[11]

"We are digitally transforming as a company," explains Mercedes-Benz lead architect Andy Krieger. "We want to be a software-driven car manufacturer, and GitHub helps us on our way forward."

"It helps us onboard new software engineers and get them productive right away. We have all our source code, issues, and pull requests in one place.

[10] This includes an IP indemnification from GitHub for unmodified suggestions from Copilot. See "Can GitHub Copilot users simply use suggestions without concern?" answer https://resources.github.com/copilot-trust-center/.
[11] https://www.youtube.com/watch?v=HdAqqNM1i9c

With GitHub Actions, we can automate and deploy our software. GitHub is a complete platform that frees us from menial tasks and enables us to do our best work." - Application Manager Fabian Faulhaber.

III. Regulation should be carefully crafted to support competition in AI

Legislative work on the recent AI Act included considerable debate on the competition effects of the regulation and how to minimize distortions. While the outcome remains to be seen, it is important for regulators to consider the competitive effects of non-AI-specific regulation on the generative AI market and its constituent layers in the AI stack. Some research, for example, finds that GDPR has increased the cost of data for EU firms by an average of 20% and led them to be less data-intensive in their operations, with smaller firms and those in software industries experiencing the most distortion.[12]

Two areas that may raise challenges for developers to lawfully build and share open source AI components are data protection and copyright. Today large language models are commonly trained on large datasets collected from the public internet. In many cases, open source developers have transparently documented their collection and data curation decisions in leading datasets like the Pile.[13] Datasets commonly include sources that may contain personal data and copyrighted content. Regulatory guidance for European data protection authorities and other national regulators may directly impact the extent to which open source AI development is viable in member states and thus affect competition in the AI market.

Outside of the EU, the UK ICO recently concluded a consultation on their interpretation of UK GDPR focused on the legitimate interest basis for personal data processing in training AI systems.[14] As framed, their interpretation may restrict the open sharing of AI model weights trained on personal data because the developer cannot constrain how such model weights are adapted and used. Given that models today are trained on vast amounts of internet content that contains personal data, this risks an effective prohibition against open source AI development. More broadly, restrictions on text and data mining to collect AI training data risk entrenching today's data-dominant firms.

[12] https://www.chicagofed.org/publications/working-papers/2024/2024-02
[13] https://pile.eleuther.ai/
[14] https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-first-call-for-evidence/

Text and data mining rules under the 2019 EU Copyright Directive permit the collection of data for AI training, requiring non-research organizations to respect machine readable opt-outs. Under the AI Act, developers of general purpose AI models will need to respect this opt-out and describe their training data in line with a Commission-drafted template. These requirements raise administrative burdens that may fall more heavily on small players, as was seen with GDPR.[15] In any case, the principle enshrined in the 2019 Directive promotes competition: by enabling text and data mining for AI training, data moats from large companies do not dictate winners in the AI market. Absent such an exemption and one that is administered nimbly, AI developers would need to go to every individual author on the net to seek a license, a herculean task.

As European institutions draft interpretive guidance for the AI Act and associated regulations, bringing clarity for open source developers and start-ups will support a competitive AI market. Across EU institutions, investing in capacity and expertise on open source software, its measurement, and its competition effects can support a more vibrant single market.

[15] See footnote 12.