HUGGING FACE

Hugging Face Comments on the Open Consultation on Copyright and Artificial Intelligence

Submitted to the United Kingdom Intellectual Property Office

About Hugging Face

Hugging Face is a community-driven company based in the U.S. and France, dedicated to democratizing responsible machine learning (ML). Our platform is the most widely used for sharing and collaborating on ML systems, fostering open-source and open-science initiatives. We host machine learning models and datasets within an infrastructure that enables efficient data processing, analysis, and research. Additionally, we provide educational resources, courses, and tooling to lower the barrier to AI participation for individuals from all backgrounds.

Supporting Open Research and Small Developer Needs in AI Copyright Frameworks

Based on our experience working with the AI community, we broadly support **Option 3's approach**—a text and data mining exception with rights reservation. Our perspective is informed by direct engagement with open-source developers, researchers, and content creators who use our platform. Below, we outline key considerations drawn from practical experience.

The Value of Open Datasets and Transparency

The Hugging Face **Datasets Hub** hosts thousands of publicly available datasets, each documentation support via **Dataset Cards** that detail their contents, licensing terms, and limitations. This transparency framework has demonstrated concrete benefits:

- **For researchers**: Ensures data provenance can be verified, facilitating proper attribution of sources. Hugging Face datasets and dataset cards have supported projects such as the Data Provenance Initiative, which in turn have been used to examine fair use arguments as it relates to AI.
- **For rights holders**: Provides visibility into where and how their content is being used, and assists rights holders in clarifying licensing information.
- **For developers**: Clear licensing documentation in open datasets by rights holders, in turn, clarifies legal boundaries for dataset usage, reducing uncertainty.

Our experience shows that **open documentation practices enhance trust without impeding innovation**. A key example is the **BigScience ROOTS dataset**, which was used to train the **BLOOM model** while maintaining transparency regarding data sources and filtering decisions via the **ROOTS Search tool.** Hugging Face's fully open and well-documented datasets such as **FineWeb**, have since powered open, benchmark leading models in their size class, such as **SmolLM** and **SmolVLM**.

Hugging Face transparency initiatives demonstrate that **proper documentation does not create undue burdens when appropriate tools are available**. In fact, research on Hugging Face datasets has shown that well-documented datasets are also the most downloaded ones—**95% of download traffic comes from datasets with documentation.**

Transparency also helps in compensation. We have observed that while licensing frameworks are being explored in the ecosystem, early attempts at compensating creators have yielded minimal financial returns for most rights holders (with one estimate pointing at payouts of $0.01 per image), while being prohibitively expensive for smaller developers to negotiate such deals. Our experience with transparent datasets and opt-out mechanisms has shown that structured transparency and open-source tooling may provide more practical and economical benefits to both creators and developers than complex licensing schemes that primarily benefit large corporations.

Finally, we recommend that transparency requirements align with existing **international frameworks**, such as the **EU AI Act**, which mandates dataset summaries for training large AI models. Public dataset documentation should be encouraged rather than restricted to internal regulatory disclosures.

Practical Implementation of Opt-Out Mechanisms

As a platform hosting thousands of datasets, we have gained **practical experience implementing rights-holder preferences**, demonstrating that effective opt-out mechanisms are feasible:

- **Dataset search tools**: Features like **Data Studio** **(with a dataset viewer and natural language dataset query)** enable creators to explore dataset contents and identify potential use of their works.
- **Integration with Spawning.ai's opt-out API**: Allows content creators to register preferences regarding AI training use that is displayed via a widget on the Hugging Face datasets webpage for datasets with image URLs.
- **BigCode's Am I In the Stack** app allows creators to remove their github repositories from the database via an easy opt-out tool.

These efforts illustrate how **platform-based solutions,** in partnership with ecosystem actors, benefit small developers and individual researchers while respecting rights holders' requirements, reducing the burden of implementing custom technical solutions. We advocate for standardized opt-out signals (e.g., machine-readable rights expressions) across platforms to prevent fragmentation and ensure accessibility for creators of all sizes.

Balancing Open Research with Copyright Protections

Text and data mining (TDM) exceptions are **critical** for fostering research and innovation, particularly for:

- **Scholarly researchers** analyzing AI capabilities and risks.
- **Open-source developers** with limited legal and financial resources.
- **Educational institutions** training future AI practitioners.
- **Community-led research initiatives** such as **Aya** and **BLOOM**.

TDM and TDM-like exceptions are already practiced in several jurisdictions, such as via DMCA exceptions in the United States and the TDM exception in the EU AI Act. An approach with Option 3 that preserves TDM will keep the UK in line with global jurisprudence and extend the benefits of TDM exceptions for researchers in the UK.

Our **BigScience initiative** showcased how open, collaborative research can **advance AI understanding while maintaining ethical data practices**. The **data governance working group** developed frameworks that **balance research needs with copyright protections**, demonstrating the feasibility of responsible AI development.

In summation, to safeguard open research, we recommend:

- Ensuring **research exemptions remain intact** for **non-commercial and educational uses**.
- Avoiding rigid licensing requirements that could disproportionately impact **smaller AI labs, academic institutions, and open-source developers**.
- Supporting **technical infrastructure** that enables researchers to **document and disclose dataset provenance** while complying with copyright law.

Conclusion & Recommendations

Based on our platform experience, **Option 3** offers the most balanced approach to supporting innovation while upholding creator rights. However, effective implementation is crucial:

1. **Transparency requirements** should build on **existing community standards** rather than introduce entirely new frameworks.

2. **Opt-out mechanisms** should be **technically supported** to be **accessible to small creators and developers**.
3. **Research exceptions** must be preserved to ensure AI's continued **advancement in education and open science**.
4. **Proportional compliance mechanisms** should be implemented to avoid **disadvantaging smaller organizations**.
5. **Dataset registration and metadata standards** should be promoted to **reduce ambiguity** and facilitate compliance.

Hugging Face remains committed to developing tools and best practices that promote **responsible AI development while respecting intellectual property rights**. We welcome continued engagement on practical implementation strategies.

Submitted by:
Avijit Ghosh, Yacine Jernite, and Irene Solaiman
With input from Bruna Trevelin
Hugging Face