Chapter 4 / Capítulo 4

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 4: One Person Company: Artificial Intelligence, Autonomous Entrepreneurship and the Future of Business
ISBN (this volume): 978-9915-9966-3-9
DOI: 10.62486/978-9915-9966-3-9.Ch04
Pages: 37-47

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Individual and organizational creators of the most-downloaded open artificial intelligence models

Creadores individuales y organizativos de los modelos de inteligencia artificial abiertos más descargados

Dennyse María Patricia Hermosa Guzmán[1]

1 Universidad de la Amazonia. Florencia, Colombia.

ABSTRACT

Introduction: Whether the most-used open artificial intelligence models are produced by organizations or by individual creators is an open question with direct consequences for funding and platform policy. Methods: The 500 most-downloaded public models on the Hugging Face Hub were retrieved and each author account was classified as an individual or an organization using the platform application programming interface. Results: Organizations authored 430 of the 500 models (86.0 percent) and 91.3 percent of their downloads, against 70 models (14.0 percent) and 8.7 percent of downloads by individuals. Ten individual creators reached the top, concentrated in speech recognition and embedding models, and their median downloads (6.55 million) matched the organizational median (6.47 million). Conclusions: The most-downloaded open models are organizationally dominated, but a small set of individual super-creators still reaches the top in specialized niches, which suggests targeted support for individuals is feasible rather than futile.

Keywords: open artificial intelligence; model creators; Hugging Face; individual developers; platform ecosystem

RESUMEN

Introducción: Si los modelos de inteligencia artificial abiertos más utilizados los producen organizaciones o creadores individuales es una pregunta abierta con consecuencias directas para la financiación y la política de plataformas. Métodos: Se recuperaron los 500 modelos públicos más descargados de Hugging Face Hub y se clasificó cada cuenta autora como individuo u organización mediante la interfaz de programación de la plataforma. Resultados: Las organizaciones crearon 430 de los 500 modelos (86,0 por ciento) y el 91,3 por ciento de sus descargas, frente a 70 modelos (14,0 por ciento) y el 8,7 por ciento de las descargas de individuos. Diez creadores individuales alcanzaron la cima, concentrados en reconocimiento de voz y modelos de incrustación, y su mediana de descargas (6,55 millones) igualó la mediana organizativa (6,47 millones). Conclusiones: Los modelos abiertos más descargados están dominados por organizaciones, pero un pequeño conjunto de supercreadores individuales sigue llegando a la cima en nichos especializados, lo que sugiere que el apoyo dirigido a individuos es factible más que inútil.

Palabras clave: inteligencia artificial abierta; creadores de modelos; Hugging Face; desarrolladores individuales; ecosistema de plataforma

INTRODUCTION

The open artificial intelligence ecosystem now supplies the models that a large share of downstream applications are built on, and Hugging Face has become its central marketplace.[1] Foundation models, the large pre-trained systems on which the ecosystem rests, have reorganized how artificial intelligence is produced and distributed,[2] and their release has become a policy question in its own right.[3] The practical stakes of open-model production are high, because the openness of a model determines who can inspect it, adapt it and build on it, and those downstream freedoms are precisely what the ecosystem's value proposition rests on. For a funder deciding where to place a grant, for a platform deciding how to govern its creators, and for an individual developer deciding whether to compete, one question matters before all others: who actually creates the models that everyone uses. The concern is that the most-downloaded models are produced almost exclusively by well-resourced organizations, which would make the openness of the ecosystem a thin veneer over concentrated production.[4] If, by contrast, individual creators can still reach the top, then the ecosystem retains the distributed character that its proponents claim.[5] The question is sharpened by the dual nature of open models: they widen access, but they also raise security and governance obligations that individual creators may be poorly placed to meet.[6] The stakes are not abstract, because the models at the top of the download ranking are the ones that downstream developers, researchers and products actually depend on, so the identity of their creators determines who holds the leverage in an ecosystem that presents itself as open.

The current answers to this question are partial, which is why they do not suffice for decisions. The literature documents the rapid growth of the platform and the mechanics of model release and documentation,[7,8] surveys the model marketplaces that have emerged around it,(9) and has begun to measure development activity on the Hub at large.(10) What it has not done is isolate the top of the download distribution and ask who sits there, individual or organization, and how much of the usage each group commands. Studies of the ecosystem tend to aggregate across the long tail, where the distribution of creators is different from the head, or to focus on governance and security rather than on the creator mix.[11,12] The decentralized and international dimensions of open artificial intelligence are debated as matters of principle,[13,14] but the empirical creator mix at the point of highest use has not been quantified. The distinction between the head and the tail is not a technical nicety: the long tail of the platform is populated by thousands of experimental and personal models, whereas the head is where the actual usage concentrates, and the two have different creators, different maintenance burdens and different policy significance. A funder or platform that reasons from the long tail will therefore draw the wrong conclusion about where the leverage lies, because the long tail over-represents individuals and under-represents the usage that the head carries. Conversely, a funder that reasons only from the very apex of the ranking will conclude the ecosystem is closed to individuals, because the apex is the most organizational part of the head. The measurement must therefore be taken at the head as a whole, not at the tail and not at the single most-downloaded model. For the purpose of deciding where scarce support should go, the head of the distribution is what matters, and it is precisely the head that has not been characterized.

The decision contexts that depend on this answer are concrete rather than rhetorical. A research funder must choose between concentrating its open-model budget on established organizations, which is safe but consolidates the status quo, and spreading small grants across individuals, which is risky unless the individual route actually works. A platform must decide how much of its tooling and promotion to devote to individual creators, whose models may be lighter-weight but whose diversity of niches is a public good. An individual developer must decide whether the open ecosystem still offers a realistic path to reach and influence, or whether it has become an organizational game. Each of these decisions is currently made without a measurement of who actually occupies the head of the download distribution, and each would be made differently depending on whether the head is closed or porous. The measurement reported here is intended to supply that missing fact, and it is framed so that each decision-maker can read the same result through their own lens.

This study supplies that characterization. The objective is to quantify the contribution of individual and organizational creators to the most-downloaded open models, and to describe the individual super-creators who nevertheless reach the top. The application scope is the Hugging Face Hub, the dominant public repository for open models.(15) The study is deliberately descriptive rather than causal: it does not ask why the split is what it is, only what the split is, and it leaves the causal mechanisms to the comparison with the existing literature in the Discussion. This is the

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

appropriate division of labour for an applied question, where the first requirement is an accurate measurement of the current state rather than a further theoretical conjecture.

METHODS

DATA SOURCE

The Hugging Face Hub public application programming interface was queried without authentication to retrieve the 500 models with the highest cumulative download counts at the time of extraction. Each model record includes its identifier, its author namespace, and its download count. The extraction is a cross-sectional snapshot of the platform's download ranking. The Hub is an established empirical object for this kind of study, and its suitability for replicable research has been assessed directly;(15) model release practices on the platform, including versioning and documentation, are themselves the subject of a growing measurement literature,[8,16] and the models retrieved are known to be real, inspectable artifacts whose supply chain and synchronization with upstream repositories have been characterized.[17,18] The download count is the platform's own cumulative usage metric, which is the closest available proxy for how heavily a model is actually pulled into downstream work, and it is the metric by which the platform's own leaderboard orders models.

CREATOR CLASSIFICATION

The author of each model was taken as the namespace preceding the first slash of the model identifier. Each unique author was then classified as an individual or an organization by querying the platform's user overview endpoint; an author for which the endpoint returned a user profile was classified as an individual, and an author for which it returned a not-found error was classified as an organization, because organizations on the platform are not served by the user endpoint. This classification follows the platform's own account type rather than an external heuristic, and it is therefore the same distinction that the platform itself draws between personal accounts and organizational accounts. The measurement is limited to what the public endpoints expose, so any model that is gated or private is not part of the ranking, and the classification inherits the platform's account typology, including the possibility that a personal account represents a small team; both points are returned to in the Discussion. The classification procedure is fully mechanical, which is a deliberate design choice: it makes the result reproducible by any reader with the same endpoint access, and it avoids the researcher judgement that a manual classification of borderline accounts would otherwise introduce into the data. Because the procedure queries each unique author only once and then reuses the cached classification, the classification is consistent across all of an author's models, which avoids the inconsistency that a per-model judgement would produce.

CREATOR CLASSIFICATION

The author of each model was taken as the namespace preceding the first slash of the model identifier. Each unique author was then classified as an individual or an organization by querying the platform's user overview endpoint; an author for which the endpoint returned a user profile was classified as an individual, and an author for which it returned a not-found error was classified as an organization, because organizations on the platform are not served by the user endpoint. This classification follows the platform's own account type rather than an external heuristic.

ANALYSIS

Models were grouped by creator type, and three quantities were computed for each group: the number of models, the share of models, and the total and median downloads. The median was used in preference to the mean for the download comparison, because download counts are heavily right-skewed and a single apex model would otherwise dominate the group average and obscure the fact that a typical top model, whether individual or organizational, is downloaded at a similar rate. The individual creators with models in the top 500 were then listed by model count

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

and total downloads to identify the super-creators. All counts are reported as absolute values followed by percentages, in keeping with the applied orientation of the study, so that a reader can see both the raw magnitude and the relative weight of each group. The choice of the top 500 rather than a larger or smaller head is a deliberate one: it is large enough to admit several distinct niches of individual success, but small enough to represent the models that actually carry the bulk of downstream usage. A smaller head would have missed the individual super-creators who sit lower in the ranking, while a larger head would have diluted the measurement with the long tail whose creators and economics are different. The analysis was performed in Python 3.14 with the standard library and pandas, and is reproducible from the public endpoints with no restricted access, so a reader who wishes to re-run it with a different head size or a different snapshot date can do so directly.

RESULTS

The download volume was heavily concentrated in a small number of models at the very top, with the single most-downloaded model accounting for 190.1 million downloads, and the head of the ranking was dominated by embedding and sentence-similarity models, which are general-purpose infrastructure used across a wide range of downstream tasks.

This concentration means that a handful of organizational releases carry a disproportionate share of the ecosystem's total usage, and the individual share, while real, is concentrated in the lower reaches of the top 500 rather than at the very apex. The practical consequence is that the apex is the most organizationally controlled part of the distribution, and the individual presence strengthens as one moves down from the apex toward the middle of the head, a gradient that the aggregate split alone would obscure.

Table 1 summarizes the split. Of the 500 most-downloaded models, 70 were authored by individuals, or 14.0 percent, and 430 by organizations, or 86.0 percent (Table 1). The download share was even more concentrated: individual-authored models accounted for 517.0 million downloads, 8.7 percent of the total, against 5.43 billion downloads, 91.3 percent, for organizational models.

Table 1. Individual and organizational creators of the 500 most-downloaded open models.

Creator type	Models	Share of models (%)	Total downloads	Share of downloads (%)	Median downloads	Unique creators
Individual	70	14.0	517005740	8.7	6554675	10
Organization	430	86.0	5429052960	91.3	6470221	36

Note: Downloads are cumulative platform download counts at the time of extraction. Share is the percentage of the column total.

The median downloads of the two groups were nearly identical, 6.55 million for individual models and 6.47 million for organizational models, which means an individual model that reaches the top is downloaded about as much as an organizational one, and the entry barrier for an individual is therefore one of placement rather than of eventual reach. The absolute gap between the groups therefore reflects how many models each side placed in the head, not how heavily a placed model is used.

The ten individual creators were also more concentrated than the thirty-six organizations, in that the top two individuals between them held thirty of the seventy individual models, or nearly half, whereas the organizational share was spread across a larger number of accounts and no single organization dominated the head to the same degree. This means the individual margin is not only thin but internally concentrated, with a small number of prolific individuals carrying most of the individual presence.

Ten individual creators accounted for the entire individual presence in the top 500, whereas 36 organizations accounted for the organizational share (Table 2). The individual creators were concentrated in two niches: speech recognition and text embeddings. The most prolific individual, with 15 models and 121.6 million downloads, built speech-recognition models, and the second, also with 15 models and 106.5 million downloads, built embedding models.

The remaining eight individuals each held five models, and their models spanned speech recognition, text generation and model conversion across several task families. The download totals of the leading individuals, 121.6 million and 106.5 million, placed them among the most-used models in the entire ranking, not merely among individuals. Figure 1 shows the split by creator

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

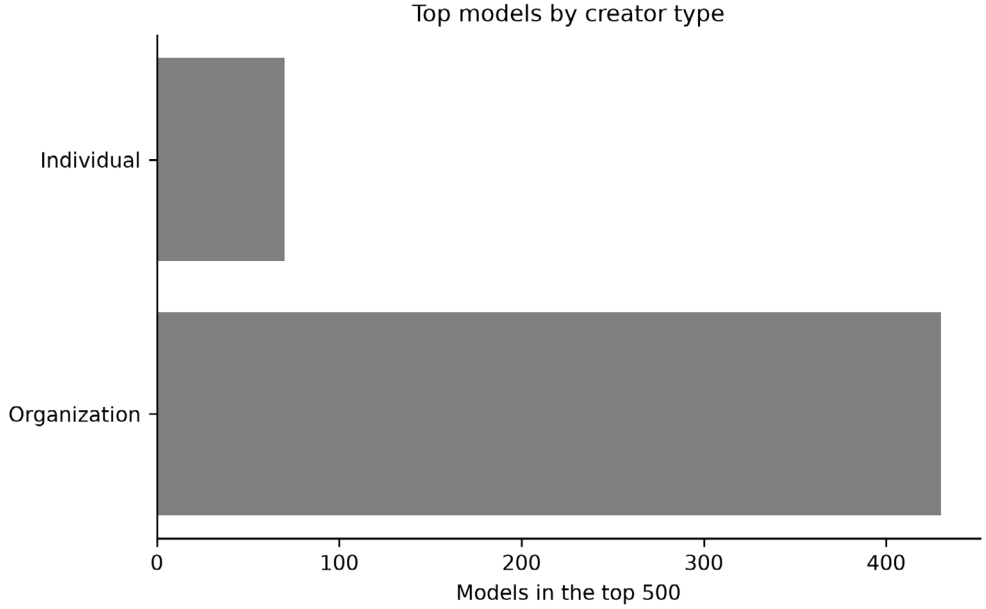
type (Figure 1).

Table 2. Individual super-creators in the 500 most-downloaded open models.

Individual creator	Models in top 500	Total downloads
jonatasgrosman	15	121615500
intfloat	15	106483055
hexgrad	5	54032405
prithivMLmods	5	45445520
Bingsu	5	43817880
audio-cpp	5	37054500
Serveurperso	5	32723530
mudler	5	28497890
SamLowe	5	25090595
Xenova	5	22244865

Note: Models = number of the creator's models in the top 500. Total downloads are cumulative across those models.

Figure 1. Distribution of the 500 most-downloaded open models by creator type.



Note: The bar shows the number of the top 500 models authored by each creator type.

For practitioners, three facts stand out. First, individual success in the top 500 is rare, ten people across the entire ecosystem. Second, when it happens it is deep rather than broad, with the top two individuals each holding fifteen models. Third, individual success clusters where the task is narrow and the data is abundant, speech and embeddings, rather than where frontier-scale compute is required. The individual models are also downloaded at a rate that matches the organizational models, which means the rarity of individuals is a matter of entry, not of reach once entered.

The concentration of individual success in ten people, two of whom hold thirty models between them, further suggests that the individual margin is winner-take-most rather than a broad base of solo creators. This implies that the individual route to the top runs through specialization and reuse, not through head-on competition with organizational labs, and that it rewards the few who find a narrow, high-demand task before others do.

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

DISCUSSION

The implication that leads is that the open ecosystem's usage is organizationally dominated, but not closed to individuals. A funder who assumes that only organizations produce the models people use would be mostly right, but a funder who concludes that supporting individuals is pointless would be wrong, because ten individuals command a non-trivial share of the head and their models are downloaded as heavily as organizational ones. This is the practical hinge of the result: individual support is not a lost cause, but it must be aimed at the niches where individuals actually win.

The magnitude of the split, 86 percent of models and 91 percent of downloads, is large enough that the ecosystem cannot be described as a level playing field, yet the survival of a measurable individual share is enough that it cannot be described as closed either; the accurate description is an organizational head with a porous individual margin.

For funders, the actionable reading is to target the speech-recognition and embedding niches. The two leading individuals built in those two areas and accumulated more than one hundred million downloads each, which suggests that a small grant placed there, or a compute credit aimed at multilingual or low-resource speech and embeddings, can buy reach that matches an organizational release.

For the platform, the actionable reading is to invest in the conversion and quantization tooling that most individual top models depend on, because several of the ten individuals reached the top by repackaging existing models rather than training new ones, and the platform's tooling is what makes that low-cost route viable. For an individual developer, the actionable reading is to compete on the narrow task with abundant data, where the organizational advantage in compute and staff is least decisive, rather than on frontier-scale training.

The funder decision can now be stated as a trade-off between reach and diversity. A grant placed with an organization buys reach, because organizations hold the bulk of the head, but it also buys more of what already exists. A grant placed with an individual in the speech or embedding niche buys less aggregate reach but more of the niche diversity that the organizational head under-serves.

The present data show that the individual option is not a lottery with no winners: ten individuals have already converted narrow niches into top-500 reach, and the two leaders command download totals that would be respectable for an organization. In practice, a funder should therefore treat the individual route as a niche strategy, funded in the speech, embedding and conversion niches where the evidence shows individuals win, rather than as a general-purpose alternative to organizational funding. This is a concrete, evidence-based allocation rule, which is what the applied orientation of this study is meant to deliver. The rule is also self-correcting: because the measurement is reproducible, a funder can re-run it on a later snapshot to see whether the individual margin has widened or narrowed, and adjust the allocation accordingly rather than relying on a static judgement.

The comparison with the existing literature sharpens these implications. The documented growth of the platform and the mechanics of model release[7,8] describe a growing supply of models, but they do not say who sits at the top; the present result fills that gap by showing the head is organizational while the individual route survives at the margin. The quantitative analysis of development activity on the Hub(10) found a broad and distributed community, which is consistent with our finding that the long tail is broad, but the head is what carries the usage, and the head is concentrated. The suitability of the Hub as an empirical object is well established,(15) and the documentation and supply-chain literature[12,18] confirms that the models we rank are real, public artifacts rather than metadata noise.

The governance implications connect to the wider debate on open artificial intelligence. The result that usage is organizationally concentrated complicates the claim that openness automatically democratizes production,[4,5] and it aligns with the observation that open science is at risk of falling behind the proprietary frontier even as open model counts grow.(19) At the same time, the survival of ten individual super-creators is evidence that the barriers are not yet prohibitive, which matters for the argument that open release is a democratizing force.[20,21] The transparency and documentation infrastructure that makes individual releases credible, model cards and datasheets, is a precondition for the individual route, and its uneven adoption is itself a governance question.(22,23)

The practical reading of the individual route can be made more precise by separating two claims that are often conflated. The first claim is that individuals can reach the top, which the data confirm for ten creators. The second claim is that individuals compete on the same terms as

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

organizations, which the data reject, because the individual presence is confined to niches where the compute and maintenance burden is low. The distinction matters because it tells a funder where individual support is high-leverage, in the low-burden niches, and where it is not, in the frontier, and it prevents the false inference that the survival of a few individuals implies a level playing field.

The comparison with the software-engineering literature adds precision to the individual route. The reuse and integration of pre-trained models in open-source projects has been shown to follow the same dependency logic as ordinary software, which means an individual model becomes valuable precisely when it can be depended on by others.(24) The synchronization between Hub models and their upstream repositories is imperfect, and the executability of the example code that individual model cards carry is uneven, both of which raise the cost of the individual route even where the model itself is competitive.[17,25] The concentration of organizational production is not simply a matter of compute; it is also a matter of the documentation, maintenance and trust infrastructure that organizations can afford and individuals cannot.(16) This is why the individual super-creators cluster in niches where the model is self-contained and the documentation burden is low, speech and embeddings, rather than in the frontier where ongoing maintenance dominates.

The concentration result also bears on the economic reading of openness. An open ecosystem whose usage is 91 percent organizational is open in the sense of access, because anyone can download and inspect the models, but concentrated in the sense of production, because the models that carry usage come from a small set of organizations. Both facts are true at once, and a policy that attends to only one of them will misread the ecosystem. The distinction between access-openness and production-openness is the one that matters for competition policy and for the claim that open models counterbalance proprietary concentration.(4) The present data show that access is genuinely open, but that the production of the highest-use models is an organizational affair with a thin individual margin, which means openness and concentration coexist rather than being opposites. This is a more precise statement than either the celebratory or the sceptical position in the existing debate, and it is the measurement that allows the two positions to be reconciled.

The analogy with traditional open-source software is instructive but must be applied carefully. In conventional open-source, individual maintainers are known to sustain a large share of critical infrastructure, often with minimal support, and the bus-factor problem of a single maintainer is a recognised governance risk. The open-model ecosystem differs in one decisive respect: the cost of producing a frontier model is orders of magnitude higher than the cost of maintaining a software package, so the individual route is compressed into the niches where a model can be trained or adapted at a cost an individual can bear. This is why the individual super-creators appear in speech and embeddings, tasks with mature recipes and abundant public data, rather than in large language models, where the compute barrier is prohibitive. The open-source-software lesson that individuals can sustain critical infrastructure therefore carries over, but only for the subset of model work that is computationally within reach of a person, which the present data delimit precisely.

The organizational side of the split is also explained by the ecosystem's own dynamics. Companies have learned to create vendor-led open-source foundations as a way to coordinate and legitimate their releases,(26) and the model supply chain is now a recognized risk surface that organizations are better equipped to manage than individuals.[18,27] The role of individuals inside organizations, including the underrepresentation of some groups in Hugging Face organizations, is itself a governance question that the present creator-level view only partially captures.(28) The gradient of release, from closed to fully open, is a deliberate strategic choice made disproportionately by organizations, and the transparency of those choices varies.[29,30] Together these observations suggest that the organizational dominance of the head is structural, resting on compute, documentation and trust, rather than a transient state that individual creators will soon overturn.

The limitation of the creator-type classification is that it follows the platform's account type, and an account held by a one-person organization or a loosely coordinated collective may be misclassified in either direction; the recommendation that overcomes it is to validate the classification against a manual sample of author profiles. The limitation of the snapshot is that download rankings drift as new models arrive, so the precise membership of the top 500 is time-dependent; the recommendation is a longitudinal re-extraction that tracks the head over successive quarters. The limitation of using downloads is that a model can be downloaded heavily without being used productively, so downloads measure reach rather than value; the

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

recommendation is to complement downloads with a usage-proxy such as likes or dependent-model counts. Each of these limitations is a measurement refinement rather than a threat to the central split, because the dominance of organizations is too large to be an artefact of any single one. A further limitation is that the platform's account typology does not capture the funding or staffing behind a nominally individual account, so an individual creator who is backed by an institution would still be counted as individual; the recommendation is to link creator accounts to disclosed funding or affiliation where available, which would sharpen the distinction between genuinely independent and institutionally supported individuals.

The platform's own role deserves a concrete recommendation, because the platform is the intermediary through which the individual route runs. The finding that several individual super-creators reach the top through model conversion and adaptation implies that the platform's conversion, quantization and hosting tooling is load-bearing for individual success, and a platform that wants a distributed creator base should invest in lowering the cost of those operations further. The platform's promotion and discovery mechanisms also matter, because an individual model that is not surfaced will not accumulate downloads regardless of its quality; a discovery policy that gives niche and individual models a path to visibility is a concrete lever the platform holds. These recommendations are directed at the platform operator specifically, in keeping with the applied orientation of the study. The platform is not a neutral host in this story, because its tooling and discovery decisions determine whether the individual route is cheap or costly, and the present data indicate that where the platform has made conversion and adaptation cheap, individual creators have responded by reaching the head.

For the purpose of policy, the result supports a two-track stance: keep the general infrastructure open, because it is what lets individuals compete, and add targeted support for the niches where individuals demonstrably win. The ecosystem's openness is doing real work precisely where the ten super-creators operate, and a policy that treated open models as a purely organizational affair would forfeit that margin. The generalisability of the result is bounded by its snapshot and by the platform itself: the Hugging Face Hub is the dominant but not the only repository for open models, and a different platform or a different point in time could return a different creator mix, so the precise percentages should be read as a measurement of this platform at this time rather than as a universal law of open-model production. The mechanism that generalises, however, is the distinction between the head and the tail, and between the niches where individuals win and the frontier where organizations dominate, which is portable to any open-model marketplace.

CONCLUSIONS

Organizations authored 430 of the 500 most-downloaded open models and 91.3 percent of their downloads, while ten individual super-creators still authored 70 models concentrated in speech recognition and embeddings, with median downloads matching the organizational level, a split that is concentrated at the apex but porous at the margin. The most-downloaded open artificial intelligence models are therefore organizationally dominated but not closed to individuals, and the individual route to the top runs through specialized niches rather than through head-on competition with organizational labs.

The practical consequence is that support for individual creators should be targeted at the specialized niches where they demonstrably win, speech recognition, embeddings and model adaptation, rather than abandoned as futile or spread thinly across the long tail of experimental models. The ecosystem's openness is doing measurable work for the ten individuals who occupy the head, and the task of funders and platforms is to widen that margin rather than to treat the head as an organizational monopoly.

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

REFERENCES

Pol UR. Hugging Face: Revolutionizing AI and NLP. IJRASET. 2024;12(8):1121-1124. doi: 10.22214/ijraset.2024.64023
Bommasani et al. On the Opportunities and Risks of Foundation Models [Internet]. 2021. Available from: https://arxiv.org/abs/2108.07258
Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021:610-623. doi: 10.1145/3442188.3445922
Luitse D, Denkena W. The great Transformer: Examining the role of large language models in the political economy of AI. Big Data & Society. 2021;8(2):20539517211047734. doi: 10.1177/20539517211047734
Widder, West and Whittaker. Open (for business): Big tech, concentrated power, and the political economy of open AI [Internet]. 2023. Available from: https://arxiv.org/abs/2308.10160
Al-Kharusi Y, Khan A, Rizwan M, Bait-Suwailam MM. Open-Source Artificial Intelligence Privacy and Security: A Review. Computers. 2024;13(12):311. doi: 10.3390/computers13120311
Banh L, Strobel G. Generative artificial intelligence. Electron Markets. 2023;33(1):63. doi: 10.1007/s12525-023-00680-1
Ajibode A, Bangash AA, Cogo FR, Adams B, Hassan AE. Towards semantic versioning of open pre-trained language model releases on hugging face. Empir Software Eng. 2025;30(3):78. doi: 10.1007/s10664-025-10631-3
Qian M, Musa AA, Biswas M, Guo Y, Liao W, Yu W. Survey of Artificial Intelligence Model Marketplace. Future Internet. 2025;17(1):35. doi: 10.3390/fi17010035
Osborne C, Ding J, Kirk HR. The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub. J Comput Soc Sc. 2024;7(2):2067-2105. doi: 10.1007/s42001-024-00300-8
Lin Z, Li Z, Liao X, Wang X. Characterizing the Ecosystem of Open-Source Uncensored Large Language Models: Supply Chain, Governance, and Risks. Proceedings of the 2026 ACM Internet Measurement Conference. 2026:495-513. doi: 10.1145/3777912.3839799
Stalnaker T, Wintersgill N, Chaparro O, Heymann LA, Penta MD, German DM, et al. An Empirical Analysis of Machine Learning Model and Dataset Documentation, Supply Chain, and Licensing Challenges on Hugging Face. ACM Trans Softw Eng Methodol. 2026;35(8):1-49. doi: 10.1145/3776739
Keršič V, Turkanović M. A review on building blocks of decentralized artificial intelligence. ICT Express. 2025;11(3):486-506. doi: 10.1016/j.icte.2025.04.001
Emery-Xu N, Jordan R, Trager R. International governance of advancing artificial intelligence. AI & Soc. 2025;40(4):3019-3044. doi: 10.1007/s00146-024-02050-7
Ait A, Cánovas Izquierdo JL, Cabot J. On the suitability of hugging face hub for empirical studies. Empir Software Eng. 2025;30(2):57. doi: 10.1007/s10664-024-10608-8
Lu H, Han J, Huang Y, Liu J, Bao L, Han S, et al. What Makes a Good Model Card?. J Software Evolu Process. 2026;38(10):e70170. doi: 10.1002/smr.70170
Ajibode A, Bangash AA, Sghaier OB, Adams B, Hassan AE. On the synchronization between Hugging Face pre-trained language models and their upstream GitHub repository. Empir Software Eng. 2026;31(5):117. doi: 10.1007/s10664-026-10826-2
Chen Z, Chen Z, Wu S, Chen B, Song W, Huang Y, et al. A First Look at Model Supply Chain: From the Risk Perspective. Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering. 2026:2913-2924. doi: 10.1145/3744916.3773171
Pei G, Huang H. Open science falling behind in the era of artificial intelligence. Front Res Metr Anal. 2025;10:1595824. doi: 10.3389/frma.2025.1595824
Osborne and Weiser. Open Models, Closed Minds? On the Democratic Promise and Limits of Open-Source Generative AI [Internet]. 2024. Available from: https://arxiv.org/abs/2312.14471
Chen M, Wu K, Pradhan P, Allen C, Nativi S, Hubacek K, et al. Steering open-source AI to accelerate the sustainable development goals. Nat Commun. 2026;17(1):4959. doi: 10.1038/s41467-026-73866-8
Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019:220-229. doi: 10.1145/3287560.3287596
Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, III HD, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86-92. doi: 10.1145/3458723
Yasmin J, Jiang W, Davis JC, Tian Y. Software dependencies 2.0: An empirical study of reuse and integration of pre-trained models in open-source projects. Empir Software Eng. 2026;31(6):173. doi: 10.1007/s10664-026-10894-4
Yao S, Zhao J, Zhang Y, Wu Y, Wang T, Zhang X. Understanding the Executability of Example Code in Hugging Face Model Cards: An Empirical Study. Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering. 2026:443-445. doi: 10.1145/3774748.3795671
Yenişen Yavuz E, Riehle D, Mehrotra A. Why do companies create and how do they succeed with a vendor-led open source foundation. Empir Software Eng. 2025;30(1):40. doi: 10.1007/s10664-024-10588-9
Yun S, Zhang Y, Shi Z, Wang L, Wang Y, Bai Y. A Large-Scale Empirical Study of Secret Key Leakage in Hugging Face Spaces. Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering. 2026:1520-1531. doi: 10.1145/3744916.3773117
Salinas MT, González A, Martínez-Fernández S. Exploring the Role of Women in Hugging Face Organizations. Anais do XXVIII CongressoIbero-Americano em Engenharia de Software (CIbSE 2025). 2025:75-89. doi: 10.5753/cibse.2025.35293

https://doi.org/10.62486/978-9915-9966-3-9.Ch04

Solaiman. The Gradient of Generative AI Release: Methods and Considerations [Internet]. 2023. Available from: https://arxiv.org/abs/2302.04844
Bommasani, Klyman, Longpre et al. The Foundation Model Transparency Index [Internet]. 2023. Available from: https://arxiv.org/abs/2310.12941

https://doi.org/10.62486/978-9915-9966-3-9.Ch04