

Karakteristik Yuridis Pelanggaran Hak Cipta dalam Penggunaan Data untuk Pelatihan *Large Language Model* (LLM) Generatif

Asep Supriyadi¹, Aturkian Laia²

Ilmu Hukum, Hukum, Universitas Dirgantara Marsekal Suryadarma, Jakarta Timur, Indonesia¹

Ilmu Hukum, Hukum, Universitas Dirgantara Marsekal Suryadarma, Jakarta Timur, Indonesia²

*Email Korespodensi: asep.nts30@gmail.com

Sejarah Artikel:

Diterima 07-07-20267
Disetujui 14-07-2026
Diterbitkan 16-07-2026

ABSTRACT

The rapid development of generative artificial intelligence, particularly Large Language Models (LLMs), has created new copyright issues concerning the use of copyrighted works as training data without the authors' permission. This study aims to examine the legal characteristics of copyrighted data used in LLM training, identify potential copyright infringements under Indonesian law, and analyze the regulatory challenges surrounding generative AI. The research employs a normative legal method using statutory and conceptual approaches, based on Law Number 28 of 2014 on Copyright, the Electronic Information and Transactions Law, and relevant academic literature. The findings indicate that data scraping, reproduction, and data processing for AI training may infringe the exclusive rights of copyright holders because such activities do not fall within the scope of fair use. The absence of explicit regulation on text and data mining creates legal uncertainty. Therefore, Indonesia should establish specific copyright exceptions, collective licensing mechanisms, and fair compensation to balance AI innovation with copyright protection.

Keywords: *generative artificial intelligence; copyright; large language model; text and data mining*

ABSTRAK

Perkembangan kecerdasan artifisial generatif berbasis *Large Language Model* (LLM) menimbulkan persoalan hukum terkait penggunaan karya berhak cipta sebagai data pelatihan tanpa izin pencipta. Penelitian ini bertujuan menganalisis karakteristik yuridis penggunaan data berhak cipta dalam pelatihan LLM, mengidentifikasi potensi pelanggaran hak cipta menurut hukum Indonesia, serta mengkaji tantangan pengaturannya. Penelitian menggunakan metode yuridis normatif dengan pendekatan perundang-undangan dan konseptual, berdasarkan Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta, Undang-Undang Informasi dan Transaksi Elektronik, serta berbagai literatur ilmiah. Hasil penelitian menunjukkan bahwa proses *data scraping*, reproduksi, dan pengolahan data berpotensi melanggar hak eksklusif pencipta karena tidak termasuk penggunaan wajar. Ketidadaan pengaturan eksplisit mengenai *text and data mining* menimbulkan kekosongan hukum. Oleh karena itu, diperlukan pengaturan khusus mengenai pengecualian hak cipta, mekanisme lisensi kolektif, dan kompensasi yang adil guna mendukung inovasi AI sekaligus melindungi hak pencipta.

Katakunci: *kecerdasan artifisial generatif; hak cipta; large language model; text and data mining.*

PENDAHULUAN

Perkembangan kecerdasan artifisial (*artificial intelligence/AI*), khususnya teknologi *Large Language Model* (LLM) generatif seperti GPT, Claude, Gemini, dan LLaMA, telah mengubah cara manusia mengakses, mengolah, dan menghasilkan informasi secara signifikan. Model-model tersebut dikembangkan melalui proses pelatihan (*training*) yang memerlukan korpus data dalam jumlah sangat besar yang diperoleh melalui teknik *web crawling* serta *text and data mining* (TDM) dari berbagai sumber digital, seperti situs berita, forum diskusi, repositori kode program, ensiklopedia daring, dan karya sastra digital (Lemley & Casey, 2021). Sebagian besar data tersebut merupakan karya cipta yang memperoleh perlindungan hukum sejak ciptaan diwujudkan berdasarkan prinsip deklaratif sebagaimana diatur dalam Pasal 1 angka 1 juncto Pasal 40 Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta. Kondisi tersebut menimbulkan persoalan hukum karena penggunaan karya berhak cipta sebagai data pelatihan AI pada umumnya dilakukan tanpa persetujuan maupun pemberian kompensasi kepada pencipta atau pemegang hak cipta.

Praktik penghimpunan data melalui *web scraping* untuk pelatihan AI generatif telah memunculkan perdebatan hukum di berbagai negara. Sejumlah sengketa, seperti *Authors Guild v. OpenAI*, *The New York Times Company v. Microsoft Corporation and OpenAI*, serta *Getty Images v. Stability AI*, menunjukkan bahwa penggunaan karya berhak cipta sebagai data pelatihan AI tanpa izin menjadi isu hukum yang semakin kompleks. Walaupun perkara tersebut berada dalam yurisdiksi Amerika Serikat dan Inggris, persoalan serupa juga relevan bagi Indonesia mengingat berbagai karya digital milik pencipta Indonesia berpotensi menjadi bagian dari data pelatihan model AI yang dikembangkan perusahaan teknologi multinasional. Di sisi lain, hukum positif Indonesia belum memiliki pengaturan yang secara khusus mengatur legalitas penggunaan data berhak cipta dalam proses pelatihan AI generatif.

Permasalahan tersebut semakin kompleks karena Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta disusun sebelum berkembangnya teknologi AI generatif sehingga belum mengatur secara eksplisit mengenai *text and data mining* sebagai bentuk penggunaan yang dikecualikan maupun sebagai bentuk pelanggaran hak cipta. Ketentuan mengenai penggunaan wajar dalam Pasal 43 dan Pasal 44 Undang-Undang Hak Cipta masih dibatasi pada kepentingan pendidikan, penelitian nonkomersial, dan kepentingan tertentu lainnya sehingga penerapannya terhadap pelatihan model AI yang bersifat komersial menimbulkan ketidakpastian hukum. Selain itu, Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik beserta perubahannya juga memiliki keterkaitan karena mengatur akses, pengumpulan, dan pemanfaatan informasi elektronik, termasuk larangan akses tanpa hak terhadap sistem elektronik sebagaimana diatur dalam Pasal 30. Ketidadaan pengaturan yang spesifik mengenai penggunaan data untuk pelatihan AI mengakibatkan adanya dilema antara perlindungan hak ekonomi pencipta dan kebutuhan pengembangan inovasi teknologi berbasis AI di Indonesia.

Berbagai penelitian sebelumnya telah membahas hubungan antara hak cipta dan kecerdasan artifisial, namun masih berfokus pada aspek yang berbeda. Sitorus dan Anggraeny (2021) mengkaji perlindungan hak cipta atas karya yang dihasilkan AI dan menyimpulkan bahwa AI belum diakui sebagai subjek hukum pemegang hak cipta dalam sistem hukum Indonesia. Ramli dan Nugroho (2022) menitikberatkan kajian pada pertanggungjawaban hukum penggunaan *big data* dalam pengembangan AI dari perspektif perlindungan data pribadi. Selanjutnya, Wijayanti (2023) membahas doktrin *fair use* dalam hukum hak cipta Amerika Serikat sebagai perbandingan dengan rezim penggunaan wajar di Indonesia, sedangkan Prasetya (2024) menelaah legalitas *web scraping* berdasarkan Undang-Undang Informasi dan Transaksi Elektronik tanpa menghubungkannya dengan rezim hak cipta. Kajian internasional yang dilakukan Sag

(2023) mengenai *text and data mining* dalam perspektif hukum hak cipta komparatif juga memberikan landasan konseptual yang penting, namun masih berorientasi pada sistem hukum Amerika Serikat dan Uni Eropa.

Berdasarkan penelitian-penelitian tersebut, terlihat bahwa kajian mengenai hubungan antara AI dan hak cipta masih bersifat parsial. Sebagian penelitian berfokus pada status kepemilikan karya yang dihasilkan AI, sebagian lainnya membahas perlindungan data pribadi atau legalitas *web scraping*, namun belum terdapat kajian yang secara komprehensif menganalisis penggunaan karya berhak cipta sebagai data pelatihan *Large Language Model* generatif berdasarkan perspektif hukum positif Indonesia. Oleh karena itu, penelitian ini menawarkan kebaruan dengan menempatkan fokus kajian pada tahap *input* pengembangan AI, yaitu proses pengumpulan, reproduksi, dan pengolahan karya berhak cipta untuk pelatihan LLM generatif melalui pendekatan yang mengintegrasikan rezim Undang-Undang Hak Cipta, Undang-Undang Informasi dan Transaksi Elektronik, serta prinsip-prinsip perlindungan kekayaan intelektual. Pendekatan tersebut diharapkan mampu memberikan pemetaan yang lebih komprehensif mengenai karakteristik yuridis pelanggaran hak cipta dalam pengembangan AI generatif sekaligus memberikan arah bagi pembentukan regulasi yang adaptif terhadap perkembangan teknologi.

Berdasarkan uraian tersebut, penelitian ini mengkaji tiga permasalahan utama, yaitu: (1) bagaimana karakteristik penggunaan data berhak cipta dalam pelatihan *Large Language Model* generatif; (2) bagaimana potensi pelanggaran hak cipta dalam penggunaan data tersebut menurut hukum Indonesia; dan (3) bagaimana tantangan pengaturan perlindungan hak cipta terhadap penggunaan data dalam pengembangan AI generatif. Sejalan dengan rumusan masalah tersebut, penelitian ini bertujuan untuk menganalisis karakteristik yuridis penggunaan data berhak cipta dalam pelatihan LLM generatif, mengidentifikasi potensi pelanggaran hak cipta berdasarkan hukum positif Indonesia, serta merumuskan tantangan pengaturan hukum yang dapat menjadi dasar pengembangan kebijakan perlindungan hak cipta yang seimbang dengan kebutuhan inovasi kecerdasan artifisial di Indonesia.

METODE PENELITIAN

Penelitian ini menggunakan metode penelitian hukum yuridis normatif (*legal doctrinal research*), yaitu penelitian yang mengkaji hukum sebagai norma atau kaidah yang berlaku dalam masyarakat dan menjadikan bahan pustaka atau data sekunder sebagai sumber data utama, sebagaimana dikemukakan Soekanto dan Mamudji (2015) bahwa penelitian hukum normatif meneliti hukum dari perspektif internal dengan objek penelitian berupa norma hukum. Pendekatan yang digunakan meliputi pendekatan perundang-undangan (*statute approach*), yaitu dengan menelaah seluruh regulasi yang bertautan dengan isu hukum yang dikaji; pendekatan konseptual (*conceptual approach*), guna memahami doktrin dan konsep hukum hak cipta serta kecerdasan artifisial; dan pendekatan perbandingan hukum (*comparative approach*) secara terbatas untuk memperkaya analisis mengenai pengaturan TDM di yurisdiksi lain seperti Uni Eropa dan Jepang.

Bahan hukum yang digunakan terdiri atas bahan hukum primer, yaitu Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta, Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik sebagaimana diubah dengan Undang-Undang Nomor 19 Tahun 2016 dan Undang-Undang Nomor 1 Tahun 2024, Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi, Surat Edaran Menteri Komunikasi dan Informatika Nomor 9 Tahun 2023 tentang Etika Kecerdasan Artifisial, serta Persetujuan *TRIPs* (*Agreement on Trade-Related Aspects of Intellectual Property Rights*) sebagai

instrumen hukum internasional yang telah diratifikasi Indonesia melalui Undang-Undang Nomor 7 Tahun 1994. Bahan hukum sekunder meliputi buku teks hukum kekayaan intelektual, jurnal ilmiah nasional dan internasional, hasil penelitian terdahulu, serta dokumen kebijakan terkait regulasi AI. Bahan hukum tersier berupa kamus hukum dan ensiklopedia untuk menjelaskan istilah teknis.

Teknik pengumpulan bahan hukum dilakukan melalui studi kepustakaan (*library research*) dengan menelusuri peraturan perundang-undangan, putusan pengadilan yang relevan, dan literatur akademik baik cetak maupun digital. Analisis bahan hukum dilakukan secara kualitatif dengan teknik deskriptif-analitis, yakni menguraikan norma hukum yang berlaku kemudian menganalisisnya secara kritis dikaitkan dengan karakteristik teknis pelatihan model AI generatif, sehingga diperoleh konstruksi hukum yang argumentatif mengenai status penggunaan data berhak cipta dalam konteks tersebut.

HASIL DAN PEMBAHASAN

1. Karakteristik Penggunaan Data Berhak Cipta dalam Pelatihan Large Language Model Generatif

Large Language Model merupakan model pembelajaran mesin berbasis arsitektur transformer yang dilatih menggunakan metode pembelajaran mendalam (*deep learning*) atas korpus data tekstual berskala sangat besar untuk mempelajari pola statistik bahasa, sehingga mampu menghasilkan luaran berupa teks, kode program, maupun modalitas lain yang menyerupai hasil karya manusia (Bommasani et al., 2021). Proses pelatihan tersebut secara garis besar terdiri atas empat tahapan yang memiliki relevansi hukum berbeda-beda, yaitu: (a) tahap pengumpulan data (*data collection/harvesting*), (b) tahap penyimpanan dan penyusunan korpus (*dataset curation*), (c) tahap reproduksi ke dalam representasi numerik melalui proses *tokenisasi* dan *embedding*, dan (d) tahap pelatihan model itu sendiri melalui optimisasi parameter berbasis data tersebut.

Dari perspektif hukum hak cipta, karakteristik yang membedakan penggunaan data untuk pelatihan AI dari bentuk pemanfaatan karya cipta konvensional terletak pada sifatnya yang massal, otomatis, dan non-ekspresif pada tahap pemrosesan. Karya cipta yang dihimpun melalui web crawling dapat mencakup hampir seluruh jenis ciptaan yang dilindungi Pasal 40 UUHC, yaitu karya tulis, program komputer, karya fotografi, karya seni, dan basis data, sehingga rentang objek yang berpotensi terdampak jauh lebih luas dibandingkan pelanggaran hak cipta konvensional yang biasanya hanya melibatkan satu atau beberapa ciptaan tertentu.

Secara teknis, proses pengumpulan data pelatihan LLM dilakukan melalui web crawler atau bot otomatis yang menyalin (*copy*) seluruh konten dari suatu laman web ke dalam server penyimpanan pengembang model, tanpa melalui proses seleksi manual terhadap status hak cipta konten tersebut. Tindakan penyalinan ini, meskipun bersifat sementara atau bertujuan teknis (*transient copying*), secara doktrinal tetap memenuhi unsur reproduksi ciptaan sebagaimana didefinisikan dalam Pasal 1 angka 12 UUHC, yaitu penambahan jumlah suatu ciptaan, baik secara keseluruhan maupun bagian yang substansial, dengan menggunakan bahan yang sama ataupun tidak sama, termasuk mengalihwujudkan secara permanen atau temporer. Frasa "*temporer*" dalam definisi tersebut penting untuk digarisbawahi karena menunjukkan bahwa hukum hak cipta Indonesia pada dasarnya telah mengantisipasi bentuk penggandaan digital yang bersifat sementara, sehingga argumen bahwa penyalinan data untuk pelatihan AI bukan merupakan reproduksi karena sifatnya yang transien tidak dapat dibenarkan secara yuridis.

Karakteristik kedua yang khas adalah sifat transformatif dari pemrosesan data. Berbeda dengan penggandaan konvensional yang menghasilkan salinan identik, data pelatihan LLM mengalami

transformasi menjadi representasi vektor numerik (*embedding*) yang secara kasat mata tidak lagi menyerupai ciptaan asal. Karakteristik inilah yang kerap dijadikan argumen oleh pengembang model AI bahwa penggunaan tersebut bersifat transformatif dan karenanya termasuk penggunaan wajar. Namun demikian, dari sudut pandang hukum hak cipta, transformasi tersebut terjadi pada tahap pemrosesan internal model (di dalam proses pelatihan), sedangkan pelanggaran jika ada telah terjadi lebih dahulu pada tahap reproduksi awal ketika data disalin secara utuh ke dalam server, terlepas dari apa yang terjadi pada tahap selanjutnya. Doktrin ini sejalan dengan pandangan Sag (2023) bahwa *transformative use* dalam konteks *machine learning* perlu dibedakan antara transformasi pada level ekspresi (*expression*) dan transformasi pada level fungsi (*function*), yang keduanya memiliki konsekuensi hukum berbeda.

Karakteristik ketiga adalah ketiadaan hubungan kontraktual antara pengembang model dan pencipta. Pada umumnya, pemanfaatan karya cipta oleh pihak ketiga misalnya penerbitan ulang, adaptasi, atau lisensi komersial didahului oleh perjanjian lisensi yang mengatur kompensasi dan syarat penggunaan. Dalam konteks pelatihan LLM, data dihimpun secara otomatis dari sumber terbuka (*open web*) tanpa negosiasi individual dengan setiap pencipta, sehingga menimbulkan apa yang oleh Lemley dan Casey (2021) disebut sebagai *fair use by scale*, yaitu argumen bahwa skala penggunaan yang demikian besar membuat proses lisensi individual menjadi tidak praktis, meskipun argumen ini tidak serta-merta menghapuskan kewajiban hukum untuk memperoleh izin. Ketidadaan hubungan kontraktual inilah yang menjadi akar persoalan hukum dalam penelitian ini, karena UUHC pada dasarnya menganut prinsip bahwa setiap bentuk pemanfaatan hak ekonomi pencipta oleh pihak lain wajib didasarkan pada izin (Pasal 9 ayat (2) dan (3) UUHC), kecuali ditentukan lain oleh undang-undang.

2. Potensi Pelanggaran Hak Cipta dalam Penggunaan Data untuk Pelatihan LLM Generatif Menurut Hukum Indonesia

Pasal 9 ayat (1) UUHC memberikan pencipta atau pemegang hak cipta hak eksklusif untuk melaksanakan sendiri atau memberikan izin kepada pihak lain untuk melakukan penerbitan, penggandaan, penerjemahan, adaptasi, pendistribusian, pertunjukan, pengumuman, komunikasi, dan penyewaan ciptaan. Ayat (2) dan (3) pasal yang sama menegaskan bahwa setiap orang yang melaksanakan hak ekonomi tersebut tanpa izin pencipta atau pemegang hak cipta merupakan pelanggaran hak cipta, kecuali ditentukan lain dalam undang-undang. Berdasarkan struktur norma tersebut, dapat dikonstruksikan bahwa tindakan penyalinan data berhak cipta ke dalam dataset pelatihan LLM tanpa izin berpotensi memenuhi unsur pelanggaran, sepanjang tidak termasuk dalam pengecualian yang diatur secara limitatif dalam Pasal 43 dan Pasal 44 UUHC.

Pasal 43 UUHC mengecualikan sejumlah perbuatan dari kategori pelanggaran, antara lain pengumuman, pendistribusian, atau komunikasi ciptaan untuk kepentingan pendidikan dan ilmu pengetahuan yang tidak merugikan kepentingan wajar pencipta, serta penggunaan untuk kepentingan pertunjukan atau pementasan tanpa tujuan komersial. Adapun Pasal 44 mengatur pengecualian bagi perbanyakan atau penggandaan untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah, dan kegiatan sejenis, sepanjang sumbernya disebutkan atau dicantumkan secara lengkap. Ketentuan ini pada dasarnya merupakan padanan doktrin *fair dealing* yang bersifat tertutup (*closed list*), berbeda dengan doktrin *fair use* di Amerika Serikat yang bersifat terbuka dan diuji berdasarkan empat faktor: tujuan dan sifat penggunaan, sifat ciptaan yang digunakan, jumlah dan substansialitas bagian yang digunakan, serta dampak penggunaan terhadap pasar potensial ciptaan.

Apabila dianalisis secara cermat, pelatihan LLM generatif oleh entitas komersial yang produknya kemudian dijual atau dimonetisasi melalui skema berlangganan maupun *application programming interface* (API) sulit dikategorikan sebagai kegiatan pendidikan atau penelitian nonkomersial sebagaimana disyaratkan Pasal 44 UUHC. Sifat komersial pengembangan LLM oleh perusahaan teknologi besar menjadikan penerapan pengecualian tersebut problematis, karena maksud dan tujuan norma pengecualian dalam UUHC secara historis ditujukan untuk kepentingan akademik, kritik, dan pendidikan yang tidak berorientasi laba, bukan untuk pengembangan produk komersial berskala industri. Dengan demikian, karakteristik yuridis pelanggaran yang muncul bersifat particular pada dua unsur: pertama, tidak adanya izin (*unauthorized use*) pada tahap reproduksi data; dan kedua, tidak terpenuhinya syarat pengecualian penggunaan wajar karena sifat komersial dan skala industri dari kegiatan pelatihan model.

Selain UUHC, UU ITE turut relevan dalam mengkualifikasi legalitas proses pengumpulan data. Pasal 30 UU ITE melarang setiap orang mengakses komputer dan/atau sistem elektronik milik orang lain dengan cara apa pun tanpa hak atau melawan hukum, termasuk dengan cara melanggar, menerobos, melampaui, atau menjebol sistem pengamanan. Apabila proses *web scraping* dilakukan dengan melampaui ketentuan *Terms of Service* suatu platform atau menerobos mekanisme pembatasan teknis seperti robots.txt atau CAPTCHA, maka tindakan tersebut berpotensi memenuhi unsur akses ilegal menurut Pasal 30 UU ITE, sebagaimana dianalisis Prasetya (2024) dalam konteks yang lebih umum. Namun demikian, perlu dicatat bahwa Pasal 30 UU ITE mengatur legalitas cara memperoleh akses terhadap sistem elektronik, bukan legalitas pemanfaatan atas konten berhak cipta yang diperoleh melalui akses tersebut, sehingga kedua rezim hukum ini UUHC dan UU ITE bersifat komplementer dan bukan substitutif. Suatu tindakan scraping data dapat sekaligus melanggar UU ITE (dari aspek cara memperoleh akses) dan UUHC (dari aspek pemanfaatan hak ekonomi atas konten), atau melanggar salah satunya saja, bergantung pada modus dan objek yang diperoleh.

Karakteristik lain yang perlu diperhatikan adalah persoalan pembuktian dan atribusi. Berbeda dengan pelanggaran hak cipta konvensional yang objeknya dapat diidentifikasi secara langsung melalui perbandingan ciptaan asli dengan ciptaan yang diduga melanggar, pelanggaran dalam konteks pelatihan AI bersifat tersembunyi (*latent*) karena data yang telah diproses menjadi parameter model tidak lagi dapat ditelusuri secara langsung ke sumber aslinya (*model opacity/black box problem*). Kesulitan pembuktian ini menjadi tantangan tersendiri dalam penegakan hukum, sebagaimana akan dibahas lebih lanjut pada subbab 3. Meski demikian, terdapat fenomena yang disebut *regurgitation* atau *memorization*, yaitu kondisi ketika model AI generatif menghasilkan luaran yang secara substansial identik atau sangat mirip dengan data pelatihan aslinya, sebagaimana ditemukan dalam berbagai eksperimen akademik. Fenomena inilah yang menjadi bukti tidak langsung (*circumstantial evidence*) bahwa data berhak cipta memang telah direproduksi dan disimpan dalam parameter model, sehingga memperkuat argumen adanya pelanggaran hak reproduksi meskipun mekanisme internalnya bersifat kompleks.

3. Tantangan Pengaturan Perlindungan Hak Cipta terhadap Penggunaan Data untuk Pengembangan AI Generatif

Tantangan pertama yang dihadapi adalah tidak adanya norma pengecualian yang secara eksplisit mengatur *text and data mining* dalam UUHC. Ketiadaan ini berbeda dengan pengaturan di Uni Eropa yang telah mengadopsi *Directive* (EU) 2019/790 on *Copyright in the Digital Single Market*, khususnya Pasal 3 dan Pasal 4, yang memberikan pengecualian TDM bagi lembaga riset dan warisan budaya (Pasal 3) serta pengecualian TDM yang lebih luas bagi siapa pun sepanjang pemegang hak tidak menyatakan keberatan

melalui mekanisme *opt-out (rights reservation)* yang dinyatakan secara mesin-terbaca (*machine-readable*) sesuai Pasal 4. Demikian pula Jepang, melalui amandemen Pasal 30-4 *Copyright Act*, memberikan pengecualian yang relatif luas bagi penggunaan karya cipta untuk keperluan analisis data termasuk pembelajaran mesin, sepanjang tidak bertujuan untuk menikmati langsung nilai ekspresif ciptaan tersebut (*non-enjoyment purpose*). Kekosongan norma serupa di Indonesia menimbulkan ketidakpastian hukum, baik bagi pencipta yang tidak memiliki mekanisme hukum yang jelas untuk menyatakan keberatan (*opt-out*) atas pemanfaatan karyanya, maupun bagi pengembang AI domestik yang tidak memiliki kepastian mengenai batas legalitas penggunaan data untuk riset dan pengembangan produk.

Tantangan kedua bersifat teknis-yuridis, yaitu kesulitan penegakan hukum akibat sifat lintas batas (*borderless*) dari proses pelatihan model AI. Banyak pengembang LLM besar berkedudukan di luar yurisdiksi Indonesia, sementara data yang digunakan dapat berasal dari, atau memuat karya cipta warga negara Indonesia. Prinsip teritorialitas hukum hak cipta (*territoriality principle*) yang selama ini menjadi dasar penegakan hukum kekayaan intelektual menjadi kurang relevan ketika proses pengumpulan data, pelatihan model, dan pemanfaatan hasilnya terjadi di berbagai yurisdiksi yang berbeda secara simultan. Choi dan Wu (2023) mencatat bahwa fenomena ini menciptakan *choice of law problem* yang kompleks, karena sulit menentukan hukum negara mana yang berlaku terhadap suatu proses pelatihan model yang dilakukan di server yang berlokasi di satu negara, menggunakan data dari pencipta yang berkewarganegaraan negara lain, untuk kemudian dipasarkan secara global melalui internet.

Tantangan ketiga adalah persoalan pembuktian dan transparansi algoritma (*algorithmic opacity*). Sebagaimana disinggung pada subbab sebelumnya, sifat *black box* dari LLM menyebabkan pencipta yang merasa dirugikan mengalami kesulitan signifikan dalam membuktikan bahwa karyanya benar-benar digunakan dalam proses pelatihan suatu model tertentu, karena pengembang model umumnya tidak mempublikasikan secara rinci komposisi dataset pelatihannya dengan alasan kerahasiaan dagang (*trade secret*) atau keunggulan kompetitif. Ketiadaan kewajiban transparansi ini menimbulkan asimetri informasi yang signifikan antara pencipta dan pengembang model, sehingga upaya penegakan hak melalui jalur litigasi menjadi sangat berat dari sisi pembuktian, sebagaimana tercermin dalam berbagai gugatan di Amerika Serikat yang pada tahap awal terkendala oleh kesulitan pembuktian keterkaitan langsung antara karya pengugat dengan output model tergugat.

Tantangan keempat menyangkut penentuan bentuk kompensasi yang adil (*fair remuneration*) apabila kelak dirumuskan norma pengecualian TDM disertai kewajiban kompensasi. Skema lisensi individual antara setiap pencipta dan setiap pengembang model dinilai tidak praktis mengingat skala data yang digunakan dapat mencapai miliaran dokumen. Alternatif yang banyak diusulkan dalam literatur internasional adalah skema lisensi kolektif wajib (*extended collective licensing*) yang dikelola oleh lembaga manajemen kolektif (LMK), sebagaimana telah dikenal dalam praktik hak cipta musik di Indonesia melalui Lembaga Manajemen Kolektif Nasional (LMKN). Skema serupa berpotensi diadaptasi untuk konteks data pelatihan AI, namun memerlukan penyesuaian kelembagaan yang signifikan mengingat LMKN saat ini hanya berwenang mengelola royalti di bidang lagu dan musik sebagaimana diatur Peraturan Pemerintah Nomor 56 Tahun 2021 tentang Pengelolaan Royalti Hak Cipta Lagu dan/atau Musik, sehingga diperlukan perluasan kewenangan kelembagaan atau pembentukan LMK baru khusus data digital untuk kepentingan AI.

Tantangan kelima bersifat kebijakan, yaitu ketegangan antara kepentingan melindungi hak pencipta dan kepentingan mendorong inovasi teknologi nasional. Surat Edaran Menteri Komunikasi dan Informatika Nomor 9 Tahun 2023 tentang Etika Kecerdasan Artifisial baru mengatur prinsip etis pengembangan AI

secara umum, seperti transparansi, akuntabilitas, dan keadilan, namun belum memberikan panduan teknis mengenai legalitas penggunaan data berhak cipta untuk pelatihan model. Kebijakan yang terlalu restriktif berpotensi menghambat pengembangan model AI berbahasa Indonesia yang justru dibutuhkan untuk kedaulatan digital dan pelestarian bahasa serta budaya nasional dalam ekosistem AI global, sementara kebijakan yang terlalu longgar berpotensi mengorbankan hak-hak ekonomi pencipta lokal. Diperlukan pendekatan regulasi yang proporsional (*proportionality approach*), misalnya dengan membedakan perlakuan antara penggunaan data untuk riset nonkomersial oleh lembaga pendidikan dan penggunaan data untuk pengembangan produk komersial oleh korporasi, sebagaimana pola yang mulai diterapkan Uni Eropa melalui AI Act (*Regulation* (EU) 2024/1689) yang mewajibkan penyedia model AI generatif tujuan umum (*general-purpose AI*) untuk menyusun ringkasan memadai mengenai konten yang digunakan untuk pelatihan model (Pasal 53 AI Act), sebagai bentuk kompromi antara transparansi dan kerahasiaan dagang.

Berdasarkan uraian tersebut, dapat disimpulkan bahwa tantangan pengaturan hak cipta terhadap penggunaan data untuk pengembangan AI generatif di Indonesia bersifat multidimensional, mencakup dimensi normatif (kekosongan norma pengecualian TDM), dimensi yurisdiksional (sifat lintas batas), dimensi pembuktian (*opacity algoritma*), dimensi kelembagaan (ketiadaan mekanisme lisensi kolektif yang relevan), dan dimensi kebijakan (ketegangan antara perlindungan hak cipta dan pendorong inovasi). Kelima dimensi tantangan ini perlu direspons secara simultan melalui reformasi legislasi yang komprehensif, bukan sekadar penafsiran ekstensif terhadap norma UUHC yang ada saat ini.

KESIMPULAN

Berdasarkan hasil dan pembahasan di atas, dapat ditarik beberapa kesimpulan. Pertama, karakteristik penggunaan data berhak cipta dalam pelatihan *Large Language Model* generatif bersifat massal, otomatis, non-kontraktual, dan melibatkan proses reproduksi awal yang bersifat utuh sebelum ditransformasikan menjadi representasi numerik pada tahap pelatihan model, sehingga tidak dapat serta-merta dikategorikan sebagai penggunaan transformatif (*transformative use*) yang menghapuskan kewajiban memperoleh izin pencipta pada tahap reproduksi. Kedua, penggunaan data berhak cipta tersebut berpotensi memenuhi unsur pelanggaran hak cipta menurut Pasal 9 UUHC karena dilakukan tanpa izin pencipta dan tidak memenuhi syarat pengecualian penggunaan wajar (*fair use/fair dealing*) yang diatur secara limitatif dalam Pasal 43 dan 44 UUHC, mengingat sifat komersial dan skala industri dari kegiatan pengembangan LLM oleh sebagian besar pelaku usaha teknologi; sementara itu, proses pengumpulan datanya sendiri berpotensi pula melanggar Pasal 30 UU ITE apabila dilakukan dengan melampaui sistem pengamanan elektronik, sehingga kedua rezim hukum tersebut bersifat komplementer dalam mengkualifikasi legalitas suatu tindakan *scraping* data. Ketiga, tantangan pengaturan perlindungan hak cipta terhadap penggunaan data untuk pengembangan AI generatif bersifat multidimensional, meliputi kekosongan norma pengecualian *text and data mining*, kesulitan penegakan hukum akibat sifat lintas batas (*borderless*) proses pelatihan model, persoalan pembuktian akibat *opacity algoritma* (*algorithmic opacity*), ketiadaan kelembagaan lisensi kolektif yang relevan, serta ketegangan kebijakan antara perlindungan hak pencipta dan dorongan inovasi teknologi nasional.

Berdasarkan simpulan tersebut, penelitian ini merekomendasikan agar pembentuk undang-undang mempertimbangkan revisi terbatas (*limited amendment*) terhadap UUHC untuk merumuskan norma pengecualian *text and data mining* yang proporsional, disertai mekanisme *rights reservation* yang *machine-readable* bagi pencipta yang berkeberatan atas pemanfaatan karyanya, pembentukan atau perluasan

kewenangan lembaga manajemen kolektif untuk mengelola kompensasi atas penggunaan data digital bagi kepentingan pelatihan AI, serta kewajiban transparansi terbatas bagi penyedia model AI generatif skala besar mengenai kategori sumber data pelatihan yang digunakan, sebagai jalan tengah antara kepentingan melindungi hak ekonomi pencipta dan kepentingan mendorong pengembangan ekosistem kecerdasan artifisial nasional yang berkelanjutan dan berkeadilan. Penelitian lanjutan disarankan untuk mengkaji secara empiris implementasi mekanisme lisensi kolektif tersebut serta membandingkan secara lebih mendalam dengan perkembangan regulasi *AI Act* Uni Eropa yang telah mulai diterapkan secara bertahap.

DAFTAR PUSTAKA

- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Chang, D., Chen, C., Feizi, S., Li, F.-F., & Liang, P. (2021). *On the opportunities and risks of foundation models*. Stanford Institute for Human-Centered Artificial Intelligence (HAI). <https://arxiv.org/abs/2108.07258>
- Choi, J., & Wu, T. (2023). Cross-border copyright challenges in the age of generative artificial intelligence. *Harvard Journal of Law & Technology*, 36(2), 411–448.
- Copyright Act of Japan, Act No. 48 of May 6, 1970, Article 30-4 (As amended up to Act No. 30 of May 25, 2018).
- Damian, E. (2019). *Hukum hak cipta* (Edisi Kedua). Alumni.
- Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. *Official Journal of the European Union*, L 130/92.
- Kementerian Komunikasi dan Informatika Republik Indonesia. (2023). *Surat Edaran Menteri Komunikasi dan Informatika Nomor 9 Tahun 2023 tentang Etika Kecerdasan Artifisial*.
- Lemley, M. A., & Casey, B. (2021). Fair learning. *Texas Law Review*, 99(4), 743–785.
- Peraturan Pemerintah Nomor 56 Tahun 2021 tentang Pengelolaan Royalti Hak Cipta Lagu dan/atau Musik. Lembaran Negara Republik Indonesia Tahun 2021 Nomor 136.
- Prasetya, A. B. (2024). Legalitas web scraping dalam perspektif Undang-Undang Informasi dan Transaksi Elektronik. *Jurnal Hukum dan Teknologi Informasi*, 8(1), 55–71.
- Ramli, T. S., & Nugroho, F. A. (2022). Tanggung jawab hukum pemanfaatan big data dalam pengembangan kecerdasan artifisial di sektor bisnis. *Jurnal Hukum Bisnis*, 41(3), 210–228.
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *Official Journal of the European Union*, L 2024/1689.
- Sag, M. (2023). Copyright safety for generative AI. *Houston Law Review*, 61(2), 295–347.
- Saidin, O. K. (2015). *Aspek hukum hak kekayaan intelektual (intellectual property rights)*. Rajawali Pers.
- Sitorus, R. H., & Anggraeny, I. (2021). Perlindungan hak cipta atas karya yang dihasilkan oleh kecerdasan artifisial di Indonesia. *Jurnal Hukum & Pembangunan*, 51(4), 902–919. <https://doi.org/10.21143/jhp.vol51.no4.3292>
- Soekanto, S., & Mamudji, S. (2015). *Penelitian hukum normatif: Suatu tinjauan singkat*. Rajawali Pers.
- Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik. Lembaran Negara Republik Indonesia Tahun 2008 Nomor 58. sebagaimana telah beberapa kali diubah terakhir dengan Undang-Undang Nomor 1 Tahun 2024 tentang Perubahan Kedua atas Undang-Undang

- Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik. Lembaran Negara Republik Indonesia Tahun 2024 Nomor 1.
- Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi. Lembaran Negara Republik Indonesia Tahun 2022 Nomor 218.
- Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta. Lembaran Negara Republik Indonesia Tahun 2014 Nomor 266.
- Undang-Undang Nomor 7 Tahun 1994 tentang Pengesahan *Agreement Establishing the World Trade Organization* (Persetujuan Pembentukan Organisasi Perdagangan Dunia) [Termasuk di dalamnya *Agreement on Trade-Related Aspects of Intellectual Property Rights* (TRIPs)]. Lembaran Negara Republik Indonesia Tahun 1994 Nomor 57.
- Wijayanti, C. D. (2023). Doktrin fair use dalam hukum hak cipta Amerika Serikat dan relevansinya bagi riset kecerdasan artifisial di Indonesia. *Jurnal Hukum Ius Quia Iustum*, 30(1), 88–110. <https://doi.org/10.20885/iustum.vol30.iss1.art5>