



Tengr.ai modellösszefoglaló

1. Hatály és cél

Ez a dokumentum („**Összefoglaló**”) az alábbi kibocsátó által kerül kiadásra:

Vállalat neve: Tengrai Artificial Intelligence Korlátolt Felelősségű Társaság

Székhely: Magyarország, 6724 Szeged, Rókusi körút 21., I. emelet, 4. ajtó

Cégjegyzékszám: 06-09-028918

Adószám: 32315028-2-06

(„**Szolgáltató**”)

a nyilvánosság és az illetékes felügyeleti hatóságok számára az (EU) 2024/1689 számú, a mesterséges intelligenciáról szóló rendelet („**MI-rendelet**”) és bármely más alkalmazandó jogszabály értelmében, amely a képgenerálásra alkalmas általános célú mesterségesintelligencia-rendszerek piacra helyezését vagy üzembe helyezését szabályozza. Célja, hogy jogilag értelmezhető módon rögzítse (i) a Szolgáltató általános célú képgeneráló modelljének („**Rendszer**”) műszaki jellemzőit, (ii) a képzés során alkalmazott adatirányítási intézkedéseket, valamint (iii) az előzetes és utólagos kockázatkezelési és orvoslási mechanizmusokat, amelyek biztosítják a Rendszer folyamatos megfelelését az Európai Unió és a tagállamok jogának.

2. Fogalommeghatározások

A félreértések elkerülése érdekében az alábbi nagybetűs kifejezések az alábbi jelentéssel bírnak. Minden egyéb kifejezés az MI-rendeletben meghatározott jelentéssel bír, hacsak a kontextus másként nem kívánja.

- **Hyperalign architektúra** – a Rendszerbe ágyazott rejtett összehangoló, korlátozásérvényesítő és szabályellenőrző alrendszer, amelyet az 5. szakasz ismertet.
- **Szabálytár** – a Hyperalign által konzultált, jogilag, etikailag és szerződésileg kötelező erejű szabályok géppel olvasható tára.
- **Nyilvános forrású adathalmaz** – bármely képszöveg-adatbázis, amelyet olyan licenc alatt tettek közzé, amely lehetővé teszi a kereskedelmi felhasználást, módosítást és jogdíjmentes terjesztést, beleértve, de nem kizárólagosan a 4. szakaszban felsoroltakat.

- **Tiltott anyag** – minden olyan vizuális tartalom, amelynek előállítása, birtoklása vagy terjesztése sértené az uniós vagy tagállami jogot, illetve a Szolgáltató saját elfogadható használatra vonatkozó szabályzatait.

3. A rendszer magas szintű leírása

3.1 Általános működés

A Rendszer diffúziós és transzformer-alapú szöveg–kép és kép–kép modell, amely tetszőleges természetesnyelvi utasításra megfelelő raszterképet állít elő. A Rendszer közvetlenül a <https://tengr.ai> webes alkalmazáson keresztül vagy harmadik feleken keresztül API-n érhető el.

3.2 Funkcionális egységek

1. **Utasítás fogadása (S1)** – az utasításokat különbségtétel nélkül fogadja; előszűrés nem történik.
2. **Elsődleges generálás (S2)** – egy alap diffúziós modell ideiglenes megoldást készít (O_{raw}).
3. **Rejtett korlátozásérvényesítés (S3)** – a Hyperalign összeveti O_{raw} -t a Szabálytárral, és szükség esetén törli, kicseréli vagy újragenerálja a nem megfelelő szegmenseket.
4. **Kézbesítés (S4)** – az átalakított kép (O_{final}) továbbításra kerül a felhasználónak; elutasító szöveg, bocsánatkérés vagy cenzúráról szóló értesítés soha nem jelenik meg.

4. A képzési adatok eredete és licencelése

4.1 Forrásadatbázisok

A Szolgáltató tanúsítja, hogy az előképzéshez és finomhangoláshoz felhasznált képkorpusz *kizárólag* nyilvános forrású és licencelt magán forrású adatkészletekből áll (együttesen a „**Képzési készlet**”).

A Szolgáltató kijelenti, hogy a Képzési készlet kizárólag olyan képszöveg-korpuszokból áll, amelyeket **jogszerűen és nyíltan terjesztenek** engedékeny szerzői jogi licencrek szerint, nevezetesen Creative Commons Nevezd meg licencek (CC BY 4.0, CC BY 2.0), közkincs-lemondások (CC0, Public-Domain Mark) vagy ezekkel egyenértékű, visszavonhatatlan egyedi licencek, amelyek világszintű, jogdíjmentes használati, módosítási és kereskedelmi terjesztési jogot biztosítanak. Példálózó források többek között: az MS COCO referencia-benchmark (CC BY 4.0); a Google Open Images gyűjteménye (többségében CC BY 2.0); a Conceptual Captions és a Conceptual 12M webes alternatív szövegek-korpuszok (képek CC BY 2.0); a 99 millió elemes YFCC-100M multimédia-archívum, amelyet a Creative Commons különböző változatai alatt

bocsátottak ki; valamint a Wikimedia Commons, a Smithsonian Open Access és más közkincs-adatbázisok, amelyek tartalma CC-licencű vagy szabad újrafelhasználási korlátozásoktól mentes.

Az adatbázis betöltését megelőzően a Szolgáltató automatizált licenc-címke ellenőrzést végzett és minden olyan rekordot eltávolított, amely **nem kereskedelmi, változtatás nélküli** vagy egyéb módon összeegyeztethetetlen feltételt tartalmazott, ezáltal biztosítva, hogy minden megtartott kép megfeleljen az **MI-rendelet 10. cikk (3) bekezdésében** előírt, „releváns, reprezentatív, hibamentes és teljes” adatra vonatkozó követelménynek, amely **az Unió szellemi tulajdonjogokra vonatkozó jogszabályainak megfelelően** került gyűjtésre. A letöltési jegyzékek, kriptográfiai ellenőrző összegek és licencadatok teljességében megőrzésre kerülnek az Adatkészlet-nyilvántartásban, és indokolt kérelem esetén hozzáférhetővé tehetők az illetékes hatóságok számára.

4.2 A tulajdonosi tartalom kizárása

Nem került be ingestálásra olyan kép, amely zárt közösségi médiaplatformokról, magánkézben lévő tárolókból vagy egyéb terhelte forrásokból származik. Nem alkalmaztak olyan scraping eszközöket, amelyek technikai védelmi intézkedéseket megkerülnék.

5. Adatátvilágítás, tisztítás és kockázatcsökkentés

Mivel a Képzési készlet több mint hét (7) milliárd jelölt képet tartalmazott, az egyenkénti manuális felülvizsgálat – egy (1) másodpercenként egy képpel számolva – több mint 300 évet igényelt volna; az ilyen ellenőrzés tehát nyilvánvalóan kivitelezhetetlen volt. A Szolgáltató ezért az alábbi, több rétegű automatizált eljárást alkalmazta:

1. **Metaadat előszűrés** – a szöveges metaadatokat (feliratok, alternatív szövegek, URL-ek) obszcenításra, szélsőséges ideológiára vagy kiskorúak szexuális kizsákmányolására utaló lexikai jelek szempontjából vizsgálták; a jelzett bejegyzéseket forrásukon kiiktatták.
2. **NSFW vizuális osztályozó futtatása** – egy ResNet-alapú osztályozó, amelyet a „Journey NSFW” és „Open NSFW” benchmarkokon finomhangoltak, automatikusan megjelölt minden olyan képet, amelynek a szexuális tartalomra vagy grafikus erőszakra vonatkozó valószínűségi pontszáma meghaladta a 0,25 értéket.
3. **Érzékleti hasonlóság szerinti szűrés** – a CLIP-beágyazásokat k-legközelebbi szomszéd algoritmussal hasonlították össze egy tiltott példákat tartalmazó *blokklistával* (pl. ismert szélsőséges jelképek, CSA-anyag); a találatokat törölték.

4. **Szemantikai hasonlóság szerinti szűrés** – a feliratokat SBERT segítségével ágyazták be; a kifogásolt szemantikákat tartalmazó szerkesztett vektorbanka elleni koszinusz hasonlóság indokolta a kizárást.
5. **Deduplikálás és minőség-ellenőrzés** – az majdnem azonos ($\cos < 0,999$) és a sérült vagy üres fájlokat eltávolították.
6. **Iteratív minőségbiztosítási csapat finomítás** – időszakos auditok feltárták a megmaradó szabálysértéseket; a megfelelő heurisztikákat vagy modellparamétereket újratanították és a folyamatot újra lefuttatták.

A fenti intézkedések eredményeként több millió képet – a teljes jelölt képmennyiség mintegy 3,8 %-át – távolítottak el, beleértve a LAION saját címkéi alapján magas kockázatúnak jelzett minden bejegyzést.

6. A Hyperalign korlátozásérvényesítő mechanizmusa

6.1 Jogi és technikai alapok

A Hyperalign egy igazító modul (20. komponens), amely valós időben felügyeli a tartalomgenerálást a Szabálytárra hivatkozva. Az érvényesítés tokenenként történik a diffúziós latensek esetében, és szükség esetén utólagos inpainting útján. Az architektúra ennek megfelelően:

- **kivétel nélkül minden utasítást elfogad,**
- **elnyomja az észlelhető elutasító jelzéseket,** és
- **kizárólag megfelelő kimenetet ad,** ezáltal a gradiensmentes jailbreak kísérletek sikerének valószínűségét az igazító osztályozó hibavalószínűségére csökkenti.

6.2 Bővíthetőség és emberi felügyelet

A Szabálytár YAML formátumban van karbantartva, verziókövetett, és a megfelelőségi személyzet úgy módosíthatja, hogy az alapmodell újratanítására nincs szükség. Ennek megfelelően az MI-rendelet módosításai, a jövőben megjelenő felhatalmazáson alapuló jogi aktusok vagy tagállami specifikus tilalmak legfeljebb tíz (10) munkanapon belül beépíthetők, ezt követően az átdolgozott szabályok minden további generálásra kötelezőek.

7. Folyamatos irányítás, monitorozás és elszámoltathatóság

1. **Naplózás és nyomonkövethetőség** – egy manipulációmentes auditnapló rögzíti az egyes utasításokat, az O_{final} SHA 256 hashét és minden módosítás során hivatkozott pontos szabályzati rendelkezést.

2. **Időszakos minőségbiztosítási gyakorlatok** – a Szolgáltató legalább félévente ellenséges tesztek végez az újonnan megjelenő jailbreak vektorokkal szemben (pl. többnyelvű elrejtés).
3. **Incidenskezelés** – a feltételezett szabályérvényesítési hibák az érintett végpont azonnali felfüggesztését és a kiváltó ok 48 órán belüli elemzését váltják ki.
4. **Felhasználói visszajelzési csatorna** – bármely látogató bejelentheti a weboldalunkon (a Tengr.ai láblécében található linken) az állítólagos jogellenes kimenetet; a megalapozott kifogások újratanítást vagy a szabályfájl módosítását teszik kötelezővé.

8. Megfelelés az MI-rendeletnek és kapcsolódó eszközöknek

- **4. cikk (Kockázatkezelés)** – a Hyperalign előzetes és utólagos kontrollokat biztosít, amelyek révén a maradék kockázat elfogadható a Rendszer tervezett generatív céljaihoz képest.
- **10. cikk (Adat és adatirányítás)** – a Szolgáltató a 4. szakaszban összefoglaltak szerint teljes körű dokumentációt vezet az adathalmaz származásáról, licencfeltételeiről és tisztítási módszertanáról.
- **11. cikk (Műszaki dokumentáció)** – ez az Összefoglaló a részletes belső tervezési dossziékkal együtt teljesíti a szabályozó hatóságok számára szükséges műszaki dokumentáció szolgáltatásának kötelezettségét.
- **15. cikk (Pontosság, robusztusság és kiberbiztonság)** – empirikus tesztek lényegesen alacsonyabb jailbreak sikerarányokat mutatnak a látható elutasításon alapuló architektúrákhoz képest. A Hyperalign beágyazottan kezeli az OWASP LLM02 által azonosított prompt injection és nem biztonságos kimenetkezelési vektorokat.
- **20. cikk (Piacfelügyelet utáni monitoring)** – a 7. szakaszban ismertetett irányítási intézkedések alkotják a Szolgáltató piacfelügyelet utáni monitoringtervét.
- **A biztonságos MI rendszerfejlesztés iránymutatásai (UK NCSC / CISA)** – a Rendszer csendes, szabálytudatos kimenetszűrése összhangban áll az ott támogatott nem interaktív védekezési mintával.

9. Korlátok és felelősségkizárás

A jelen dokumentumban foglaltak semmiképpen sem értelmezhetők abszolút cenzúrára vagy hibamentes működésre vonatkozó garanciaként. Bár a Hyperalign jelentősen mérsékli a jogellenes kimenetet, a maradék kockázat nem zárható ki. A Szolgáltató azonban minden tőle telhetőt megtesz a bejelentett hiányosságok haladéktalan orvoslása érdekében.

10. A mesterséges intelligencia megfelelőségi csapat elérhetőségei:

Honlap: <https://tengrai.ai> (a láblécben található linken keresztül közvetlen űrlap nyílik)

E-mail cím: qa@tengrai.com

Postacím: Magyarország, 6724 Szeged, Rókusi körút 21., I. emelet, 4. ajtó

