Kubernetes alapú referencia architektúra gyártási MI ügynökök skálázható telepítéséhez: Hierarchikus keretrendszer peremhálózati-adatközponti együttműködéssel a Model Context Protocol alkalmazásával

A Kubernetes-Native Reference Architecture for Sovereign Edge AI in Manufacturing: Hierarchical Agents with Continuous Human-in-the-Loop Learning via the Model Context Protocol

ing. TAMÁS-PÉTER József[1]

[1] Debreceni Egyetem Műszaki Kar, Debrecen, Ótemető u. 2-4, 4028 Magyarország, +4 0770124697, tamas.peter.jozsef@eng.unideb.hu, https://eng.unideb.hu/

Abstract

Our earlier work achieved a 32.4% throughput improvement using Model Context Protocol (MCP) based digital twins with edge artificial intelligence (AI) in ready-mix concrete scheduling. Building on that, we present a Kubernetes-native reference architecture addressing three limitations: (1) lack of experiment reproducibility, (2) absence of hierarchical edge-to-datacenter reasoning, and (3) no continuous learning from operator feedback. The framework deploys as pre-configured templates. A lightweight sovereign edge model (Gemma3:4B) handles real-time scheduling, escalating only complex states via MCP to a datacenter-tier expert (70B+). A human-in-the-loop pipeline lets the edge agent adapt through Retrieval-Augmented Generation (RAG) without retraining. On heterogeneous hardware, the architecture shows robust performance gains, cross-model escalation consistency, and measurable behavioral adaptation, aligned with Industry 5.0 and sovereign AI.

Keywords: Model Context Protocol, digital twin, edge AI, Kubernetes, human-in-the-loop

Kivonat

Korábbi munkánkban 32.4%-os teljesítményjavulást értünk el Model Context Protocol (MCP) alapú digitális ikrekkel és peremhálózati mesterséges intelligenciával (MI) transzportbeton üzem ütemezésében. Erre építve Kubernetes natív referencia architektúrát mutatunk be, amely három korlátot kezel: (1) a kísérletek reprodukálhatóságának, (2) a hierarchikus peremhálózati és adatközponti következtetésnek, és (3) az operátori visszajelzésből való tanulásnak a hiányát. A keretrendszer előre konfigurált sablonokként telepíthető. Könnyű, helyben futó modell (Gemma3:4B) végzi a valós idejű ütemezést, és csak a komplex helyzeteket továbbítja MCP-n keresztül adatközponti szakértőhöz (70B+). Ember a hurokban visszajelzés és visszakereséssel bővített szöveggenerálás (RAG) teszi lehetővé az adaptációt újratanítás nélkül. Heterogén hardveren a rendszer robusztus teljesítményjavulást, modellek közötti eszkalációs konzisztenciát és mérhető adaptációt mutat, az Ipar 5.0 és a szuverén MI elveivel összhangban.

Kulcsszavak: Model Context Protocol, digitális iker, peremhálózati MI, Kubernetes, ember a hurokban

1. BEVEZETÉS

Az MI-ügynökök gyártási alkalmazása 2023 óta felgyorsult, a többügynökös rendszerek ígéretes utat mutatnak a termelési környezetek operatív fejlesztésére[10]. Az erőforrás korlátozott KKV-k számára azonban a kísérleti fázisból egy megbízhatóan működő és folyamatosan fejlődő rendszerhez való eljutás továbbra is komoly kihívás. A korlátozott informatikai infrastruktúra[11] és a tipikus integrációs probléma, hogy minden szimulációs platformhoz saját adaptereket kell fejleszteni[7], együttesen magas bevezetési költségeket okoz.

Az alapvető nehézség nyilvánvaló: a szuverén MI helyi telepítést követel[8, 13], de a peremhálózati hardver[4] nem képes kellően nagy modelleket futtatni az összetett következtetésekhez, miközben a 70B+ modellek olyan adatközponti erőforrásokat igényelnek, amelyek a peremhálózaton ritkán állnak rendelkezésre. Ráadásul a reprodukálhatósági problémák[12] miatt a kisebb szervezetek nehezen tudják ellenőrizni a publikált eredményeket.

Korábbi munkánk[14] bemutatta, hogy a Model Context Protocol (MCP)[2] hatékonyan képes szabványosítani a gyártási MI ügynökök interfészeit digitális ikrekkel, 32.4%-os teljesítményjavulást elérve egy 4B paraméteres peremhálózati modellel transzportbeton ütemezésben. Az a prototípus azonban három fontos korláttal rendelkezett: (1) monolitikus telepítés, infrastruktúra reprodukálhatóság nélkül, (2) egyszerű, sík ügynöktopológia eszkalációs mechanizmus nélkül, valamint (3) teljesen statikus viselkedés az operátori visszajelzésből való tanulás képessége nélkül.

Bár a digitális ikrek egyre inkább előíró szerepet töltenek be[1][15], a legtöbb továbbra is felhőfüggő és monolitikus. Az MCP[2] szabványosítási megközelítésként növekvő ipari érdeklődést kelt[5][6], de a valós gyártási alkalmazások száma még mindig alacsony. Kifejezetten a betonütemezésben a meglévő kollaboratív[17] és vegyes egészértékű lineáris programozási (MILP)[16] módszerek MI-ügynökök nélkül kezelik az üzemeltetési komplexitást. A hibrid ügynökalapú MI-vel[3] és az európai szuverén MI-kezdeményezésekkel[9] foglalkozó kutatások kiemelik a helyi hierarchikus rendszerek értékét, de egyetlen publikált keretrendszer sem oldja meg egyszerre mindhármat, a reprodukálhatóságot, a hierarchikus következtetést és a folyamatos tanulást, szuverén, peremhálózaton telepíthető architektúrában.

Ezt a hiányt három fő hozzájárulással kezeljük. Először teljes infrastruktúra mint kód (IaC) megoldást biztosítunk: a teljes rendszert konténer vezénylő platformon (Kubernetes) telepítjük, a helyi üzemi környezetekhez jobban illeszkedő k3s változattal, előre elkészített telepítési sablonokkal (Helm). Ez kézi konfiguráció nélküli telepítést és több véletlenszám maggal végzett, statisztikailag megalapozott kísérleteket tesz lehetővé heterogén, KKV szintű hardveren is. Másodszor hierarchikus szuverén ügynöktopológiát vezetünk be: egy könnyű peremhálózati modell (Gemma3:4B) végzi a valós idejű ütemezést, és csak a komplex állapotokat továbbítja MCP-n keresztül egy adatközponti szakértő MI-hez (ezek 70 milliárd paraméter felettiek). Az eszkalációs konzisztenciát sűrű transzformer és Mixture-of-Experts (MoE) modellek között validáljuk. Harmadszor egy gyakorlatias, visszakereséssel bővített szöveggenerálás (RAG) alapú ember a hurokban visszajelzési mechanizmust valósítunk meg. Ez lehetővé teszi, hogy a peremhálózati ügynök az operátori korrekciókat beépítse jövőbeli döntéseibe a modell újratanítása nélkül.

Az eredmény egy KKV szintű hardveren megbízhatóan futó, reprodukálhatóan működő és az operátoroktól folyamatosan tanuló gyártási MI-rendszer, amely minden valós idejű ütemezést helyben tart.

2. ARCHITEKTÚRA ÉS MEGVALÓSÍTÁS

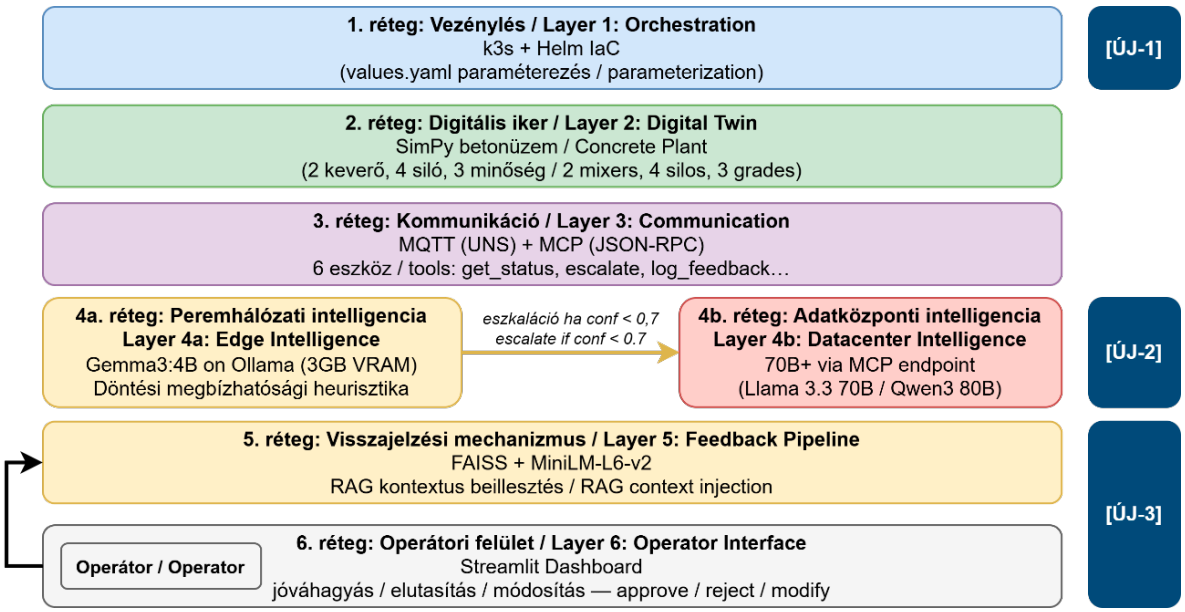
A javasolt architektúra a korábbi v1 digitális iker [14] hat rétegre szervezett változata (1. ábra), három új elemmel bővítve: **[ÚJ-1]** reprodukálható telepítés (1. réteg), **[ÚJ-2]** hierarchikus peremhálózati és adatközponti döntéshozatal (4. réteg), valamint **[ÚJ-3]** operátori visszajelzésből való tanulás (5-6. réteg).

1-3. réteg: Vezénylés, digitális iker és kommunikáció [ÚJ-1]. Az 1. réteg a k3s könnyűsúlyú változatát (tanúsított Kubernetes disztribúció) használja, Helm chartokkal az IaC (infrastruktúra mint kód) megvalósításához. A kísérleti konfigurációk teljes egészében a values.yaml fájlban vannak paraméterezve, így a több véletlenszám maggal végzett futtatások kódmódosítás nélkül reprodukálhatók. A 2. réteg egy SimPy (Python könyvtár diszkrét esemény szimulációhoz) alapú szimuláció, amely transzportbeton üzemet modellez: két keverő (8 m³ és 24 m³), négy siló, három betonminőség (C25, C30, C40) és változó teherautó érkezések. A 3. réteg kettős protokollú kommunikációt biztosít: MQTT-t (könnyűsúlyú ipari üzenetküldő protokoll) a valós idejű eseménystreaminghez az Egységesített Névtéren (UNS, Unified Namespace) [13] keresztül, valamint MCP-t [2] (JSON-RPC alapú távoli eljáráshívás) a szabványosított eszközhíváshoz. Az MCP szervert a korábbi munkánkhoz képest három új eszközzel bővítettük: escalate_to_central, log_feedback és get_feedback_examples.

4. réteg: Hierarchikus intelligencia [ÚJ-2]. A 4a réteg a Gemma3:4B modellt futtatja Ollama (nyílt forráskódú modellkiszolgáló szoftver) környezetben a valós idejű ütemezéshez. Egy egyszerű determinisztikus eszkalációs szabályt adtunk hozzá. A döntés automatikusan az adatközponti szakértőhöz kerül, ha a peremhálózati modell döntési megbízhatósága 0.7 alá esik. Ilyen helyzet áll elő például, ha tíznél több teherautó várakozik a sorban, kettőnél több betonminőség vár feldolgozásra, vagy a keverő kihasználtsága 50% alá csökken. A 4b réteg egy MCP kompatibilis, 70B+ paraméteres végpont, amelynek pontos

elérhetőségét a Helm konfiguráció határozza meg. Mivel az MCP egységes JSON-RPC interfész mögé rejti a modell identitását, az architektúra független attól, hogy a szakértő helyben, szerverparkban vagy távolról fut.

5-6. réteg: Visszajelzés és operátori felület [ÚJ-3]. Az operátori döntéseket (jóváhagyás, elutasítás vagy módosítás megjegyzéssel) a könnyűsúlyú all-MiniLM-L6-v2 (szövegbeágyazó modell) mondatbeágyazásokká alakítja, amelyeket egy FAISS (Facebook AI Similarity Search, hatékony vektor hasonlóságkereső könyvtár) vektortárban indexelünk. Minden új ütemezési döntésnél a peremhálózati ügynök lekérdezi a szemantikailag leghasonlóbb korábbi visszajelzéseket, és RAG kontextusként beilleszti azokat az LLM promptba. Ez viselkedési adaptációt tesz lehetővé újratanítás nélkül. Ha nem áll rendelkezésre releváns visszajelzés, az ügynök az alapértelmezett promptjára támaszkodik. Egy Streamlit (Python alapú interaktív műszerfal keretrendszer) műszerfal valós idejű üzemvizualizációt és döntés felülvizsgálati lehetőséget biztosít az operátorok számára.



1. ábra. *Hatrétegű referencia architektúra a három fő hozzájárulás feltüntetésével*

Validáció KKV szintű infrastruktúrán. Az architektúrát minimális hardveren validáltuk a gyakorlati megvalósíthatóság igazolására: egy vezénylő és szimulációs csomópont (ThinkPad P53, i7-9850H CPU, 64GB RAM) és egy peremhálózati MI csomópont (ThinkPad P16, RTX A2000 GPU, 8GB VRAM). A Helm alapú kialakítás a horizontális skálázást további GPU csomópontokra egyszerűvé teszi, így elég a konfigurációs értékeket frissíteni.

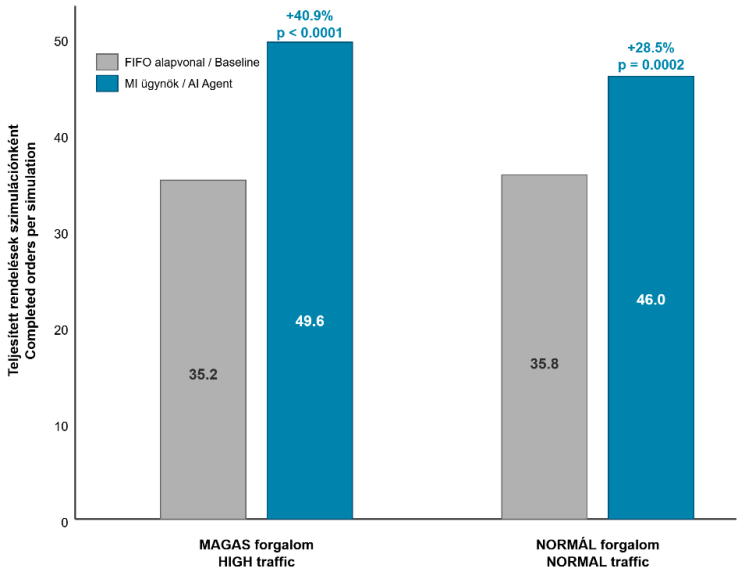
3. KIÉRTÉKELÉS

Az egyes hozzájárulások elkülönített vizsgálatára három célzott kísérletet terveztünk, amelyek közös szimulációs környezeten futnak, de más-más architekturális változót izolálnak. Az első a teljesítmény robusztusságát méri több véletlenszám maggal, a második az eszkalációs döntések modellfüggetlenségét vizsgálja, a harmadik pedig a visszajelzési mechanizmus viselkedésmódosító hatását értékeli.

3.1 Reprodukálható, több véletlenszám maggal végzett teljesítmény validáció

A reprodukálható telepítési keretrendszer lehetővé tette statisztikailag releváns kísérletek futtatását öt véletlenszám generátor maggal (42, 123, 456, 789, 1024) és hat konfigurációval (két alapvonal és négy ügynökvariáns). Ez összesen 30 teljes munkanapot szimuláló futtatást eredményezett (egyenként 480 időegység). Az eredményeket a 2. ábra foglalja össze.

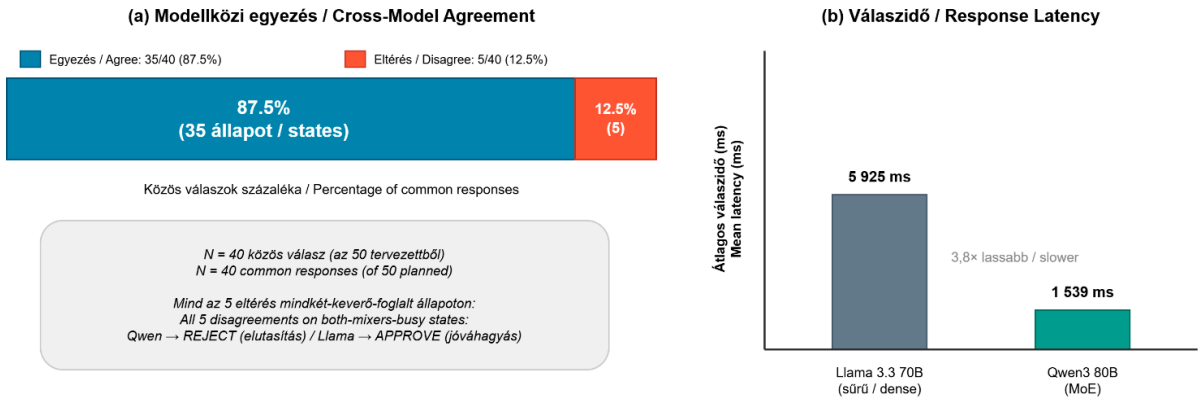
Magas forgalom mellett az MI ügynök 40.9%-os teljesítményjavulást ért el ($p < 0.0001$), felülmúlva a korábbi prototípus egyetlen maggal kapott eredményét. Az alacsony szórások megerősítik, hogy a javulás robusztus, nem egy szerencsés véletlenszám realizáció eredménye. A várakozási idők ennek megfelelően csökkentek: 152.9-ről 116.3 percre magas forgalom mellett ($p = 0.001$), illetve 112.3-ról 91.9 percre normál forgalom mellett ($p = 0.029$).



2. ábra. *Többszörös véletlenszám maggal végzett teljesítmény validáció (5 mag × 6 konfiguráció, Welch-féle t-próba, hibasávok: ±1 szórás)*

3.2 Architektúrák közötti hierarchia validáció

A hierarchikus topológia validálásához az eszkalációs konzisztenciát ötven gondosan tervezett forgatókönyvvel teszteltük (sor ≥ 8, legalább két betonminőség és legalább egy foglalt keverő). Ezeket a forgatókönyveket két architektúrájában eltérő, 70B+ modell értékelte ki adatközponti szakértőként: a Llama 3.3 70B Instruct (sűrű transzformer) és a Qwen3 80B (MoE). Azonos promptokat használtunk, mindkét modellnek ugyanazt az 50 forgatókönyvet adtuk, majd összehasonlítottuk a jóváhagyás/elutasítás döntéseket és a válaszidőket. Alkalmi következtetési időtúllépések miatt a Llama 41, a Qwen pedig 49 esetben adott választ. Az egyezést a 40 közös válaszon számoltuk (3. ábra).



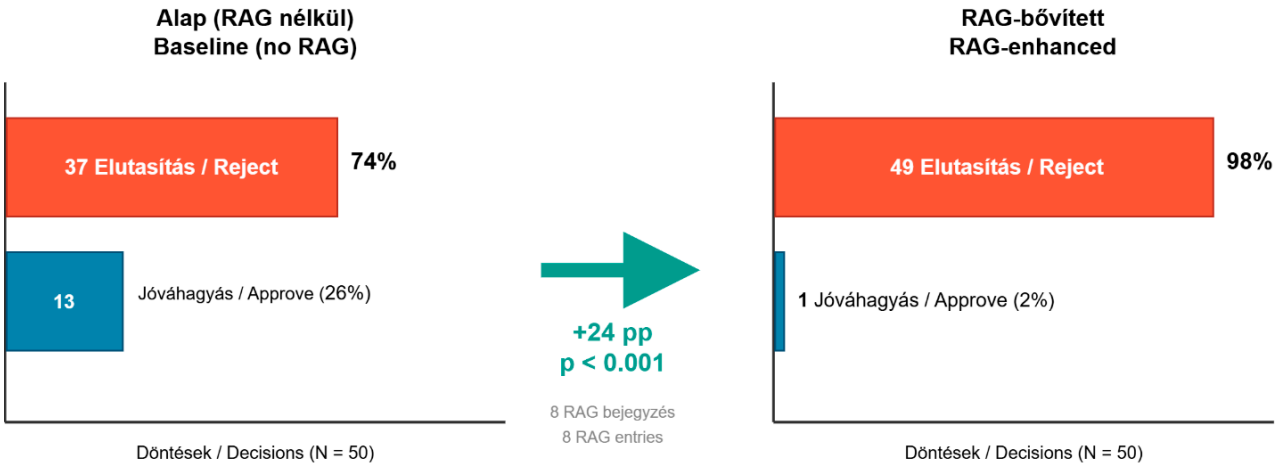
3. ábra. *Architektúrák közötti hierarchia validáció: egyezés és válaszidő 40 közös válaszon az 50 tervezettből (10 kizárva időtúllépés miatt)*

A két modell a 40 közös eset 87.5%-ában egyezett. Ez arra utal, hogy az MCP interfész és a prompt felépítése, nem pedig a konkrét modell az, ami a döntési konzisztenciát meghatározza. Mind az öt eltérés kizárólag mindkét keverő foglalt helyzetekben fordult elő: a Qwen következetesen elutasította az átütemezést a fennakadás kockázatára hivatkozva, míg a Llama az átbocsátóképesség maximalizálását részesítette előnyben. A válaszidők jelentősen eltértek (5925 ms a sűrű Llama modellnél, 1539 ms a MoE Qwen modellnél), ami jól szemlélteti a MoE architektúra hatékonysági előnyét.

3.3 Ember a hurokban visszajelzési validáció

Az ember a hurokban visszajelzési mechanizmus egyértelmű és mérhető változást idézett elő a peremhálózati ügynök viselkedésében. Ugyanazt az 50 forgatókönyvet a Gemma3:4B peremhálózati modellel

két feltétel mellett értékeltük ki: alapvonal korábbi kontextus nélkül, illetve RAG-gal kiegészített feltétel nyolc előre meghatározott operátori visszajelzéssel. Ezek a visszajelzések két tapasztalt operátori szabályt kódoltak: az átütemezés elutasítását, ha a várakozási arány meghaladja a 2.0-t a határidő túllépésének kockázata miatt, valamint az elutasítást, ha mindkét keverő foglalt a fennakadás kockázata miatt. A 4. ábra mutatja a kapott viselkedésváltozást.



4. ábra. *Ember a hurokban visszajelzés hatása: 8 operátori bejegyzés által kiváltott elutasítási arány változás (McNemar-féle egzakt teszt, $p < 0.001$)*

A nyolc visszajelzés RAG-on keresztüli beillesztése 24 százalékpontos elutasítási arány növekedést eredményezett ($p < 0.001$). A RAG mechanizmus a peremhálózati modellt a tapasztalt operátori megítéléssel összhangban konzervatívabb irányba tolta a magas terhelésű helyzetekben.

4. TÁRGYALÁS

A három elem a gyakorlatban egymást erősíti. Mivel a kísérletek reprodukálhatók, a hierarchia teljesítményadatai statisztikailag megalapozottak. Minden operátori korrekció azonnal elérhetővé válik a peremhálózati ügynök számára a RAG táron keresztül.

A KKV-k számára az architektúra több gyakorlati előnyt kínál. A gyártói kötöttség elkerülhető: az adatközponti szakértő cseréjéhez elegendő egyetlen URL módosítása a Helm konfigurációban. Az operátori visszajelzések használat közben automatikusan gyarapodó tudásbázist építenek. Végül a teljes rendszer sikeresen validálva lett két hagyományos laptopon, ami azt bizonyítja, hogy a megoldás elérhető adatközponti erőforrásokkal nem rendelkező kutatócsoportok és vállalatok számára is.

Kíméletes degradáció. Ha az adatközponti szint elérhetetlenné válik, a peremhálózati ügynök az eredeti v1 képességeivel folytatja az autonóm ütemezést. Az egyetlen korlát a komplex esetek eszkalálásának hiánya. Hasonlóképpen, ha nem áll rendelkezésre releváns visszajelzés, a rendszer egyszerűen visszatér az alapértelmezett prompthoz. A determinisztikus eszkalációs szabály biztosítja, hogy a hibamódok kiszámíthatók maradjanak.

Korlátok. A visszajelzésből eredő viselkedésváltozást szándékosan nehéz forgatókönyveken mértük. A mindennapi üzemeltetés valószínűleg kisebb eltolódást mutatna. Az előre meghatározott korrekciók kísérleti kontrollt biztosítottak, de nem helyettesíthetik a valós operátoroktól hosszú távon gyűjtött adatokat. Emellett a SimPy szimuláció egyelőre nem modellezi a berendezés meghibásodásokat és a külső tényezőket, például az időjárást. Az architektúrák közötti egyezést 50 forgatókönyvből 40-en értékeltük ki válaszidő túllépések miatt. Szélesebb körű validációt helyben futtatott modellekkel tervezünk.

Jövőbeli munka. Az előre meghatározott korrekciókat valós, idővel gyűjtött operátori visszajelzésekkel kívánjuk felváltani és a keletkező adathalmazon helyi LoRA finomhangolást tervezünk. Emellett többcsomópontos k3s telepítéseket tesztelünk további gyártási területeken és mérjük, hogy a visszajelzéssel kalibrált peremhálózati döntések hosszú távon csökkentik-e az adatközponti eszkalációk gyakoriságát.

5. KÖVETKEZTETÉS

Jelen tanulmányban egy Kubernetes natív referencia architektúrát mutattunk be, amely az MCP alapú digitális ikreket bővíti a szuverén MI elvek és az összetett gyártási döntésekhez szükséges mély következtetés közötti szakadék áthidalására. A reprodukálható kísérletek megbízható alapvonalakat biztosítottak. A kétszintű hierarchia különböző modelleken konzisztensnek bizonyult és nyolc operátori korrekció is elegendő volt a mérhető viselkedésváltozáshoz újratanítás nélkül. A validált eredmények, köztük a 40.9%-os átbocsátóképesség javulás, a 87.5%-os architektúrák közötti egyezés és a mérhető visszajelzés alapú adaptáció, megerősítik, hogy a szuverén gyártási MI KKV szintű infrastruktúrán is megvalósítható, összhangban az Ipar 5.0 célkitűzéseivel.

IRODALMI HIVATKOZÁSOK

[1] Alfaro-Viquez D., Zamora-Hernandez M., Fernandez-Vega M., Garcia-Rodriguez J., Azorin-Lopez J. A Comprehensive Review of AI-Based Digital Twin Applications in Manufacturing. Electronics, 2025, 14(4), art. 646.

[2] Anthropic. Model Context Protocol Specification, v. 2025-11-25. https://modelcontextprotocol.io/specification/2025-11-25 (Utolsó letöltés: 2026. 03. 15).

[3] Farahani M. A., Khan M. I., Wuest T. Hybrid Agentic AI and Multi-Agent Systems in Smart Manufacturing. arXiv:2511.18258, 2025, https://arxiv.org/abs/2511.18258 (Utolsó letöltés: 2026. 02. 15).

[4] Gauttam H., Nain G., Pattanaik K. K., Mendes P. Edge-AI: A systematic review on architectures, applications, and challenges. J. Netw. Comput. Appl., 2026, 245, art. 104375.

[5] Guo H., Hao Y., Zhang Y., Xu M., Lv P., Chen J., Cheng X. A Measurement Study of Model Context Protocol Ecosystem. arXiv:2509.25292, 2025, https://arxiv.org/abs/2509.25292 (Utolsó letöltés: 2026. 02. 15).

[6] Hou X., Zhao Y., Wang S., Wang H. Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. arXiv:2503.23278, 2025, https://arxiv.org/abs/2503.23278 (Utolsó letöltés: 2026. 02. 15).

[7] Memon U., Mayer W., Selway M., Stumptner M. Interoperability of AI-enhanced digital twins. J. Ind. Inf. Integr., 2025, 48, art. 100961. https://www.sciencedirect.com/science/article/abs/pii/S2452414X25001840 (Utolsó letöltés: 2026. 03. 15).

[8] Misra S., Barik K., Kvalvik P. Digital Sovereignty in the Era of Industry 5.0: Challenges and Opportunities. Procedia Comput. Sci., 2025, 254, 108–117. DOI: 10.1016/j.procs.2025.02.069.

[9] OECD. Progress in Implementing the EU Coordinated Plan on AI (Volume 2): AI in Manufacturing. OECD Publishing, 2026. https://www.oecd.org/en/publications/progress-in-implementing-the-european-union-coordinated-plan-on-artificial-intelligence-volume-2_3ac96d41-en.html (Utolsó letöltés: 2026. 02. 15).

[10] Ren Y., Liu Y., Ji T., Xu X. AI Agents and Agentic AI — navigating a plethora of concepts for future manufacturing. J. Manuf. Syst., 2025, 83, 126–133. DOI: 10.1016/j.jmsy.2025.08.017

[11] Schwaeke J., Peters A., Kanbach D. K., Kraus S., Jones P. The new normal: The status quo of AI adoption in SMEs. J. Small Bus. Manag., 2025, 63(3), 1297–1331. DOI: 10.1080/00472778.2024.2379999.

[12] Semmelrock H., Ross-Hellauer T., Kopeinik S., Theiler D., Haberl A., Thalmann S., Kowald D. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. AI Magazine, 2025, 46(2), art. e70002. DOI: 10.1002/aaai.70002.

[13] Tamás-Péter J., Tamás-Péter T. Az Automatizálástól a Digitalizációig és a Mesterséges Intelligenciáig: Az Ipar 3.0, 4.0 és 5.0 Összehasonlító Elemzése. OGÉT XXXIII, 2025.

[14] Tamás-Péter J., Pócs G. MCP-alapú Digitális Ikrek Fejlesztési Keretrendszere Gyártási MI Ügynökök Számára. Dunakavics XIV évfolyam 2026 (megjelenés alatt)

[15] Tao F., Zhang H., Liu A., Nee A. Y. C. Digital Twin in Industry: State-of-the-Art. IEEE TII, 2019, 15(4), 2405–2415. DOI: 10.1109/TII.2018.2873186.

[16] Tibaldo A. S., Montagna J. M., Fumero Y. Efficient mixed-integer linear programming model for integrated management of ready-mixed concrete production and distribution. Autom. Constr., 2025, 173, art. 106074. DOI: 10.1016/j.autcon.2025.106074.

[17] Yin J., Huang R., Sun H., Lin T. A collaborative scheduling model for production and transportation of ready-mixed concrete. Math. Biosci. Eng., 2023, 20(4), 7387–7406. DOI: 10.3934/mbe.2023320.