AI CYBERSECURITY SUMMIT

Büszkeség és bizonyíték: Három magyar nyelvmodell, újramérve

2026. 09. 08. 17:20 | Innovation Hall | HUN · 20 perc

Dr. Ligeti-Nagy Noémi
tudományos munkatárs, kutatócsoport-vezető
· ELTE Nyelvtudományi Kutatóközpont

Ligeti-Nagy Noémi, a tudományos munkatárs és kutatócsoport-vezető az ELTE Nyelvtudományi Kutatóközpontban, a magyar nagy nyelvi modellfejlesztés jelenlegi állapotát, módszertanát és jelentős technikai korlátait vizsgálja.

MIRŐL VOLT SZÓ

A prezentáció három fő megközelítést vizsgál a magyar nyelvű modellek létrehozásában: az OTP projektet, amely egy kétnyelvű angol-magyar alapmodellre összpontosít, a kutatóközpontban kifejlesztett Puli Modell Családot, és a Macska modellt. Ezek a modellek eltérő képzési filozófiákkal rendelkeznek, amelyek a magyar környezetben „született” modell felnevelésétől az adott szakértők speciális képzési fázisokon keresztüli finomhangolásáig terjednek, például a Macska modell egy 4 milliárd paraméteres Qwen 3 alapmodellt használ speciális magyar tokenizátorral.

Ligeti-Nagy Noémi kiemeli a fejlesztési folyamat kritikus kérdéseit, például a képzési döntések ellenőrzésére szolgáló ablációs tanulmányok hiányát és a tesztkészletek képzési adatokba való beépítésének gyakoriságát. Azt is megjegyzi, hogy jelentős „felejtési” hatás lép fel, amikor a modellek képességeket, például a kódolást elveszítik, miközben a magyar nyelvtudás optimalizálására törekszenek, például a Racka modell elveszíti kódolási képességét annak ellenére, hogy egy képes alapmodell finomhangolt változata. Ezek a technikai akadályok azt sugallják, hogy a jelenlegi fejlesztési módszerek nem rendelkeznek szigorú tudományos validálással és elegendő magas minőségű magyar nyelvű adattal.

KIEMELT MEGÁLLAPÍTÁSOK

- A Racka modell 11-ből 9 feladatban drámai teljesítménycsökkenést mutat, amikor a „gondolkodó” módja aktív.
- Adatkontamináció az a széles körben elterjedt probléma, amikor a modelleket olyan kérdésekre tesztelik, amelyek már jelen voltak a képzési adatokban.
- A magyar nyelvi folyékonyság finomhangolása gyakran a nem nyelvi készségek, például a kódolás és a matematikai gondolkodás elvesztésével jár.

ÜZLETI HATÁS

A magyar modellek fejlesztését a helyi adatszintű biztonság és a vállalati környezetekben a kulturális relevancia iránti igény hajtja. Azonban a hardver magas költségei és a hatalmas szénlábnyom, például a Macska modell esetében kibocsátott több mint egy tonna CO2, jelentős gazdasági és környezeti kihívásokat jelentenek. A szabványosított referenciaértékek hiánya megnehezíti a vállalatok számára, hogy objektíven mérjék és összehasonlítsák a különböző modellek teljesítményét a gyakorlati üzleti alkalmazásokhoz.

KÖVETKEZTETÉS

Ligeti-Nagy Noémi szerint bár léteznek magyar modellek, a jelenlegi fejlesztési folyamat nem rendelkezik a szükséges tudományos szigorúsággal és a csúcsteljesítmény eléréséhez szükséges magas minőségű adatokkal. Javasolja, hogy kerüljék a szennyezett adathalmazokon történő tesztelést, az edzési adatok arányait átláthatóbb módon dokumentálják, és a jövőbeli megerősítéses tanulás és emberi visszajelzési ciklusok támogatása érdekében elsőbbséget élvezzenek a magas minőségű magyar adathalmazok létrehozása.

AI CYBERSECURITY SUMMIT

Pride and Proof: Three Hungarian Language Models Re-Evaluated

2026. 09. 08. 17:20 | Innovation Hall | HUN · 20 perc

Dr. Ligeti-Nagy Noémi
Research Fellow, Research Group Leader ·
ELTE Nyelvtudományi Kutatóközpont

Noémi Ligeti-Nagy, Research Fellow and Research Group Leader at ELTE Nyelvtudományi Kutatóközpont, examines the current state, methodologies, and systemic flaws in developing Hungarian large language models.

WHAT WAS DISCUSSED

Noémi Ligeti-Nagy describes three primary approaches to creating Hungarian language models: the OTP model, which is a bilingual English-Hungarian base model; the Puli Model Family, which includes Llama and Qwen variants fine-tuned for Hungarian; and the Macska model, a 4-billion parameter Qwen 3 variant using LoRA fine-tuning and a specific Hungarian tokenizer. For example, the Puli models are described as experts who have spent significant time in Hungary to acquire local knowledge, whereas the Macska model represents a faster, cheaper method of providing partial training to an existing expert.

The presentation highlights significant technical issues in the development process, including a lack of ablation studies to prove the effectiveness of specific training choices and the absence of clear documentation regarding data proportions or scheduling. Noémi Ligeti-Nagy notes that models often suffer from catastrophic forgetting, where they lose skills like coding while learning Hungarian. For example, a comparison of six benchmarks showed that the Hungarian models lost proficiency in mathematics and coding compared to their original base models. These issues suggest that current development methods lack the scientific rigor necessary to ensure high-quality results.

KEY TAKEAWAYS

- The Macska model required over one ton of carbon dioxide emissions and significant electricity consumption during its development.
- Testing models on data that might have been included in the training set leads to contaminated and unreliable results.
- The original Qwen 3 model outperformed several Hungarian-specific models on international benchmarks despite not being designed for the Hungarian language.

BUSINESS IMPACT

The development of sovereign Hungarian models is driven by a need for local data processing and security, yet the high costs and lack of high-quality Hungarian datasets pose significant hurdles. Noémi Ligeti-Nagy points out that while there is a pressing demand for Hungarian-centric AI in local companies, the current lack of structured data and the high GPU costs make it difficult to produce high-quality models quickly.

CONCLUSION

Noémi Ligeti-Nagy concludes by urging developers to move away from testing models on exam questions that are part of the training data and to instead use independent validation sets. She recommends documenting data proportions clearly and emphasizes the urgent need to create high-quality Hungarian datasets to move beyond current limitations.