AKADEMIA E FORCAVE
TË ARMATOSURA

Hyrje në Modele të Mëdha Gjuhësore

Paratrajnimi, shkallëzimi dhe sjellja gjeneruese në MMGj

Dr. Fiorela Ciroku

Pse MMGj janë kaq të fuqishme?

- Çfarë mëson realisht një model gjatë trajnimit fillestar (pretraining)?
- Pse shkalla (të dhënat + parametrat) ka rëndësi?
- Çfarë do të thotë sjellje emergjente?
- Pse MMGj-të duken sikur “dinë” fakte, procedura dhe modele arsyetimi?

Ky modul është thelbësor për të kuptuar:

Çfarë është trajnimi fillestar?

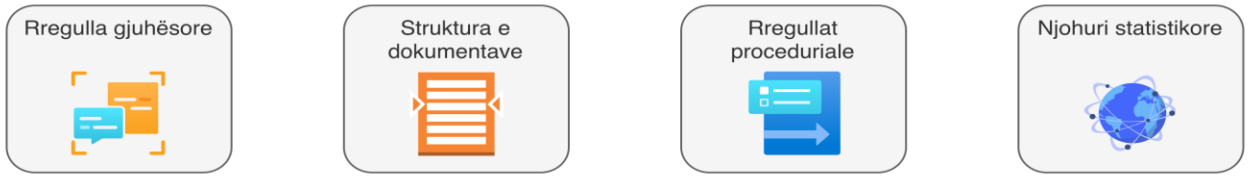
- **Trajnimi fillestar (pretraining) është mësim i vetë-mbikëqyrur (self-supervised learning).**
- Modeli trajnohet mbi një detyrë të thjeshtë:

> Duke pasur një sekuencë tokenash, të parashikojë tokenin pasues.

- Pa përdorur etiketa, pa njerëz që anotojnë kuptimin, pa detyra të përcaktuara në mënyrë eksplicite.
- Duke përdorur vetëm *libra, artikuj, faqe web, manuale, kod dhe dokumente të strukturuara*.
- Modeli arrin të mësojë gramatikën, semantikën dhe strukturën e dokumenteve.

Çfarë ofron trajnimi fillestar?

Çfarë mëson trajnimi fillestar:



- Nuk ofron të vërtetën e bazuar (grounded truth), autoritet ligjor, informacion të përditësuar apo garanci te domeinit.
- Një model mund të mësojë si shkruhen procedurat e emergjencës, por jo cila procedurë është ligjërisht e vlefshme në Shqipëri.
- Ky dallim është thelbësor për të kuptuar *hallucinacionet*.

Kufizimet e trajnimit fillestar

- Vetëm trajnimi fillestar (pretraining) nuk është i mjaftueshëm për vendosje të sigurt.
- **Problemet**: informacion i vjetëruar, procedura të sajuara, hallucinacione të sigurta (të shprehura me vetëbesim).

- Një model vetëm i trajnuar fillimisht mund të shpikë rregulla emergjence, të përziejë protokollet e vendeve të ndryshme, të tingëllojë autoritar, por të jetë i gabuar.

- Kjo është e papranueshme në mbrojtjen civile, fushën ushtarake, mjekësi, juridik, etj.

- Motivon përdorimin e RAG dhe përshtatjen e modelit (fine-tuning).

Nga token pasardhës në arsyetim

Si mund të çojë parashikimi i fjalës pasuese në arsyetim?

Parashikimi i tokenit pasues kodon në mënyrë të nënkuptuar kauzalitetin, rendin kohor dhe kufizimet logjike. Për të parashikuar saktë, modeli duhet të ndërtojë përfaqësime të brendshme.

Për të parashikuar *“Bashkia duhet të ___ banorët”, modeli duhet të nxjerrë përfundime rreth:*

- përgjegjësisë.
- rolit institucional,
- kontekstit të emergjencës.

Scaling Laws

Ligjet e shkallëzimit (scaling laws) tregojnë një rezultat befasues:

Nëse rriten së bashku madhësia e modelit, madhësia e të dhënave dhe fuqia llogaritëse, performanca përmirësohet në mënyrë të vazhdueshme dhe të parashikueshme.

Kjo do të thotë:

- nuk kërkohet ndonjë ndryshim i papritur i arkitekturës
- modelet më të mëdha performojnë vazhdimisht më mirë
- “më shumë nga e njëjta gjë” funksionon

Modelet e mëdha

- **Ndërsa modelet shkallëzohen, ato nuk thjesht memorizojnë më shumë.**
- Ato ndërtojnë hapësira semantike më të pasura, kompresojnë rregullsi komplekse, përgjithësojnë më mirë ndër detyra.
- Për shembull: Një model i vogël mund të memorizojë fraza. Ndersa, një model i madh mëson modele procedurash.
- Kjo shpjegon pse MMGj-të e mëdha mund të përmbledhin, përkthejnë, u përgjigjen pyetjeve, ndjekin udhëzime, **pa trajnim të drejtpërdrejtë për secilën detyrë.**

Aftësitë emergjente

- **Emergjenca (emergence)** i referohet aftësive që nuk ekzistojnë në shkallë të vogël dhe që shfaqen papritur pasi tejkalohet një prag madhësi.

- Shembuj**:** arsyetim me shumë hapa, ndjekje udhëzimesh, mësim brenda kontekstit (in-context learning), qëndrueshmëri bazë në aritmetikë.
- Këto **nuk ja**në të ekstrapoluara në mënyrë lineare nga modelet e vogla.

Prandaj:

- rritjet e aftësive duken “magjike”
- sjellja ndryshon papritur
- vlerësimi bëhet më i vështirë

Pse ndodh emergjenca?

Emergjenca ndodh sepse:

- kapaciteti i modelit kalon një prag
- përfaqësimet bëhen mjaftueshëm shprehëse
- mekanizma të shumtë të brendshëm përafroren

- Është e ngjashme me ngrirjen e ujit në 0°C, shfaqjen e lidhshmërisë në rrjete

- **Para pragut** sjellja është e paqëndrueshme. **Pas pragut** sjellja stabilizohet.

Përshtatja ndaj fushës së veprimit

MMGj-të e trajnuara mbi të dhëna globale:

- nuk njohin strukturën institucionale të Shqipërisë.
- nuk njohin hierarkitë e komandimit të NATO-s.
- nuk dinë cilat procedura janë të detyrueshme.

Përshtatja ndaj domenit shton dokumente të sakta, burime autoritative dhe njohuri të kontrolluara.

Kjo mund të bëhet përmes:

- përshtatjes së modelit (fine-tuning)
- përshtatjes me udhëzime (instruction tuning)
- rikthimit të informacionit (retrieval – RAG)

RAG si zgjidhje kundër limiteve të trajnimit fillestar

- Gjenerimi i Përforcuar nga Kërkimi (Retrieval-Augmented Generation – RAG) ndryshon paradigmën:
- **Në vend të: “Përgjigju nga kujtesa”, modeli detyrohet të: “Përgjigju duke përdorur këto dokumente”.**
- Ky proces ankoron përgjigjet, përmirëson saktësinë faktike, dhe mundëson citime.
- **RAG nuk e zëvendëson trajnimin fillestar — ai e kontrollon atë.**

Trajnimi fillestar dhe RAG

- Trajnimi fillestar → motor arsyetimi
- RAG → burim i besueshëm njohurish

- Pa trajnimin fillestar, RAG nuk ka inteligjencë.
- Pa RAG, inteligjenca halucinon.

- Sistemet moderne i kombinojnë gjithmonë të dyja.

Çfarë do të thotë kjo në praktikë?

Në praktikë:

- Mos u besoni modeleve të trajnuara vetëm fillimisht (raw pretrained models).
- Gjithmonë ankoroni përgjigjet në burime të besueshme.
- Monitoroni mënyrat e dështimit.

- Veçanërisht në fusha veprimi si mbrojtja, mjeksia, juridiku, etj.

- **MMGj-të nuk e zëvendësojnë ekspertin e fushës, ama një ekdpert që përdor MMGj-të e zëvendëson një ekspert që nuk përdor MMGj.**

Çfarë do të trajtojmë në vazhdimësi?

- ✓ Hyrje
- ✓ Tokenizimi & Embeddings
- ✓ Arkitektura Transformer
4. Bazat e të dhënave vektoriale (Vector Databases)
5. RAG (Gjenerim i përforcuar nga kërkimi)
6. Promptimi
7. Agjentët (Agents)
8. Përshtatja e modelit (Fine-Tuning)
9. Vlerësimi (Evaluation)
10. Siguria & Përafrimi (Safety & Alignment)
11. Vendosja e modelit (Deployment)
12. Integrimi i Projektit Final