AKADEMIA E FORCAVE
TË ARMATOSURA

Hyrje në Modele të Mëdha Gjuhësore

Dr. Fiorela Ciroku

Çfarë janë Modelet e Mëdha Gjuhësore(MMGj)?

- MMGj-të janë rrjeta nervore (neural networks) të trajnuara mbi sasi shumë të mëdha teksti për të parashikuar tokenin pasues në një sekuencë.

- I referohen modeleve gjuhësore të përgjithshme, të cilat mund të trajnohen paraprakisht (pre -trained) dhe me pas të përshtaten (fine-tuning) për qëllime specifike.

- Duke u trajnuar, modelet brendësojnë gramatikën, marrëdhëniet semantike, modelet e arsyetimit, lidhjet faktike.

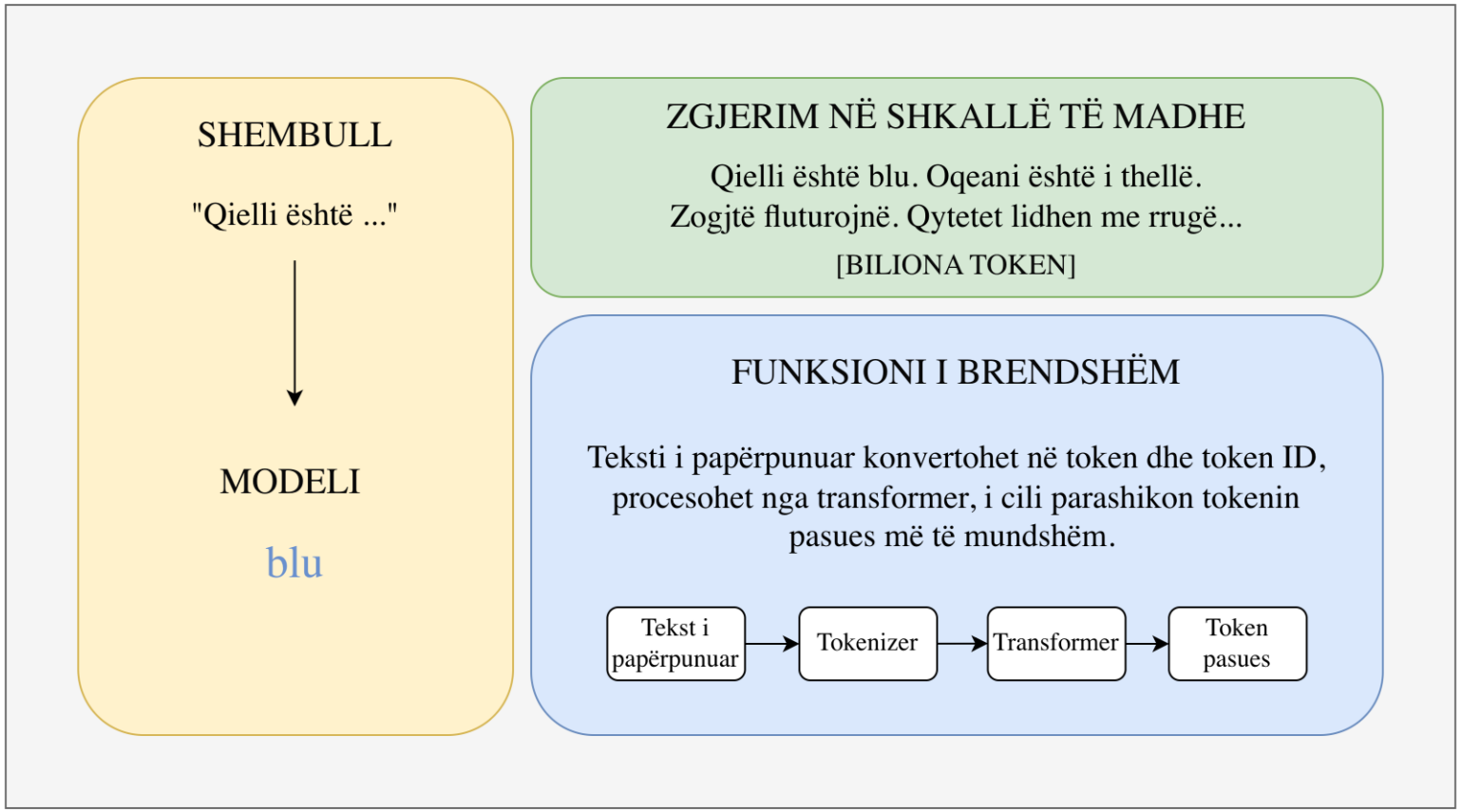
Klasifikim teksti

Sisteme pyetje përgjigje

Permbledhje dokumenti

Gjenerim teksti

Çfarë janë Modelet e Mëdha Gjuhësore?



AI për Mbrojtjen Civile

Një AI i fuqizuar nga MMGj-të për Mbrojtjen Civile mund të:

- përmbledhë dokumente të reagimit të emergjencave
- shndërrojë tekste komplekse të mbrojtjes civile në udhëzime të qarta dhe të zbatueshme
- t'i japë përgjigje pyetjeve si:
  “Cilat janë hapat e parë për t'u ndërmarrë pas një tërmeti?”

Përse janë të rëndësishme MMGj-të?

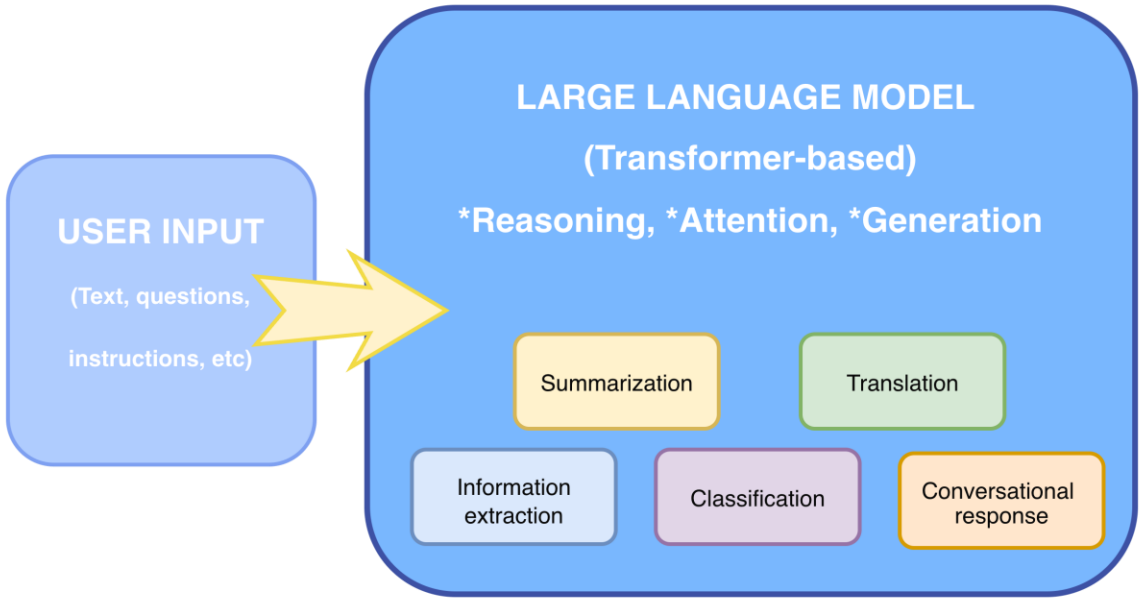
MMGj-të mundësojnë:

- Performancë të lartë në detyra të ndryshme të NLP-së
- Përgjithësim në shumë detyra njëkohësisht, si përkthim, përmbledhje, klasifikim, marrje informacioni.
- Përshtatje të shpejtë përmes promptimit

Përse janë të rëndësishme MMGj-të?

Proçesi tradicional i një NLP (Shumë modele)

Procesi i unifikuar në MMGj



AI për Mbrojtjen Civile

Në vend që të ndërtojmë:

1. një sistem të bazuar në rregulla për procedurat e përmbytjeve
2. një klasifikues për fazat e emergjencave
3. një motor kërkimi për termat e emergjencës
4. ...

Ndërtojmë **1 sistem MMGj + RAG** për të mbështetur të gjitha këto detyra.

Karakteristikat e MMGj-ve

TË MËDHENJ

Datasete trajnimi të mëdha

Numra të mëdhenj parametrash

TË PËRGJITHSHËM

Elementet e përbashkëta të gjuhës njerëzore

Kufizime të burimeve

TË PËRSHTATSHËM

Paratrajnime të modeleve të mëdha gjuhësore për një qëllim të përgjithshëm

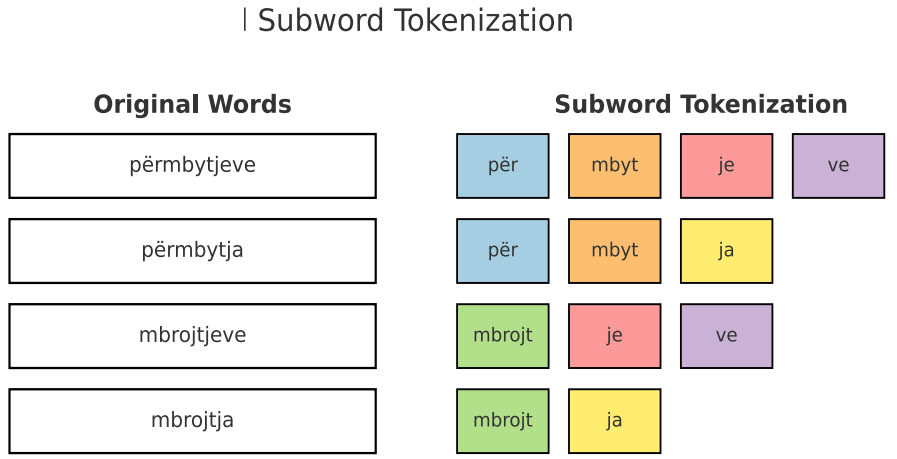
Përshtatja për qëllime specifike me një dataset shumë më të vogël.

Çfarë janë tokenat?

Tokenat janë njësitë atomike të tekstit për MMGj-të.

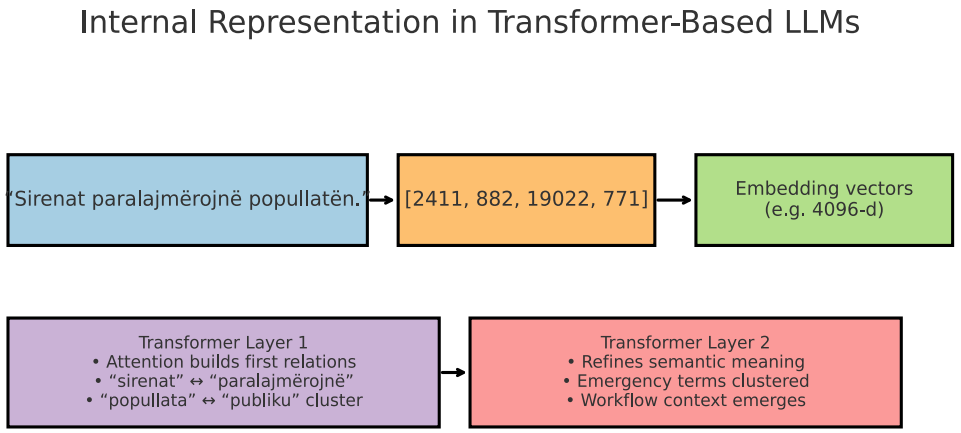
Pse nuk përdorim fjalë të plota?

- Shumë fjalë (morfologjia e shqipes)
- Fjalët e rralla → paraqitje e dobët
- Nënfjalët (subwords) lejojnë përgjithësim



Si e shohin MMGj-të tekstin?

- **Tekst → token → ID e token**
- ID-të lidhen me embeddings
- **Modeli sheh VETËM numra.**
- Modeli nuk sheh fjalë, shkronja, kuptim, gramatikë.
- Kuptimi shfaqet përmes mekanizmit të vëmendjes (attention) në momentin kur ekzaminohet lidhja mes embeddings.

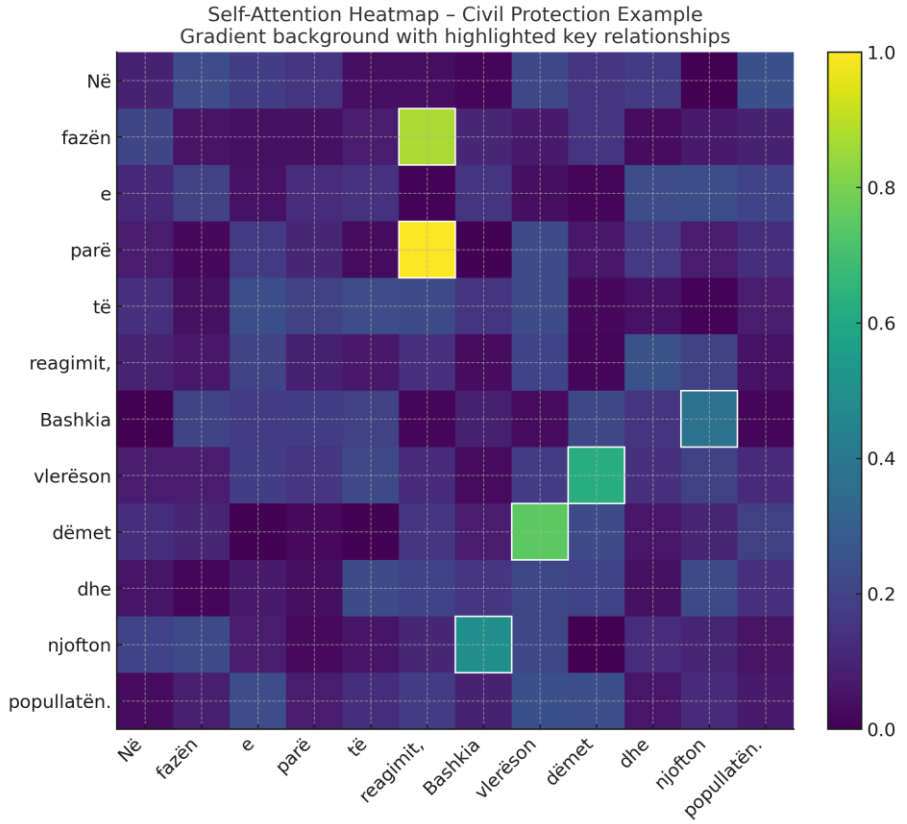


AI për Mbrojtjen Civile

Në fjalinë “Në fazën e parë të reagimit, Bashkia vlerëson dëmet dhe njofton popullatën.”, modeli lidh:

- “Bashkia” ↔ “njofton”
- “faza e parë” ↔ “reagimit”
- “vlerëson” ↔ “dëmet”

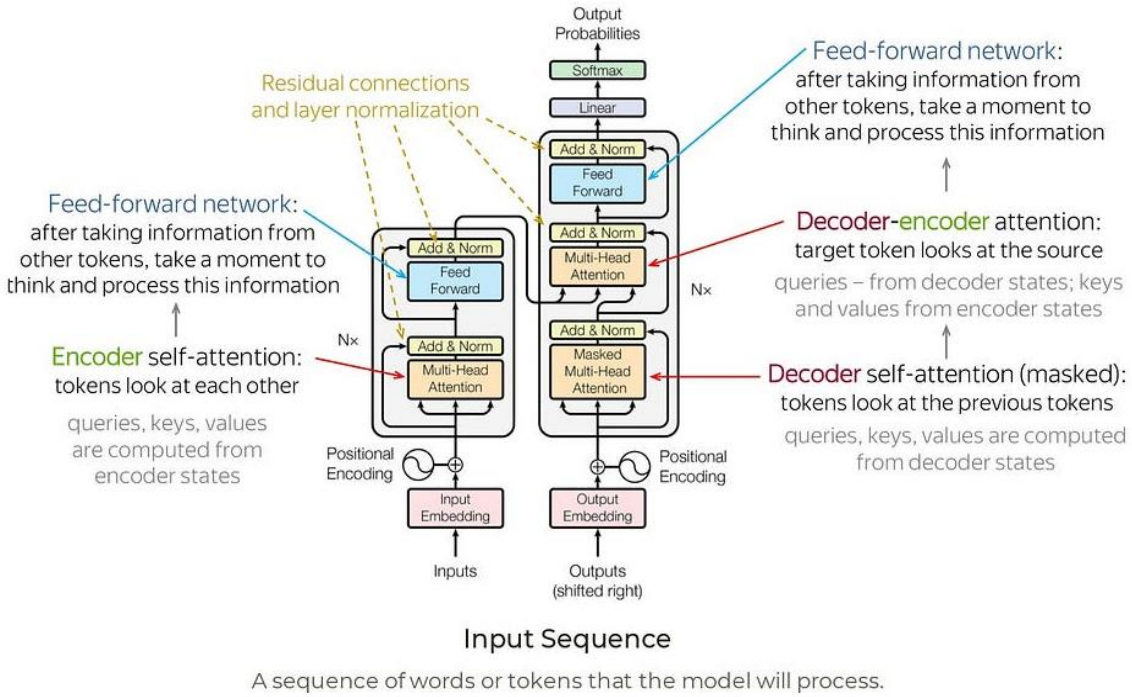
Vëmendja (attention) lejon çdo fjalë të lidhet me të gjitha të tjerat për të nxjerrë kuptimin.



Arkitektura e transformerit

Elementët kryesorë të arkitekturës Transformer:

- Self-attention (vetë-vëmendja)
- Multi-head attention (vëmendja me shumë koka)
- Shtresat feed-forward
- Normalizimi i shtresave
- Lidhjet residuale



Si trajnohen MMGj-të?

- Modelet e Mëdha Gjuhësore kërkojnë "pak" të dhëna trajnimi specifike në rastet kur duhet të zgjidhin probleme të veçanta. Këto modele arrijnë performancë të mirë edhe me pak të dhëna trajnimi në fushën e operimit (few shot, zero shot).
- Performanca e MMGj rritet në mënyrë të vazhdueshme kur shtohen të dhëna dhe parametra të rinj.

MMGj

- JO ekspert
- JO shembuj trajnimi
- JO trajnosh një model
- Ideim i prompteve

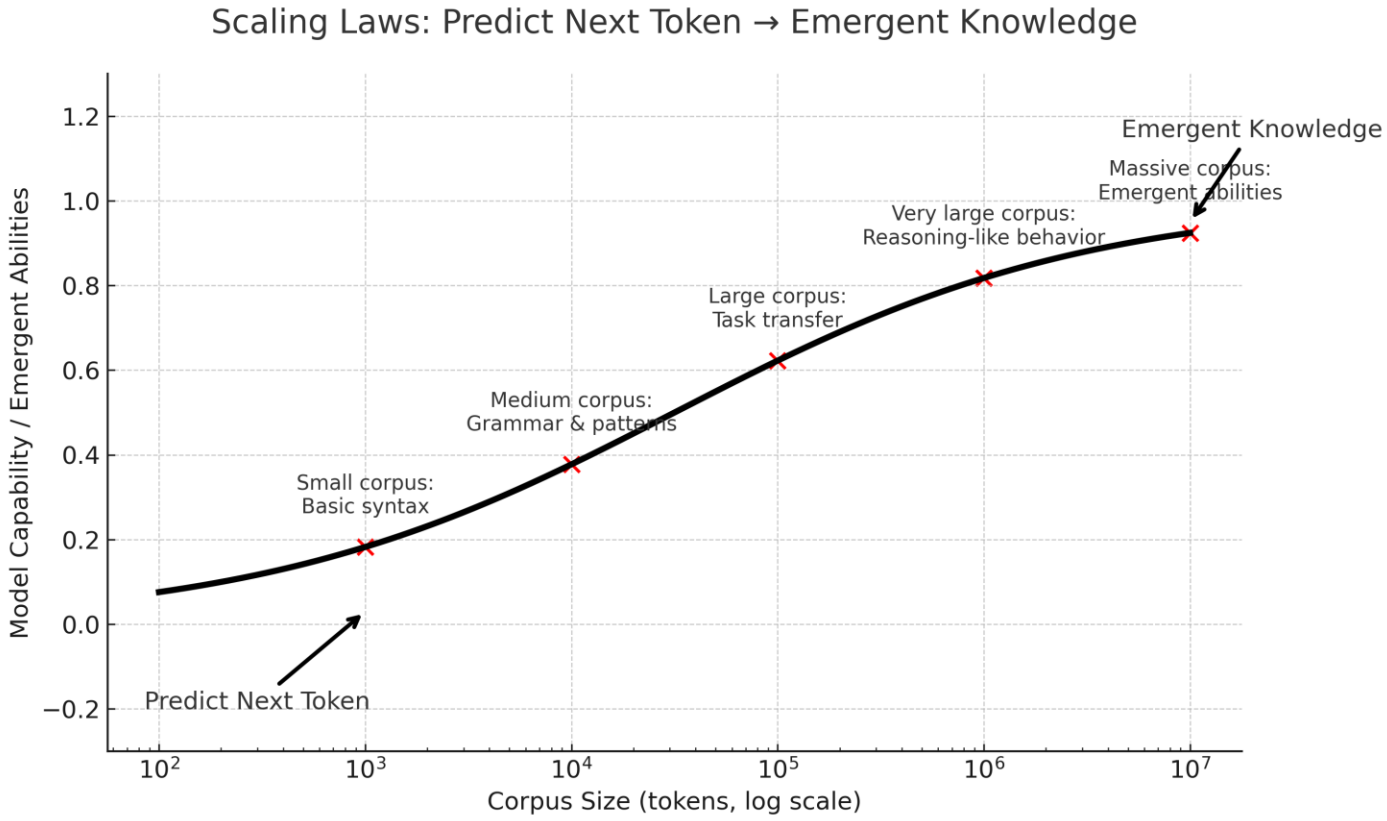
Zhvillimi tradicional i kodit

- Ekspert
- Shembuj trajnimi
- Kohë ekzekutimi
- Pajisje

Si trajnohen MMGj-të?

Përmes ekspozimit ndaj miliarda shembujve, modeli mëson:

- Shkak-pasojën
- Sekuencat kohore
- Rrjedhën procedurale
- Përkufizimet
- Rolet dhe përgjegjësitë



AI për Mbrojtjen Civile

Edhe pa trajnim specifik për Shqipërinë, modeli njeh struktura procedurale si:

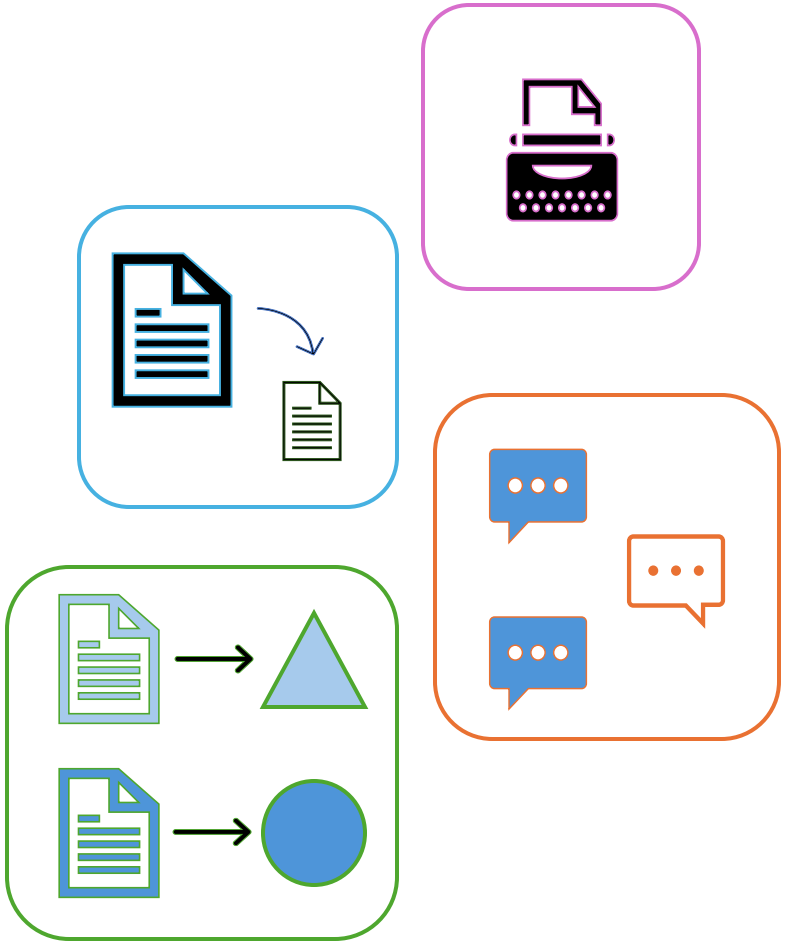
- “Së pari vlerësoni situatën…”
- “Njoftoni autoritetet përkatëse…”
- “Merrni masa sigurie…”

Më vonë do të fusim njohuri specifike për Shqipërinë përmes RAG.

Çfarë munden të bëjnë MMGj-të?

MMGj-të mund të:

- Përmbledhin tekste të gjata
- Nxjerrin hapa dhe procedura
- I japin përgjigje pyetjeve
- Riformulojnë dhe thjeshtojnë tekstin
- Përkthejnë midis gjuhëve
- Kryejnë arsyetim përmes modeleve të të dhënave



Çfarë nuk munden të bëjnë MMGj-të?

MMGj-të nuk mund të:

- Garantojnë saktësinë
- Zëvendësojnë ekspertët e fushës
- Marrin vendime operacionale
- Aksesojnë të dhëna në kohë reale
- Interpretojnë të dhëna numerike nga sensorët pa ndihmë
- Kuptojnë botën si njerëzit

AI për Mbrojtjen Civile

Asistenti ynë

duhet të fokusohet në:

- Interpretim dokumeti
- Suport trajnimi
- Navigim procedurial

Asistenti ynë

NUK duhet KURRË të:

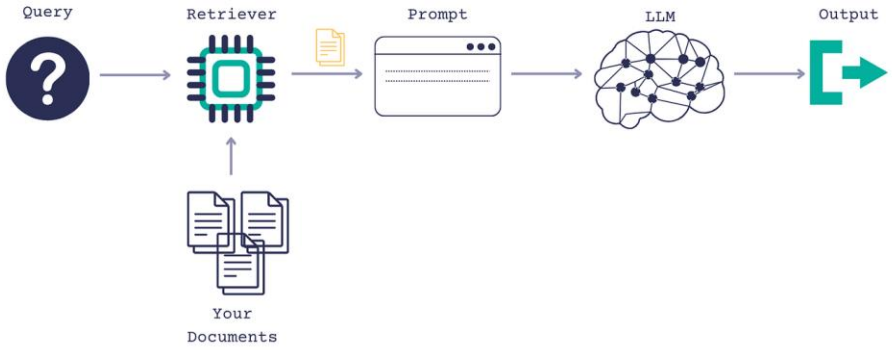
- Japë komanda operacionale
- Rekomandojë vendime taktike
- Nxjerrë përfundime për rrezik në kohë reale
- Sugjerojë rrugë evakuimi

Asistenti ynë duhet të bëjë kujdes në:

- Arsyetim kur mungon informacioni
- Në verifikim faktesh

Roli i të dhënave në sjelljen e MMGJ-ve

- **Të dhënat e trajnimit fillestar formësojnë kompetencën e përgjithshme.**
- Të dhënat e domenit nevojiten për specializim.
- RAG shton njohurinë që mungon.
- Cilësia e të dhënave > madhësia e modelit.



AI për Mbrojtjen Civile

Modeli nuk di:

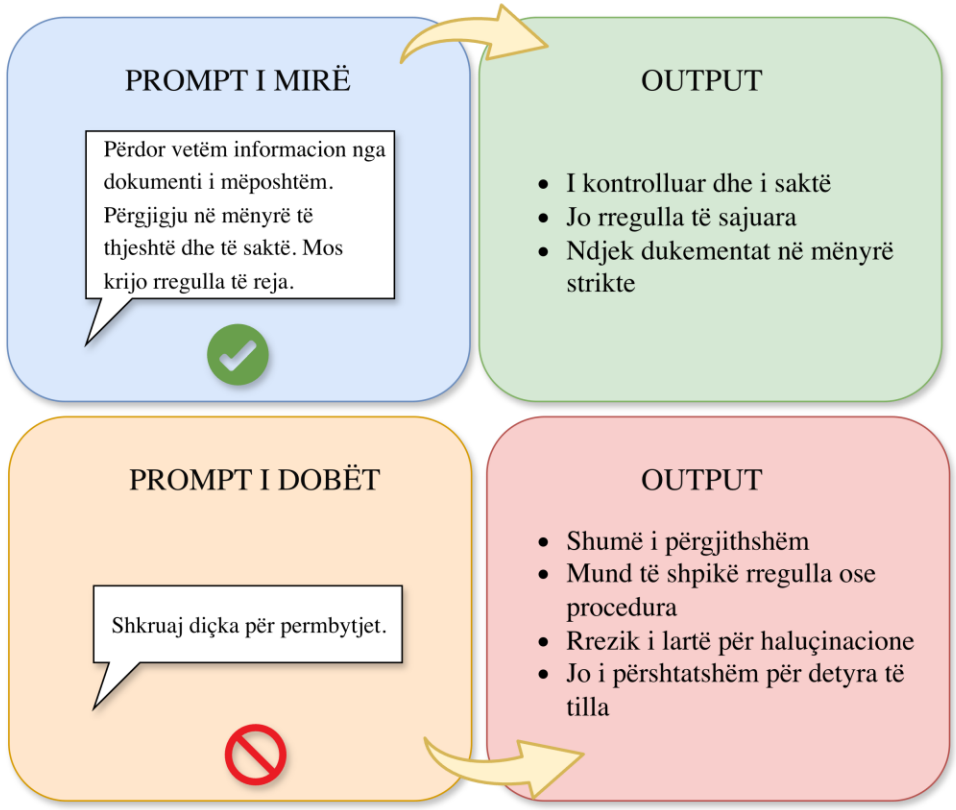
- strukturën institucionale të Shqipërisë
- procedurat lokale të emergjencave
- mandatin ligjor
- terminologjinë e mbrojtjes civile në shqip

Prandaj: RAG është thelbësor sepse shton:

- protokollet zyrtare të emergjencave në Shqipëri
- planet emergjente bashkiake/shtetërore

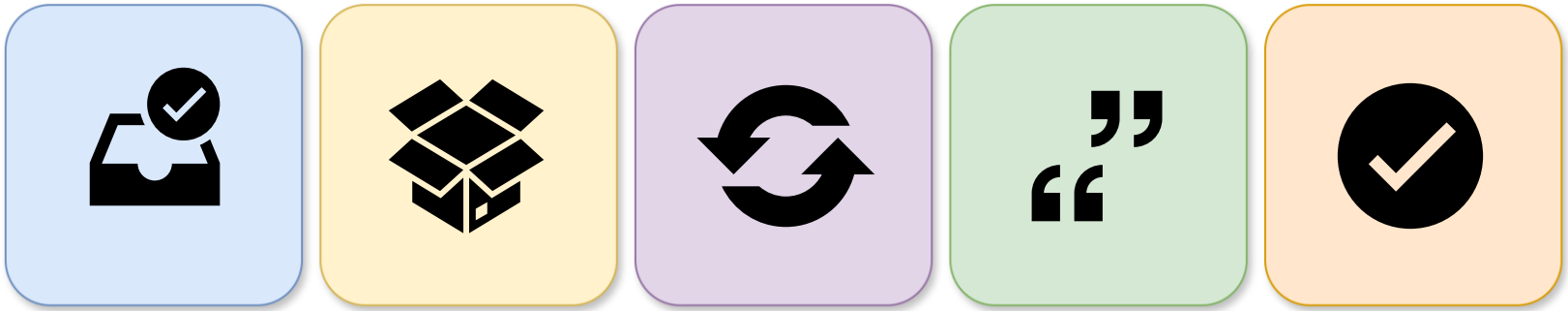
Prompting: Si të komunikojmë me MMGj-të?

- **Promptet = udhëzime**
- Ideimi i promptit është proçesi i krijimit të një prompti të përshtatur për detyrën specifike që i kërkohet sistemit të kryejë.
- Promptet kontrollojnë tonin, strukturën, veprimet e lejuara, kufizime sigurie, output të dëshiruar, etj.



Pse MMGj-të nuk janë mjaftueshëm?

RAG (Retrieval Augmented Generation) ofron:



Baza faktike

Transparencë

Informacion të përditësuar

Citim të dokumenteve

Më pak hallucinacione

Çfarë do të trajtojmë në vazhdimësi?

1. Hyrje
2. Tokenizimi & Embeddings
3. 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

Çfarë do të realizojmë?

Në përfundim të këtij kursi do të kemi:

- Një pipeline për përthithjen e dokumenteve (document ingestion pipeline)
- Një motor kërkimi vektorial (vector search engine)
- Një sistem RAG
- Template për promptet
- Mekanizma sigurie
- Një UI të vendosur për asistentin
- Një paketë vlerësimi

Projekti 1: Asistent doktrine

Ndërtoni një asistent AI që ndihmon oficerët dhe nënoficerët (NCO) të gjejnë shpejt:

- Procedurat Standarde të Operimit (SOP)
- Rregullat e sigurisë
- Manualet e trajnimit
- Udhëzime të lidhura me NATO-n
  (STANAG jo të klasifikuara, përmbledhje doktrinore)

Projekti 1: Asistent doktrine - Detyrat

Asistenti duhet të:

- Sqarojë procedurat, *p.sh. “Cili është zinxhiri i raportimit për incidentin X?”*
- Përmbledhë dokumente PDF të gjata
- T’u përgjigjet pyetjeve të tipit: *“ku mund ta gjej…?”*
- Ofrojë lista kontrolli për detyra të përsëritura, *p.sh. kontrolle sigurie të automjeteve, procedura shërbimi roje, përgatitje trajnimi.*

Projekti 1: Asistent doktrine

Datasete publike:

- Dokumente publike të NATO-s, *p.sh. përmbledhje doktrinore të NATO-s, manuale, materiale informuese*
- Manuale të misioneve paqeruajtëse të BE-së ose OKB-së
- Shembuj gjenerikë të SOP-ve ushtarake nga vende të tjera që janë publikisht të aksesueshme

Datasete lokale:

- SOP sintetike ose të redaktuara të shkruara për qëllime mësimore
- Slide dhe manuale të shkollave të trajnimit që konsiderohen tashmë publike ose edukative
- Rregulla sigurie, *p.sh. siguria në poligonin e armëve, siguria e automjeteve, shëndeti dhe siguria në punë (të përshtatura ose fiktive)*

Projekti 2: Asistent për anglishten ushtarake

Ndërtoni një asistent AI që ndihmon ushtarakët shqiptarë të:

- Përmirësojnë anglishten operacionale (NATO English)
- Kuptojnë fraza të komunikimeve
- Praktikojnë dialogë të radio-komunikimit
- Korrigjojnë raportet dhe email-et
- Simulojnë skenarë stërvitorë (patrulla, postbllok, negocim paqeruajtës)

Projekti 2: Asistent për anglishten ushtarake

Asistenti duhet të:

- Shpjegojë fraza të anglishtes ushtarake në shqip dhe anglisht
- Simulojë komunikimin radio *(kode të shkurtra – brevity codes, fraza standarde)*
- Gjenerojë dialogë praktike, *p.sh. pikë kontrolli, patrullë, briefing.*
- Jepë feedback për anglishten e shkruar në *email-e, raporte të shkurtra*.
- Luajë role skenarësh, *p. sh. paqeruajtje, patrullë e përbashkët me një vend tjetër të NATO-s*

Projekti 2: Asistent për anglishten ushtarake

Datasete publike:

- Fjalorë dhe publikime publike të NATO-s me frazeologji
- Manuale komunikimi të misioneve paqeruajtëse të OKB-së
- Të dhëna mësimore për anglishten *(korpuse ESL, sete dialogësh)*

Të krijuara:

- Lista frazash paralele shqip–anglisht për: pika kontrolli, patrulla, situata emergjente, briefing-e
- Shembuj skriptesh komunikimi radio
- Shembuj raportesh

MMGj të rekomanduara

1. **Modele bazë:**
   - `mistralai/Mistral-7B-Instruct`
   - `meta-llama/Meta-Llama-3-8B-Instruct`
   - `microsoft/Phi-3-mini-4k-instruct`

2. **Për embeddings (RAG):**
   - `sentence-transformers/all-MiniLM-L6-v2`
   - `BAAI/bge-base-en`

3. **Modele API:**
   - OpenAI `gpt-4.1` / `gpt-4o-mini` as a baseline for comparison.

Bibliografi

1. Vaswani et al., 2017 — “Attention Is All You Need.” , https://arxiv.org/abs/1706.03762
2. Kaplan et al., 2020 — “Scaling Laws for Neural Language Models.”, https://arxiv.org/abs/2001.08361
3. Brown et al., 2020 — “Language Models are Few-Shot Learners.”, https://arxiv.org/abs/2005.14165
4. Wei et al., 2022 — “Emergent Abilities of Large Language Models.”, https://arxiv.org/abs/2206.07682
5. Bender & Koller, 2020 — “Climbing Towards NLU: On Meaning, Form, and Understanding.”, https://aclanthology.org/2020.acl-main.463/
6. Eisenstein, Jacob (2019). *Introduction to Natural Language Processing*. MIT Press.
7. Goldberg, Yoav (2017). *Neural Network Methods for NLP*. Morgan & Claypool.
8. EU High-Level Expert Group on AI — “Ethics Guidelines for Trustworthy AI” (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai