A modell a szövegek statisztikai gyakorisága alapján bontja fel a szavakat, nem pedig nyelvtan szerint."

🔹 1. Nyelvtani bontás vs. statisztikai bontás

- **Nyelvtani bontás**:
  Ha nyelvtan alapján bontanánk szét a szavakat, akkor a modell tudná, hogy a „számítógép" egy szó, és mindig egyben kezeli.
- **Statisztikai bontás (amit ChatGPT használ)**:
  A modell először óriási mennyiségű szöveget olvasott be (pl. könyveket, weboldalakat), és megnézte:
  - mely betűkombinációk fordulnak elő gyakran együtt.
  - Ezekből épít „tokeneket".

Példa:

- A „számít" nagyon gyakori magyar szó → ezért külön tokenként létezik.
- Az „ógép" ritkábban fordul elő → de amikor valaki mégis írja a „számítógép"-et, a rendszer így tudja lefedni: „számít" + „ógép".

🔹 2. Miért fontos ez?

- A modell **nem tudja, hogy mi egy szó**, és mi nem.
- Csak azt tudja, hogy „bizonyos betűcsoportok gyakran jönnek elő együtt", ezért azokat egy tokenbe teszi.
- Ezért lehet, hogy:
  - Rövid szavak (pl. „kutya") → 1 token.
  - Hosszabb szavak (pl. „számítógép") → több token.
  - Idegen szavak (pl. „computer") → lehet, hogy szintén több tokenre törnek, ha ritkán fordulnak elő.

🔹 3. Analógia

Képzeld el, hogy a modell nem tud magyarul, hanem csak azt figyeli:

- „Milyen betűkombinációkat láttam már gyakran együtt?"
- Ha a „számít" és az „ógép" sokszor fordult elő együtt, akkor ő összerakja, de nem „érti", hogy az egy teljes szó.

✅ **Tehát**:
Amikor azt írtam, hogy **„nem nyelvtan szerint"**, az azt jelenti, hogy a modell **nem tud**

szavakat felismerni a nyelvtani szabályok alapján, hanem csak a **betűsorok gyakoriságából tanulja meg**, hogyan érdemes feldarabolni a szöveget kisebb részekre (tokenekre).

🔹 Példa: „számítógép”

A modell nem tudja, hogy ez **egyetlen magyar szó** → ezért így szedi szét:

- `" számít"`
- `"ógép"`

👉 Vagyis **kettő token** lesz belőle.
Azért van így, mert:

- A **„számít”** önmagában nagyon gyakori magyar szó (sokszor előfordul szövegekben → külön tokent kapott).
- Az **„ógép”** önmagában ritka, de így ki tudja egészíteni a gyakori darabot, ha valaki leírja, hogy „számítógép”.

🔹 Másik példa: „programozni”

Darabolás:

- `"Program"`
- `"ozni"`

Itt is kettéválik, mert a modell „megtanulta”, hogy a „Program” gyakori, és az „ozni” ragszerű végződésként is előfordulhat.

🔹 Rövid szavaknál

- „kutya” → általában **1 token**, mert gyakori rövid szó.
- „asztal” → szintén 1 token.

🔹 Mit jelent ez nekünk?

Ezért van az, hogy **magyar szövegek több tokent fogyasztanak**, mint az angol szövegek:

- Angolban: „computer” → gyakori → sokszor 1 token.
- Magyarban: „számítógép” → kettő token (vagy több).