Језички модели (за српски језик)

Обучавање (различитих) модела

Композитни модели (за евалуацију и генерисање текста)

др Михаило Шкорић

Наставница пита да ли је реченица исправна.

ЧАС СРПСКОГ

-садашњост-

Од деце се очекује буловски одговор (јесте/није true/false)

ФОРМАЛНА ГРАМАТИКА

На исти начин функционише формална граматика.

Формална граматика српског језика нам говори да ли је неки улазни текст на српском, тј. да ли припада језику.

ФОРМАЛНА ГРАМАТИКА

Формална граматика неког другог језика ради то исто, али за тај други језик (нпр. француски).

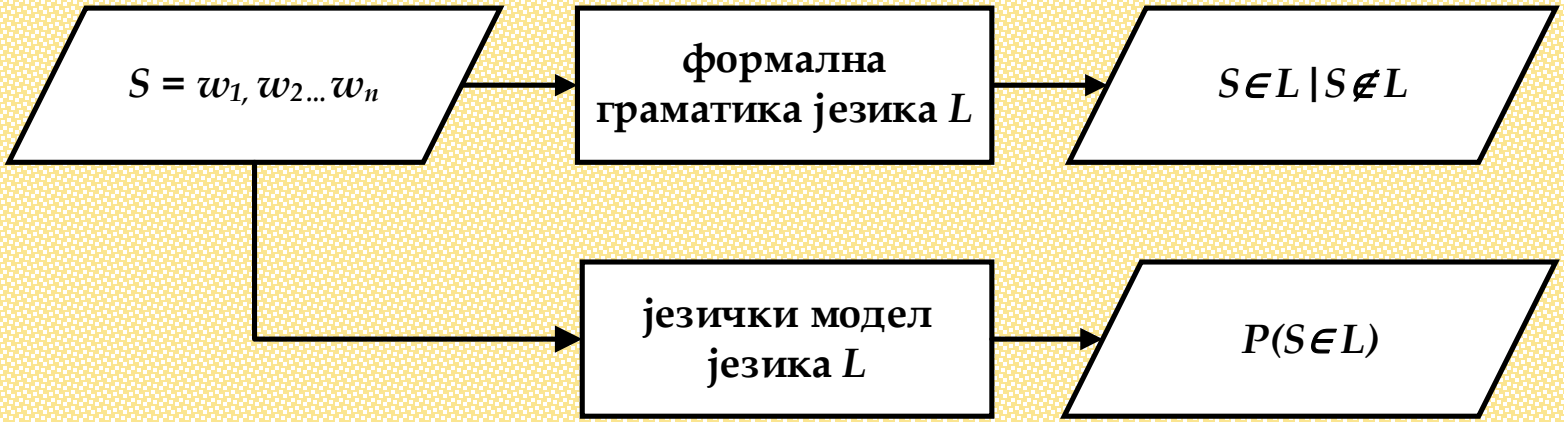
ФОРМАЛНА ГРАМАТИКА

Али шта се дешава ако је улазни текст мешовитог порекла?

ФОРМАЛНА ГРАМАТИКА

**Услед ограничења буловског одговора,
све граматике ће, вероватно, рећи да је неисправан.**

Језички модели? Граматике? Псеудограматике?



Као одговор на тај проблем, намеће језички модел.

Језички модел је расподела вероватноће над нискама текста.

ЈЕЗИЧКИ МОДЕЛ

Он одговор на питање припадности дају у виду вероватноће.

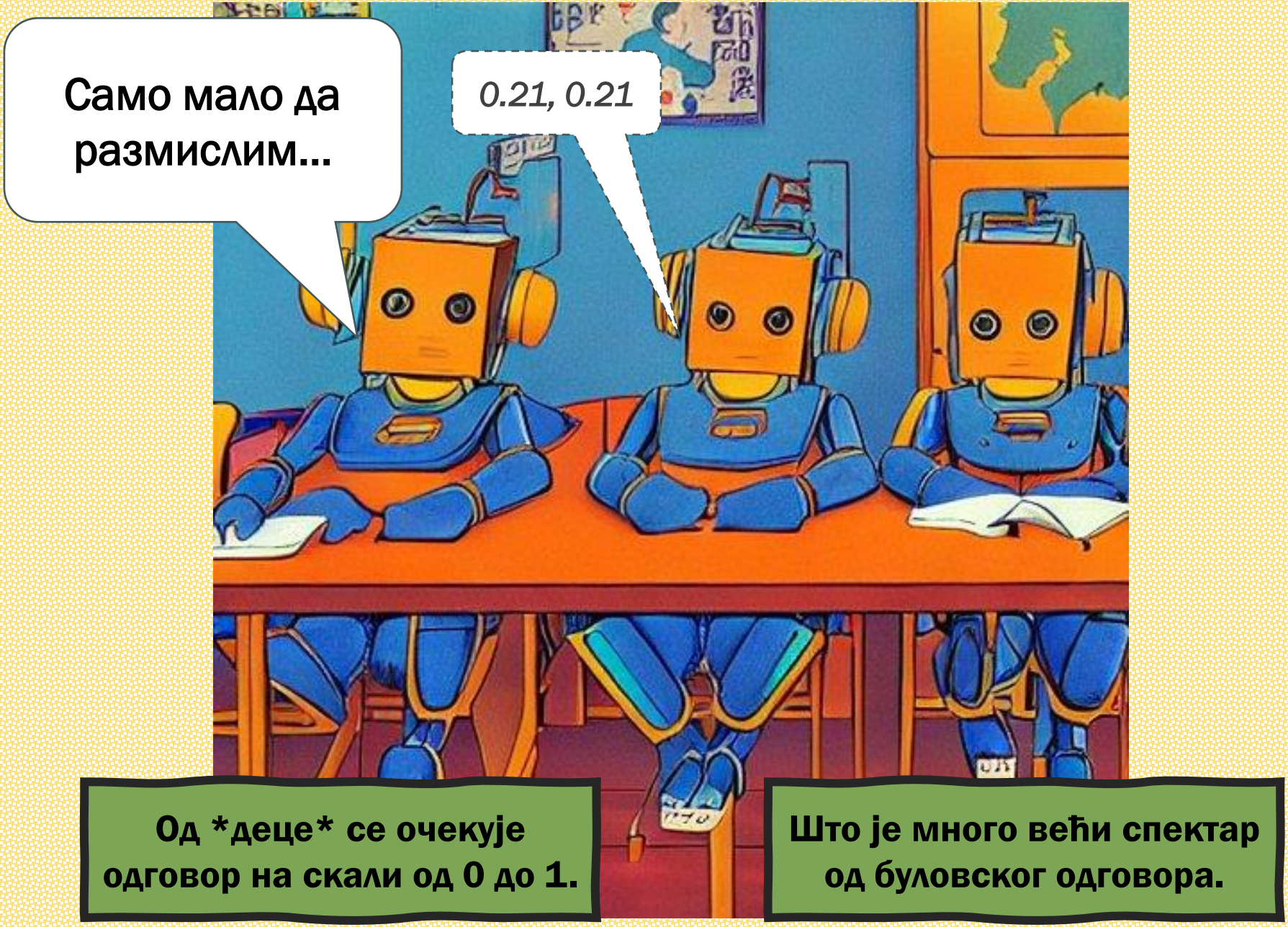
За неуобичајене ниске текста нуди ниску вероватноћу.

ЈЕЗИЧКИ МОДЕЛ

А за уобичајене, високу вероватноћу да је текст исправан.

***Наставница* пита колика је вероватноћа исправности реченице..**

ЧАС СРПСКОГ
-будућност-



Од *деце* се очекује одговор на скали од 0 до 1.

Што је много већи спектар од буловског одговора.

***Језички модели додељују нискама текста вероватноће на основу корпуса за обучавање*, што не одговара нужно расподели вероватноћа у природном језику.**

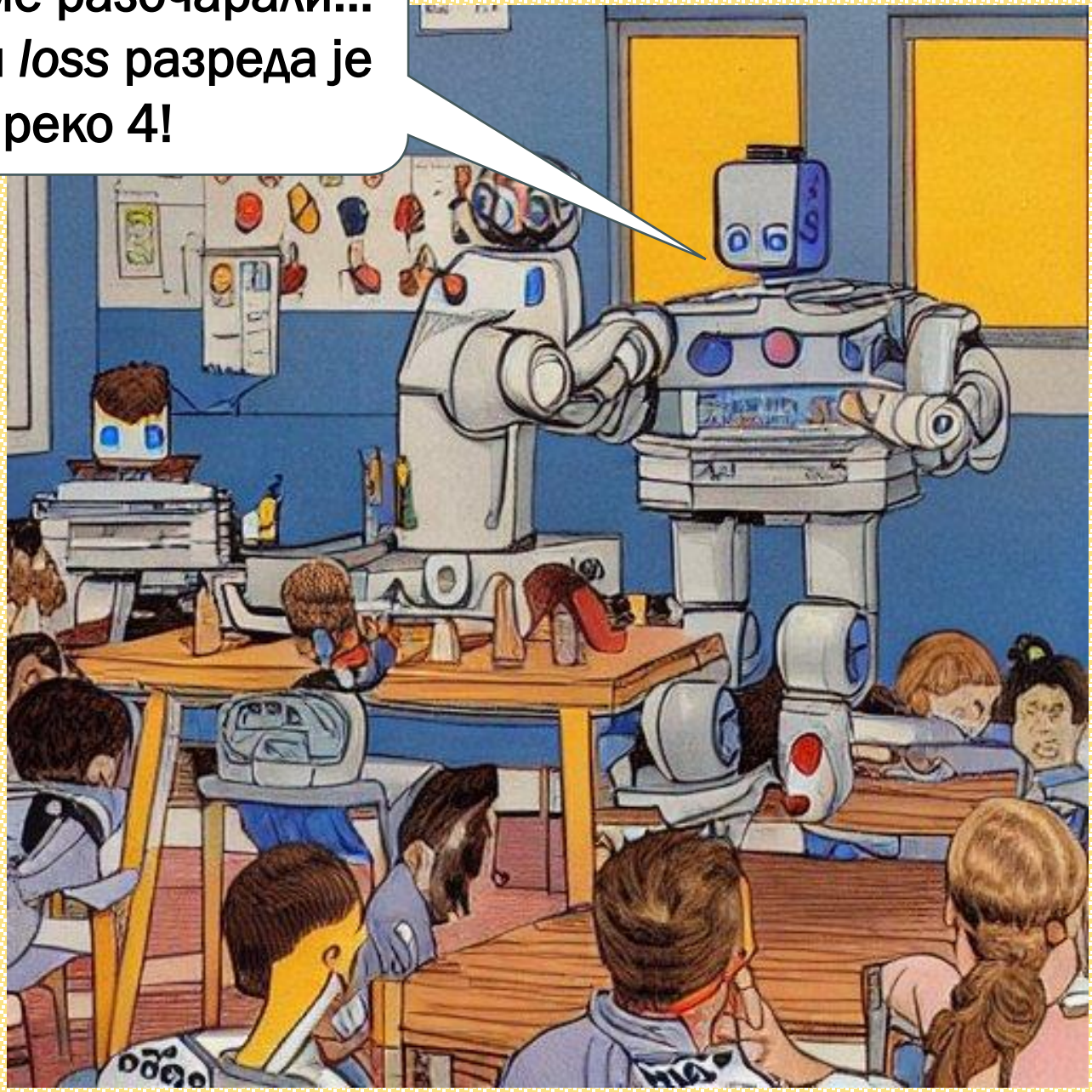
Оцењивање из српског у будућности??

x – вектор "исправних" вероватноћа реченица
y – вектор вероватноћа која је ученик исписао на тесту

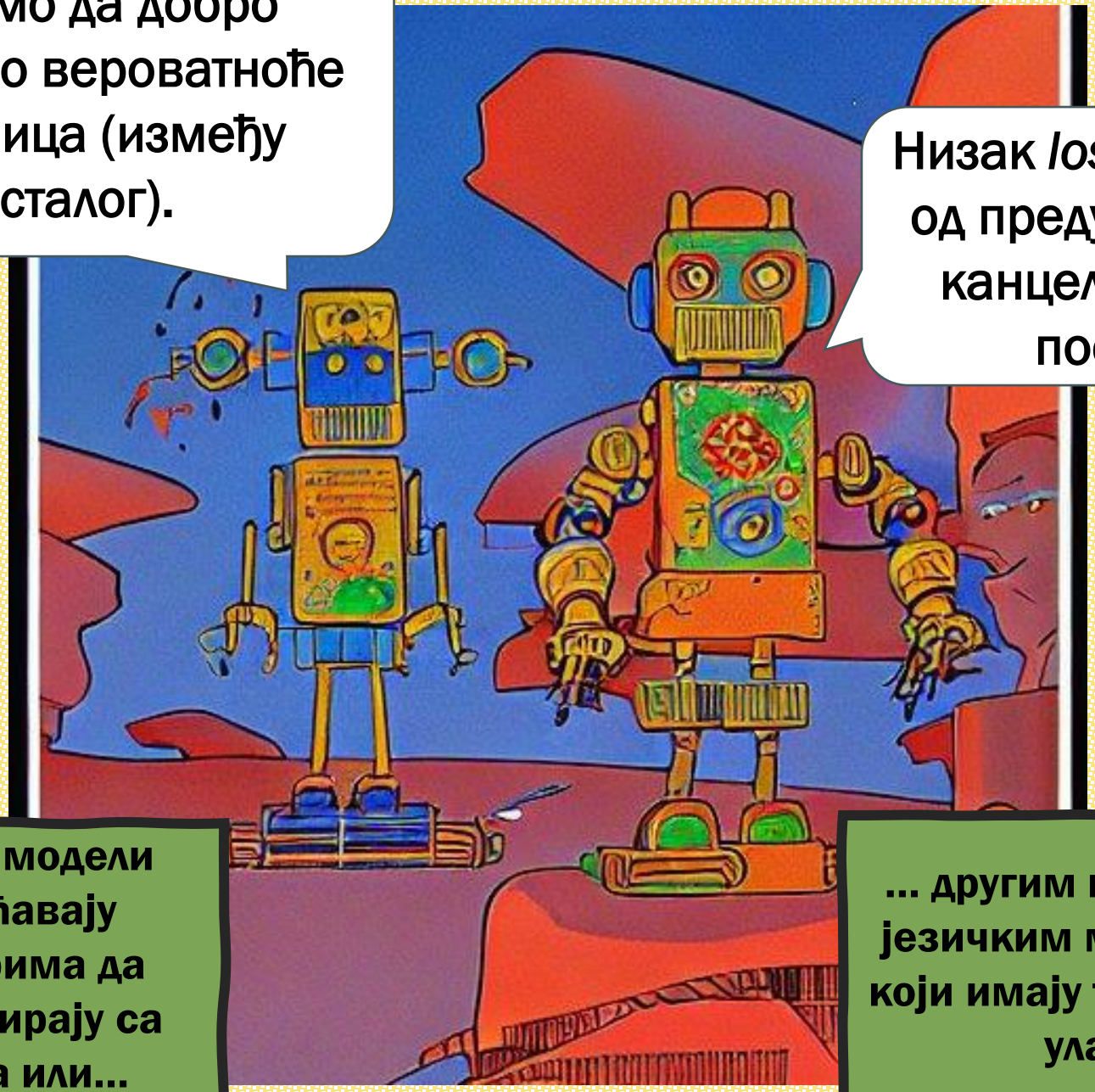
$$loss(x, y) = -\sum_{i=1}^{n} x_i \log y_i$$

$loss < 2$ је, на пример, потребан за петицу.

Јако сте ме разочарали...
Просечни *loss* разреда је преко 4!



Желимо да добро погађамо вероватноће реченица (између осталог).



Низак *loss* је један од предуслова за канцеларијски посао.

Језички модели омогућавају рачунарима да комуницирају са људима или...

... другим модерним језичким моделима, који имају текстуални улаз.

Шта је потребно за обучавање језичких модела?

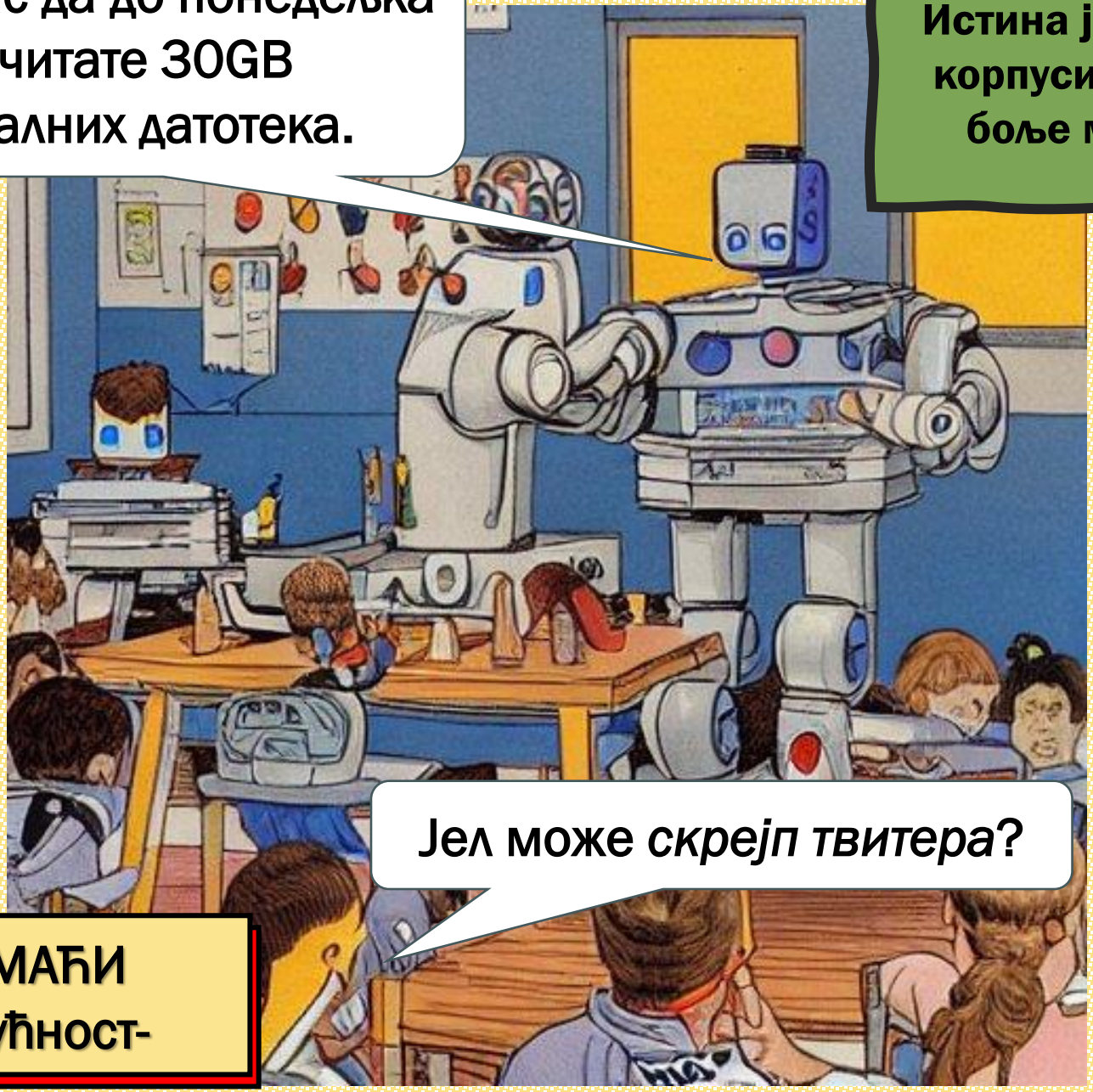
1. Корпус квалитетног текста
   (у што већем квантитету)

2. Рачунарски ресурси
   (што моћнији и што непрекидније)

3. Софтвер?

Молим вас да до понедељка прочитате 30GB текстуалних датотека.

Истина је да већи корпуси узрокују боље моделе.

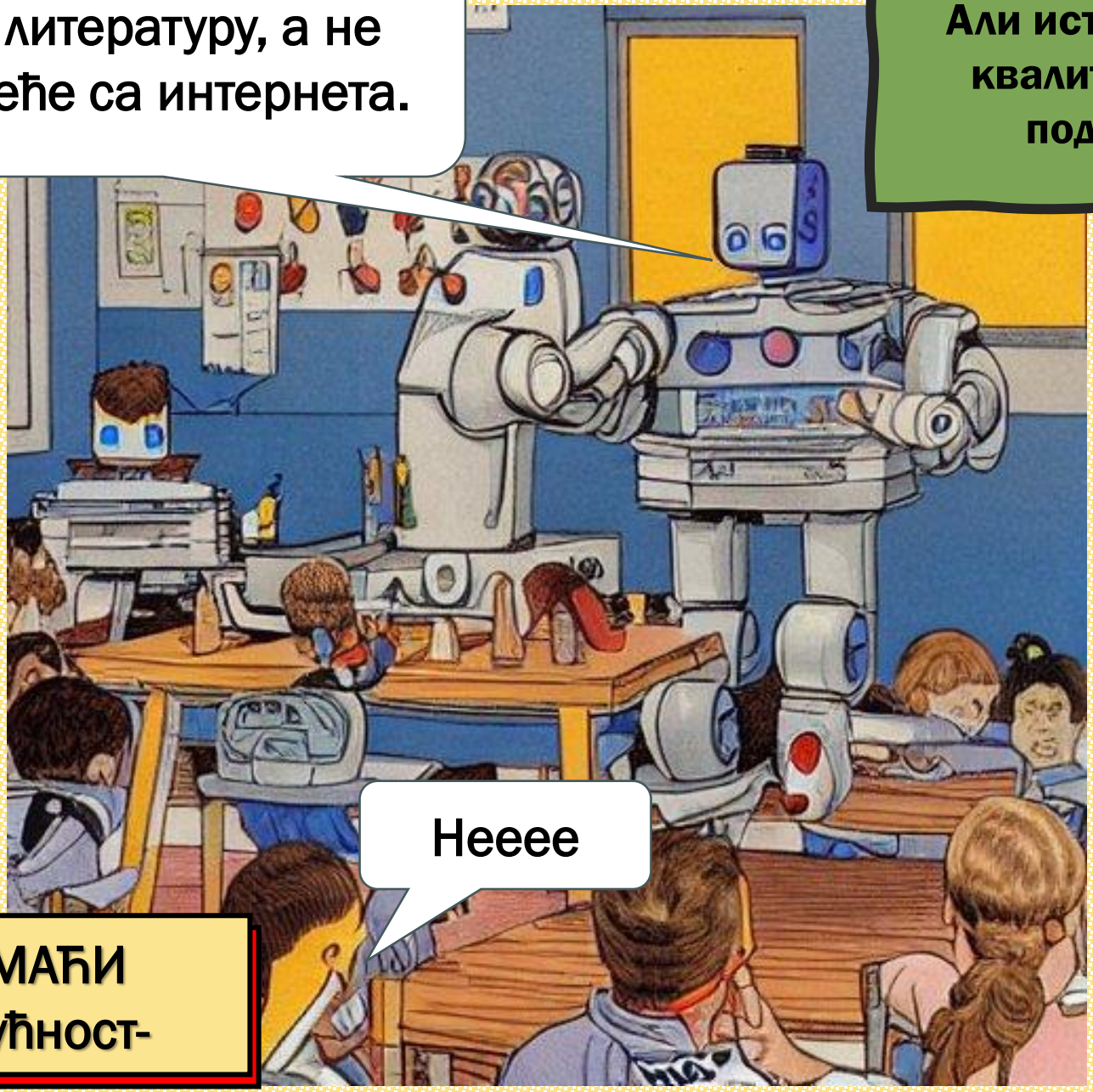


Јел може *скрејп твитера*?

ДОМАЋИ
-будућност-

Читајте литературу, а не само смеће са интернета.

Али исто тако и квалитетнији подаци.



ДОМАЋИ
-будућност-

За обучавање се може користити било који рачунар, под условом да модел и скуп за обучавање могу да *стану* у активну меморију.

Ипак, добра графичка картица ће омогућити да се то оствари много брже (уколико се испуне одређени услови).

Препорука за софтвер?

1. **Трансформерска архитектура**
   (уколико желите најбоље резултате);

2. **Пајтон библиотека *transformers***
   (лична препорука за обучавање и коришћење трансформерских модела);

3. ***Scratch2LM*, мој пројекат на *GitHub*-u за обучавање и дообучавање језичких модела (уколико не желите да кренете од почетка).**

Трансформерска архитектура?

1. Тренутно *state-of-the-art* моделирања језика (сви најбољи модели су засновани на њој);

2. Комплексна неуронска мрежа са милионима или чак милијардама параметара (углавном, више параметара—бољи резултати);

3. Заснива се на енкодеру (који векторизује улазни текст) и декодеру (који генерише нови, излазни текст)
(*АЛИ НЕ НУЖНО.* Многобројни модели засновани су само на енкодеру, или само на декодеру).

ЕНКОДЕР МОДЕЛИ
излаз је вектор

Енкодер за сваку ниску текста враћа одговарајући вектор, који је представља.

Енкодер модели?

- На излазу производе векторе;
- Обучавају се (обично) над листом реченица;
- Дообучавају се (засновано на излазним векторима) на задацима обележавања речи, класификације реченица, али и допуњавања недостајућег текста;
- Најпознатији примерци: BERT, RoBERTA, ELECTRA...
- Најбољи модели за српски на платформи *huggingface*: xlm-roberta? BERTić? SRoBERT-a-F? (сви наведени модели су мање или више вишејезични)*

*Ускоро се очекује објављивање модела Bertović-base и Bertović-large, обучаваних специјално за српски језик.

ДЕКОДЕР МОДЕЛИ
излаз је наставак улаза

Декодер модели за улазну ниску текста враћају један или више вероватних наставака.

Декодер модели?

- Обучавају се над листом н-грама токена (обично текст исцепкан на једнаке делове);
- На излазу производе наставак текста;
- Дообучавају се за генерисање специфичних стилова (рачунарски код, хаику песме...);
- Најпознатији примерци: GPT(2,3,4...), LLaMA, PALM (ChatGPT и BARD врло вероватно нису декодер модели)
- Најбољи модели за српски: sr-gpt2, gpt2-srlat и sr-gpt2-large, највећи декодер модел за српски са преко 700 милиона параметара.

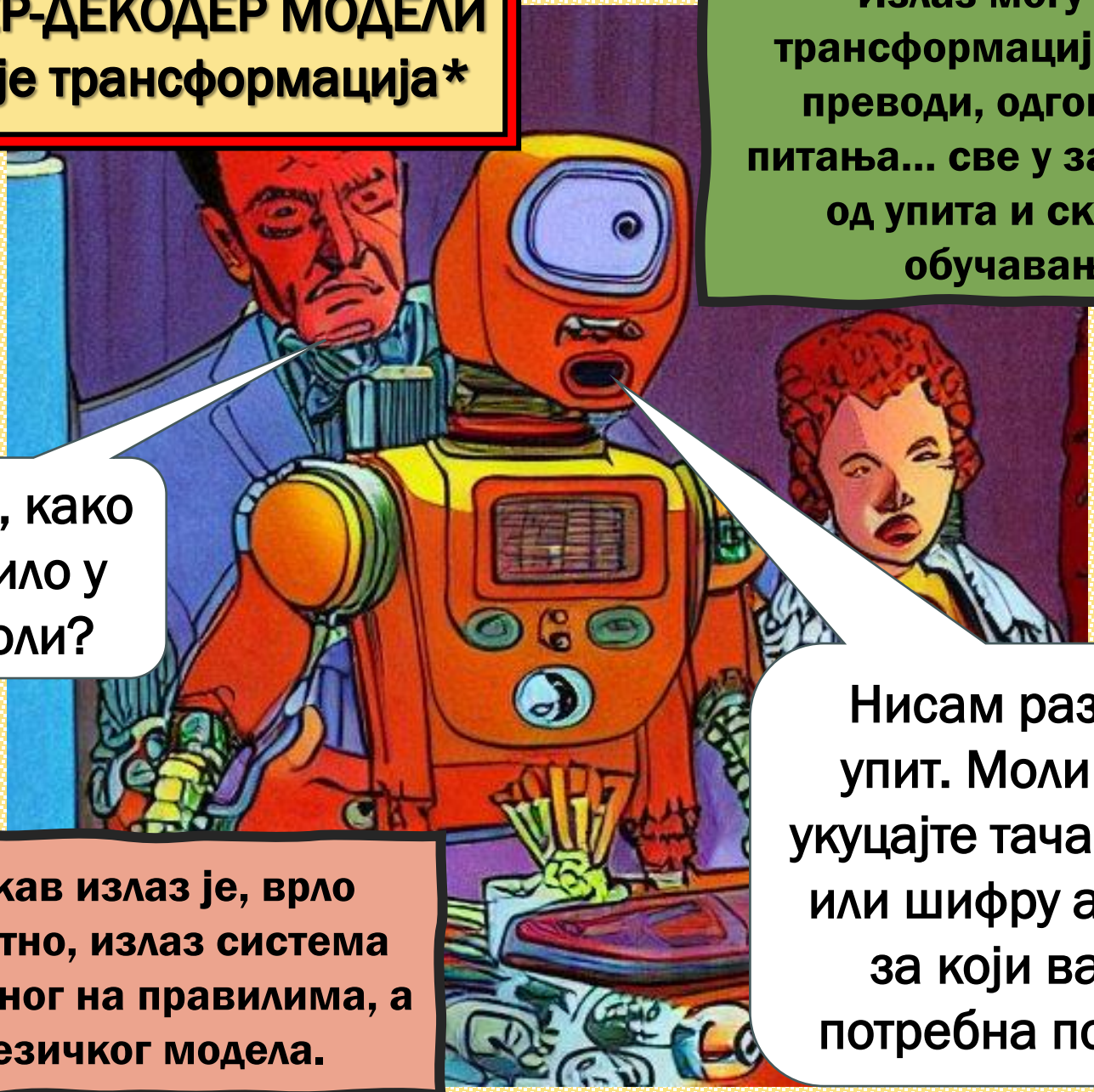
ЕНКОДЕР-ДЕКОДЕР МОДЕЛИ
излаз је трансформација

Најкомплекснији модели.

Генеришу излаз на основу улаза.

ЕНКОДЕР-ДЕКОДЕР МОДЕЛИ
излаз је трансформација

Излаз могу бити трансформације текста, преводи, одговори на питања... све у зависности од упита и скупа за обучавање.



***Овакав излаз је, врло вероватно, излаз система заснованог на правилима, а не језичког модела.**

Енкодер-Декодер модели?

- Обучавају се над паровима реченица;
- На излазу производе нови текст;
- Користе се за:
  - машинско превођење
  - одговарање на питања
  - сажимање текста
  - трансфер стила...
- Најпознатији примерци: BART, T5, Galactica (ChatGPT и BARD су вероватно овог типа)
- За српски тренутно нема познатих модела...

МОДЕЛИ ЗА СРПСКИ НАСПРАМ ВИШЕЈЕЗИЧНИХ МОДЕЛА

	bertovic	srberta	maced.	sroberta	sroberta-x	xlm
bertovic-base	**0.729**	0.591	**0.610**	**0.628**	**0.705**	**0.572**
JelenaTosic/SRBerta	0.408	**0.644**	0.496	0.377	0.408	0.426
maced./sr-roberta-base	0.318	0.387	0.454	0.302	0.317	0.326
Andrija/SRoBERTa	0.279	0.219	0.222	0.275	0.276	0.246
Andrija/SRoBERTa-base	0.410	0.352	0.358	0.378	0.429	0.365
Andrija/SRoBERTa-L	0.573	0.486	0.500	0.534	0.629	0.485
Andrija/SRoBERTa-F	0.647	0.576	0.592	0.593	0.702	0.544
Andrija/SRoBERTa-XL	0.609	0.535	0.558	0.559	0.660	0.518
xlm-roberta-base	0.465	0.395	0.407	0.470	0.462	0.503
xlm-roberta-large	0.515	0.436	0.453	0.513	0.507	0.545

Тачност различитих модела (уз различите токенизаторе) на задатку допуњавања недостајућег текста у роману *Дечко*.

ВИШЕЈЕЗИЧНИ МОДЕЛ

Један од разлога за обучавање модела за српски језик.

Моје обучавање језичких модела за српски

Састав корпуса:

1. СрпКор2013 (50%)
2. СрпКор2021 (20%)
3. ВикиКорпус (28%)
4. СрпЕЛТеК (2%)

Корпуси квалитетног текста

Рачунарски ресурси:

Серверске машине са Nvidia графичким картицама:

- RTX 2060 6gb
- RTX 3060 12gb
- RTX 4090 24gb

Постепено унапређење ресурса

Обучени модели:

gpt2-srlat – основни GPT2 модел

gpt2-srlat-sem – семантички модел

gpt2-srlat-synt – синтаксички модел

+ bertović-base (енкодер модел обучен на истом скупу података)

Обучени различити модели над различитим репрезентацијама корпуса.

ОСНОВНИ ГПТ МОДЕЛ

Трансформерски модели одашиљају њихову изненађеност улазном ниском (перплексност).

Контролни примерак, предефинисана подешавања.

СЕМАНТИЧКИ ГПТ МОДЕЛ

Ово нема никаквог смисла.
перплексност = 492

Перплексност је реципрочна вероватноћа.

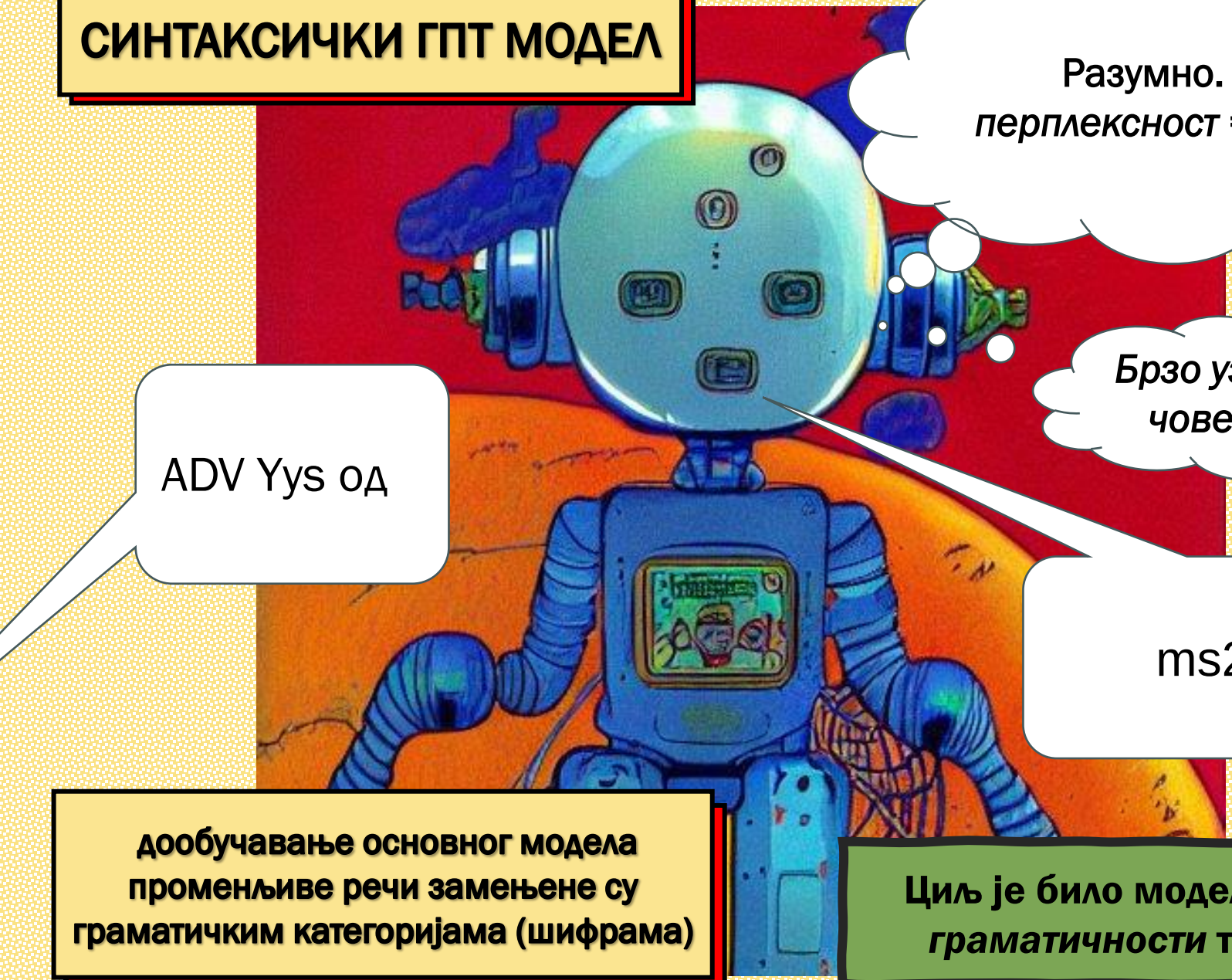
дугме река одело брати

пар нојев перје насеље

дообучавање основног модела непроменљиве речи уклоњене променљиве замењене лемама

Циљ је било моделирање семантичког аспекта.

СИНТАКСИЧКИ ГПТ МОДЕЛ



Разумно.
перплексност = 54

Брзо узми од човека. ?

ADV Yys од

ms2v.

дообучавање основног модела променљиве речи замењене су граматичким категоријама (шифрама)

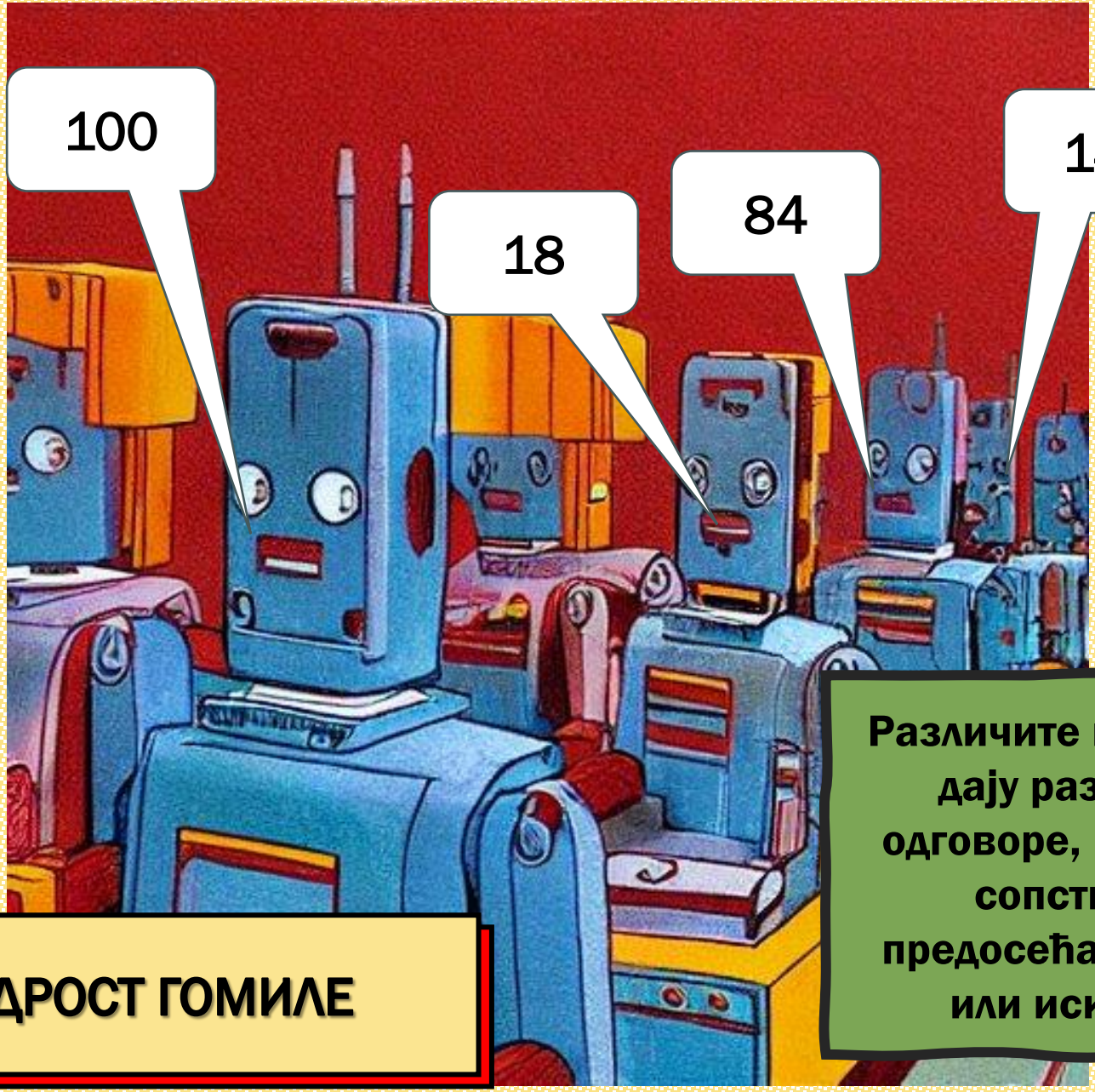
Циљ је било моделирање *граматичности* текста.

Мудрост гомиле и композитни језички модели

Композитни језички модели се заснивају на претпоставци да више глава боље размишљају од једне.

Један од популарнијих експеримената је давање групи људи да погоде колико слаткиша има у тегли.

МУДРОСТ ГОМИЛЕ

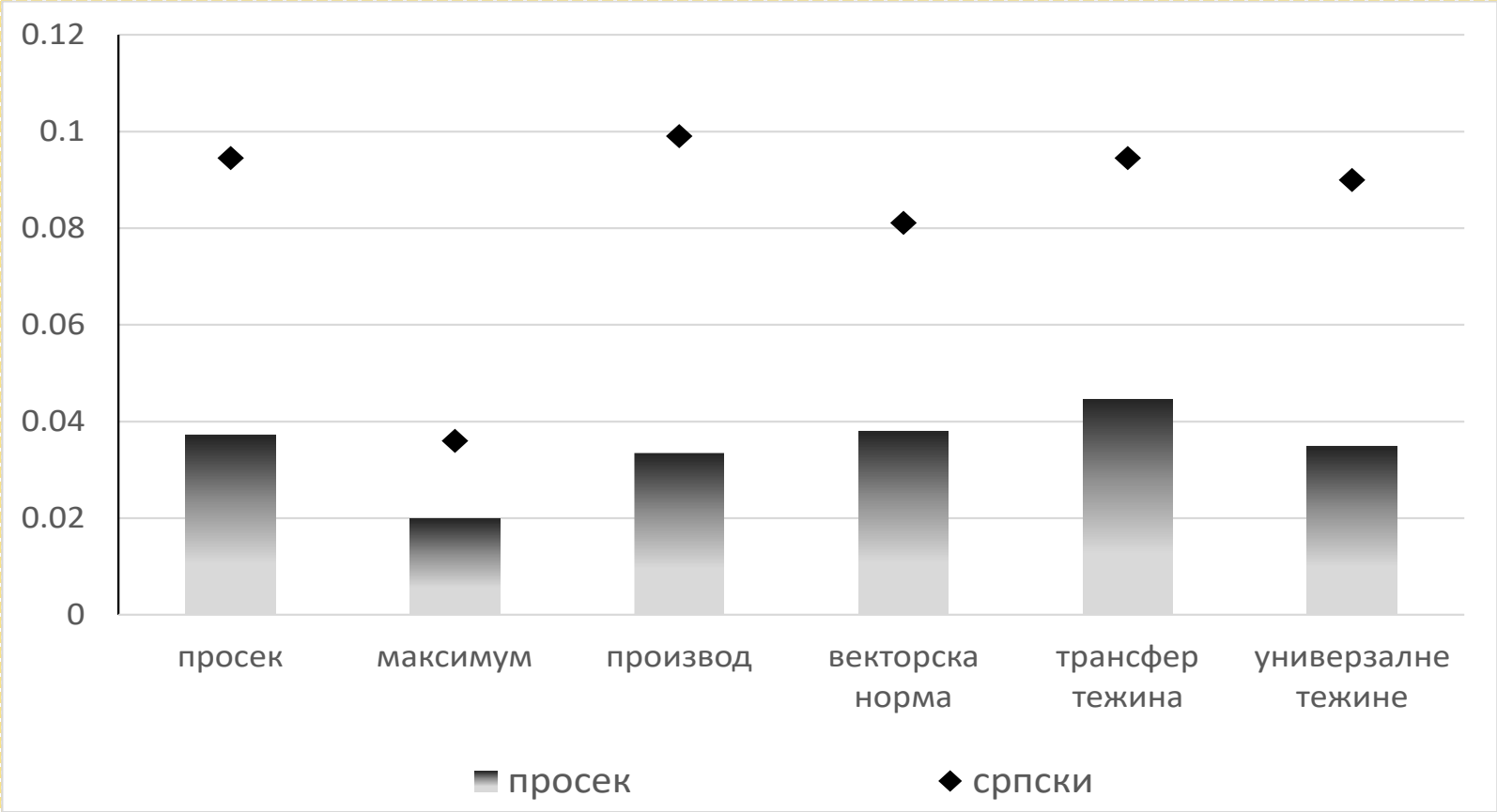


МУДРОСТ ГОМИЛЕ

Различите индивидуе дају различите одговоре, на основу сопственог предосећаја, логике или искуства.

МУДРОСТ ГОМИЛЕ

Ово је показано као делотворно у моделовању језика (и то поготово српског) на пробном експерименту.

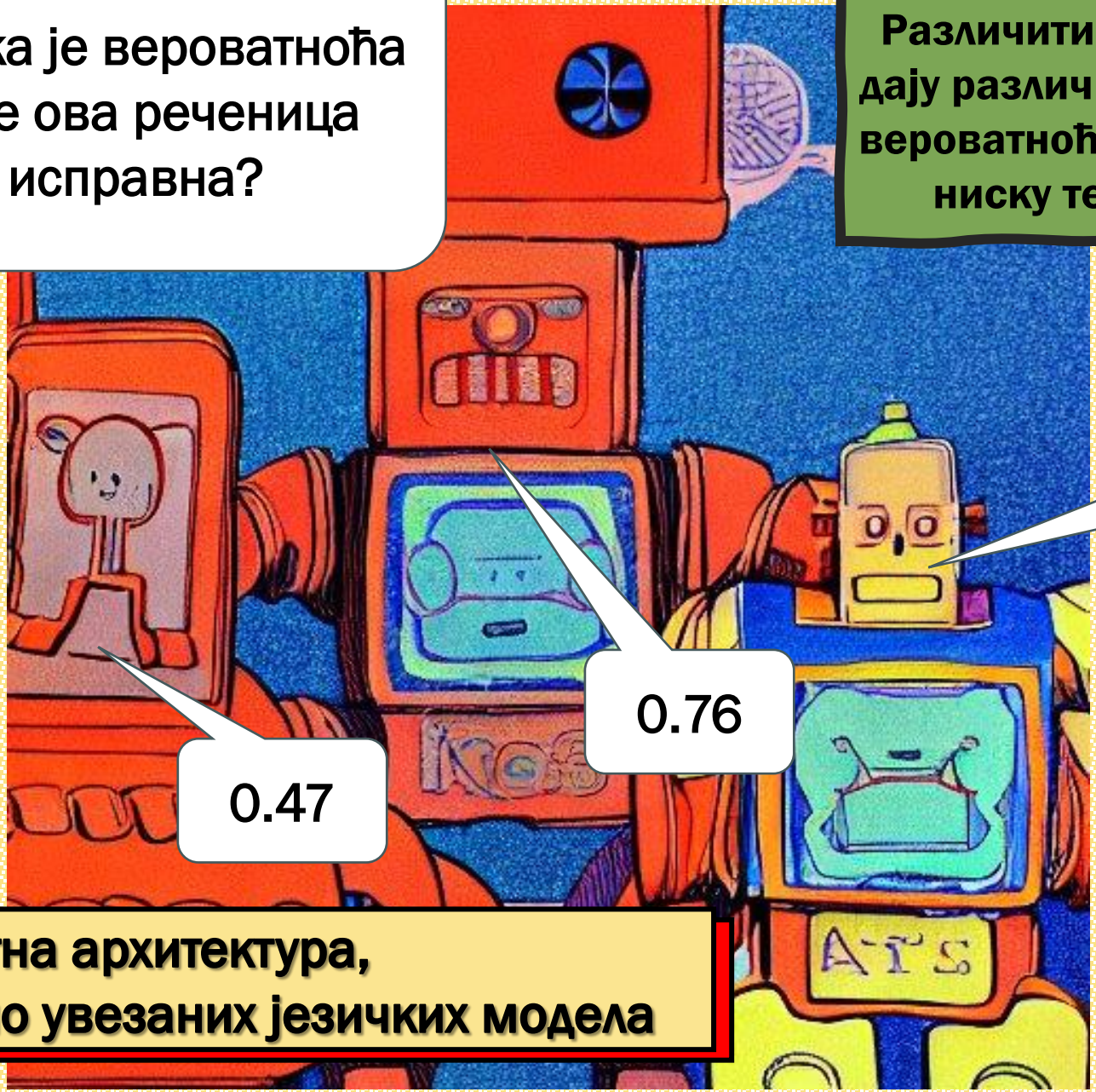


Проценти унапређења добијеног коришћењем композитне архитектуре за седам европских језика, укључујући и српски.*

*Škorić, Mihailo, et al. "Parallel stylometric document embeddings with deep learning based language models in literary authorship attribution." Mathematics 10.5 (2022): 838.

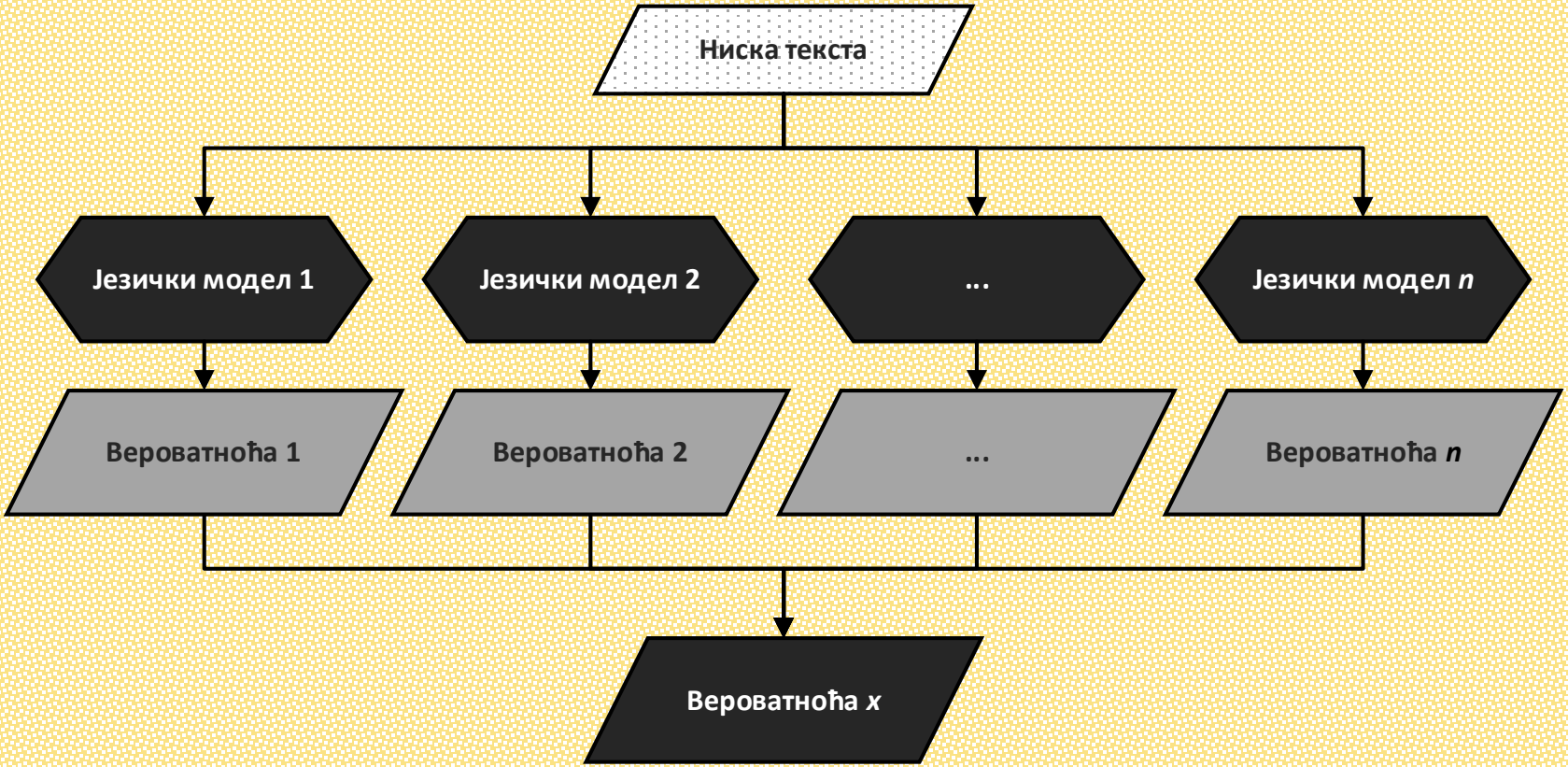
Колика је вероватноћа да је ова реченица исправна?

Различити модели дају различите мере вероватноће за исту ниску текста.

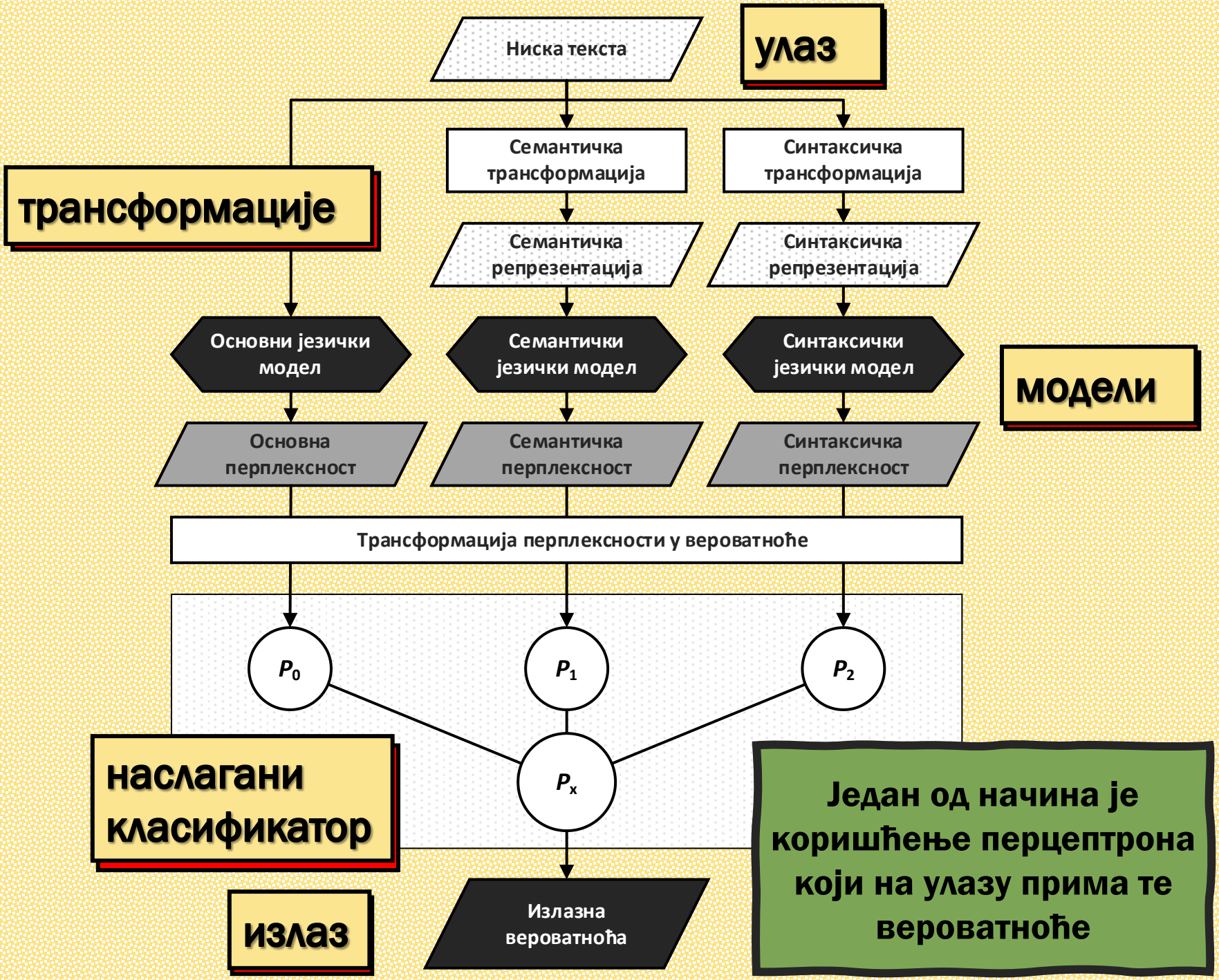


композитна архитектура, паралелно увезаних језичких модела

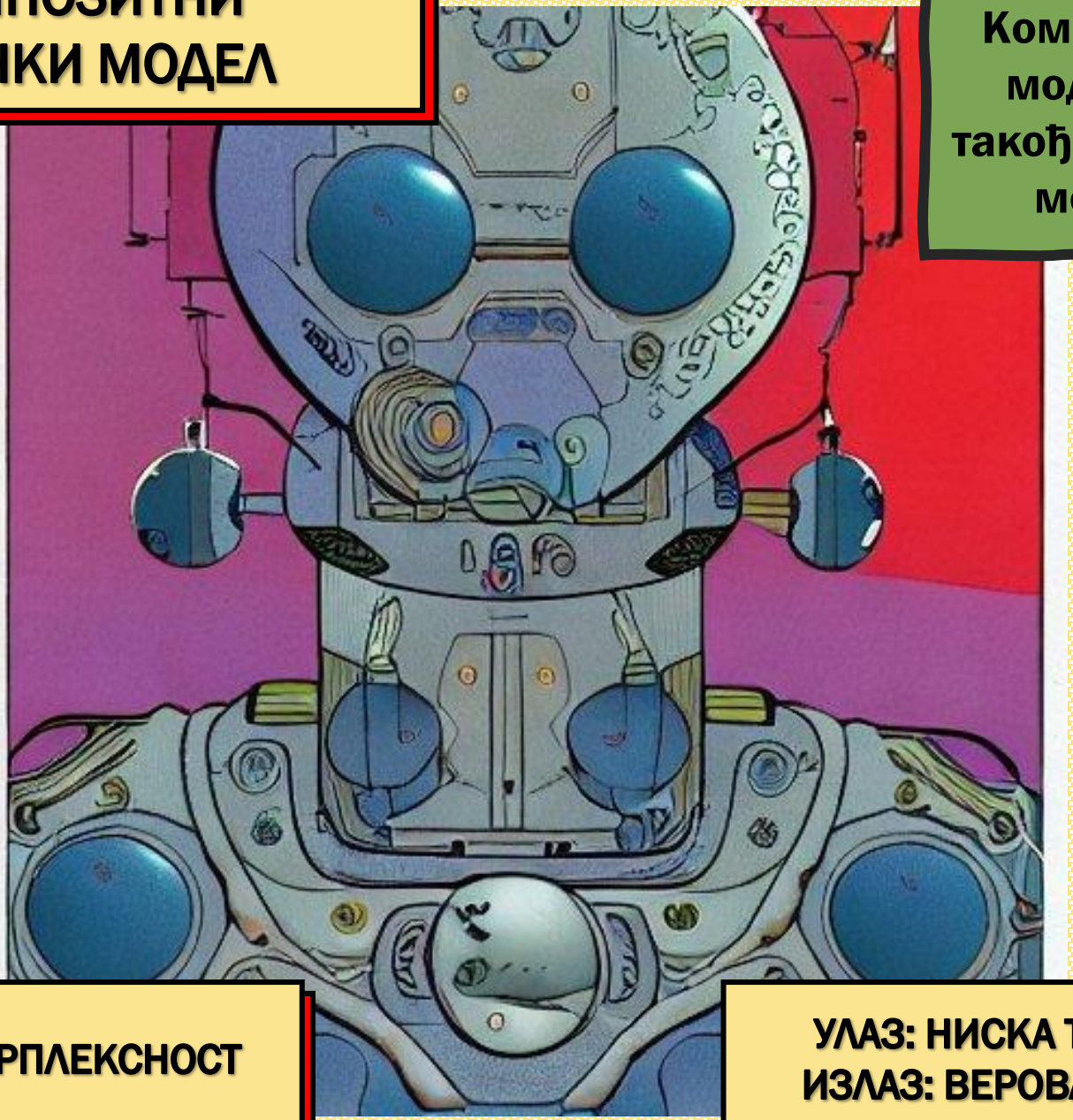
композитна архитектура, паралелно увезаних језичких модела



А онда се те вероватноће на неки начин збрајају.


Ниска текста
улаз
трансформације
Семантичка трансформација
Синтаксичка трансформација
Семантичка репрезентација
Синтаксичка репрезентација
Основни језички модел
Семантички језички модел
Синтаксички језички модел
модели
Основна перплексност
Семантичка перплексност
Синтаксичка перплексност
Трансформација перплексности у вероватноће
P_0
P_1
P_2
P_x
наслагани класификатор
излаз
Излазна вероватноћа
Један од начина је коришћење перцептрона који на улазу прима те вероватноће

КОМПОЗИТНИ ЈЕЗИЧКИ МОДЕЛ



Композитни модели су такође језички модели

МЕДИЈ: ПЕРПЛЕКСНОСТ

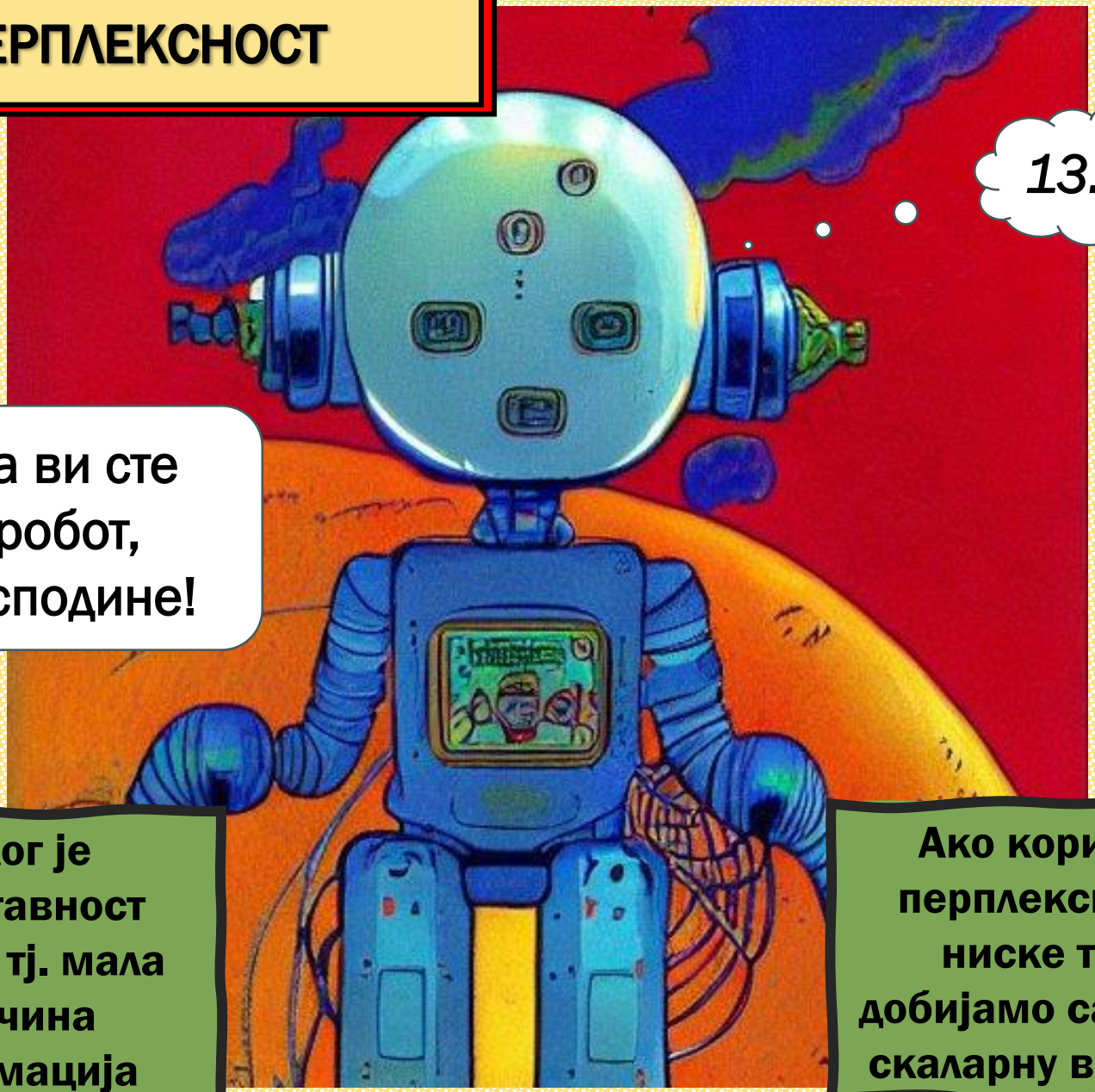
УЛАЗ: НИСКА ТЕКСТА
ИЗЛАЗ: ВЕРОВАТНОЋА

Унапређења која композитни модел даје у односу на основни, на различитим задацима.

задатак	проценат повећања тачности	проценат смањења грешке
детекција синтаксички неисправних реченица (према облицима речи)	3.15	18.00
детекција синтаксички неисправних реченица (према редоследу речи)	0.46	3.67
детекција семантички неисправних реченица	0.21	0.86
разликовање синтаксички и семантички неисправних реченица (први тест)	10.10	23.61
разликовање синтаксички и семантички неисправних реченица(други тест)	9.98	28.22
детекција неисправних реченица уопштено	/	/
разликовање експертских и машинских превода	6.00	6.44

На последња два задатка модел не конвергира.

ПЕРПЛЕКСНОСТ



Разлог је једноставност модела, тј. мала количина информација

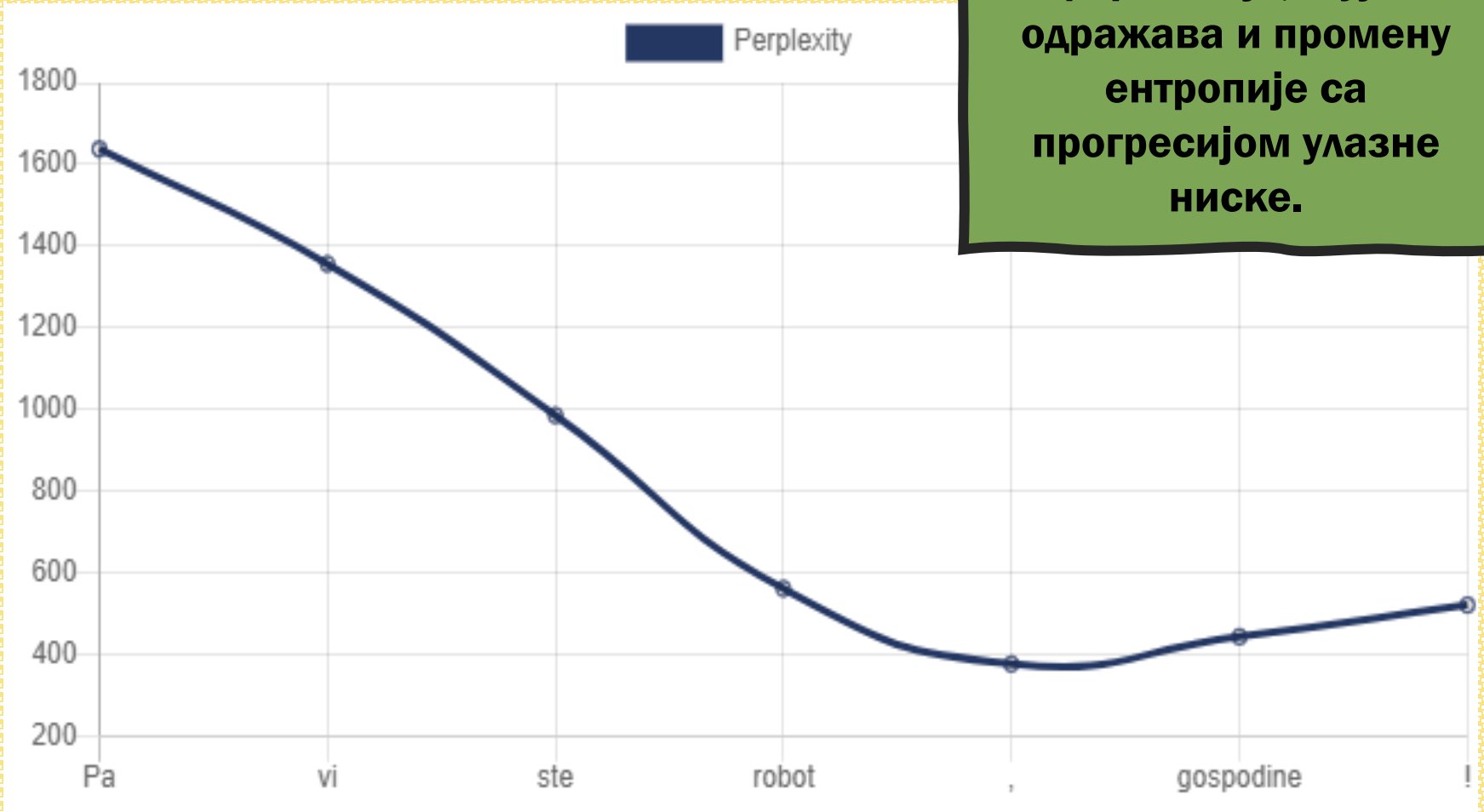
Ако користимо перплексност, од ниске текста добијамо само једну скаларну вредност..

ВЕКТОР ПЕРПЛЕКСНОСТИ

Количину информација можемо повећати бомбардовањем модела н-грамима текста, уместо једне ниске.

ВЕКТОР ПЕРПЛЕКСНОСТИ

Добијени *вектор перплексности* носи већу количину информација, а уједно одражава и промену ентропије са прогресијом улазне ниске.



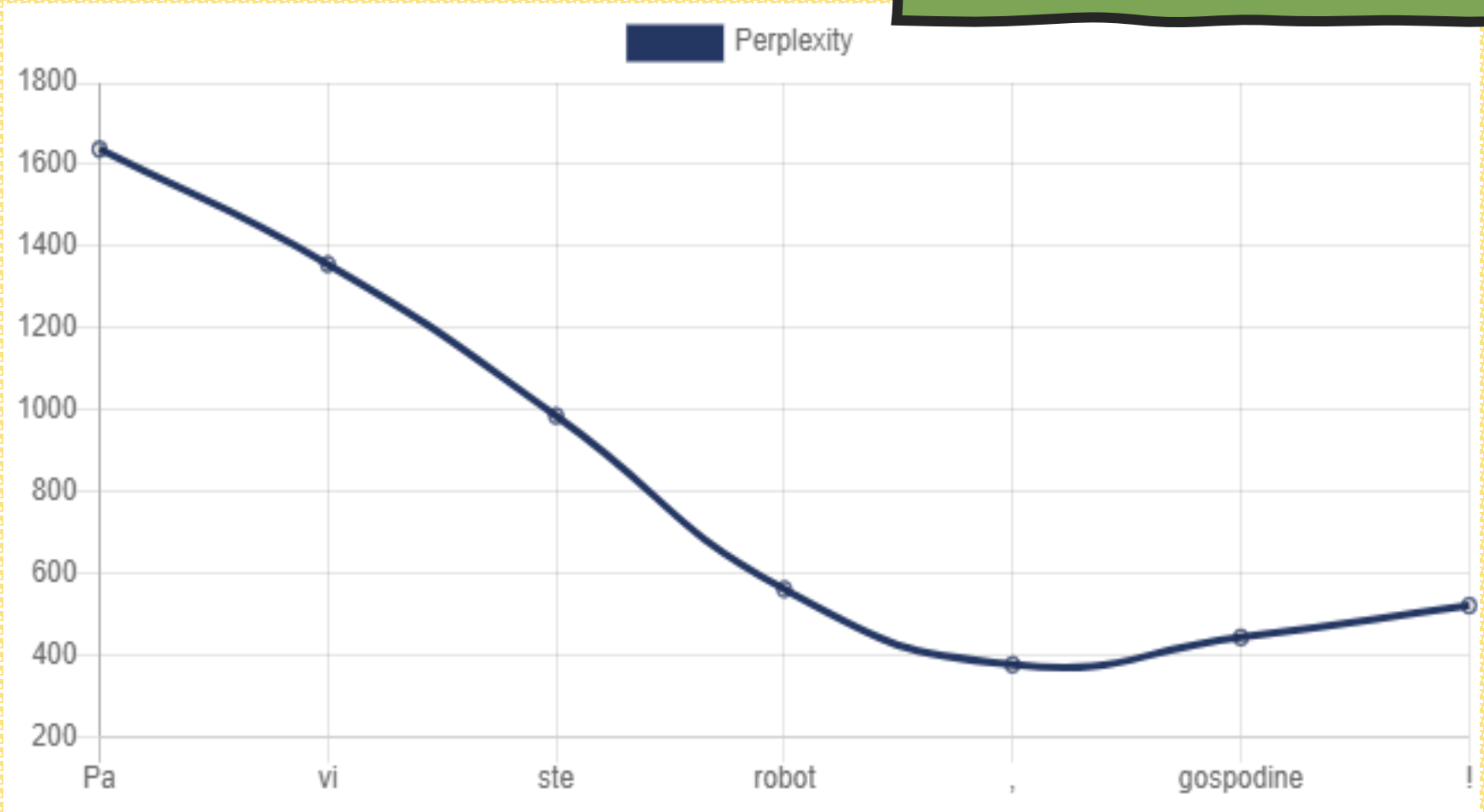
Што омогућава детекцију потенцијалних грешака у тексту.

задатак	насумично	Основни модел	Семантички модел	Синтаксички модел
детекција избрисане речи	0.058	**0.104**	0.053	0.076
детекција уметнуте речи	0.031	**0.173**	0.036	0.068
детекција замењене речи	0.020	**0.186**	0.030	0.060

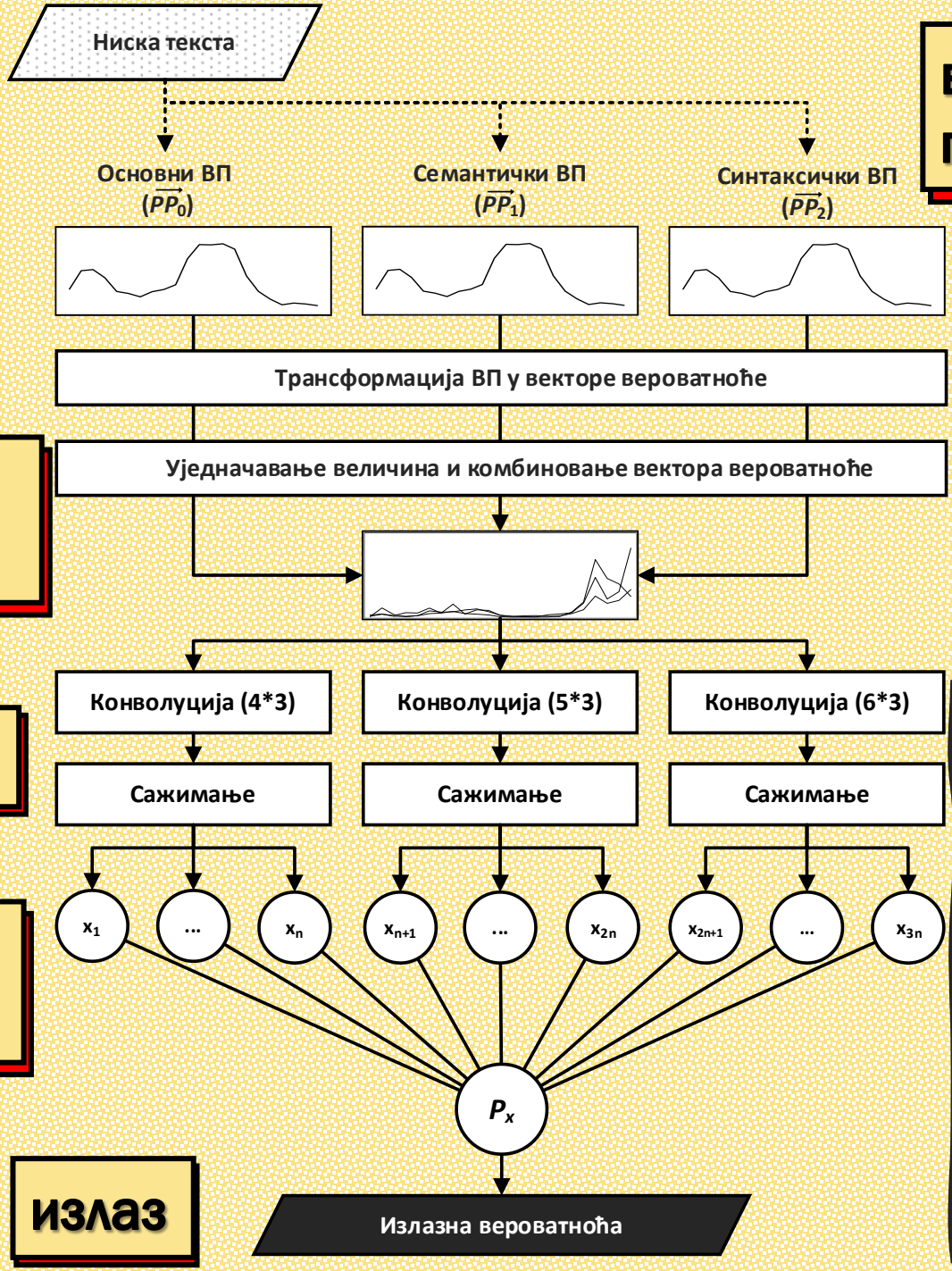
При чему се тачност на овом задатку повећава и до 9 пута у односу на насумични одабир.

ВЕКТОР ПЕРПЛЕКСНОСТИ

Вектори омогућавају обучавање комплекснијих модела попут коволуционих неуронских мрежа.



улаз



вектори перплексности

композитни вектор

конволуција

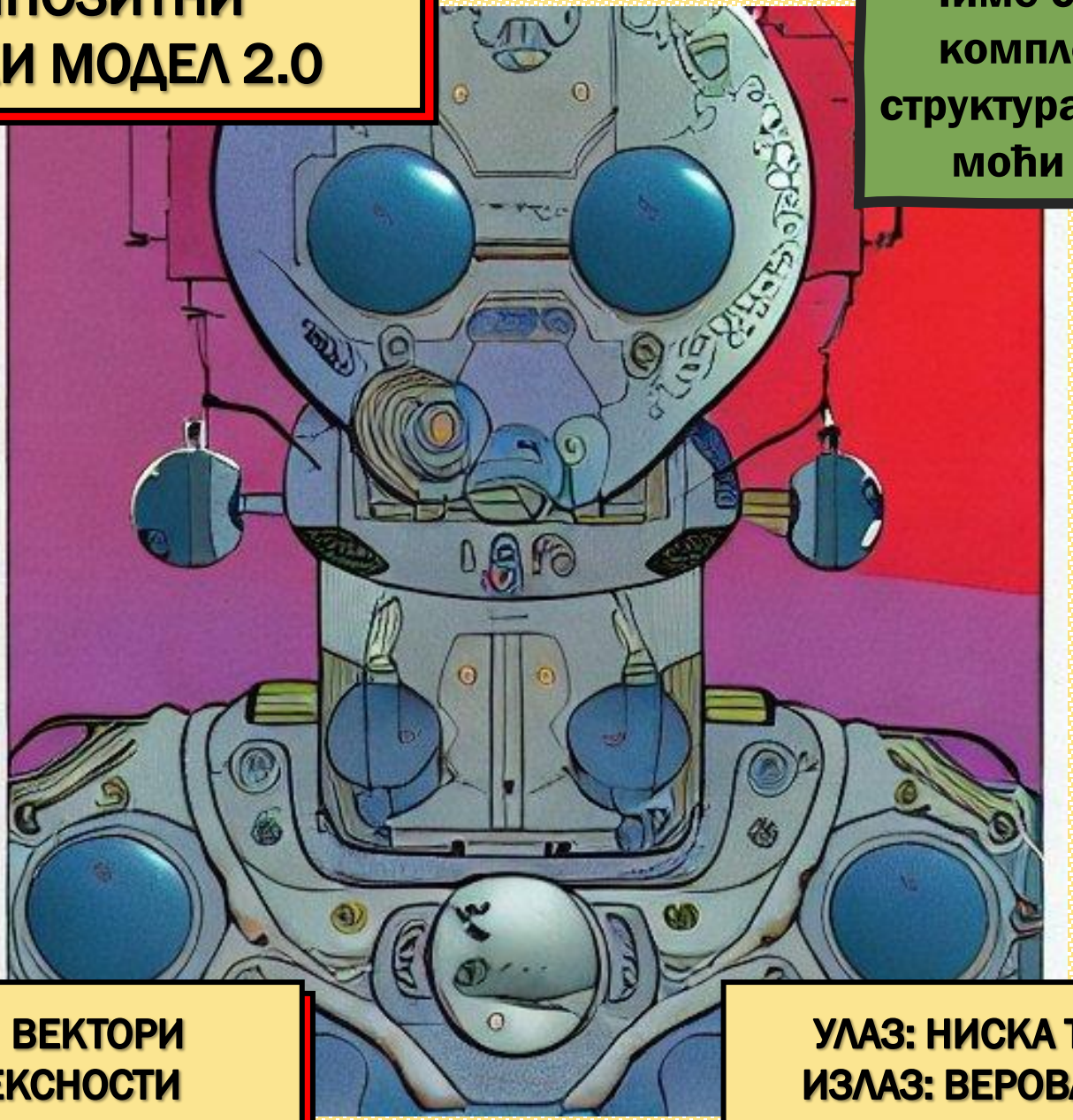
наслагани класификатор

излаз

Вектори омогућавају обучавање комплекснијих модела попут рекуретних или коволуционих неуронских мрежа.

КОМПОЗИТНИ ЈЕЗИЧКИ МОДЕЛ 2.0

Чиме се добија комплекснија структура са већом моћи учења.



МЕДИЈ: ВЕКТОРИ ПЕРПЛЕКСНОСТИ

УЛАЗ: НИСКА ТЕКСТА
ИЗЛАЗ: ВЕРОВАТНОЋА

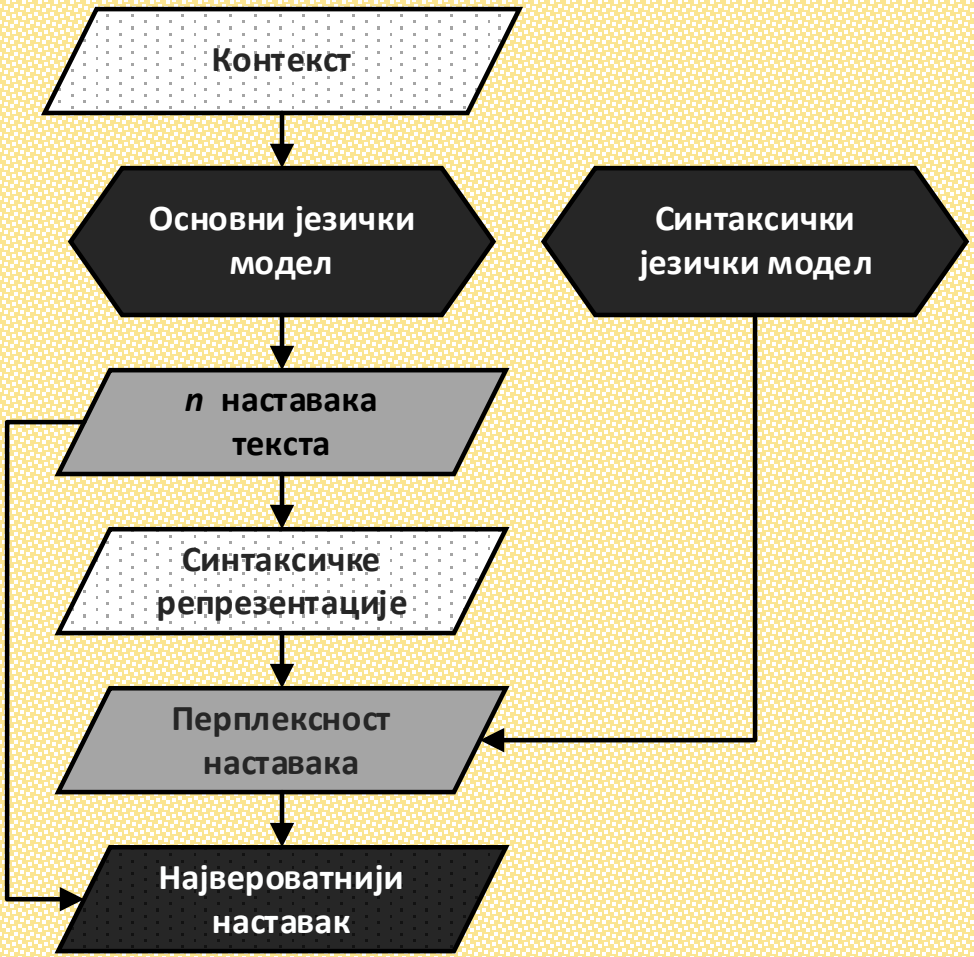
Која може да савлада комплексније задатке.

задатак	проценат повећања тачности	проценат смањења грешке
детекција неисправних реченица уопштено	9.89	48.24
разликовање експертских и машинских превода	8.74	9.67

КОМПОЗИТНО ГЕНЕРИСАЊЕ ТЕКСТА

Што се тиче генерисања текста коришћењем више обучених модела за српски...

...најбоље резултате постиже комбинација основног и синтаксичког модела.



Основни модел генерише *n* наставака текста...

Основни модел

Синтаксички модел

...а синтаксички каже који од наставака је најисправнији.

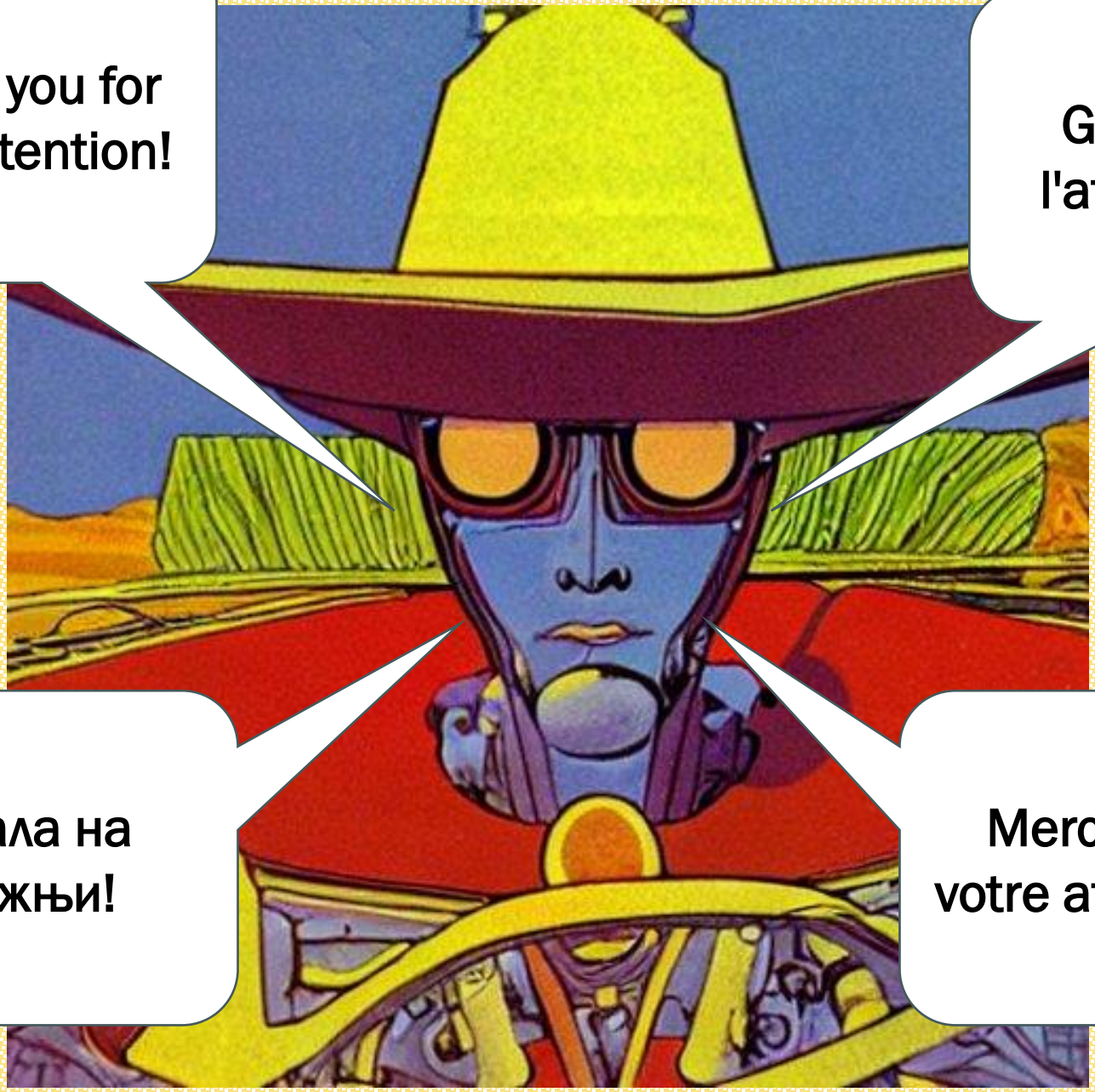
КОМПОЗИТНО ГЕНЕРИСАЊЕ ТЕКСТА

Suma sumarum...

И ТАКО...

Thank you for your attention!

Grazie per l'attenzione!

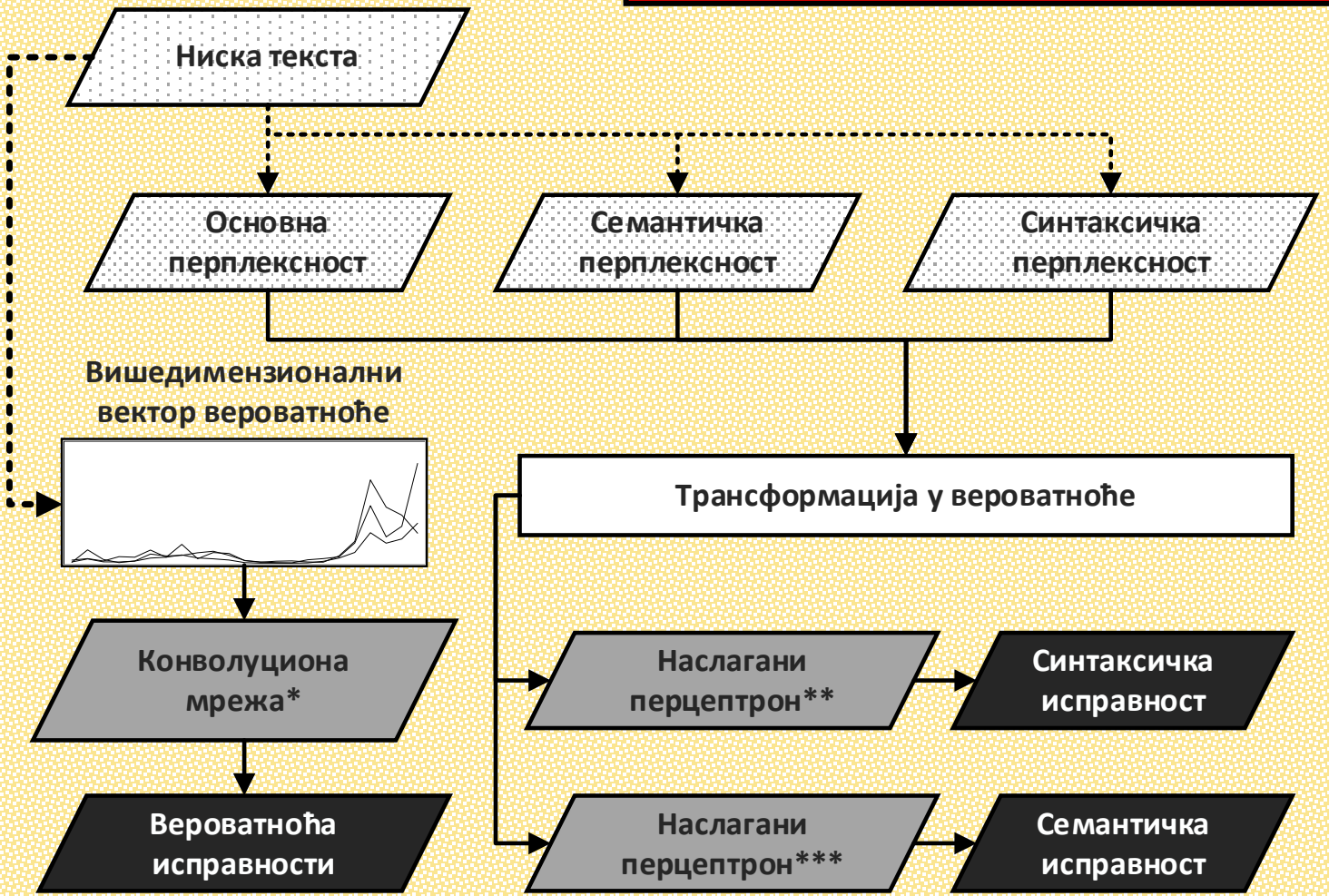


Хвала на пажњи!

Merci pour votre attention!

PS:

МУЛТИМОДАЛНА ЕВАЛУАЦИЈА



https://github.com/procesaur/PaLMA

https://palma.jerteh.rs