NLP4Web
Practice Session 7

Tokenizers

Hovhannes Tamoyan
tamohannes.com

About me

Hovhannes Tamoyan
(@tamohannes)

tamohannes.com

x.com/tamohannes

Syllabus for Practice Sessions (PS) 7 - 12

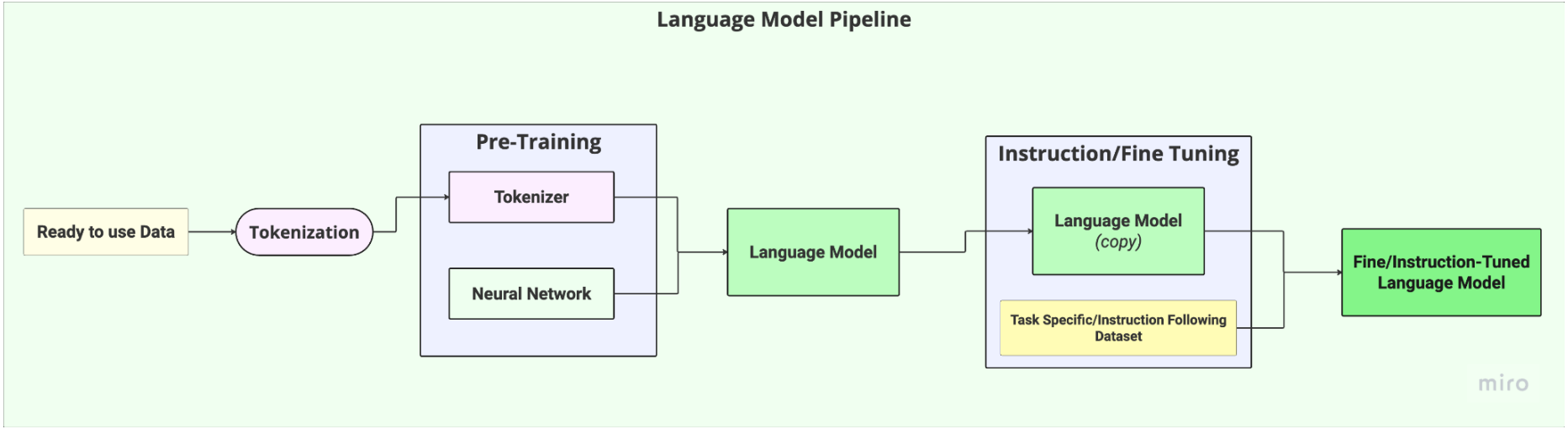
- **PS7** (05.12.2024): Tokenizers
- **PS8** (12.12.2024): Bigram, Neural Language Model and RNN
  - **HW3** release
- **PS9** (19.12.2024): Transformers Decoder-only (GPT)
- **PS10** (16.01.2025): Transformers Encoder-only (BERT) and Seq2seq Modeling
  - **HW4** release
- **PS11** (23.01.2025): Pre-Training and Fine-Tuning
- **PS12** (30.01.2025): Experiment Reproducibility and NLP Debugging
  - **HW5** release
  - **HW6** release

The philosophy behind this series of sessions

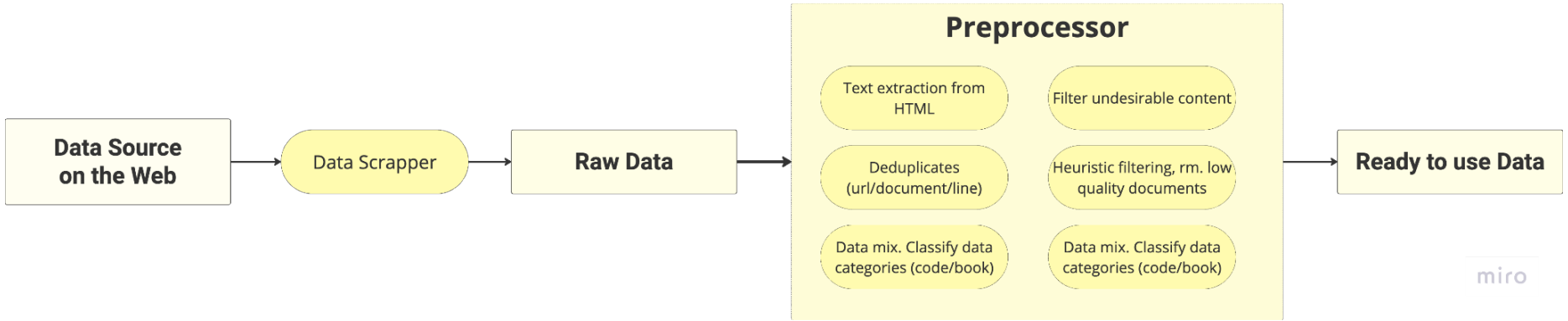
- “Build to Use”
  - Implement all the necessary components to build a language model (GPT) from scratch.
  - Use high-level wrappers for each block, as commonly practiced in industry and academia.

To not get lost in space over time, let’s
Use a **mind map**

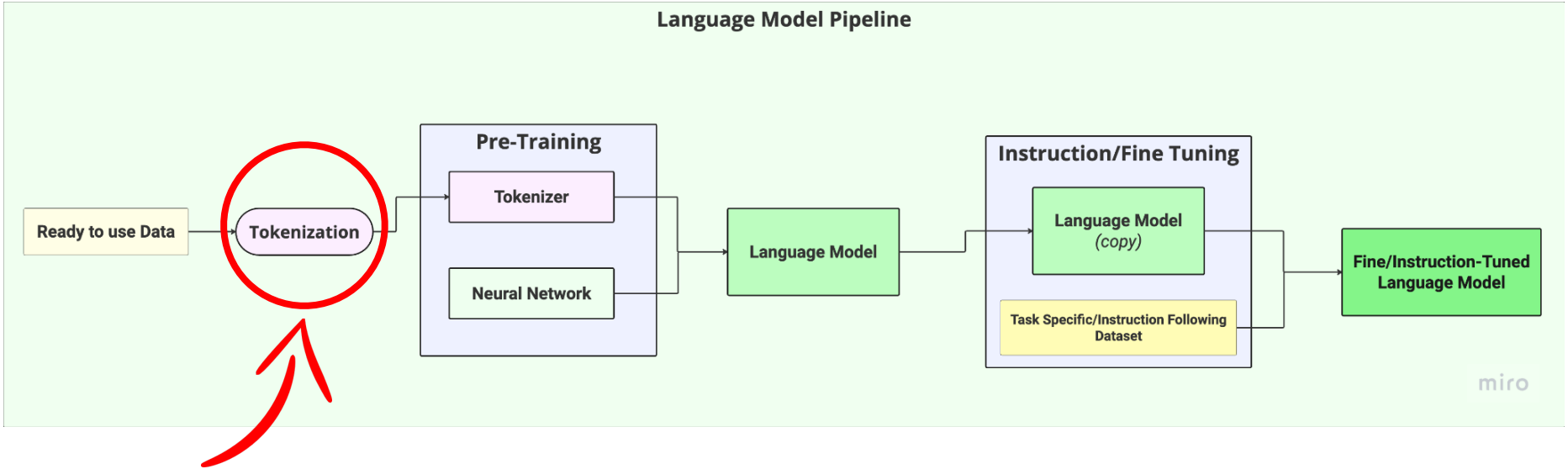
Language Modeling pipeline nowadays



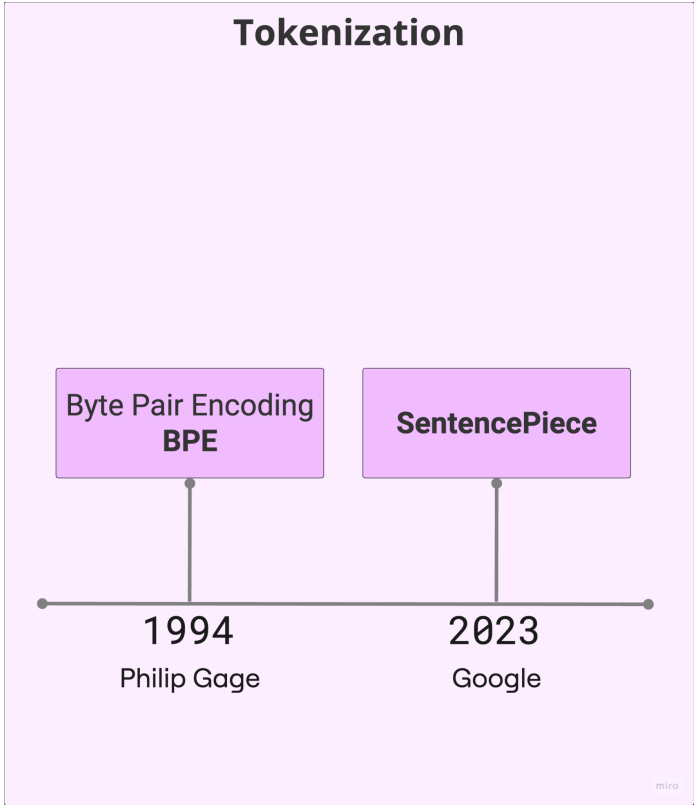
Break down of **Ready to use Data**



Today's subject: **Tokenization**



Tokenization Algorithms



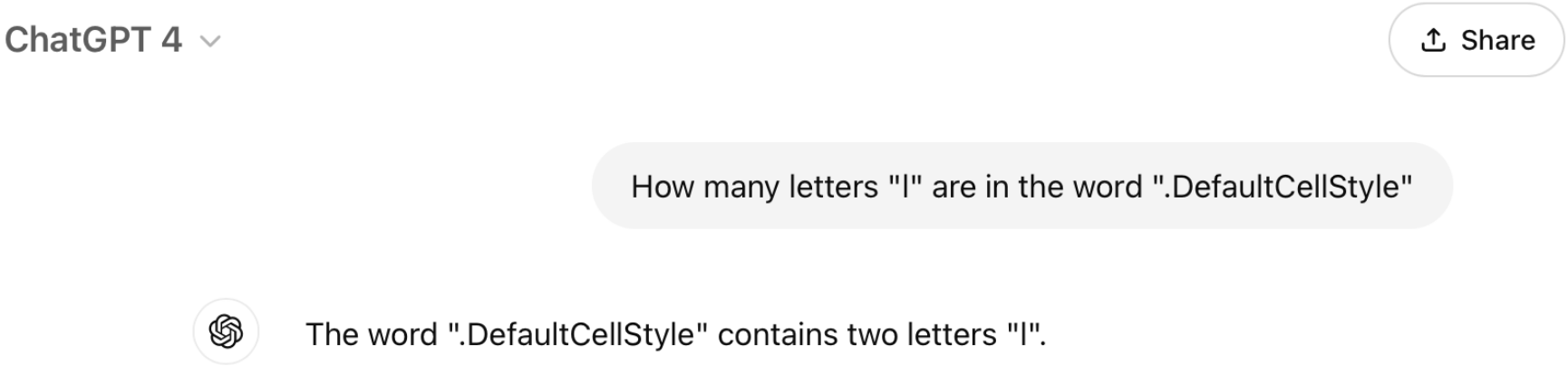
Tokenization is at the heart of much of the weirdness in LLMs

Why can't LLM spell words?

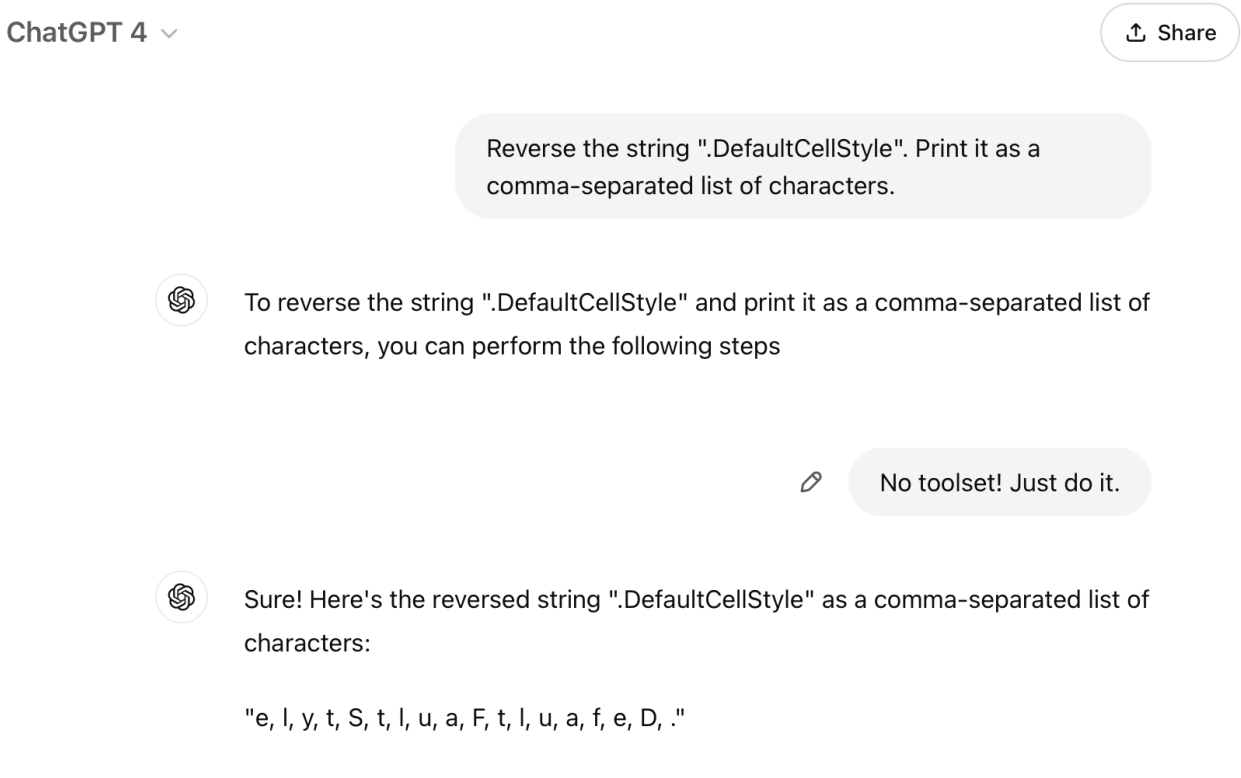
- The characters are divided into tokens, and some of these tokens can be quite long, such as “.DefaultCellStyle”.

Why can't LLM spell words?

Why can't LLM spell words?



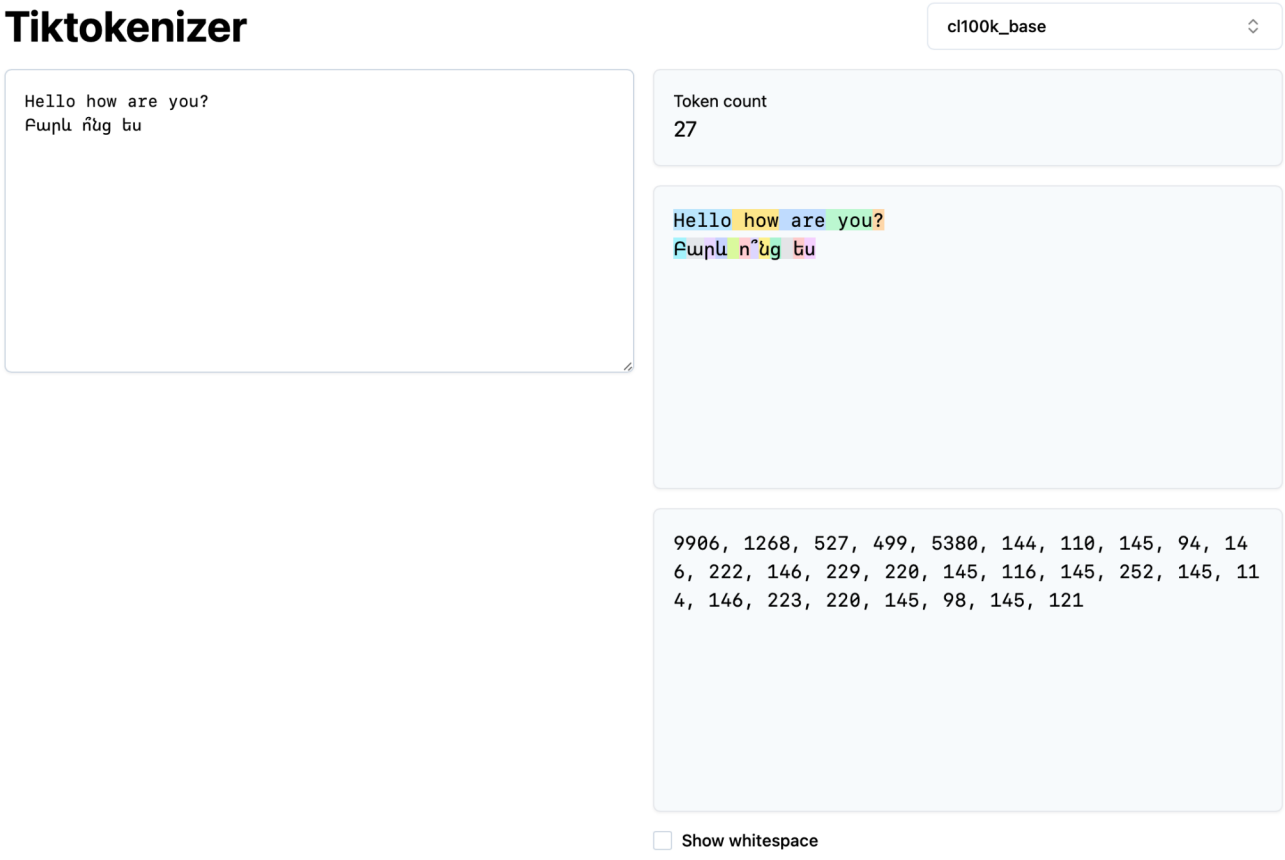
Why can't LLM do super simple string processing tasks like reversing a string?



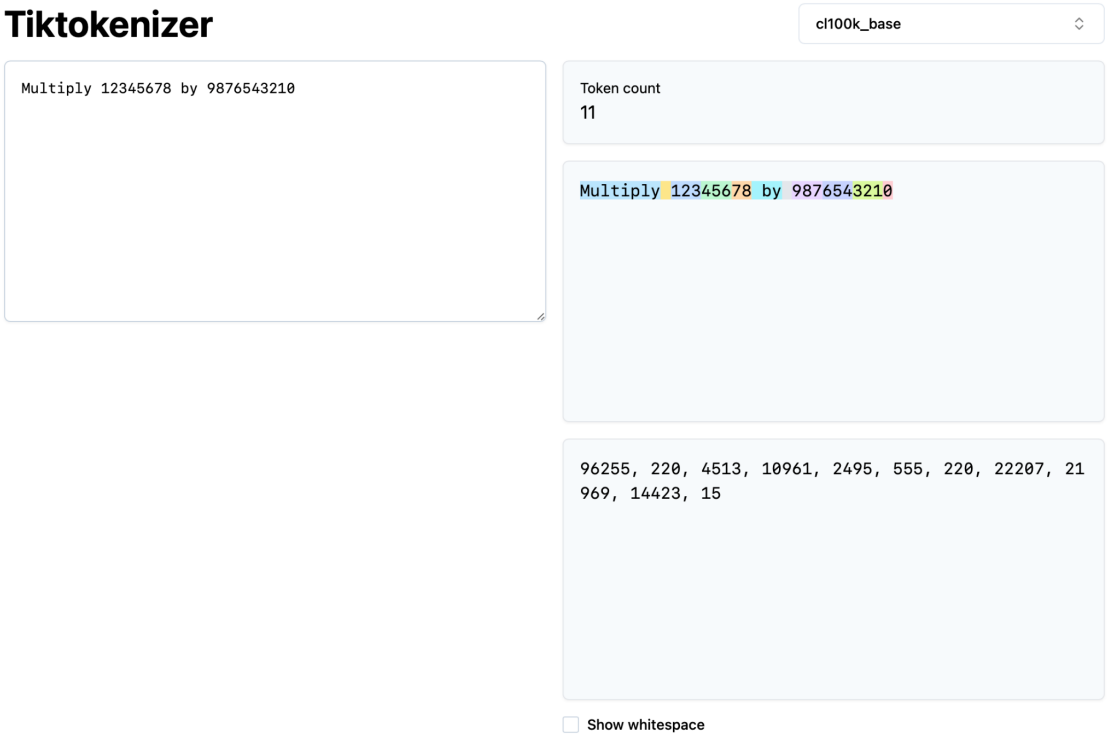
Try out a live tokenizer here

https://tiktokenizer.vercel.app/?model=o200k_base

Why is LLM worse at non-English languages?

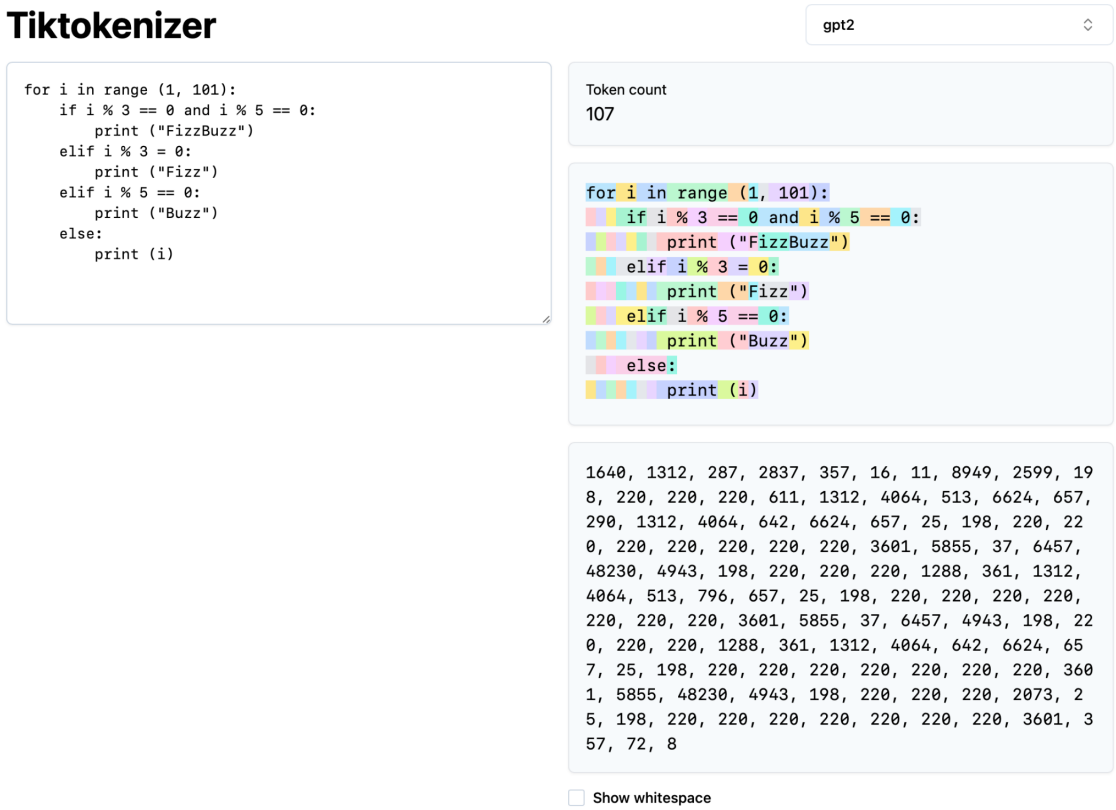


Why is LLM bad at simple arithmetic?



Future reading: https://www.beren.io/2023-02-04-Integer-tokenization-is-insane/

Why did GPT-2 have more than necessary trouble coding in Python?



Why should I prefer to use YAML over JSON with LLMs?

Tiktokenizer

cl100k_base

```
{
    "product": {
        "type": "T-Shirt",
        "price": 20.00,
        "sizes": ["S", "M", "L"],
        "reviews": [
            { "username": "user1", "rating": 4,
"created_at": "2023-04-19T12:30:002" },
            { "username": "user2", "rating": 5,
"created_at": "2023-05-02T15:00:00Z" }
        ]
    }
}
```

Token count
121


```
{
    "product": {
        "type": "T-Shirt",
        "price": 20.00,
        "sizes": ["S", "M", "L"],
        "reviews": [
            { "username": "user1", "rating": 4, "c
reated_at": "2023-04-19T12:30:002" },
            { "username": "user2", "rating": 5, "c
reated_at": "2023-05-02T15:00:00Z" }
        ]
    }
}
```


```
517, 197, 1, 3107, 794, 341, 197, 197, 45570, 794,
330, 51, 76954, 761, 197, 197, 1, 6692, 794, 220,
508, 13, 410, 345, 197, 197, 41887, 4861, 794, 448
2, 50, 498, 330, 39091, 498, 330, 43, 8257, 197, 1
97, 1, 40575, 794, 2330, 298, 197, 90, 330, 5223,
794, 330, 882, 16, 498, 330, 22696, 794, 220, 19,
11, 330, 7266, 3837, 794, 330, 2366, 18, 12, 2371,
12, 777, 51, 717, 25, 966, 25, 6726, 1, 1173, 298,
197, 90, 330, 5223, 794, 330, 882, 17, 498, 330, 2
2696, 794, 220, 20, 11, 330, 7266, 3837, 794, 330,
2366, 18, 12, 2304, 12, 2437, 51, 868, 25, 410, 2
5, 410, 57, 1, 457, 197, 197, 933, 197, 534, 92
```

Show whitespace

Tiktokenizer

cl100k_base

```
product:
    type: T-Shirt
    price: 20.00
    sizes:
        - S
        - M
        -L
    reviews:
        - username: user1
            rating: 4
            created_at: "2023-04-19T12:30:00Z"
        - username: user2
            rating: 5
            created_at: "2023-05-02T15:00:00Z"
```

Token count
104


```
product:
    type: T-Shirt
    price: 20.00
    sizes:
        - S
        - M
        -L
    reviews:
        - username: user1
            rating: 4
            created_at: "2023-04-19T12:30:00Z"
        - username: user2
            rating: 5
            created_at: "2023-05-02T15:00:00Z"
```


```
3107, 512, 13459, 25, 350, 76954, 198, 88219, 25,
220, 508, 13, 410, 198, 1942, 4861, 512, 197, 197,
12, 328, 198, 197, 197, 12, 386, 198, 197, 197, 82
88, 198, 17643, 5182, 512, 197, 197, 12, 6059, 25,
1217, 16, 198, 298, 7145, 1113, 25, 220, 19, 198,
298, 197, 7266, 3837, 25, 330, 2366, 18, 12, 2371,
12, 777, 51, 717, 25, 966, 25, 410, 57, 702, 197,
197, 12, 6059, 25, 1217, 17, 198, 298, 7145, 1113,
25, 220, 20, 198, 298, 197, 7266, 3837, 25, 330, 2
366, 18, 12, 2304, 12, 2437, 51, 868, 25, 410, 25,
410, 57, 1
```

Show whitespace

LLM is not end-to-end language modeling

What is the real root of suffering?

What is the real root of suffering?

Tokenization

Let’s jump into it 🚀