Generating my first text with a language model

written by Chun Ding on Functor Network
original link: https://functor.network/user/1/entry/1152

I followed the book *Hands-on Large Language Models* to generate my first piece of text:

```python
from transformers import AutoModelForCausalLM, AutoTokenizer

# Load model and tokenizer
model = AutoModelForCausalLM.from_pretrained(
    "microsoft/Phi-3-mini-4k-instruct",
    device_map="cuda",
    torch_dtype="auto",
    trust_remote_code=False,
)
tokenizer = AutoTokenizer.from_pretrained("microsoft/Phi-3-mini-4k-instruct")

from transformers import pipeline

# Create a pipeline
generator = pipeline(
    "text-generation",
    model=model,
    tokenizer=tokenizer,
    return_full_text=False,
    max_new_tokens=500,
    do_sample=False
)

# The prompt (user input / query)
messages = [
    {"role": "user", "content": "tell a funny joke"}
]

# Generate output
output = generator(messages)
print(output[0]["generated_text"])
```

I asked it to tell a funny joke. The output was:

```
Why don't scientists trust atoms? Because they make up everything!
```

The sentence is fluent and understandable, but at first I didn't get the punchline. When I asked the model what was funny about the sentence, it ran out of memory and couldn't explain. So I turned to ChatGPT and finally got the humor: It was an English pun.