Tokenization

"Expectation vs Reality"

Expectation

"Tokenization: Splitting text into clean, **perfect pieces."**

Reality

"Tokenization: Why did I get 'New' and 'York' as separate tokens?!"


"NLP BOT: 'I JUST
CAN'T HANDLE ALL THESE
AMBIGUOUS PRONOUNS ANYMORE!'"
imgflip.com

Evolution of NLP Models

A small, cute puppy

"Small Model" - Fast but weak, struggles with complex text

A medium-sized dog

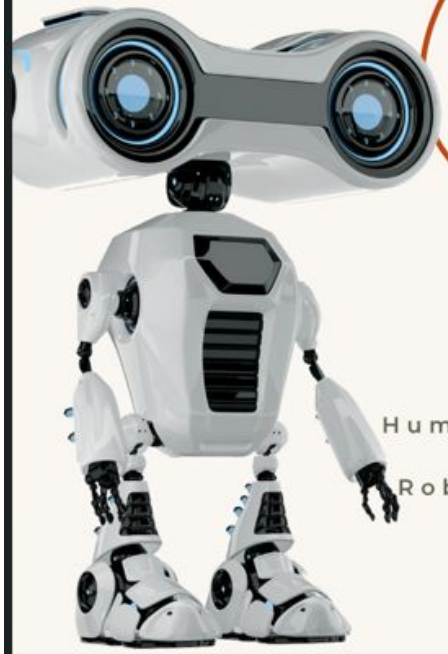
"Medium Model" - Decent, can handle some tasks well

A big, muscular dog

"Large Model" - Powerful, requires more resources

futuristic robot dog (Boston Dynamics style)

"Transformer Model" - Can do everything, but eats up all your GPU 😆



"I tried to understand human language... but then I encountered sarcasm, idioms, and ambiguous pronouns. Send help."

Human: "That exam was a piece of cake!"
Robot: "Error: Cake not found in exam data." 🍰

Human: "Oh great, another software update. because the last one was soooo perfect."
Robot: "Thank you for the positive feedback 😇

ME TRYING TO EXPLAIN TOKENIZATION TO MY GRANDMA

"WHY CAN'T THE COMPUTER JUST READ THE WHOLE SENTENCE LIKE A NORMAL PERSON?"

'UGH, WORDS SPLIT, MACHINE HAPPY."