トークン（AIは文章を小さな単位に分けて扱う）

文章を生成するAIは、入力した文をどのような形で扱っているのだろう

入力した文：　**今日は晴れています**

1 トークン（小さな単位）に区切る

今日 | は | 晴れ | て | い | ます

2 それぞれを番号（ID）に変える

今日	は	晴れ	て	い	ます
3120	25	8841	61	72	410

3 AIは番号の並びを計算する

番号の並びから、次に来るトークンの
番号の確からしさを計算し、選ぶ。
それをくり返して、答えの番号の並びをつくる。

4 番号を文字にもどして表示する

トークンについて知っておくこと

・区切り方は、AIのモデルによってちがう。1文字が1トークンとは限らない（日本語と英語でも区切り方がちがう）

・一度に読める長さや、使った量（料金）を、トークンの数で数えることが多い

・文字の数を数える・1文字ずつ並べかえる、などが苦手なことがある理由の一つになる

なぜ番号にするの？

コンピューターは数しか計算できないので、ことばを番号（さらに数の並び）に変えて扱う。

よく出てくる言葉や文字の組み合わせを1つのトークンにすると、少ない数で多くの文を表せる。

※ 模式図。区切り方と番号は架空の例。