■ Padding (in NLP)

- Used to make all input sequences the same length.
- Adds 0s (PAD tokens) to shorter sentences.
- Needed because models (like LSTM, BERT) need uniform input sizes.
- Types:

• Pre-padding: zeros before sequence → [0, 0, 12, 45]

• Post-padding: zeros after sequence → [12, 45, 0, 0]

Example:

[12, 45, 78]

[56, 34] → [56, 34, 0]

■ loc vs iloc (in Pandas)

Feature	loc	iloc
Based on	Label (name)	Index (position)
Syntax	df.loc[row_label, col_label]	df.iloc[row_index, col_index]
Example	df.loc['a', 'Age']	df.iloc[0, 1]
Range	Inclusive of end	Exclusive of end
Use When	You know row/column names	You know row/column positions

■ loc → label-based

■ iloc → integer-based

■ One Hot Encoding

- Converts categorical text data into binary numeric vectors.
- Each unique category becomes its own column.
- Only one column has 1 (the active category).

Example:

Color
Red
Blue
Green

➡■ One-Hot Encoded:

Red	Blue	Green
1	0	0
0	1	0
0	0	1

Why use it:

- Models need numeric input.

- Avoids implying order between categories.

Limitations:

- Increases dimensionality (many new columns for large category sets).

■ Quick Recap:

- Padding: equalizes sequence length in NLP.
- loc: label-based data selection.
- iloc: position-based data selection.
- One-Hot Encoding: turns categories into binary vectors for ML models.