Special tokens in NLP

- Common special tokens in NLP

token	Meaning / use
<pad>	Padding to make sequences the same lenght in a batch size
<eos>	End of sentence (marks where text stops) ⇒ stops GPT generation
<bos>	Beginning of sentence (used in some models to indicate start)
<unk>	Unknown token (used for words not in vocab)
<cls>	classification token (used in Bert for sentence lvl tasks) Ex I love Pizza ⇒ <cls> I love Pizza (bert would add cls to start) • cls represents whole sentence • after encoding, the vector for cls is used for tasks like, sentiment classification, topic classification etc.
<sep>	Seperator (splits sentences or segments in a sequence) helps seperate different segments in sentence like making Q A pairs Ex <cls> how are you? <sep> i am fine cls ⇒ represents whole input for classification, <sep> ⇒ marks boundary between Q and A (sep = segment seperator)
<mask>	Masked tokens (used for masking tokens see p 120)