# Tokenization Algorithms in Large Language Models

D. Gueorguiev 2/19/2024

## References

[Language Models are Unsupervised Multitask Learners, Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, OpenAI, 2018](https://github.com/dimitarpg13/large_language_models/blob/main/articles/Language_Models_are_Unsupervised_Multitask_Learners_Alec_Radford_OpenAI_2018.pdf)

[Neural Machine Translation of Rare Words with Subword Units, Rico Senrich et al, 2016](https://github.com/dimitarpg13/large_language_models/blob/main/articles/Neural_Machine_Translation_of_Rare_Words_with_Subword_Units_Rico_Senrich_2016.pdf)

Minimal Byte Pair Encoding Algorithm (Andrej Karpathy repo): <https://github.com/karpathy/minbpe/tree/master>

[Let's build the GPT Tokenizer with Andrej Karpathy, February 2024, youtube video](https://youtu.be/zduSFxRajkE)

[Let's build GPT: from scratch, in code, spelled out with Andrej Karpathy, February 2023, youtube video](https://www.youtube.com/watch?v=kCc8FmEb1nY)

[How ChatGPT is Trained with Ari Seff, February, 2023, youtube video](https://www.youtube.com/watch?v=VPRSBzXzavo)