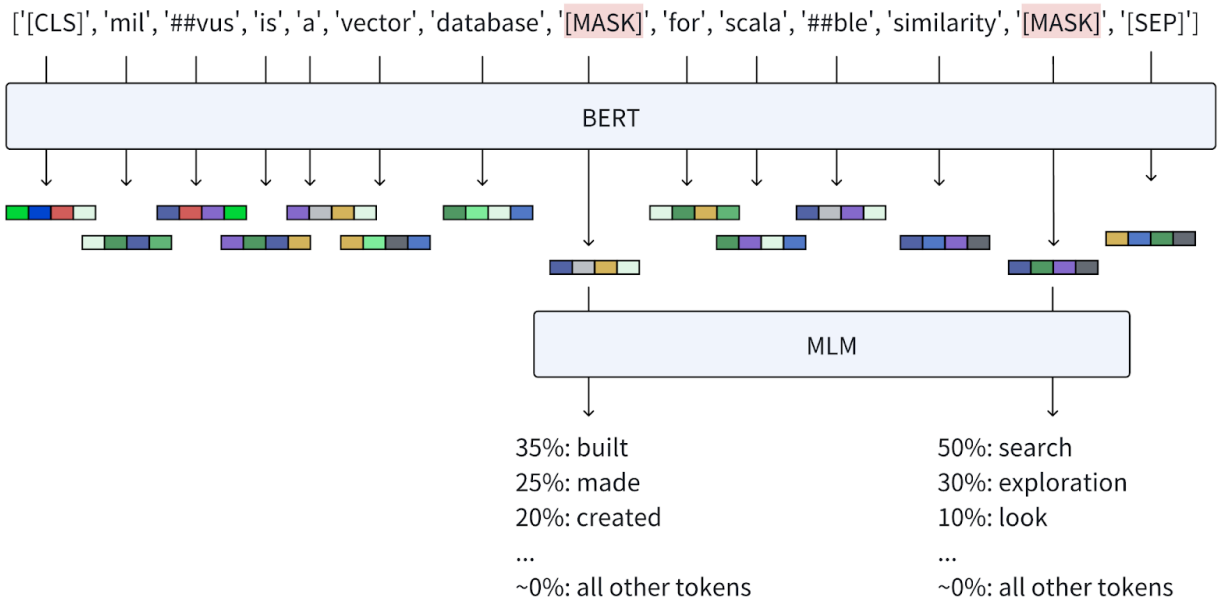
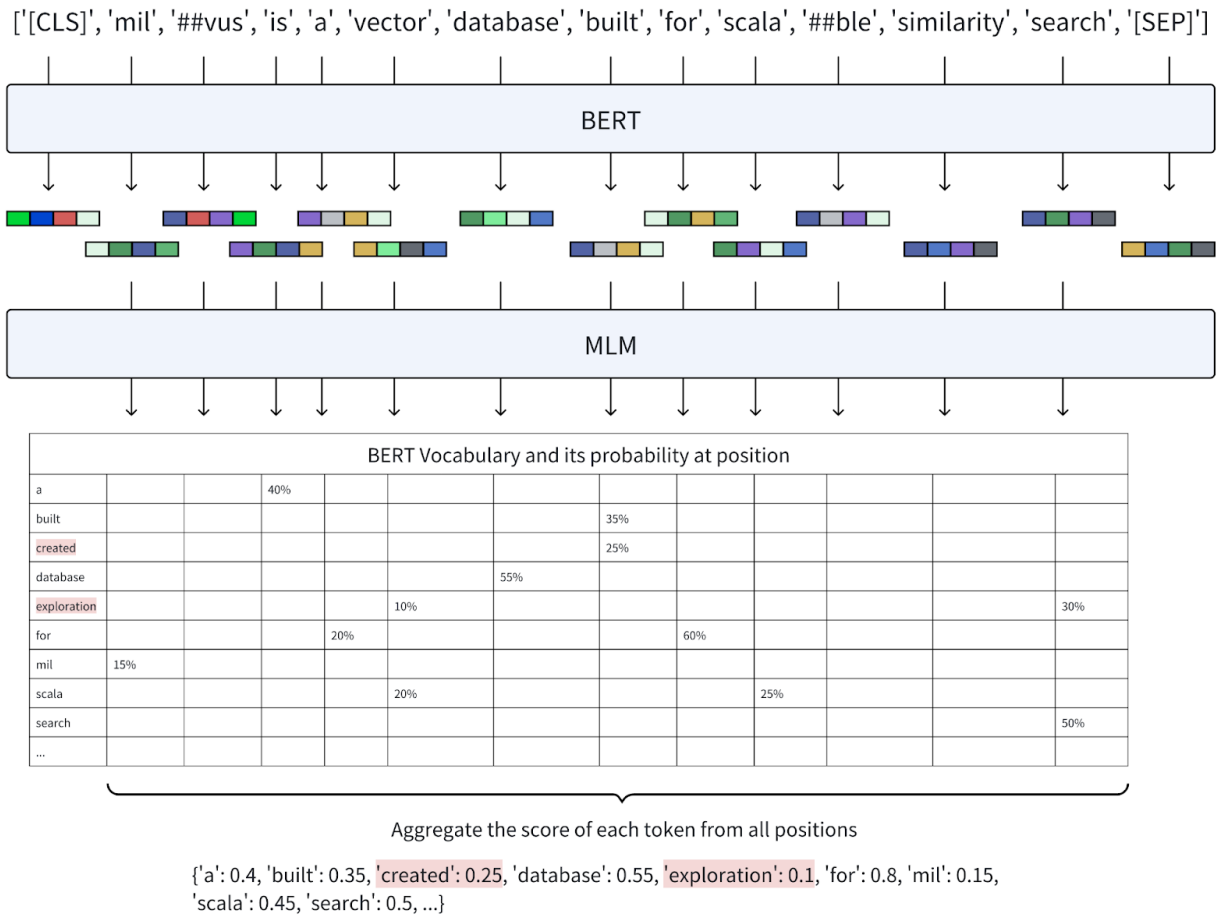
SPLADE signifies a progression in creating learned sparse embeddings by enhancing the original BERT architecture with a distinctive approach to refining embedding sparsity. To understand this approach, it's essential to revisit the fundamental training mechanism of BERT, known as Masked Language Modeling (MLM).

MLM is an effective unsupervised learning task that hides a portion of the input tokens, forcing the model to infer the missing words solely from their surrounding context. This method enhances the model's understanding of language and its structure, relying on nearby tokens to make accurate predictions and fill in the gaps.



MLM predicts the original token based on the BERT embedding of [MASK]

During pre-training, for each masked slot, the model leverages the contextualized embedding (H[i]) from BERT to generate a probability distribution (w_i). Here, (w_{ij}) represents the likelihood that a particular token from BERT's vocabulary occupies the masked position. This output vector (w_i), which has a length corresponding to BERT's extensive vocabulary (usually 30,522 words), acts as a crucial learning signal for improving the model's predictions.



MLM aggregates the score of each token from all positions.

SPLADE utilizes the strengths of Masked Language Modeling (MLM) during the encoding phase. After the initial tokenization and conversion to BERT embeddings, SPLADE applies MLM across all token positions, calculating the probability that each token corresponds to every word in BERT's vocabulary. It then aggregates these probabilities for each vocabulary word across all positions, employing a regularization technique to encourage sparsity through a log saturation effect. The resulting weights indicate the relevance of each vocabulary word to the input tokens, forming a learned sparse vector.

A key benefit of SPLADE's embedding approach is its ability to facilitate term expansion. It can identify and incorporate relevant terms that are absent from the original text. For example, in a given instance, terms like "exploration" and "created" appear in the sparse vector even though they were not included in the initial sentence. Notably, for a concise input like "milvus is a vector database built for scalable similarity search," SPLADE can enhance the context by expanding it to include 118 tokens, significantly improving exact term matching in retrieval tasks and enhancing the model's precision.

This detailed process highlights how SPLADE enriches traditional BERT embeddings, providing greater granularity and utility for tasks such as search and retrieval, where both the breadth and specificity of term relevance are crucial.