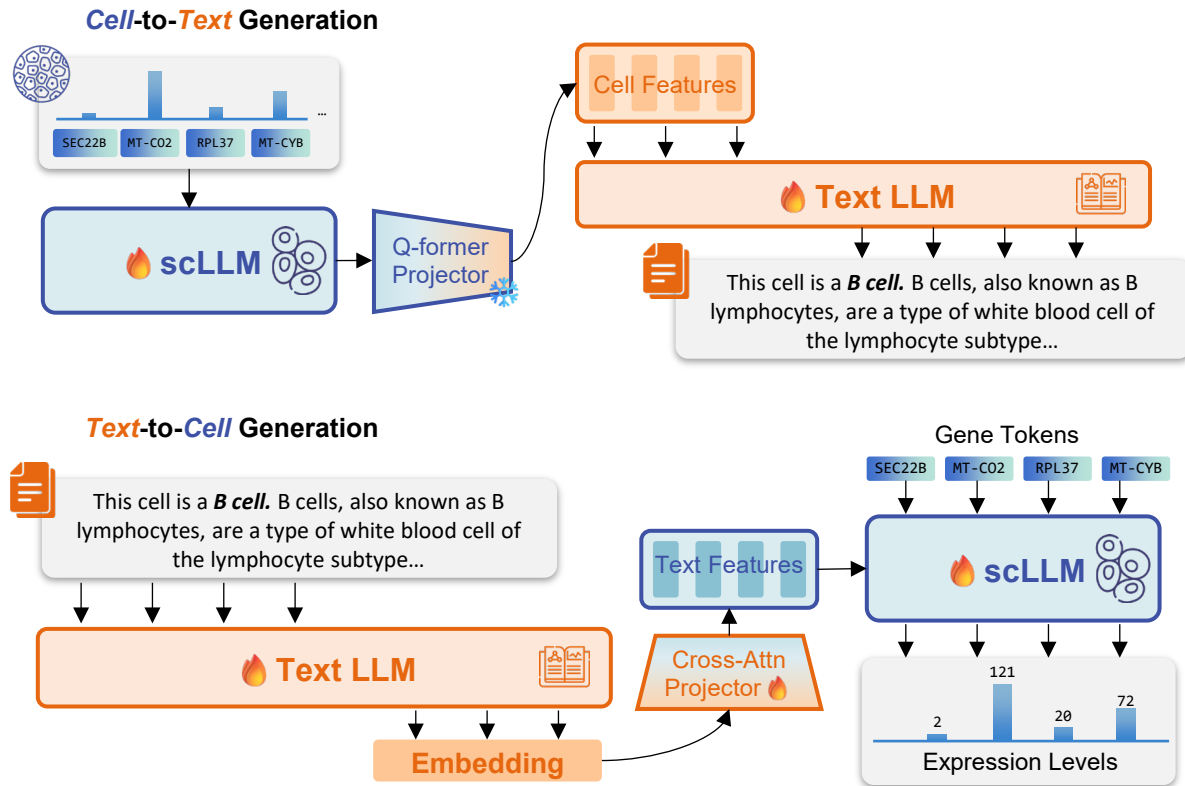


(a) Stage 1: Cross-modal *Discriminative* Pre-training



(b) Stage 2: Cross-modal *Generative* Pre-training