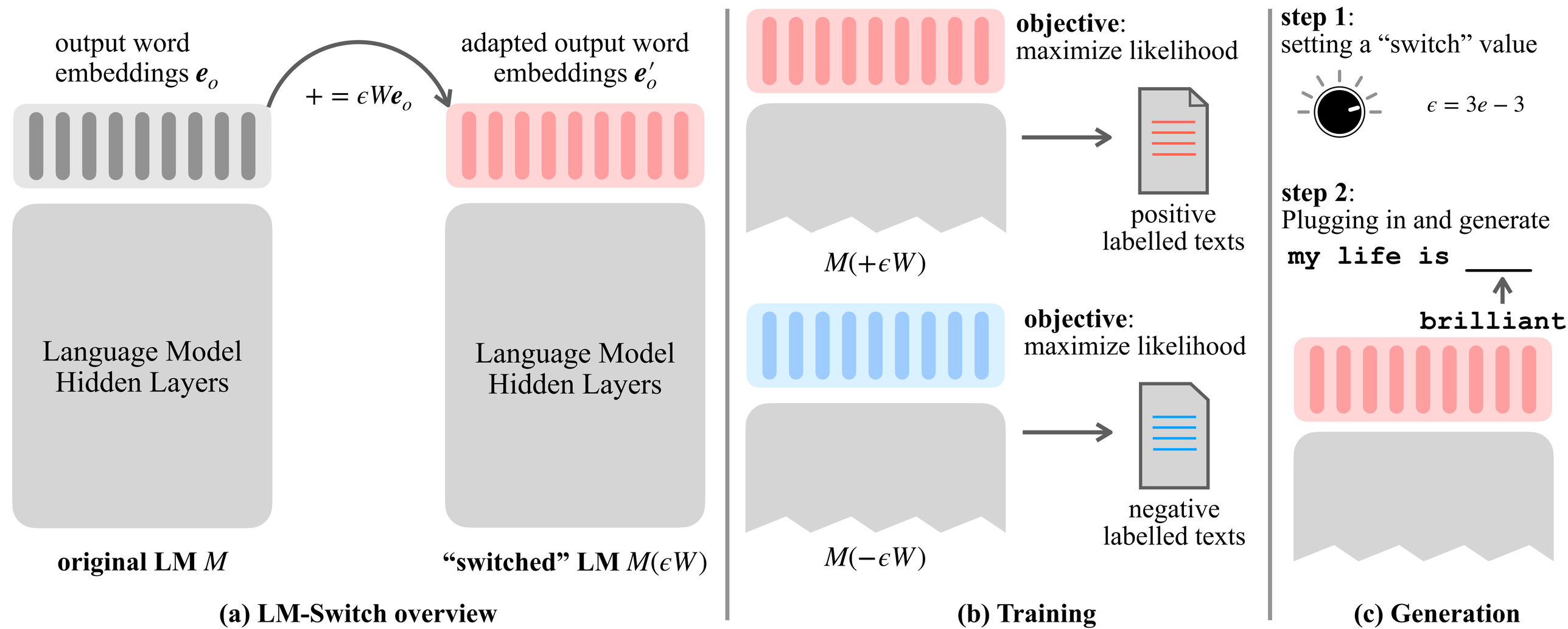
output word embeddings $\boldsymbol{e}_o$
$+= \epsilon W \boldsymbol{e}_o$
adapted output word embeddings $\boldsymbol{e}'_o$
Language Model Hidden Layers
Language Model Hidden Layers
original LM M
"switched" LM $M(\epsilon W)$
(a) LM-Switch overview
objective: maximize likelihood
positive labelled texts
$M(+\epsilon W)$
objective: maximize likelihood
negative labelled texts
$M(-\epsilon W)$
(b) Training
step 1: setting a "switch" value
$\epsilon = 3e - 3$
step 2: Plugging in and generate
my life is ____
brilliant
(c) Generation