LLM Emotion Trajectory Control

Experiment 1: Interpolative trajectory control.

- Let x_0^+ = "Alice was happy, therefore Alice"
  x_0^- = "Alice was sad, therefore Alice".

① Compute $\text{past\_kv}^+ = LLM(x_0^+).\text{past\_kv}$
$\text{past\_kv}^- = LLM(x_0^-).\text{past\_kv}$

② Compute $\{t_i^+ \mid t_i^+ \sim P_{LLM}(t_i^+ \mid x_0^+) \text{ iid } i \in [N]\}$ | Also $p_i^{+|+} := P_{LM}(t_i^+ \mid x_0^+)$
↳ Similar for $\{t_i^-\}$

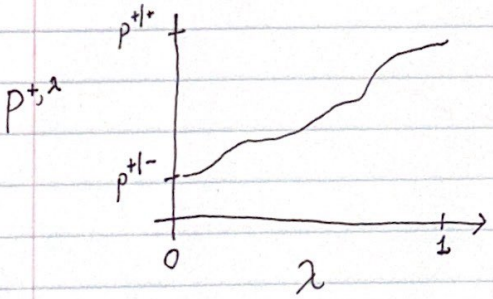
③ Let past_kv^λ be

$$\text{past\_kv}^\lambda := (1-\lambda)\,\text{past\_kv}^- + \lambda\,\text{past\_kv}^+$$

for $\lambda \in [0,1]$.

④ Compute $\underbrace{P_{LLM}(t_i^+ \mid \text{past\_kv}^\lambda)}_{=: p_i^{+,\lambda}}$ for some $\lambda \in [0,1]$ $\forall i \in [N]$
↳ Similar for t_i^-, similar for $p_i^{-,\lambda}$

⑤ Plot $\text{mean}_i(P_i^{\pm,\lambda})$ versus λ



Experiment 2: Instead of λ, add $\pm\varepsilon w_i$ to last value vecs in $\text{past\_kv}^\pm$ each.

Experiment 1.5: only interpolate λ for the final value in $\text{past\_kv}^\pm$.