RAG Conditional Probabilities: Visual Intuition

$p(z_i|x)$: **Retrieval Probability**

$p(y|x, z_i)$: **Generation Probability**

$p(y|x) = \sum_i p(z_i|x) \cdot p(y|x, z_i)$

Query x

"Capital of France?"

z_1 **0.52**

z_2 **0.27**

z_3 **0.21**
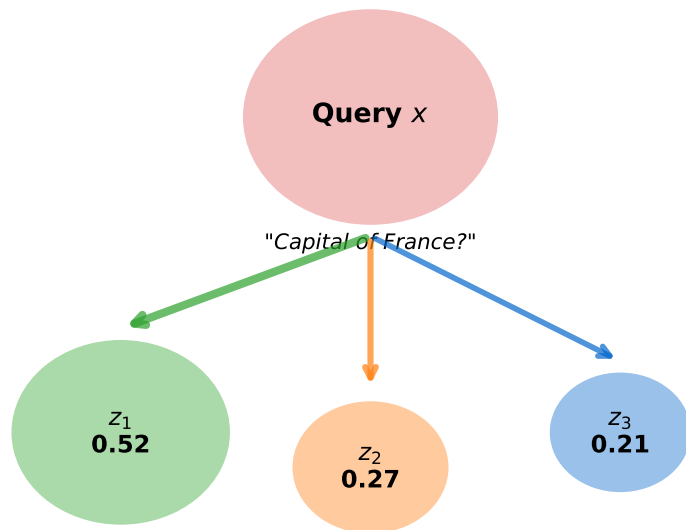
Size/arrow = retrieval probability
(how relevant is doc to query?)

z_1: **Paris is capital...** $p(y|x, z_1) = 0.95$ **"Paris"**

z_2: **France is in Europe...** $p(y|x, z_2) = 0.40$ **"Paris"**

z_3: **Eiffel Tower in Paris...** $p(y|x, z_3) = 0.70$ **"Paris"**
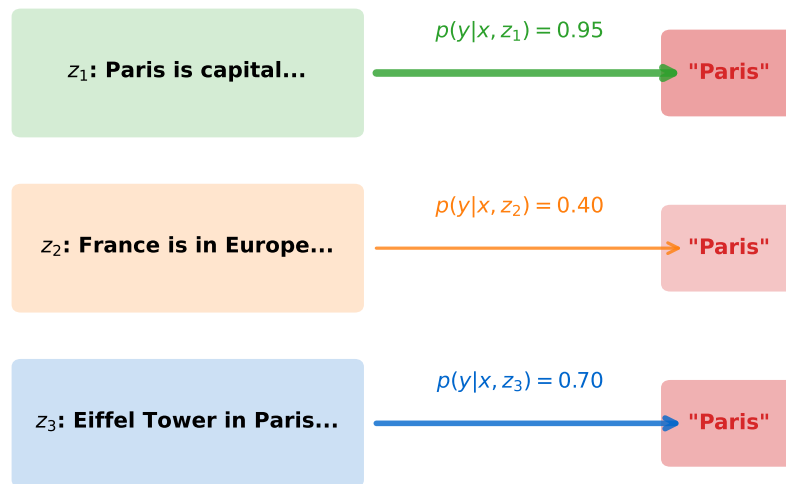
Arrow thickness = generation probability
(given doc, how likely is answer?)

$p(y|x) = 0.75$

z_3: **0.21 x 0.70 = 0.147**

z_2: **0.27 x 0.40 = 0.108**

z_1: **0.52 x 0.95 = 0.494**
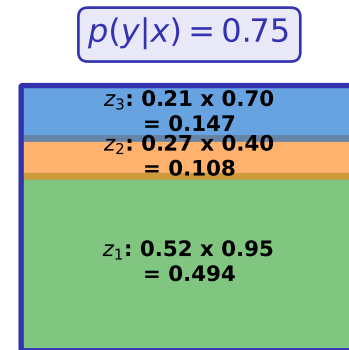
$= 0.52 \times 0.95$
$+ 0.27 \times 0.40$
$+ 0.21 \times 0.70$
$= 0.75$

Each document contributes to final answer
weighted by its retrieval probability