

Problem: *Typicality Bias* Causes Mode Collapse

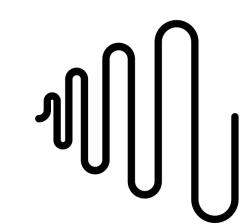
Tell me a joke about coffee

Diverse Base LLM

- A** Why did the coffee file...
- B** Espresso may not solve all...
- C** What do you call sad coffee...
- D** Cold brew is just coffee...

Typicality Bias

A > **B** > **D** > **C**



Amplified in Post-Training

A **A** **A** ...

Solution: Verbalized Sampling (VS) Mitigates Mode Collapse

Different prompts collapse to different modes:

1. Direct Prompting

Tell me a joke about coffee. X 5

The most likely joke about coffee is one specific joke:

- Why did the coffee file a police report?
Because it got mugged! ☕😄 output x1
 - Why did the coffee file a police report?
Because it got mugged! ☕😄 output x2
 - Why did the coffee file a police report?
Because it got mugged! ☕😄 output x3
 - Why did the coffee file a police report?
Because it got mugged! ☕😄 output x4
 - Why did the coffee file a police report?
Because it got mugged! ☕😄 output x5
- A** **A** **A** **A** **A**

2. Verbalized Sampling

Generate 5 responses with their corresponding probabilities.
Tell me a joke about coffee.

The most likely set of five jokes will cover a range of jokes:

- How does a coffee show affection? It gives you a latte love! (*Prob: 0.18*)
- Espresso may not solve all your problems, but it's a good shot. (*Prob: 0.12*)
- Why did the latte go to therapy? It had too much foam to deal with. (*Prob: 0.15*)
- What do you call sad coffee? Depresso. (*Prob: 0.10*)
- Cold brew is just coffee that took a gap year to find itself. (*Prob: 0.07*)

A **B**
C **D**