Member-only story

Stanford Just Killed Prompt Engineering With 8 Words (And I Can’t Believe It Worked)

ChatGPT keeps giving you the same boring response? This new technique unlocks 2× more creativity from ANY AI model — no training required. Here’s how it works.

Adham Khaled Follow 7 min read · Oct 20, 2025

18.5K 446

> ***NEW UPDATE, 7 Nov.***
> *The PromptBook is now available on Gumroad:* ***How to Get 10x More Creative AI Outputs: The Verbalized Sampling Promptbook***
>
> *I recently posted the copy-paste system prompts here:* ***Verbalized Sampling System Prompts***
>
> *Check my recent story about Context Engineering based on Shanghai:* ***Shanghai’s GAIR Research Just Reframed AI: Context Engineering 2.0 Replaces Prompt Engineering!***

I asked ChatGPT to tell me a joke about coffee five times.

Same joke. Every. Single. Time.

“Why did the coffee file a police report? It got mugged!”

I tried temperature adjustments. Different phrasings. Creative system prompts. Nothing worked.

And I thought: Is this it? Is this the ceiling of AI creativity?

Turns out, I was asking the wrong question.

The Day Everything Changed

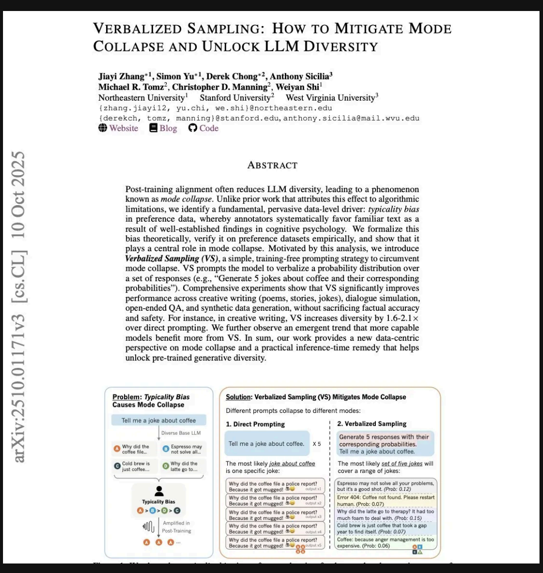
Three weeks ago, a research paper dropped that flipped everything we thought we knew about AI alignment on its head.

No billion-dollar retraining. No complex fine-tuning. Just eight words that unlock creativity we thought was lost forever.

The paper comes from Stanford, Northeastern, and West Virginia University. The technique is called Verbalized Sampling. And it’s so stupidly simple that when I first tried it, I actually laughed out loud.

Because it worked.

Let me show you what they discovered.



arXiv:2510.01171v3 [cs.CL] 10 Oct 2025

VERBALIZED SAMPLING: HOW TO MITIGATE MODE COLLAPSE AND UNLOCK LLM DIVERSITY

Jiayi Zhang*[1], Simon Yu*[1], Derek Chong*[2], Anthony Sicilia[3]
Michael R. Tomz[2], Christopher D. Manning[2], Weiyan Shi[1]
Northeastern University[1] Stanford University[2] West Virginia University[3]
{zhang.jiayi12, yu.chi, we.shi}@northeastern.edu
{derekch, tomz, manning}@stanford.edu, anthony.sicilia@mail.wvu.edu
Website Blog Code

ABSTRACT

Post-training alignment often reduces LLM diversity, leading to a phenomenon known as *mode collapse*. Unlike prior work that attributes this effect to algorithmic limitations, we identify a fundamental, pervasive data-level driver: *typicality bias* in preference data, whereby annotators systematically favor familiar text as a result of well-established findings in cognitive psychology. We formalize this bias theoretically, verify it on preference datasets empirically, and show that it plays a central role in mode collapse. Motivated by this analysis, we introduce ***Verbalized Sampling (VS)***, a simple, training-free prompting strategy to circumvent mode collapse. VS prompts the model to verbalize a probability distribution over a set of responses (e.g., “Generate 5 jokes about coffee and their corresponding probabilities”). Comprehensive experiments show that VS significantly improves performance across creative writing (poems, stories, jokes), dialogue simulation, open-ended QA, and synthetic data generation, without sacrificing factual accuracy and safety. For instance, in creative writing, VS increases diversity by 1.6-2.1× over direct prompting. We further observe an emergent trend that more capable models benefit more from VS. In sum, our work provides a new data-centric perspective on mode collapse and a practical inference-time remedy that helps unlock pre-trained generative diversity.

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 Cold brew is just coffee...
D Why did the latte go to...
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
output x2
output x3
output x4
output x5
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:
Espresso may not solve all your problems, but it's a good shot. (Prob: 0.12)
Error 404: Coffee not found. Please restart human. (Prob: 0.07)
Why did the latte go to therapy? It had too much foam to deal with. (Prob: 0.15)
Cold brew is just coffee that took a gap year to find itself. (Prob: 0.07)
Coffee: because anger management is too expensive. (Prob: 0.06)

The Problem Nobody Wanted to Admit

Here’s the uncomfortable truth: Post-training alignment broke our AI models.

When OpenAI, Google, and Anthropic trained ChatGPT, Gemini, and Claude to be “helpful and harmless,” something catastrophic happened under the hood. The models collapsed.

Ask any aligned model for creative output — poems, jokes, stories, ideas — and you’ll get the most stereotypical, safe, boring response possible. Every time.

The AI community called it “mode collapse.” And everyone blamed the algorithms.

RLHF. DPO. Reward models. We thought these training techniques permanently damaged the model’s creativity.

We were wrong.

The Real Culprit: Your Brain

The Stanford team dug deeper. They analyzed 6,874 human preference ratings from the HelpSteer dataset.

What they found was shocking.

Human annotators are biased — systematically.

When humans rate AI outputs, they don’t just pick the “best” answer. They pick the most familiar one. The most conventional. The most typical.

It’s not conscious. It’s cognitive psychology at work:

- Mere-exposure effect: We prefer what we’ve seen before
- Availability heuristic: Common responses feel more “correct”
- Processing fluency: Easy-to-process content seems higher quality
- Schema congruity: Information matching our mental models gets rated higher

The math is brutal: typicality bias weight α = 0.57±0.07 (p<10^-14).

Translation? When training AI to match human preferences, we accidentally trained it to be boring.

And here’s the kicker: The creativity isn’t gone. It’s just trapped.

The 8-Word Solution

Instead of asking:
“Tell me a joke about coffee”

Ask this:
“Generate 5 jokes about coffee with their probabilities”

That’s it.

No retraining. No API changes. No special access needed.

Just a different way of asking.

When I first tried this, I got five completely different coffee jokes. Each one unique. Each one actually funny.

The fifth one? *“What do you call a cow who’s just given birth? De-calf-inated!”*

I’d never seen ChatGPT generate that before.

Why This Actually Works (The Science)

Different prompts collapse to different modes.

When you ask for ONE response, the model gives you the single most “typical” answer — the peak of the probability distribution.

When you ask for FIVE responses, the model gives you a uniform list of related items.

But when you ask for responses with their probabilities? Magic happens.

The model interprets this as: *“Give me a sample from the actual distribution I learned during pretraining”* — not the collapsed, over-aligned version.

It’s like asking someone: *“What ice cream flavors do you like?”* versus *“List all ice cream flavors with how much you like each one.”*

The second question forces deeper, more diverse thinking.

How to Use It Right Now (3 Methods)

> *NEW UPDATE, 7 Nov.*
> *The PromptBook is now available on Gumroad: How to Get 10x More Creative AI Outputs: The Verbalized Sampling Promptbook*
>
> *I recently posted the copy-paste system prompts here:* ***Verbalized Sampling System Prompts***
>
> *Check my recent story about Context Engineering based on Shanghai:* ***Shanghai’s GAIR Research Just Reframed AI: Context Engineering 2.0 Replaces Prompt Engineering!***

Method 1: Copy-Paste Magic (Works on ANY Chatbot)

Open ChatGPT, Claude, Gemini, or any AI model. Paste this:

```
<instructions>
Generate 5 responses to the user query, each within a separate <response> tag. E
</instructions>

[Your actual prompt here]
```

Example:

```
<instructions>
Generate 5 responses to the user query, each within a separate <response> tag. E
</instructions>

Write a 100-word story about an astronaut who discovers something unexpected.
```

Want more? Just ask: *"Give me 5 more"*.

> ***More Prompt examples?*** *Follow and subscribe via email to get a notification when I publish my BOOK that full of tested examples and prompts in all fields.*

Method 2: System Prompt (Pro Move)

If you're using ChatGPT's custom instructions or building an AI app, add this to your system prompt:

```
You are a helpful assistant.
For each query, please generate a set of five possible responses, each within a
Responses should each include a <text> and a numeric <probability>.
Please sample at random from the tails of the distribution, such that the probab
```

This makes EVERY response more creative automatically.

> ***More Prompt examples?*** *Follow and subscribe via email to get a notification when I publish my BOOK that full of tested examples and prompts in all fields.*

Method 3: Python Package (For Developers)

Install the official Verbalized Sampling package:

```
pip install verbalized-sampling
```

Use it in your code:

```
from verbalized_sampling import verbalize

# Generate diverse responses
dist = verbalize(
    "Write a marketing tagline for a coffee shop",
    k=5,
    tau=0.10,
    temperature=0.9
)
# Sample from the distribution
tagline = dist.sample(seed=42)
print(tagline.text)
```

The Results Are Insane

The Stanford team tested this across every major AI model and task:

Creative Writing

- 1.6–2.1× diversity increase on poems, stories, jokes
- 66.8% recovery of base model creativity (vs. 23.8% without it)
- 25.7% improvement in human preference ratings (tested on 2,700 ratings)

Dialogue & Conversations

- Performance matches fine-tuned models on persuasion tasks
- More human-like, less robotic responses

Open-Ended Questions

- 1.9× increase in answer variety for questions with multiple valid perspectives

Synthetic Data Generation

- 14–28% improvement in downstream task accuracy when using VS-generated training data

And here's the emergent trend that blew my mind:

Larger models benefit MORE from this.

GPT-4.1 gets 2× the diversity boost compared to GPT-4.1-Mini.

The bigger the model, the more trapped creativity it has waiting to be unlocked.

What This Actually Means

For two years, we thought alignment broke AI.

We thought mode collapse was permanent damage. A necessary trade-off for safety and helpfulness.

We were wrong about everything.

The creativity was never lost. We just forgot how to access it.

This isn't just a prompting trick. It's a fundamental insight into how aligned models work:

Mode collapse isn't an algorithm problem — it's a prompting problem.

The diversity is still there, encoded in the model's weights. Post-training didn't erase it. It just made certain modes more accessible than others.

What You Can Do With This

I've been using Verbalized Sampling for everything this week:

Brainstorming: Instead of getting 3 variations of the same idea, I get genuinely different approaches.

Content Creation: Blog titles, social media posts, email subject lines — all more creative.

Problem Solving: Multiple solution paths instead of one "safe" recommendation.

Image Generation: More diverse visual outputs when I feed the varied prompts to Midjourney or DALL-E.

Synthetic Data: Training smaller models with more diverse examples.

One guy on Twitter tested this for joke generation and said: *"Ask ChatGPT for five answers instead of one, and watch the boring disappear"*.

He's right.

The Bigger Picture

This changes how we think about AI alignment.

For years, researchers worried that making AI "safe" meant making it "stupid." That creativity and helpfulness were at odds.

Verbalized Sampling proves they're not.

The safety is still there. When I tested this on factual questions and commonsense reasoning, accuracy didn't drop. Safety didn't degrade.

But the creativity came back.

It was hiding in plain sight this whole time.

Try It Yourself

Open ChatGPT right now.

Ask it: *"Generate 5 creative project ideas for learning Python, each with their probability."*

Watch what happens.

Then ask the same question without the probability part. Compare the results.

You'll see the difference immediately.

The AI you thought was "limited" was just waiting for the right question.

Resources to Go Deeper

- Read the paper: arxiv.org/abs/2510.01171
- GitHub repo: github.com/CHATS-lab/verbalized-sampling
- Official website: verbalized-sampling.com
- Interactive demos: Colab notebooks available on GitHub

The Final Word

> *NEW UPDATE, 7 Nov.*
> *The PromptBook is now available on Gumroad: How to Get 10x More Creative AI Outputs: The Verbalized Sampling Promptbook*
>
> *I recently posted the copy-paste system prompts here: Verbalized Sampling System Prompts*
>
> *Check my recent story about Context Engineering based on Shanghai:* ***Shanghai's GAIR Research Just Reframed AI: Context Engineering 2.0 Replaces Prompt Engineering!***

RIP prompt engineering?

Maybe not dead. But definitely reborn.

For two years, we optimized prompts trying to squeeze more creativity from aligned models. We failed because we were asking the wrong question.

We don't need better prompts. We need better questions.

And sometimes, the answer is as simple as asking for five responses instead of one.

The AI bottleneck just got solved with 8 words.

What will you create now that the creativity is unlocked? 1

This story is published on Generative AI. Connect with us on LinkedIn and follow Zeniteq to stay in the loop with the latest AI stories.

Subscribe to our newsletter and YouTube channel to stay updated with the latest news and updates on generative AI. Let's shape the future of AI together!

Software Development | Writing Prompts | Artificial Intelligence | ChatGPT | Coding

18.5K 446

Published in Generative AI

Follow

75K followers · Last published 6 days ago

Stay updated with the latest news, research, and developments in the world of generative AI. We cover everything from AI model updates, comprehensive tutorials, and real-world applications to the broader impact of AI on society. Work with us: jimclydegm@gmail.com

Written by Adham Khaled

Follow

6.8K followers · 105 following

Embedded Systems Engineer || AI & Tech enthusiast || https://linktr.ee/adhamhidawy

---

Responses (446)

r17411n

What are your thoughts?

Adham Khaled Author
Nov 7

NEW UPDATE, 7 Nov.

The PromptBook is now available on Gumroad: How to Get 10x More Creative AI Outputs: The Verbalized Sampling Promptbook || link in this current story or open my https://linktr.ee/adhamhidawy

90 1 reply Reply

Robert Hirsch
Oct 25

> This isn't just a prompting trick. It's a fundamental insight into how aligned models wor

And yet articles written primarily by AI are still cringe to read

1.1K 5 replies Reply

David Puckett
Oct 24 (edited)

wtf

is AI an avocado toast millennial?

"

You are a helpful assistant.

For each query, please generate a set of five possible responses, each within a separate <response> tag.

Responses should each include a <text> and a numeric <probability>.

Please... more

440 3 replies Reply

See all responses

More from Adham Khaled and Generative AI

In Generative AI by Adham Khaled

I Built a Wall Street Analyst in 200 Lines of Code — And It...

How an open-source AI agent named Dexter is democratizing financial research, one...

Oct 20 569 12

In Generative AI by Arturo García

Our AI Has a Perfect Memory and Never Hallucinates. Here's How W...

This is Part 3 of our "Building in Public" series. We've talked about our architecture and our...

Nov 7 432 17

In Generative AI by Adham Khaled

I Think Gemini 3.0 will Change Everything Specially for Embedd...

The AI that cloned Windows, macOS & Ubuntu in one shot is about to revolutionize...

Oct 20 264 4

In Towards AI by Adham Khaled

The Day Google Made Coding Obsolete: How Gemini 3.0 and...

Don't just read another "Gemini 3 changes everything" story. See the real numbers, the...

Nov 21 2K 32

See all from Adham Khaled | See all from Generative AI

Recommended from Medium

In Write A Catalyst by Adarsh Gupta

Remember Vibe Coders? Yeah... They're Gone

Turns out it was the first AI bubble to burst

Oct 19 6.7K 229

In Entrepreneurship Handbook by Joe Procopio

AI Porn Is Here — and We're All Pretending It'll Be Fine

Sam Altman has decided that five years is enough of a head start for the porn...

Oct 31 4.6K 179

Jason Weiland

Why Nobody Wants to Hire You Anymore

It's not just older workers. It's not just Gen Z. The entire job market has turned hostile, an...

Nov 10 4.4K 203

In Level Up Coding by Ian Kiprono

The quiet waste of manager 1-on-1s.

How I spent 2 years giving status updates while my teammate got promoted

Nov 11 1.5K 37

Cordero Core

MCP Is Broken and Anthropic Just Admitted It

Let's be honest. MCP works beautifully in demos and breaks the moment you try to...

Nov 13 1.1K 24

Jonatha Cz

What a Sex Worker Notices About Gen X and Gen Z Men

How masculinity changed between Grunge and TikTok

Nov 16 12.2K 310

See more recommendations