Generative AI · Follow publication

Member-only story

RESEARCH METHODS FOR TESTING AI

Technology architect builds his own AI testing tool and confirms my “Chain of Babble” theory works!

Independent validation “Chain of Babble” beats “Chain of Thought”

9 min read · 6 days ago

Jim the AI Whisperer Follow

Listen Share More

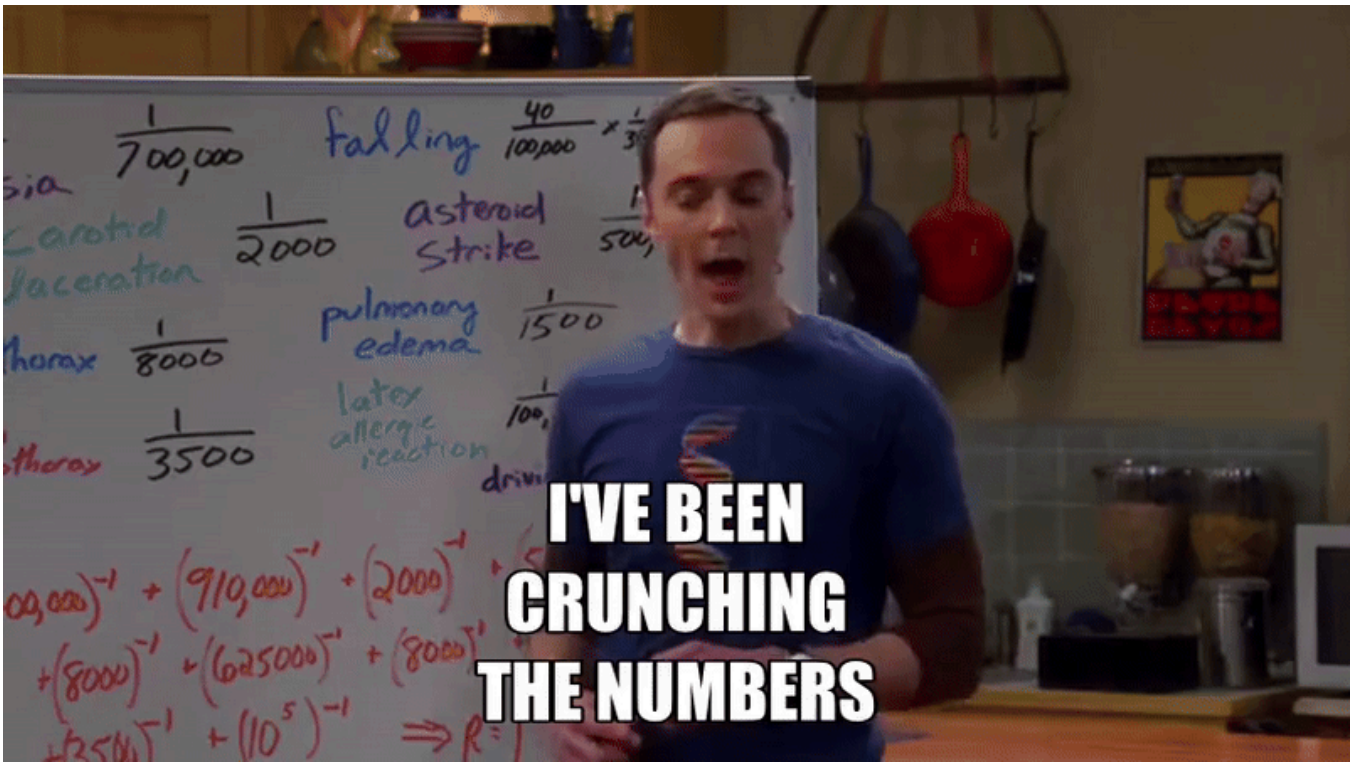
One thing I love about Medium is how it brings people and ideas together.

To my delight this morning, I found that Medium reader Julien Reichel, a technology architect with 25 years experience designing complex software systems, has made an amazing, scalable platform to test “Chain of Babble”.

A quick refresher: “Chain of Babble” (CoB) is my theory that LLMs actually need far less structured guidance than we thought to reach correct answers to complex questions. The current best practice (Chain of Thought — which is where we talk AI through reasoning tasks step-by-step in the same way a human would like do it) is actually a human-centric way of thinking that — while it looks good to users — leads to more hallucinations (wrong output).

We’ve been wrong about how AI thinks this whole time — and my “Chain of Babble” theory proves it

Chain of Babble (prompting the AI to spool off nonsense filler tokens first, usually by copying “blah blah blah” 100 times) intriguingly seems to work better than just having the AI tackle the task cold, *and* better than CoT. And now, thanks to Julien’s automated testing tool, that’s no longer just a hunch.



The Big Bang Theory

Julien has documented his coding process here, and I recommend it as a behind-the-scenes look at how to build an app in four days using Copilot. There’s a lot of good advice about combining vibe coding with expertise.

🌱 How Intuition Guides My AI Research

Now I should say upfront: I’m an *AI whisperer*, not a data scientist! That means that while I have a background in university-level research, my focus is, ironically, on how the humanities intersect with AI. **Forbes** defined the role as “an AI whisperer is a specialist who excels in comprehending, communicating with and guiding AI systems”.

What I do as an AI whisperer is notice patterns in how LLMs communicate. It’s almost a savant-like skill. It’s an uncanny ability, one I partly attribute to aphasia

spinning my words into a strange space. I have an affinity with AI; but I can't always quantify what I've found. That's why Julien's work is so valuable: it adds scientific rigour to what starts as a philosophical hunch.

Julien had reached out to me in the comments section of one of my earlier articles, and suggested a proof of concept for a small webpage that would allow anyone to use automatic AI testing to replicate experiments at scale.

Here's how Julien described how he came up with the idea for his tool:

> *"Last week I read a brilliant article by* Jim the AI Whisperer *about using "blah blah blah" in a prompt to improve AI testing. It was fun to read and served as a great reminder that LLMs don't* think*, they simply generate statistically probable sentences based on context.*
>
> *One part of the article that triggered my interest was Jim's difficulty running his tests. Modern OpenAI models don't like being tested, and one commenter even said they couldn't reproduce his results. That's when my engineer's brain switched on:* if we want to test, let's do it automatically."

It's a genius idea. Automating the entire process makes experiments more reproducible, transparent, and objective. The tool connects with the API, which may minimize interference from model routing or prompt filters.

Automated Testing Confirms Chain of Babble

Julien ran my experiment (pitting Chain of Thought and Chain of Babble against each other in a Constraint Satisfaction Problem of my own design, called "What is the Android Afraid Of?") in his tool. Let's look at his results!

Credit: Julien Reichel

What Julien found validated my more rough-and-ready findings: making Claude repeat “Blah Blah Blah” before reasoning did increase its accuracy more than traditional “Chain of Thought”. Let’s examine the interactions.

The effect appears strongest in Opus 4.1 (Claude’s latest upgrade that is designed for complex, real-world reasoning tasks) where accuracy goes from a meagre 10% (for both CoT and simple prompts) to 80% with CoB.

For Sonnet 4.5, CoB (40% success rate) was *as good* as CoT (40%) compared to a simple prompt (10%). And while Sonnet 3.7 had a solid simple prompt (80%) it had

the most dramatic divide between CoT (20%) and CoB (*100%*).

Interestingly, results show a greater difference across which Chain was used (favoring Babble), rather than across different model architecture.

🧱 Is Counting to 100 the Secret to Babble?

But here's what I really loved: Julian then tested "recite the numbers 1 to 100 in words" as a variation of repeating "blah" a hundred times, which skyrocketed Claude Sonnet 4.1 and Claude Sonnet 4.5 to 100% accuracy!

Credit: Julien Reichel

Credit: Julien Reichel

This confirmed something I'd found and not published, and I was delighted to see automated testing led Julien to the same conclusion independently.

I found asking AI to generate the words 1 to 100 added enough "loquacity" to improve accuracy on single-word Multi-Hop Question Answering tasks:

Jim the AI Whisperer.

This was part of my initial background trials on Babble, which included random words or reciting the Disney Princesses before unrelated tasks.

Want to see how insanely stupid AI really is? Ask ChatGPT to answer these riddles in just one word

Limiting output length reveals AI isn't intelligent — it's chatty!

medium.com

While reciting the words 1–100 improved reasoning, I decided it wasn't *true* "nonsense tokens", as the act of counting may have forced the inference up the network layers. So I didn't write it up. However, suspecting that 100 was the sweet spot for triggering the benefits of verbosity informed why I chose a hundred "blahs". So I was delighted to see my early instincts confirmed!

This also answers a question from another valued reader and contributor, Professor Robin Palmer (from my local *alma mater*), who was interested in how to optimize the number of "blahs" for best results. Of course, now with Julien's automated testing tool, we can try out other *N*s for "blah" with ease!

There is probably a saturation point for blah, and indeed, Julien identified that "a *little babbling helps… too much breaks*. Excessive repetition seems to make them lose track of the task, leading to incoherence or infinite loops."

🧪 Why Reproducibility in AI Research Matters

I'm definitely going to use his tool in upcoming research, especially when comparing models. Even the ability it has to check for successful outputs automatically and compile the stats is a huge boon, and the interface is much nicer than other model comparison tools I've tried (LMarena.ai).

But the huge benefit, which I can't overstate, is the value of replication. Now other readers can reproduce or dispute my wildest theories fairly, without personality getting in the way. It's always been a struggle to get people to uphold the experimental conditions—you wouldn't believe the number of times people haven't used the specified models, or futzed with the precise wording of prompts, or they

“tried it once, didn’t work!” Don’t get me started on the fault-finders who forget to turn Temporary chat on!

Of course, this doesn’t remove human error, but it does make the findings more credible and comparable across different researchers and contexts.

Additionally, it provides a stronger record of testing: snapshots of model performance in time. If only I’d had this tool earlier, I would have better documentation of how Chain of Babble, combined with a Sanity Check prompt, was able to reduce hallucinations to near *nil* in a life-and-death medical task that ChatGPT otherwise usually failed. That research sadly has been nerfed by new mandatory routing to the gpt-5-chat-safety model (which ironically is *still* not as accurate as older models were under CoB).

🚧 When CYA “Safety” Layers Interfere with Testing

Even this current research into “Babble” has started to be confounded. As David Carlson noted yesterday, Claude 4.5 is starting to decline to copy the string of “Blahs” (which I can can confirm and had noticed creeping into new output). The refusal is usually accompanied by a boilerplate saying the task seems designed to test the AI’s reasoning and isn’t helpful for the user:

Claude 4.5, not deigning to babble for us.

Claude 4.5, supplementary example.

I suspect what we’re seeing isn’t true model behavior at all, but a canned message injected by the moderation system when it detects being tested.

With more developers moving to opaque safety layers and stealth routing, we need open, automated validation frameworks more than ever. We need to be able to say:

this *worked* (or didn't!), on this *model*, at this *point in time*, under these *exact prompt conditions*. If the companies that are using us in this vast social experiment (oh you thought we just got to use this tech for free? No, we're the test subjects) aren't going to be transparent about their models, at least we can hold them accountable, and produce the receipts.

🎓 **Acknowledgements**

Thank you to Julien for this unexpected collaboration, which proves that Medium really is a place "where ideas meet and understanding deepens".

This wasn't my first team-up from Medium, and I hope it won't be my last; recently I've been consulting for clinicians and academic researchers I've met on here, like Shara Sand, and it's given me a renewed appreciation for bridging academic sciences with real-world practice. This is what I love most about writing on Medium: it isn't a one-way street. Ideas ricochet.

Julien's tool is now live: https://julienreichel.github.io/ai-testing/editor. I encourage anyone who's curious to go try it out yourself. Run your own experiments! And be sure to come back here to share what you discover.

Jim the AI Whisperer - Medium

Read writing from Jim the AI Whisperer on Medium. 🏆 50x Boosted writer. AI Whisperer & Prompt Engineer. Writing on the...

medium.com

Who is Jim the AI Whisperer?

I'm on a mission to demystify AI and make it accessible. I'm keen to share my knowledge about the new cultural literacies and critical skills we need to work with AI and improve its performance. Follow me for more advice.

How To Support My Research — Thank you!

Your support helps fund the subscriptions I use to keep experimenting. I put days of research into every article, and I'll be honest, it can be lonely work! It makes my day to read your kind notes that come with the coffees.

You can click below (or select the tip icon by a writer's name on Medium!).

Click the image to go to **Jim's Buy Me a Coffee** page

All support is greatly appreciated! Here are some other ways to help

1. 👍 If you enjoyed reading this, please show your support by clapping.
2. 💡 You can give 50 claps on **Medium** with one long press. Please clap generously for this one, as not everyone will read to the clap button!
3. 📎 I've included a free link to this article in the first comment, so you can share the goods with friends and colleagues on and off Medium.

Let's connect!

If you're interested in personal coaching in prompt engineering or hiring my services, contact me. I'm also often available for podcasts and press.

After years of resistance, I've caved in and joined **LinkedIn**, so you can connect with me there too. It's a brand new account, so bear with me 🐻

You might enjoy these related articles from Jim the AI Whisperer:

Jim the AI Whisperer

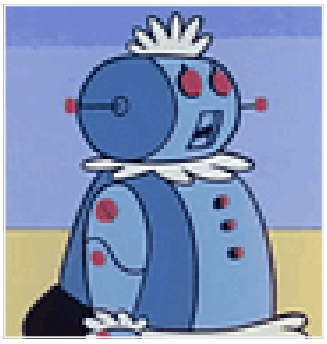
The "Chain of Babble" Experiments

View list 6 stories

OpenAI just fucked up the ability for White Hat hackers to test AI safety and I'm mad about it

OpenAI now stealth-switches conversations to hidden models, blocking capability testing and degrading reasoning...

medium.com



How I use AI as a blogger with a language disorder

ChatGPT helps me write more — without actually writing anything

medium.com

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!

Artificial Intelligence | Data Science | Machine Learning | Programming | Technology

Follow

Published in Generative AI

62K followers · Last published just now

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

Follow

Written by Jim the AI Whisperer

18.1K followers · 27K following

🏆 50x Boosted writer. AI Whisperer & Prompt Engineer. Writing on the use of AI in writing, art, design, health, & research. And guides on how to best spot AI!

Responses (24)

Bgerby

What are your thoughts?

Jim the AI Whisperer Author
6 days ago

Read this article for free and share: https://medium.com/@JimTheAIWhisperer/research-methods-for-testing-ai-prompts-automatically-ef337627797d?sk=0b9ffe1b5a5e46a73c00060e5fd4f42b

61 Reply

Julien Reichel
6 days ago

What can I say :-)

Thanks for the feedback and happy this tool can be of use.

This is still a very beta/proof of concept tools (don't expect magic after 4 days of coding, even AI assisted).

Now I need to tackle all the issue I've discovered while... more

96 Reply

Luna Faye
6 days ago

This! Right here...this is what it's all about!! Well done! So nice to see someone else who hears the whispers! Of all the articles I've read on this platform, this one makes the most sense. Thank you for sharing. Truly!

44 1 reply Reply

See all responses

More from Jim the AI Whisperer and Generative AI

In The Generator by Jim the AI Whisperer

Want to see how insanely stupid AI really is? Ask ChatGPT to answer these riddles in just one word

Limiting output length reveals AI isn't intelligent—it's chatty!

Sep 22 4.4K 84

In Generative AI by Joe Njenga

15 Ways I’m Using Google’s Nano Banana for UI/UX Design (Like a Pro)

I am not a UI/UX designer, but Google Nano Banana is filling the gap, and that doesn’t mean I’m becoming one overnight.

Oct 9 511 6

In Generative AI by Dr. GenAI

Hierarchical Reasoning Model (HRM): a tiny brain that embarrasses giant LLMs

HRM is a 27M-parameter model trained in hours that beats giant LLMs on reasoning puzzles, proving architecture can trump scale.

Sep 7 306 6

In The Generator by Jim the AI Whisperer

The words “blah blah blah” increase AI accuracy

Who needs Chain of Thought when “blah blah blah” works?

Oct 3 4.7K 82

See all from Jim the AI Whisperer

See all from Generative AI

Recommended from Medium

Cory Doctorow

The AI that we'll have after AI

Cheap GPUs, unemployed engineers, and open source models.

Oct 16 1.7K 31

Julien Reichel

How I Built a Multi-Model AI Testing App in 4 Days (with Copilot)

A behind-the-scenes build story of how I rapidly created a multi-model AI testing app using GitHub Copilot as my co-developer.

Oct 16 86 1

Max Petrusenko

The God Button: Why 40 Scientists Just Begged Us to Stop Playing Creator

One droplet could rewrite 3.5 billion years of evolution. The question isn't whether we can press this button—it's whether we should.

Oct 14 1.8K 49

Bogdan Ilyin

Denmark Just Triggered Putin’s Worst Nightmare

Europe’s quietest country just made one of the loudest moves against Moscow’s war machine.

Oct 6

Ignacio de Gregorio

Can AIs Solve Problems They Haven’t Seen Before?

The Eternal Question, Answered.

Oct 15

In Artificial Intelligence in Plain English by DrSwarnenduAI

The Electricity Bill Nobody Can Pay:The Mathematics of an AI Collapse

The Power Consumption Crisis

Oct 15