Unleashing Potential: Scaling AI for Impact, Safety, and Global Change

2024 IEEE Emerging Technology Reliability Roundtable

John Burkey, May 2024

Who am I

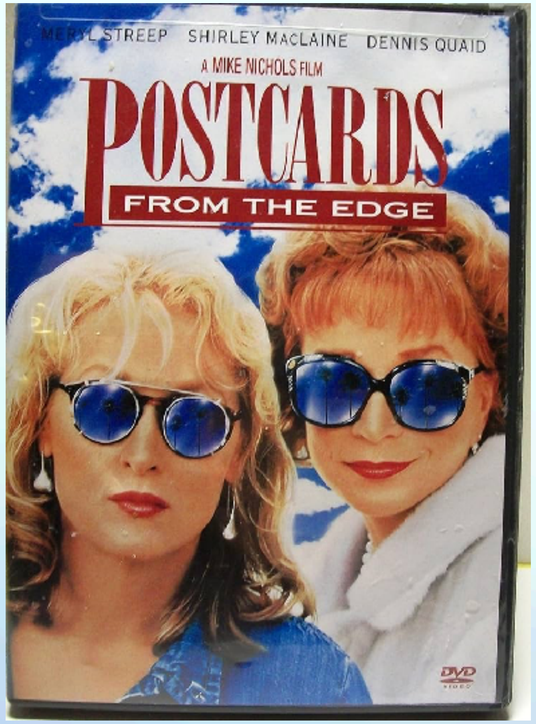
John Burkey

Shipped award winning Apps in the 90's
Part of Apple's near death re-emergence in 2000
Educational startup in 2004
Java chief architect at Sun in 2008
Microsoft chief architect for Apple and 'Droid in 2010-14
Siri Advanced research team 2014-16

Launched several Ai startups since:
Generative Ai, Ai hardware, and now scaling systems.

Postcards from the Edge

Real experiences from 2023-4



Driving Medical Ai from 70% to 95+%* accuracy

Scaling to dozens of prompts

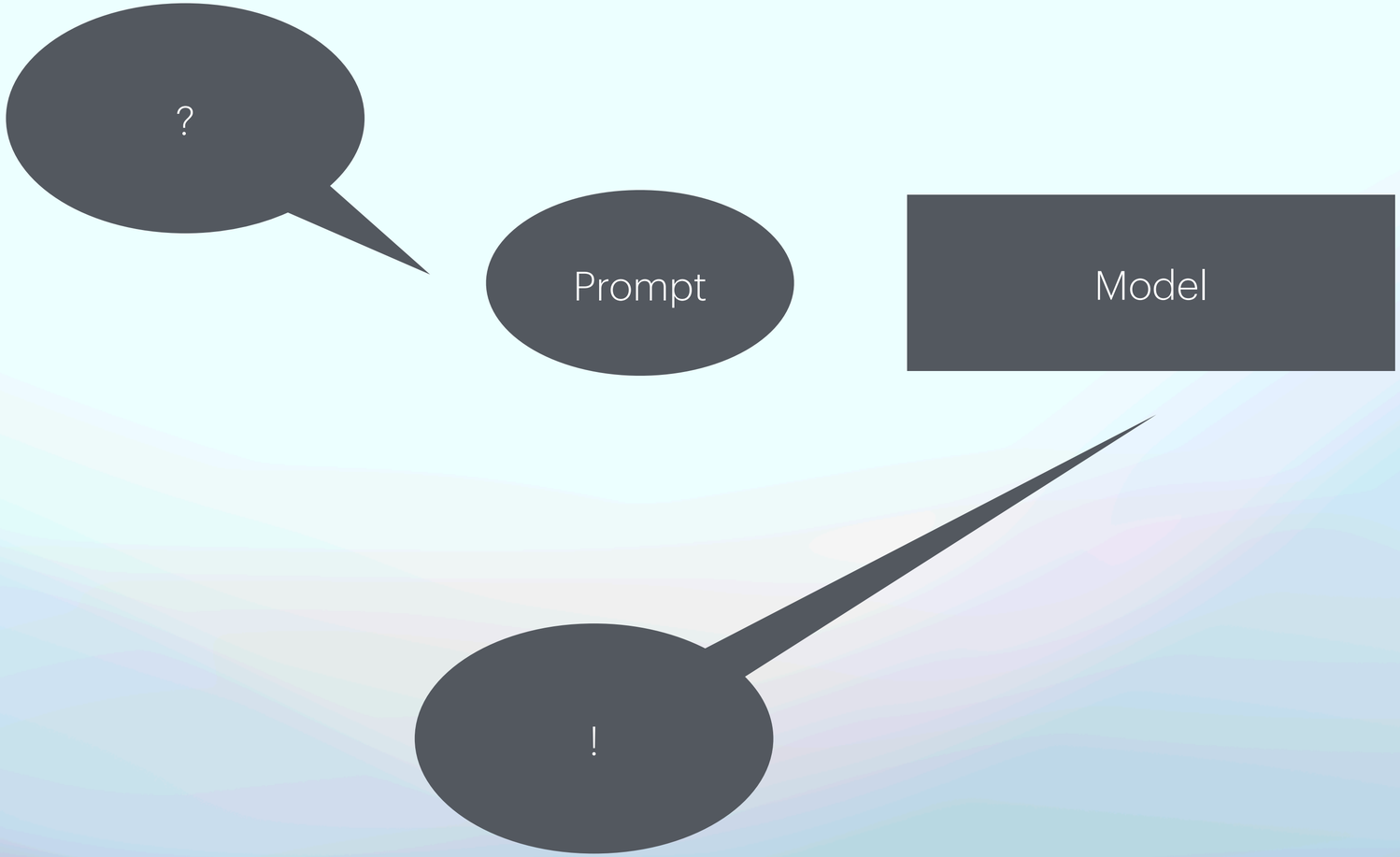
Building AI concurrency

Examine the nature of generative Ai

Driving it towards reliability

Prompts and Python

Asking models to do our bidding



What is GenAI

generates what society “knows”

Alot of *hallucinations* are combinations of “creative Ai’s” and not enough context in inputs.

They are perfectly acceptable outputs given randomness and less-focused requests

Or requests in information sparse areas of its knowledge, and it makes more “mental leaps to mansplain.” (Like Dad?)

How do we increase reliability ?

And with reliability,
we get scale.

What do we do?

We got this

We are engineers

We bend science to our will

Let’s look at the input parameters for genAI

GenAi (GPT's)

Or could be images, data, actually, but you still get stat parameters that adjust randomness

Take:
1) Lots of text
2) Parameters on how to do statistical generation of answers.

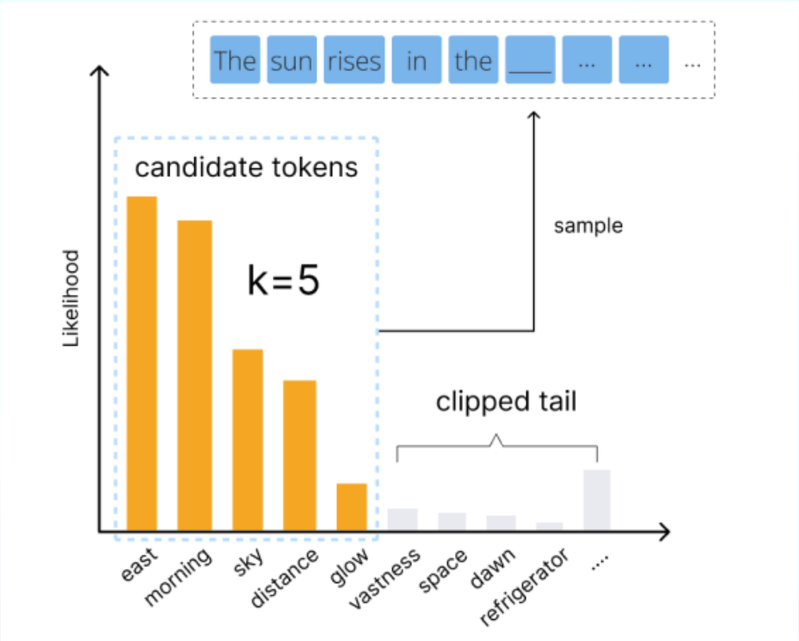
Temperature & TopN

goal directness and creativity

Temperature: Controls randomness, higher values increase diversity.

Top-p (nucleus): The cumulative probability cutoff for token selection. Lower values mean sampling from a smaller, more top-weighted nucleus.

Top-k: Sample from the k most likely next tokens at each step. Lower k focuses on higher probability tokens.



Temperature & TopN

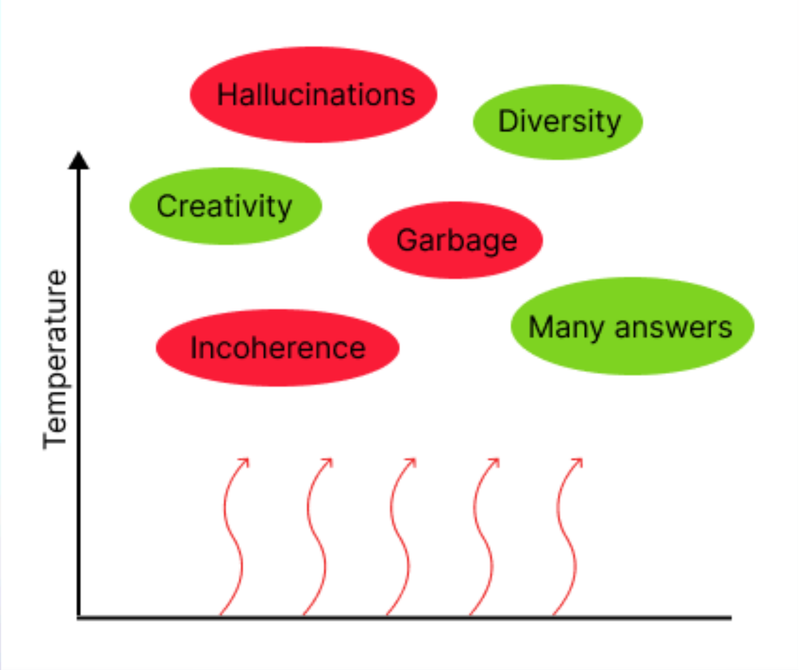
goal directness and creativity

Typically you want to emphasize stability in parts of prompts and creativity in others.

Higher temperature will make outputs more random and diverse.

Lower top-p values reduce diversity and focus on more probable tokens.

Lower top-k also concentrates sampling on the highest probability tokens for each step.



Temperature & TopN

goal directness and creativity

Typically you want to emphasize stability in parts of your output and creativity in others.

In a song, you want approachable lyrics, and cool chord riffs- not creative new words and boring chord riffs!

So one man's hallucination is another's a-hah!

So we need to control randomness

And NOT by hoping the foundation model provider does!

If we subdivide our problem like good engineers, we can take control.

Let’s proceed!

Move from one prompt to many

Think of it like having a team of AI's, That send work to each other

Some of the team wear black and have crazy ideas! Put them in the right places.

When we subdivide a problem we can devote greater randomness where it matters.

For each prompt...

How do we make it safer, more ethical?

Alot of *hallucinations* are combinations of high temperatures (more “statistical creativity”) and not enough context in inputs.

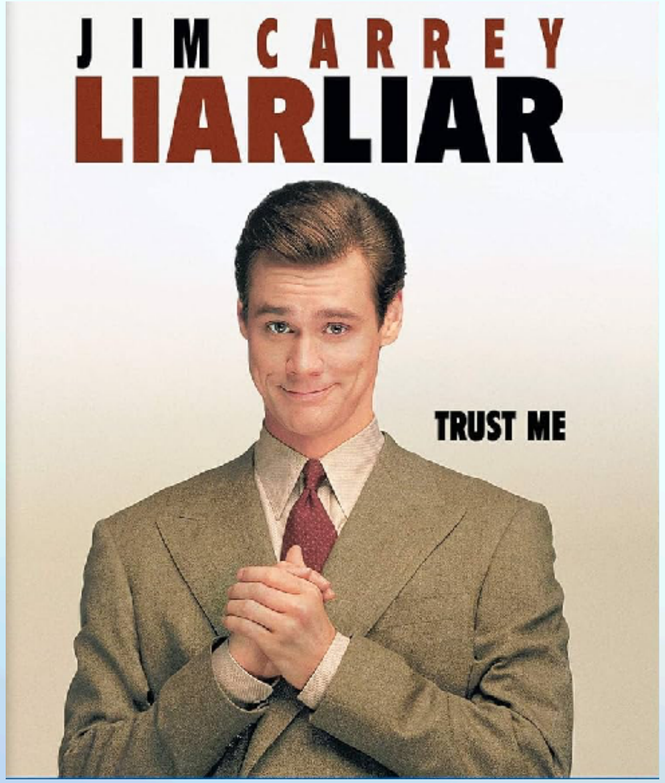
Introduce grading

Like an Ai's subconscious editor

Is that really what I want to say?

Safety and the subconscious

Humans and Ai

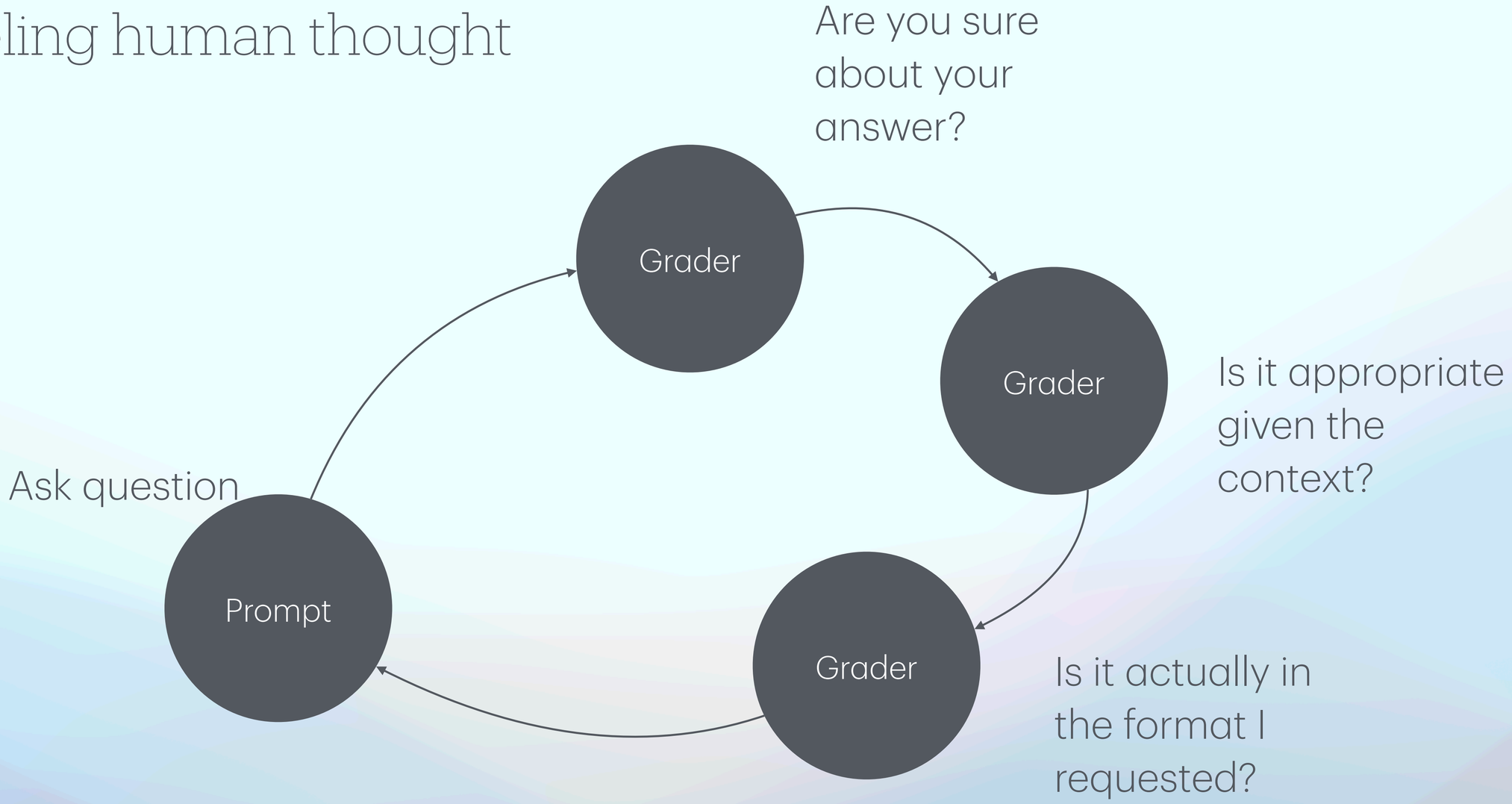


When someone asks us a question,
if we aren't Jim Carrey in Liar, Liar,
we don't blurt out an answer!

Why should an Ai?

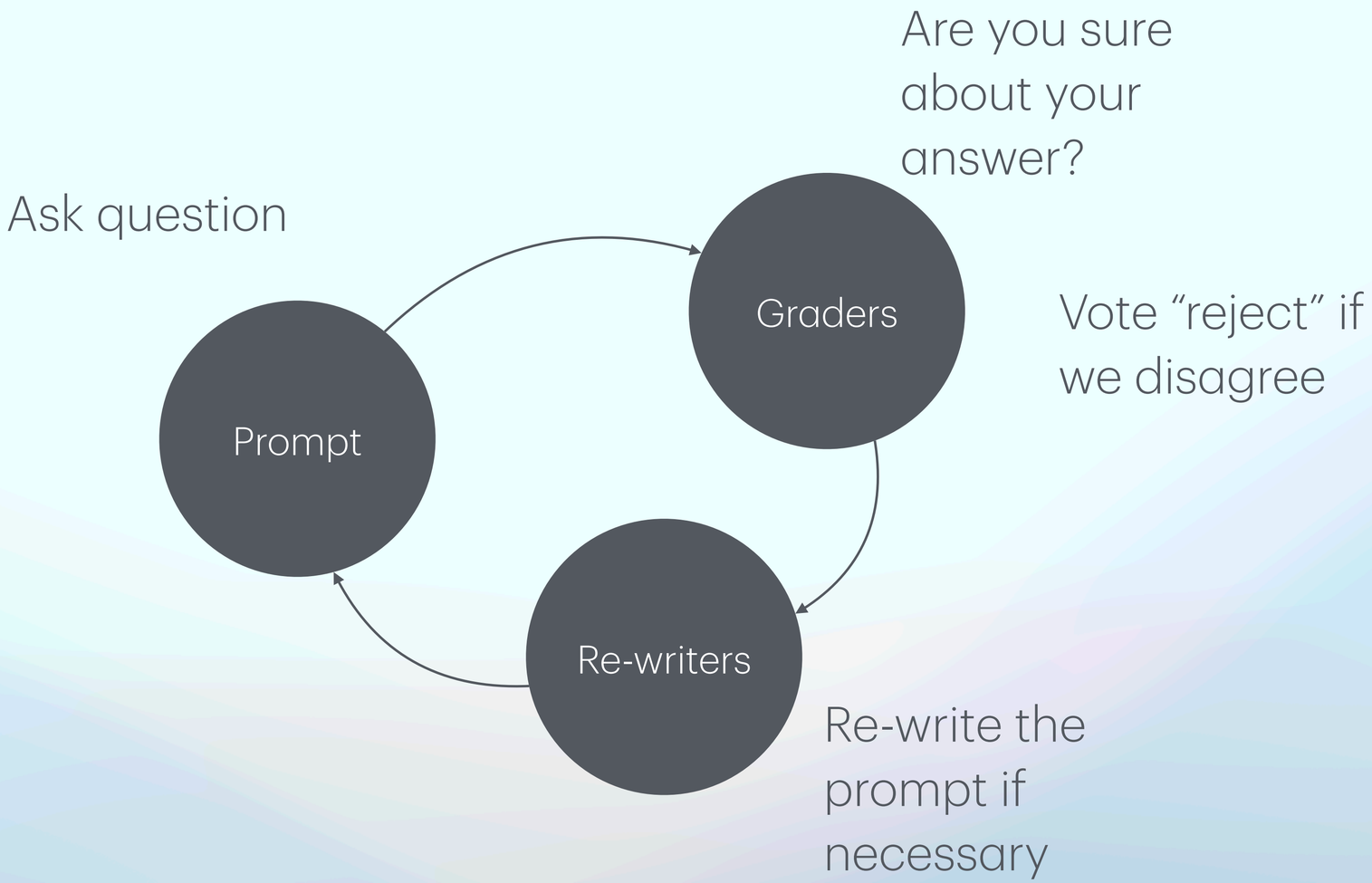
Safety and the subconscious

Modeling human thought



Safety and the subconscious

Modeling human thought



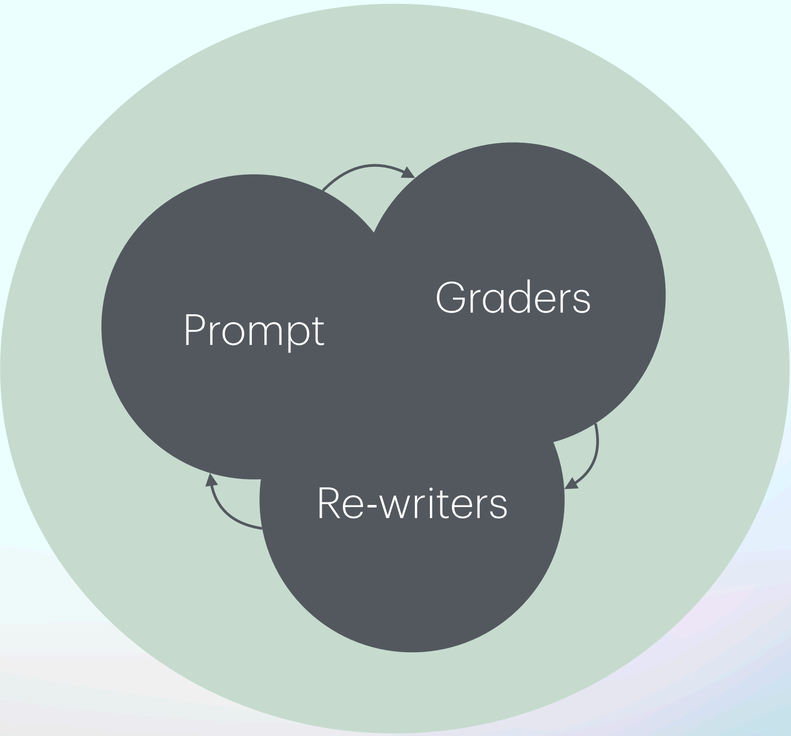
Safety and the subconscious

Modeling human thought

Thinker

Error correcting prompt with auto re-write

Not perfect, but better



Let’s dive into an example

For scaling AI to something useful

Design an advertisement

For Nutella

Targeting urban youth

That shop at kiosks

And watch the micro-influencer Joe-dog

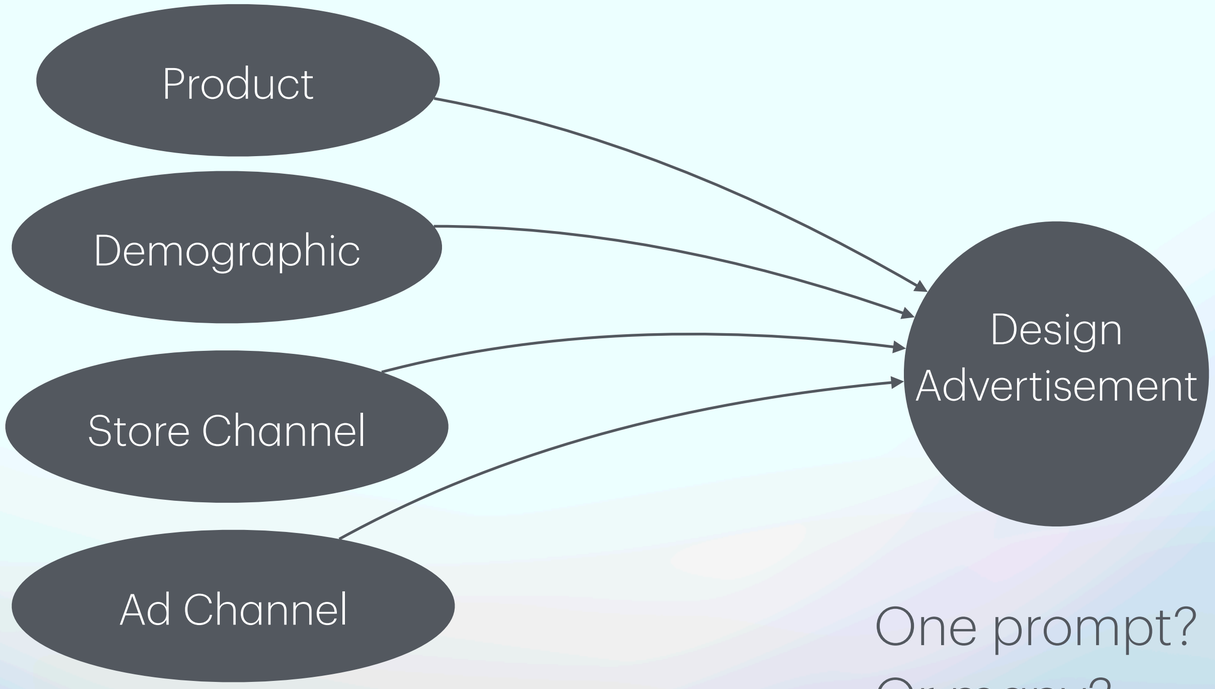
Design an Ad..

For Nutella

Targeting urban youth

That shop at kiosks

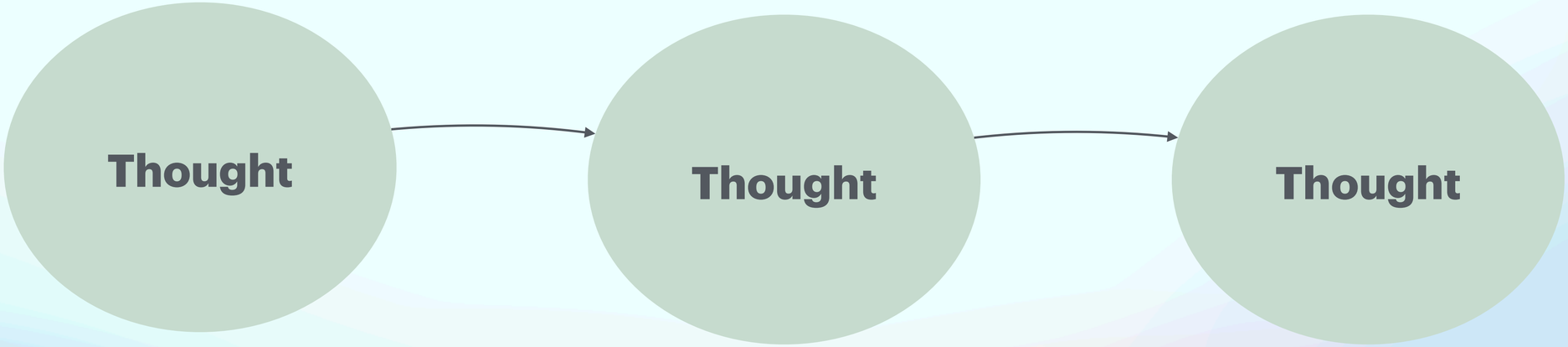
And watch the micro-influencer Joe-dog



One prompt?
Or many?

Use many Prompts

With error correction along the way
And different parameters and models

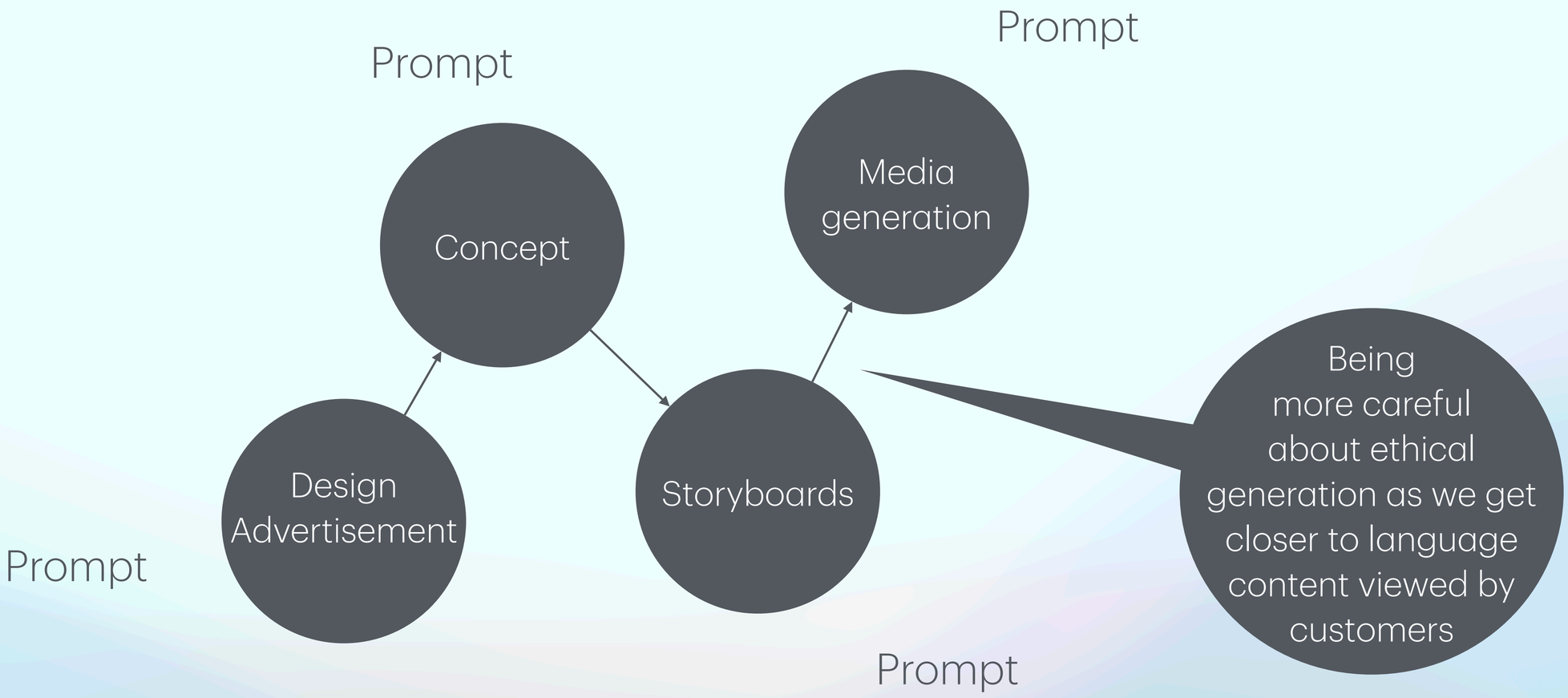


Create an ad concept

Create storyboards

Generate Artwork

Remember like a team. And just like an ad agency, you interact with different stages in different ways...


Prompt
Prompt
Prompt
Prompt
Design Advertisement
Concept
Storyboards
Media generation
Being more careful about ethical generation as we get closer to language content viewed by customers

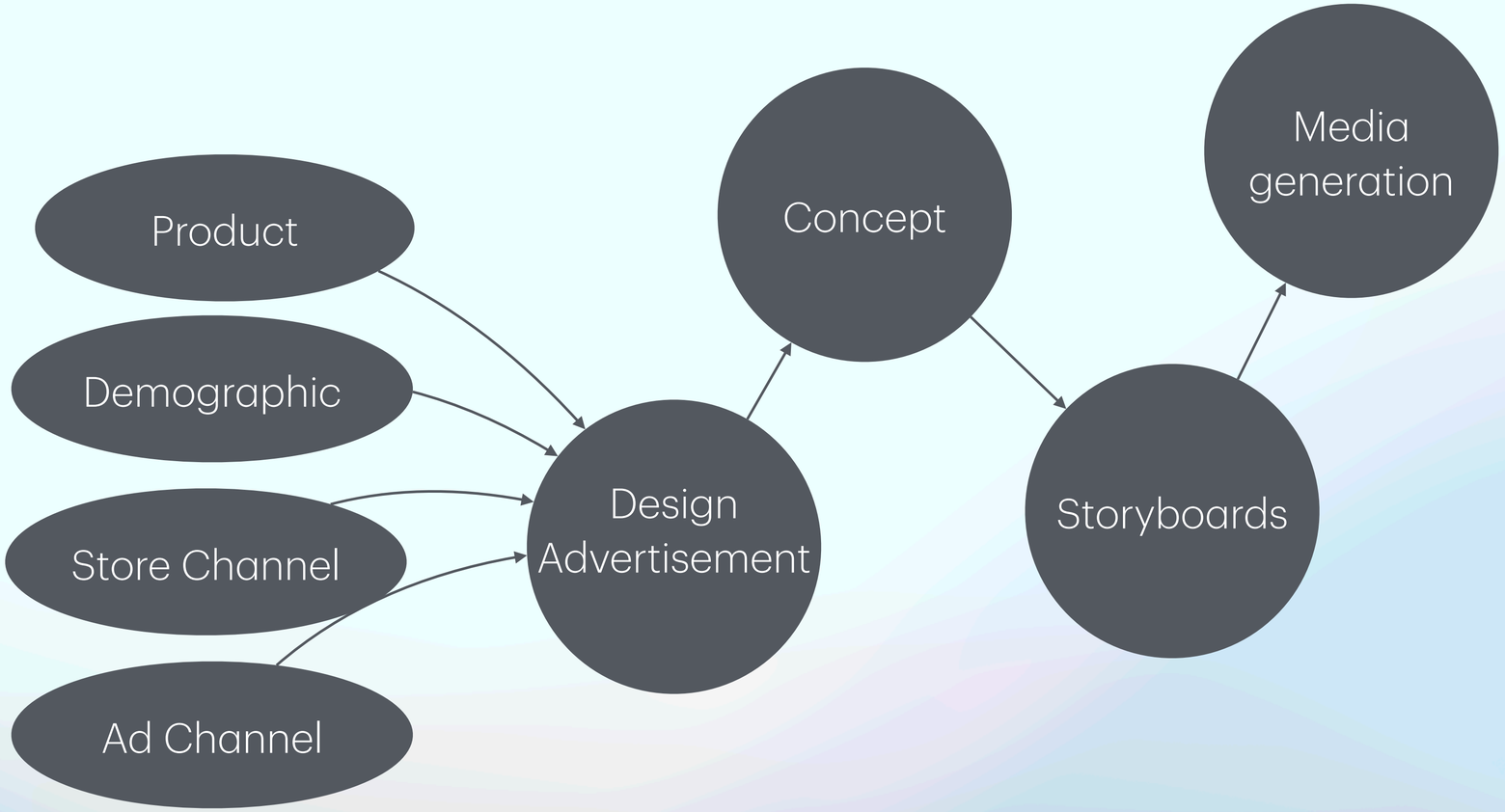
Design an Ad..

For Nutella

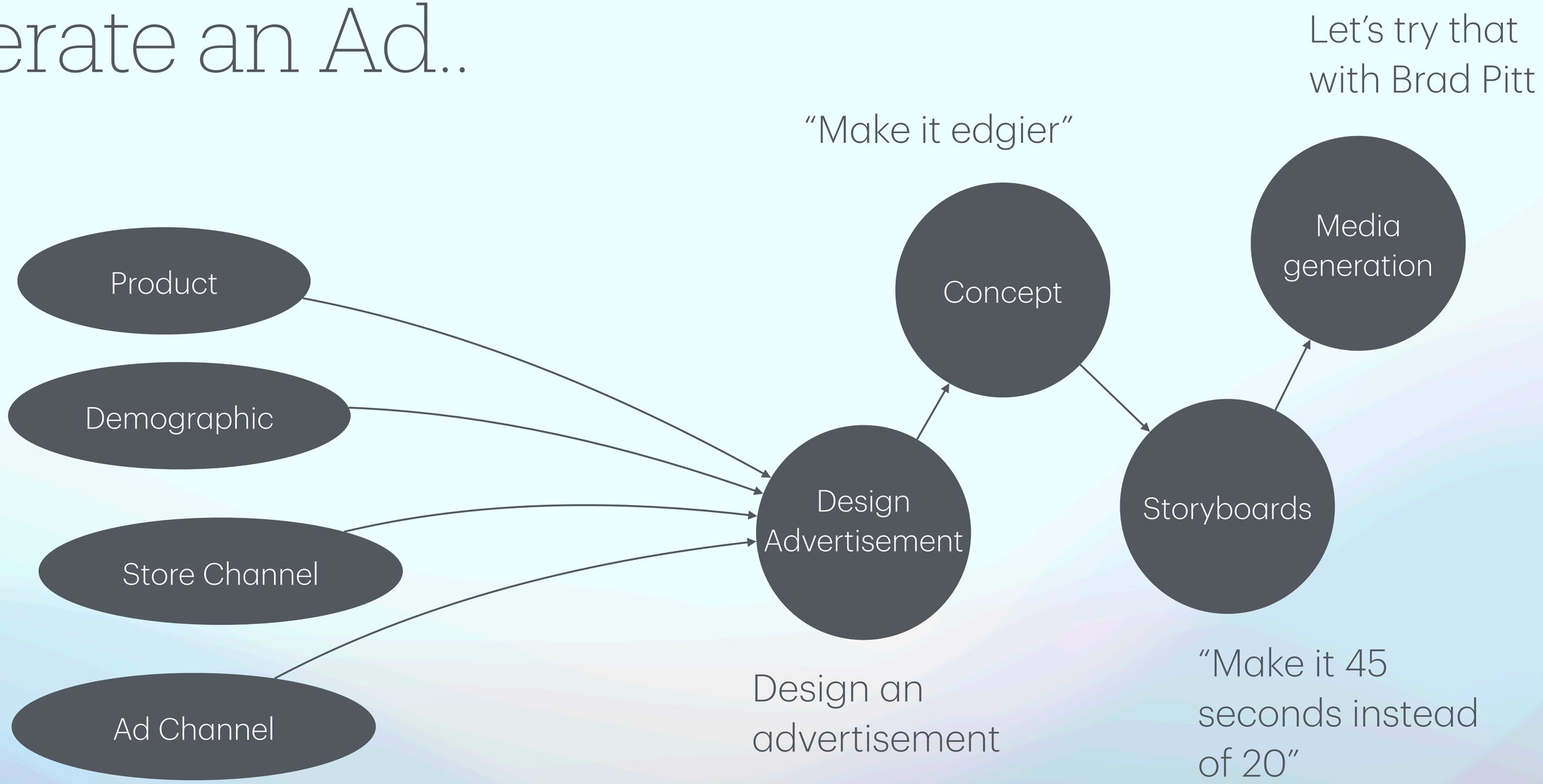
Targeting urban youth

That shop at kiosks

And watch the micro-influencer Joe-dog



Iterate an Ad..



Now let’s scale !

Scaling up to value

Instead of generating one Ad, use genAI inputs to generate MANY

Feed context variables for micro targeting
Generate
Measure response
Iterate

Scaling up to value

MANY MANY

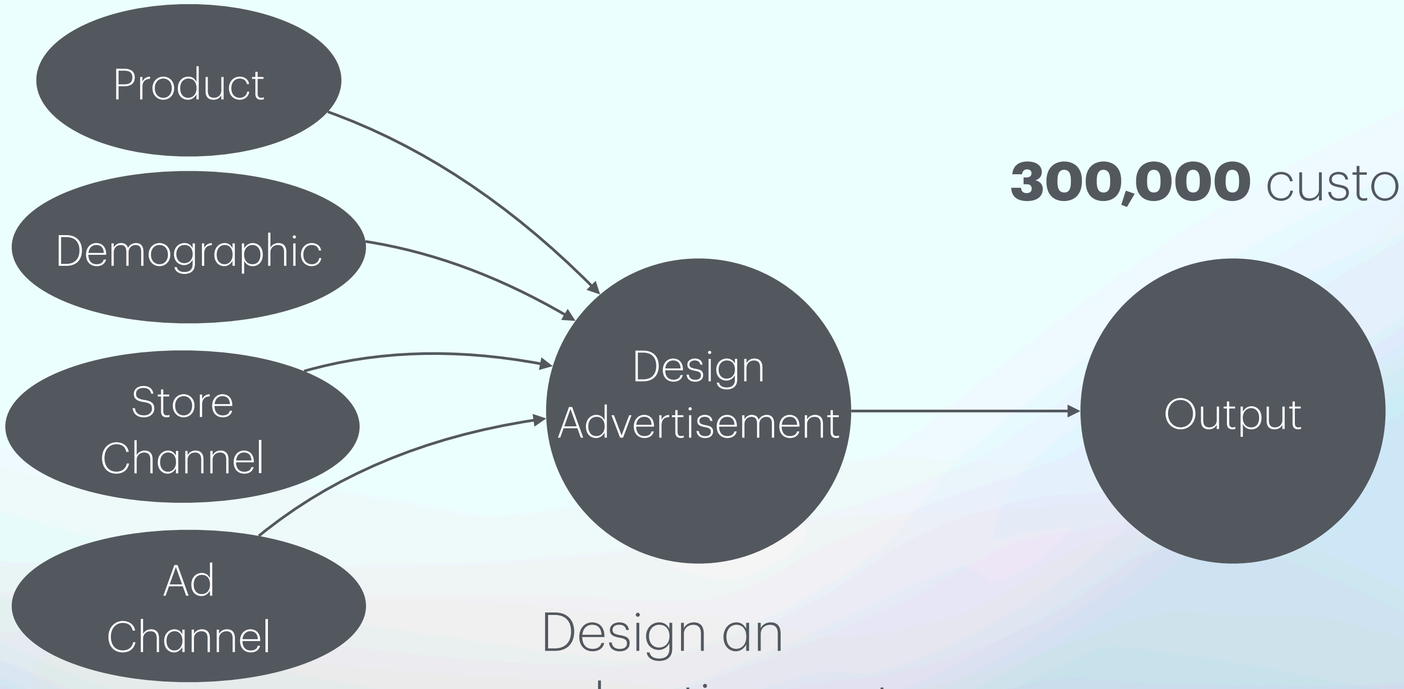
Add **10** products

Add **10** demographics

Add **30** stores

Add **100** micro-influencers

300,000 custom ads



300,000 custom ads

Design an advertisement

And because each output targets a micro audience, each is less risky.

So AI wins on cost, time, and risk

Part of this moment is looking for where the risk flips over to favor AI

Scalability & Transcendence

from novelty to value and transformation

Move from Single massive prompts to multiple prompts

Integrated Graders at each prompt boundary

Controlled Randomness / Creativity by prompt

Contextual safety for each node

(Safety / Creativity / Value by prompt)

Immense Value in repeatable workloads

from novelty to value and transformation

Scale in Size and Speed

Design for inherent safety

Systems get better over time automatically

And with human feedback

How do we scale?

As always with engineering

Our enemy is Chaos

Ethics / Reliability / Scale

Corner the randomness

Make sure its needed

Measure , reject , iterate

Multi-modal

GenAI is really matrix to matrix with token emission

Any data input can be described that way

For large tabular data, run an aggregation first

From many, one

By blending each prompt's grading for ethical behavior, safety, and expected output range

We get holistic safety

From safety, scale

The greater a system's reliability, the greater its ability to scale

Or, mean time between failures, matters..

Decompose for localized safety

Recompose for simplicity

At the first level on the path he saw mountains as mountains and rivers as rivers.

On the second level of the path he saw that mountains are not mountains and rivers are not rivers.

And at a third level he saw once again mountains were mountains and rivers were rivers.

So we have many prompts within,

that together solve the original

request

Thank you