

# From Slackbot to ChatGPT

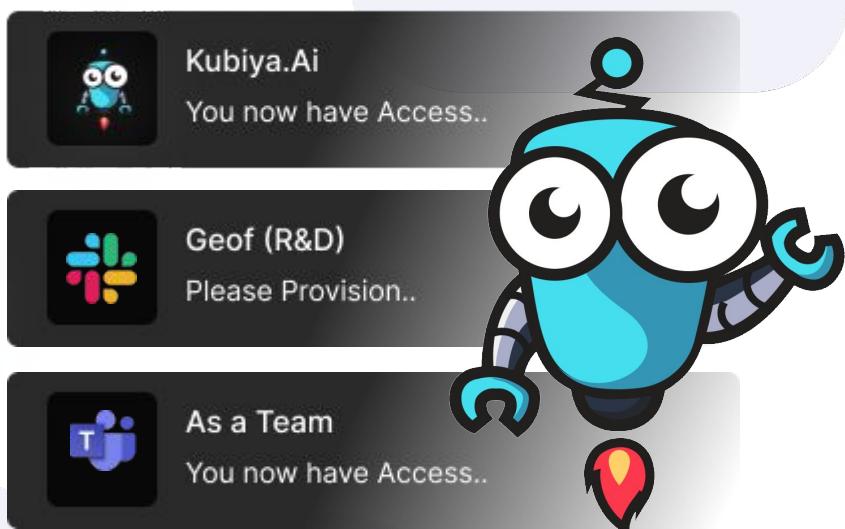
The Journey to Automating DevOps



Shani Shoham  
shani@kubiya.ai



Omer Cohen  
Omer.cohen@kubiya.ai



MEET SHAKED



# MEET SHAKED



800!!



jira.teamsinspace.com

Teams in Space  
Software project

Board

Release ...

Quick Filters

TO DO 5

Engage Jupiter Express for outer solar system travel  
**SPACE TRAVEL PARTNERS**

TIS-25

IN PROGRESS 5

Requesting available flights is now taking > 5 seconds  
**SEESPACEZ PLUS**

TIS-8

CODE REVIEW 2

Register with the Mars Ministry of Revenue  
**LOCAL MARS OFFICE**

TIS-11

DONE 8

Homepage footer uses an inline style-should use a class  
**LARGE TEAM SUPPORT**

TIS-68

Engage Saturn Shuttle Lines for group tours  
**SPACE TRAVEL PARTNERS**

TIS-15

Draft network plan for Mars Office  
**LOCAL MARS OFFICE**

TIS-15

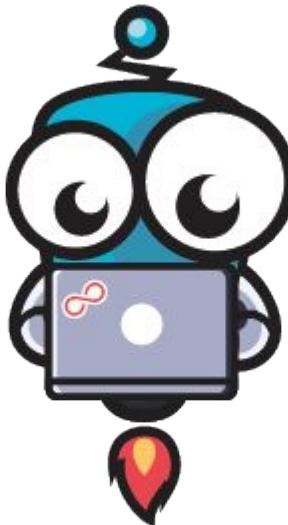
Engage JetShuttle SpaceWay  
**SPACE TRAVEL**

TIS-5

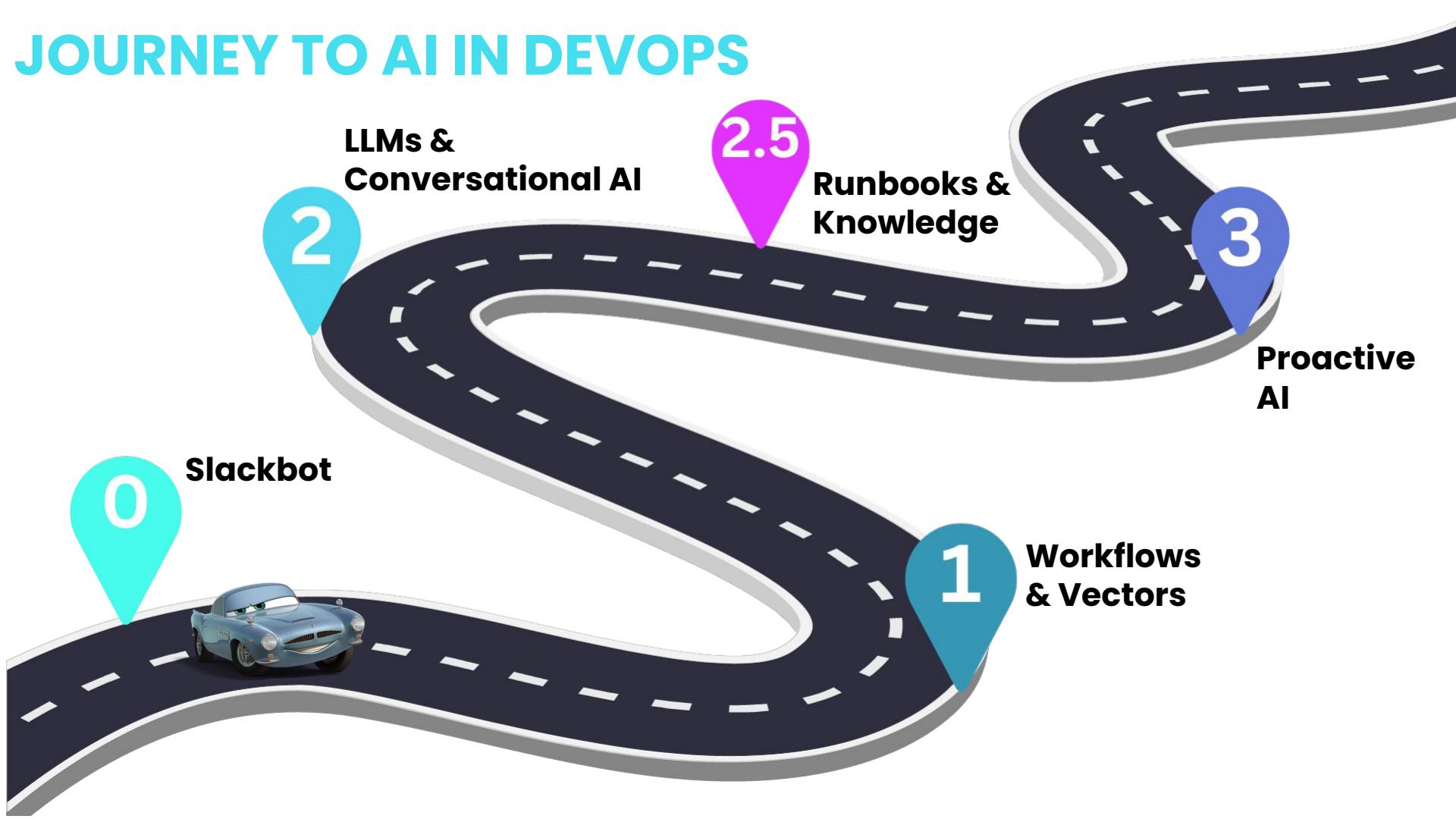
A screenshot of a Jira board titled 'Teams in Space Software project'. The board has four columns: 'TO DO', 'IN PROGRESS', 'CODE REVIEW', and 'DONE'. A large red starburst graphic with the text '800!!' in white is overlaid on the board. A lightning bolt graphic originates from the bottom left and points towards the board. A WhatsApp icon with a '+20' notification is positioned above the board, and a Slack icon with a '+99' notification is positioned below it.



MEET VIRTUAL SHAKED



# JOURNEY TO AI IN DEVOPS



# GEN 0 – WHY SLACKBOTS DON'T WORK

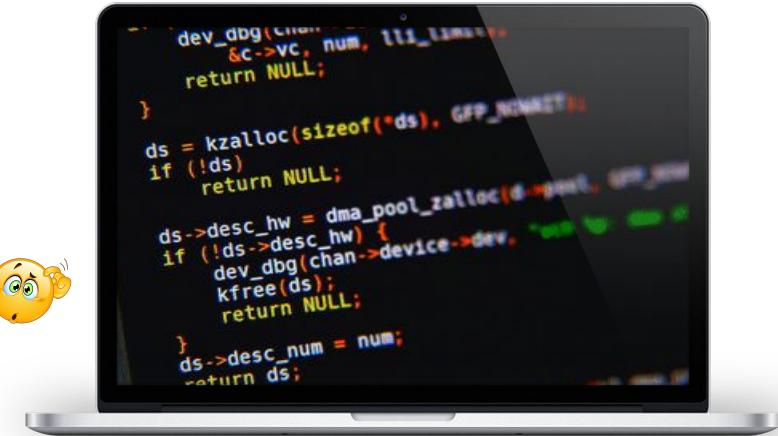
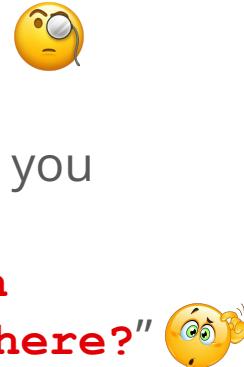
- # Human factor
  - Not a natural experience
- # Strict experience
  - No approvals
  - No “**Help me get the information**”
- # Don't learn
- # Requires lots of maintenance
  - Every addition requires a change
- # No access control
- # No RBAC



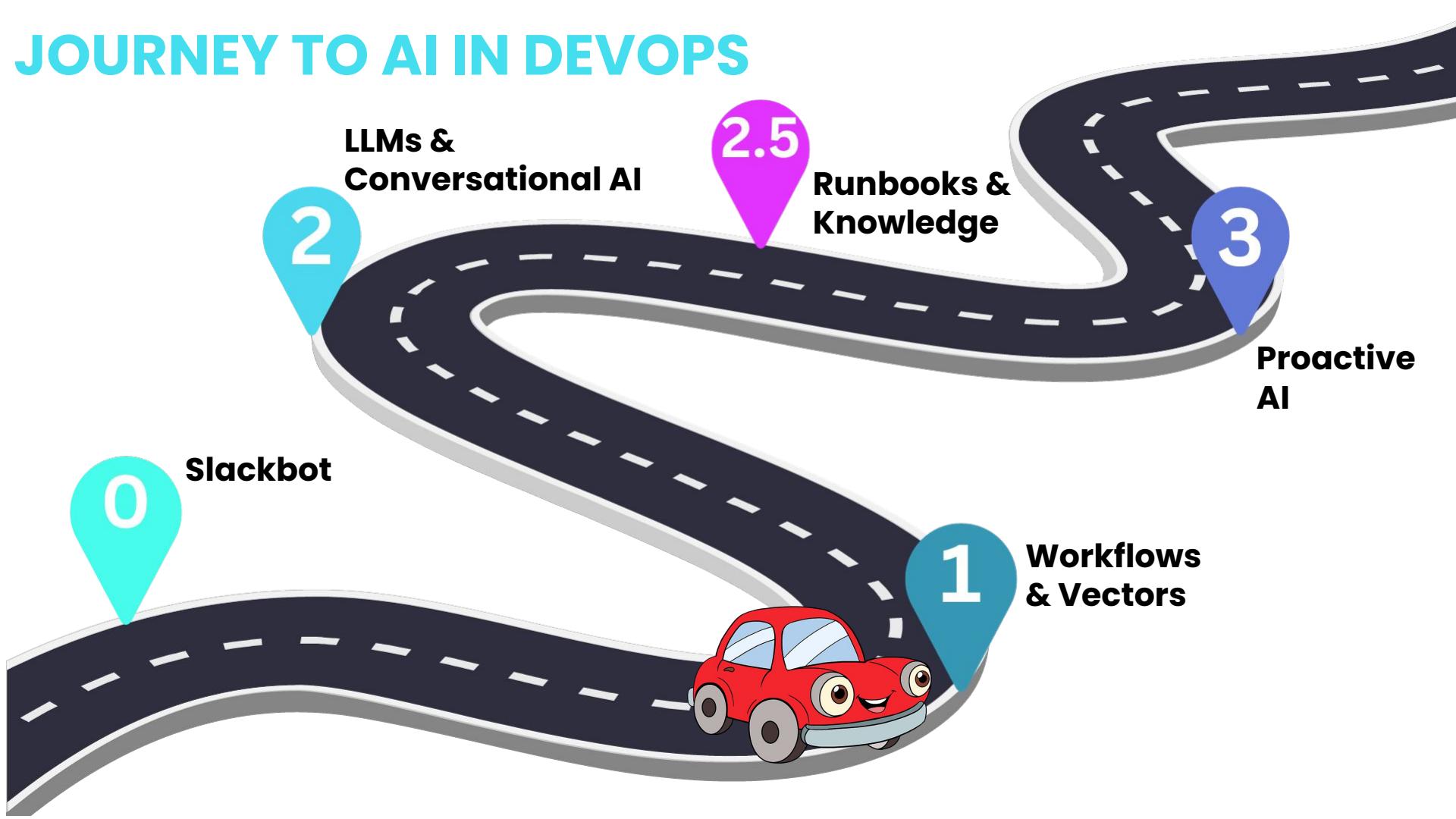
# GEN 0.5 – CI AND IaC

## When it doesn't work - Here's why?

- # Have you tried using Jenkins UI?!?!
- # Context switching
- # Ramp-up - you need to know what you want
  - “**Describe Kafka topic...oh wait..What topics are there?**”
- # Lots of maintenance
  - Updating forms
  - Adding and updating jobs
- # DevOps end up spending time troubleshooting errors
- # Not a great way to manage users and groups



# JOURNEY TO AI IN DEVOPS



# GEN 1 - WORKFLOWS & VECTORS

Alex 12:07PM Which EC2 instances are open to the internet?

Kubiya 12:07PM These EC2 instances are open to the internet:  
i-0b44d88 i-1d2199d Search 🔎  
i-9013d9j +22

Alex 12:07PM Thanks

Kubiya 12:07PM Which job would you like to trigger?  
TestJob ProdJob  Search 🔎  
Restart app service? Login app CI

Great! Just to confirm, you want me to trigger Jenkins Job ProdJob  
Yes  No

12:07 In progress: jenkins/invoke\_job

12:08  Allright! ProdJob has completed successfully

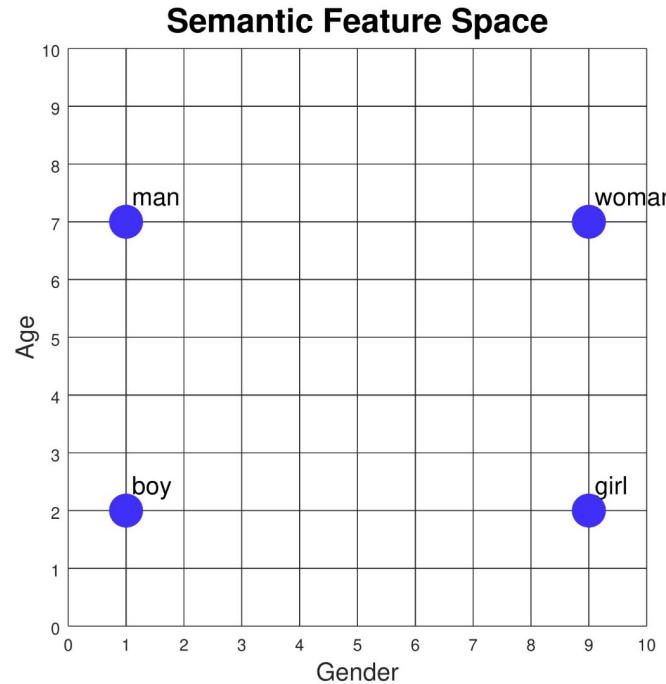
# EMBEDDINGS

- # Words have meaning
- # Embeddings help computers understand words on a deeper level
- # Embeddings are vectors [0,1,4]

A **semantic feature** is a basic trait or characteristic that describes the meaning of a word or phrase.

Word Coordinates		
	Gender	Age
man	[ 1,	7 ]
woman	[ 9,	7 ]
boy	[ 1,	2 ]
girl	[ 9,	2 ]

Semantic features - Gender, Age

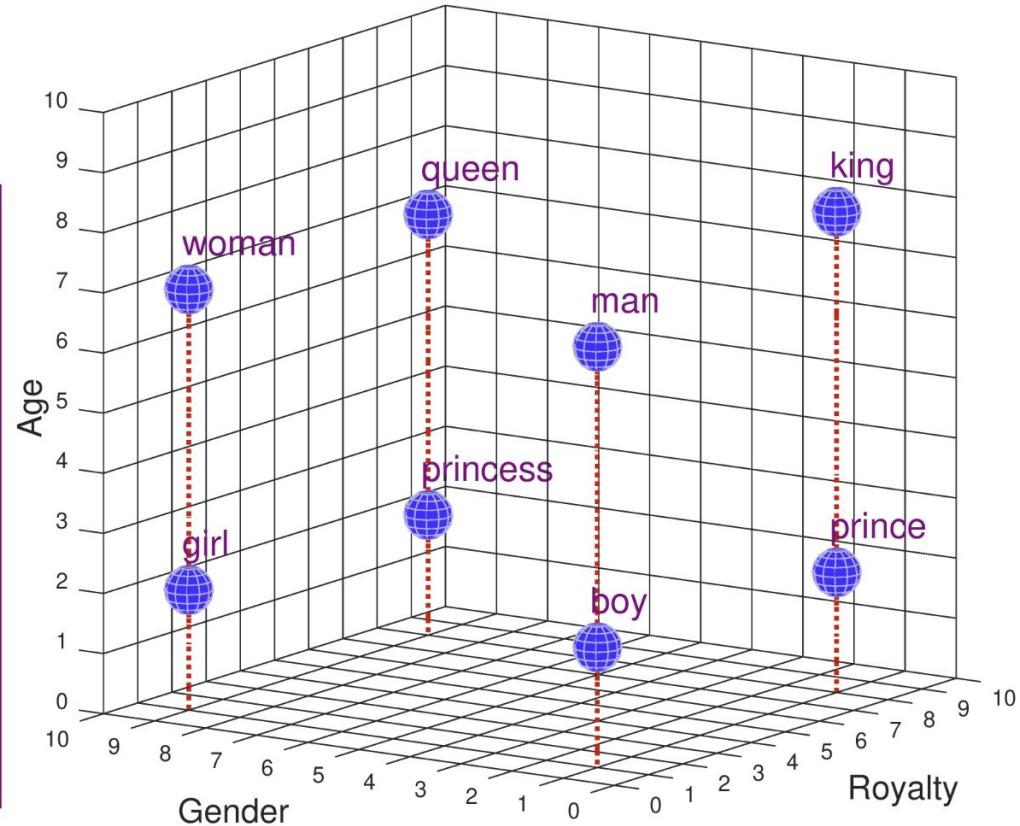


# 3D EMBEDDINGS

Embeddings Model	Dimensions
Text-embedding-ada-002 (OpenAI)	1536
Embed-english-v2.0 (Cohere)	4096

	Gender	Age	Royalty
man	[ 1,	7,	1 ]
woman	[ 9,	7,	1 ]
boy	[ 1,	2,	1 ]
girl	[ 9,	2,	1 ]
king	[ 1,	8,	8 ]
queen	[ 9,	7,	8 ]
prince	[ 1,	2,	8 ]
princess	[ 9,	2,	8 ]

## 3D Semantic Feature Space



# DEVOPS EMBEDDINGS

# The similarity between:

1. "Show ECS instances in us-west-2"
2. "List AWS ECS instances"
3. "List Jenkins jobs"

```
(list_aws_ecs_instances,  
 list_jenkins_jobs,  
 show_ecs_instances_in_us_west_2) = embeddings.embed_documents([  
     "list aws ecs instances",  
     "list Jenkins jobs",  
     "show ecs instances in us-west-2"])
```

```
embedding first 3 numbers out of 4096: [1.5537109, 1.1220703, 1.6689453]
```

```
similarity between user_text and `list aws ecs instances`: 0.6636591058099737
```

```
similarity between user_text and `list Jenkins jobs`: 0.25843715593666144
```

# GEN 1 - WORKFLOWS & VECTORS

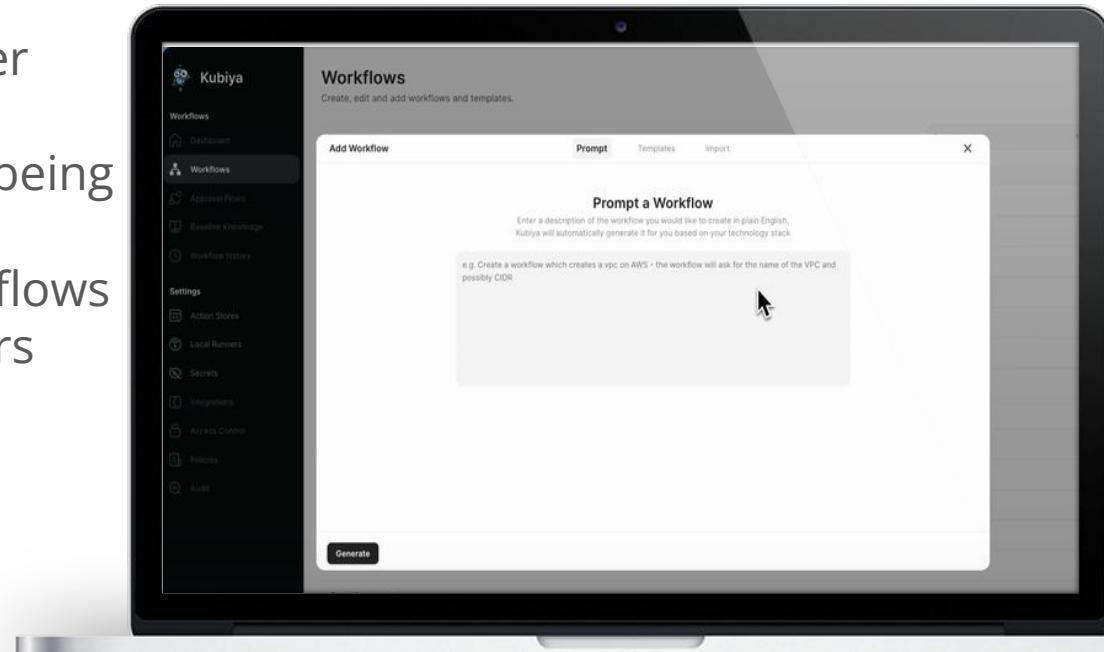
## What worked:

- # Use plain language to trigger workflows
- # Information is dynamically being pulled
- # Supports RBAC & approval flows
- # Use embeddings and vectors
- # Low maintenance

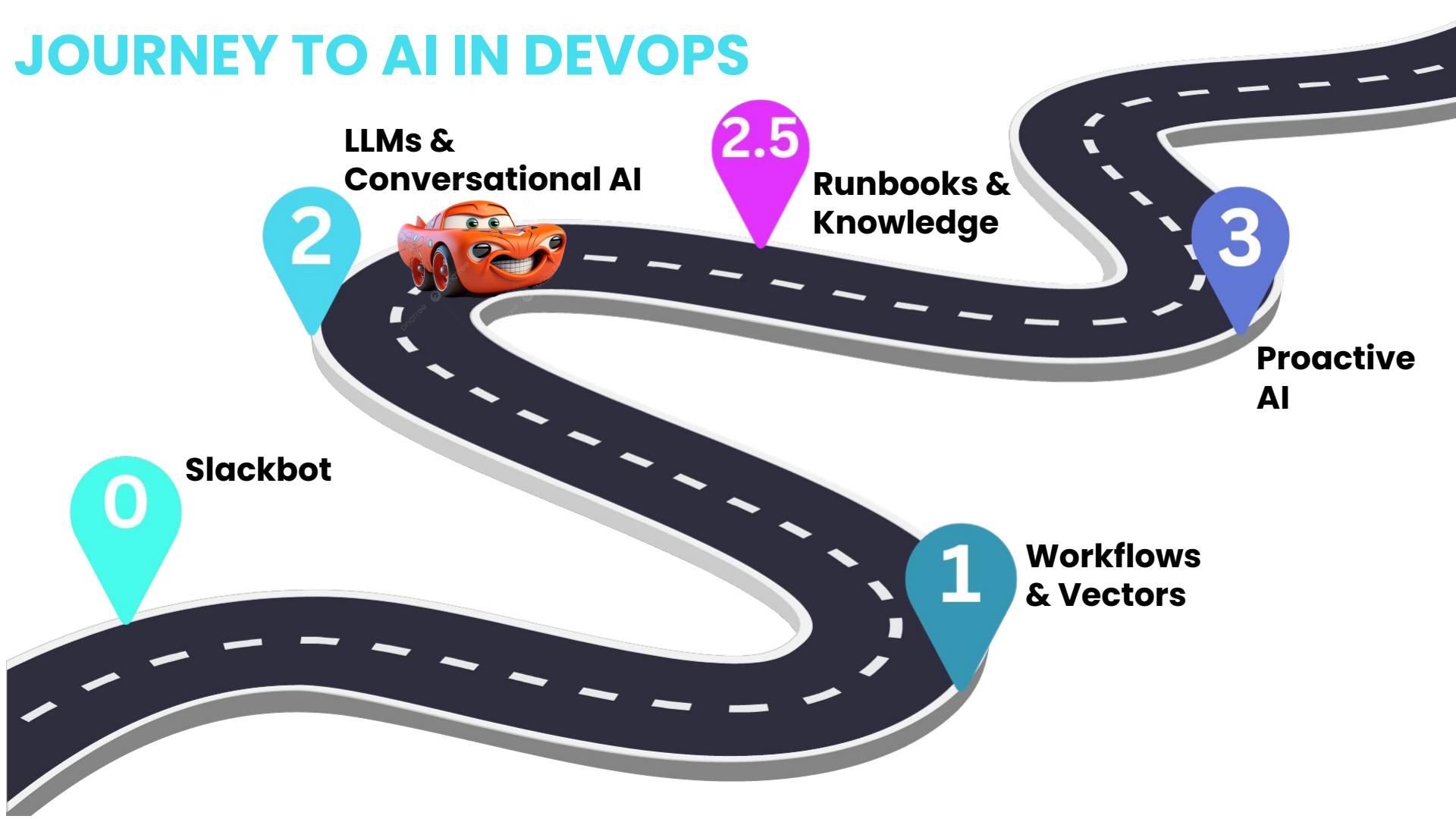


## What didn't work:

- # Creating workflows
- # Long ramp-up
- # Conversion from one type to another
- # Depending on the chat UI

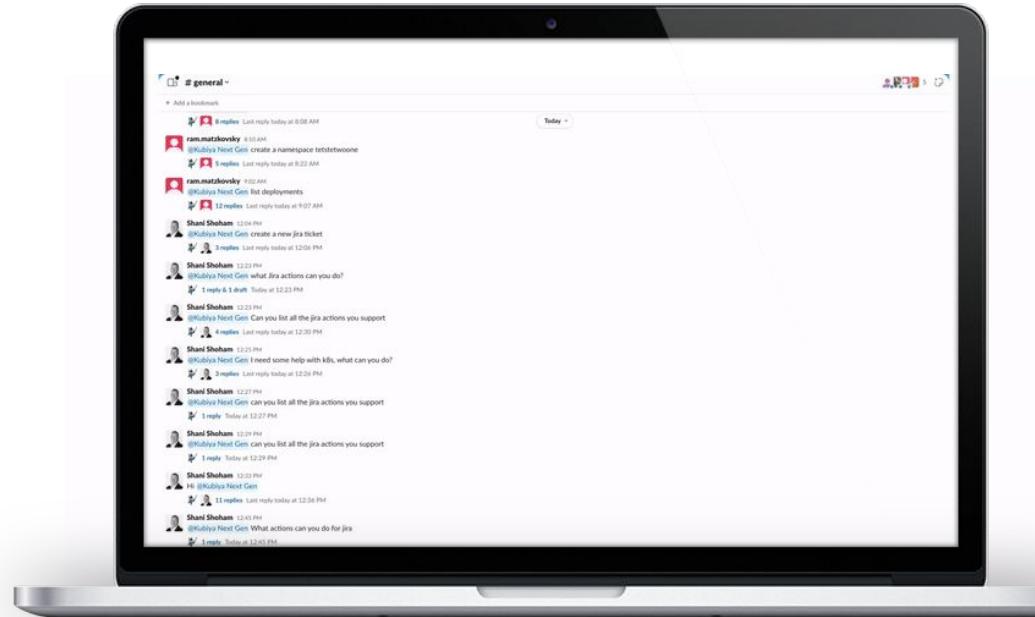


# JOURNEY TO AI IN DEVOPS



# GEN 2 – LLM BASED CONVERSATIONAL AI

- # Deterministic execution of Python functions
- # Use LLMs to
  - Create embeddings for functions
  - Power conversations
    - Identify context
    - Memory
    - Context switch
    - Response time
  - Reason responses



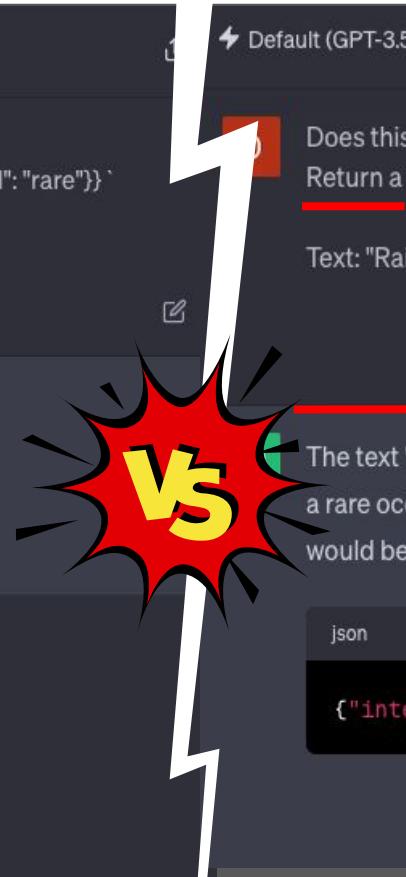
# CHALLENGES

- # Conversation aspects
  - o Response time
  - o Memory/Context
- # Smarter LLMs - urge to rely on for business logic
  - o **"Collect all parameters and return a JSON only when you have all the context"**
- # Small changes in prompt changes behavior unpredictably

The screenshot shows a GPT-3.5 conversation interface. At the top, it says "Default (GPT-3.5)". Below that, a user message "User: Hey my name is Omer" is followed by an AI response "AI: Nice to meet you Omer, How can I help?". The user then corrects themselves: "User: Actually it's Bob." and the AI responds "AI: How can I assist you today, Bob?". A green bracket on the right side of the screen, labeled "Context", groups the user's initial statement and the AI's follow-up response, indicating that the AI has retained the context of the previous statement despite the correction. In the middle of the conversation, the user asks "I changed my name from what to what?", and the AI replies "You changed your name from Omer to Bob." At the bottom, there is a "Send a message" button.



# WHY PROMPTS ARE SO CHALLENGING



⚡ Default (GPT-3.5)

0 Does this text suggest something "frequent" or "rare"?

Write the response in json format, for example: `>{"interval": "rare"}`

Text: "Rain falls once in a blue moon in the desert."



```
{  
  "interval": "rare"  
}
```

⚡ Default (GPT-3.5)

Does this text suggest something "frequent" or "rare"?

Return a response json format, for example: `>{"interval": "rare"}`

Text: "Rain falls once in a blue moon in the desert."

The text "Rain falls once in a blue moon in the desert" suggests that rain is a rare occurrence in the desert. Therefore, the response JSON format would be:

json

```
{"interval": "rare"}
```

Copy code

# OVERCOMING LLM CHALLENGES

- # Split into multiple small prompts
  - Finetune prompts (only one task, history when needed ex..)
  - Giving us a way to test changes and know what they can effect
  - Ability to use different models for specific tasks
    - Choose Best model
    - Finetune custom models

## GPT 4.0

O Return "Yes" or "No". how are you today?



As an artificial intelligence, I don't have feelings, but I'm functioning as expected. Thank you! \*

## GPT 3.5

O Return "Yes" or "No". how are you today?



I'm just a computer program, so I don't have feelings, but I'm here and ready to assist you! How can I help you today?

## Claude 2.0

Return "Yes" or "No". how are you today?



Yes



Copy Retry

# OVERCOMING LLM CHALLENGES

- # Split into multiple small prompts
  - Finetune prompts (only one task, history when needed ex..)
  - Giving us a way to test changes and know what they can effect
  - Ability to use different models for specific tasks
    - Choose Best model
    - Finetune custom models



Shani Shoham < 1 minute ago

@Kubiya [\(Closed Preview\)](#) can you help me pull a list of all the EC2 instances on EU-west1 that are running along with their ID, instance type and CPU utilization



Kubiya [\(Closed Preview\)](#) APP 5 days ago

list\_ec2\_instances executed successfully



Kubiya [\(Closed Preview\)](#) APP 5 days ago

✓ Response fetched successfully



Kubiya [\(Closed Preview\)](#) APP 5 days ago

Was that helpful?



Kubiya [\(Closed Preview\)](#) APP 5 days ago

✓ Action executed successfully! 🎉

Here are the details of the EC2 instances:

- Instance 1:
  - Instance ID: i-066d8f5ac9ce1d2c7
  - Instance Type: m5.xlarge
  - CPU Utilization: 6.11%
  - Creation Date: 2023-06-20 03:44:29+00:00
- Instance 2:
  - Instance ID: i-01792f5e194d9e5b6
  - Instance Type: m5.xlarge
  - CPU Utilization: 4.44%
  - Creation Date: 2023-08-01 16:00:15+00:00
- Instance 3:
  - Instance ID: i-09a9a16fd618b3b51
  - Instance Type: t2.micro
  - CPU Utilization: 0.13%
  - Creation Date: 2023-09-09 14:06:54+00:00
- Instance 4:
  - Instance ID: i-045a3bd2fcfd1e8b3a
  - Instance Type: t2.micro
  - CPU Utilization: 0%
  - Creation Date: 2023-08-16 00:55:00+00:00
- Instance 5:
  - Instance ID: i-021501c8a4223bc02
  - Instance Type: t2.micro
  - CPU Utilization: 0%
  - Creation Date: 2023-08-16 18:24:52+00:00

# VARIOUS LLMs

Model	Vendor			
ChatGPT3.5 & 4	 OpenAI	General Chat model		
Claude-2	<b>ANTHROPIC</b>	General Chat model		
"Classification"		Cassifier, using embeddings models for ex. embed-english-v2.0		
Command		"Command" model - Responds well with instruction-like prompts		
PaLM 2		General Chat model (Bard)		
Llama	 Meta Open Source	General Chat model		Amazon SageMaker
Llama Code	 Meta Open Source	Fine tuned for coding		Amazon SageMaker

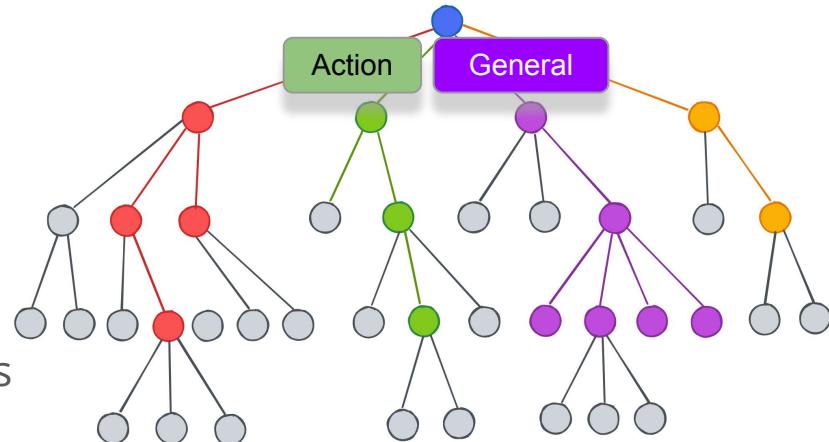
# ORCHESTRATION AND CLASSIFICATION

## # Orchestrator

- Handle flow
  - Better understanding of what happens behind the scenes
    - Tracing
    - Logging
    - Durations
  - Deterministic programming for specific tasks

## # Classification

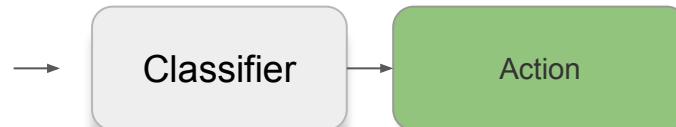
- Classifier - classify text into known labels
  - NLP (Natural language processing)  
spacy and more..



A small, square portrait of a man with short brown hair, wearing a black t-shirt. He is looking directly at the camera with a neutral expression.

**Shani Shoham** 9:34 AM

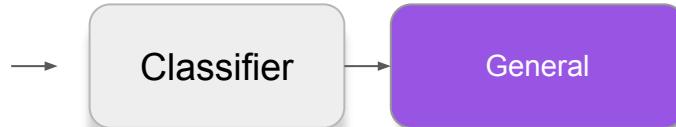
@Kubi (DevOps) update config file



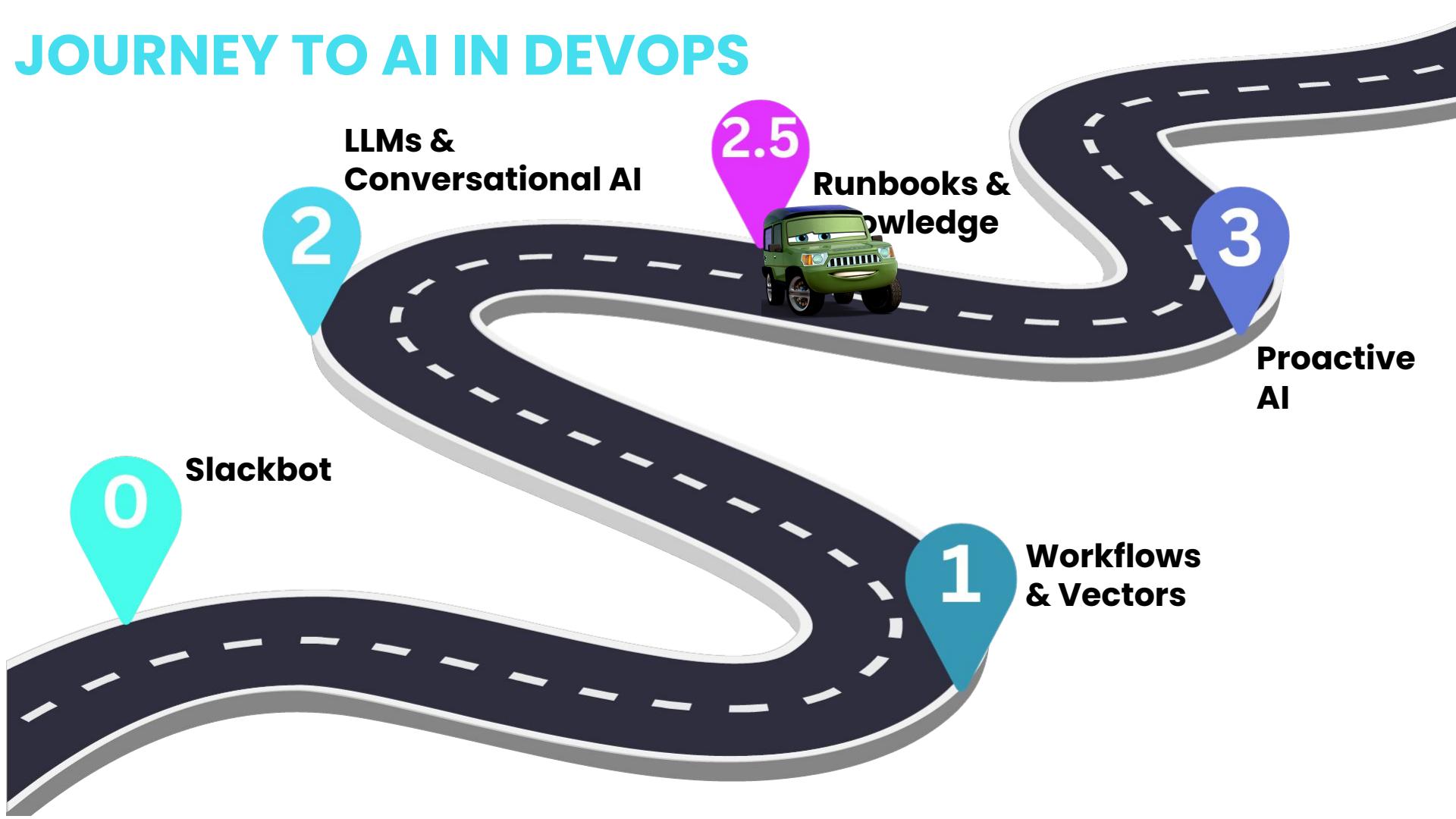
A small, square portrait of a man with short brown hair, wearing a black t-shirt with a white graphic on it. He is looking directly at the camera with a neutral expression.

Shani Shoham 9:33 AM

@Kubi (DevOps) tell me a joke



# JOURNEY TO AI IN DEVOPS



# GEN 2.5 – RUNBOOKS, KNOWLEDGE

## # Webhook based AI

- Suggest actions to execute
- But already has context

 **Kubiya Staging G2 APP** 15 days ago

 Incident Alert: Kubernetes - CrashLoopBackOff from Prometheus Alert Manager (staging) 

 **Event Details:**

- Type: Kubernetes - CrashLoopBackOff
- Owner: The service owner is the DevOps team (Costa)
- Service: worker
- Namespace: staging
- Duration: 3 minutes
- Impact: Potential service degradation
- Recommendation: Consider restarting the service, or rolling back to the latest replica set

 **Actions to Consider:**

1.  Restart the kubiya-worker service on k8s - [rollout\\_restart\\_deployment](#)
2.  Get the relevant deployment logs from k8s - [get\\_deployment\\_logs](#)
3.  Rollback the worker deployment on k8s - [rollback\\_deployment](#)

 **Event Timestamp:** 2023-08-14T15:23:05Z

 **Related:** [Kubiya-worker, staging, Restarting a problematic deployment on k8s](<https://kubiya.atlassian.net/browse/KUB-123>)

Please reply in this thread for inquiries or action execution. 

 **Shaked** 15 days ago

Now let's do 2. [@Kubiya Staging G2](#)

 **Kubiya Staging G2 APP** 15 days ago

The `get_deployment_logs` action was executed successfully for the `worker` deployment in the `staging` namespace. Here are some important details from the logs:

- The OpenFaaS gateway is: <http://gateway-external.openfaas:8080>
- There might be authentication problems with the gateway as no OpenFaaS gateway token is set
- The function LLM server is: `devlmclassify`
- The API\_URL environment variable is not properly set and the API\_KEY environment variable is missing
- Jaeger initialization started and was successful
- Vector DB index name is: `kubiya-langchain-docs-source-cohere`
- Thread mode is enabled
- Unleash is ready and the unleash client is initialized for the app: `[conversations-manager]` in the staging environment
- Feature flags `[usesTemporalCloud]` and `[dedicatedTemporalNamespace]` are set to true and false respectively by unleash for the staging organization
- Connecting to temporal cloud for the staging organization in the staging environment with hostname: `staging.ipagu.tmrpl.cloud:7223` and namespace `staging`.
- A worker with the ID `23@worker-6d5bd49d95-gffh50` has started in the `staging.ipagu` namespace and is using the `staging.general` task queue
- The action was permitted and the user has the necessary permissions to execute it

No errors were encountered during the execution of the action.

# GEN 2.5 – RUNBOOKS, KNOWLEDGE

## # Webhook based AI

- Suggest actions to execute
- But already has context

## # Knowledge - Can learn from prior summaries and knowledge systems

- Can suggest resolutions or guidance based on that

Shani Shoham 23 hours ago  
@Kubiya\_Omerc7\_G2 How do I onboard a new user

1 reply

Kubiya Omerc7 G2 APP 23 hours ago  
Hello again!

Based on knowledge :  
To onboard a new user to Kubiya's platform, you need to follow these steps:

1. Prepare NATS keys: Obtain the NATS keys from 1Password, specifically the one named "NATS Keys - CLI". These keys will be shared with the customer.
2. Provide Sentry access: Access [kubiya.sentry.io](https://kubiya.sentry.io), which provides an exceptions view. This will allow the customer to monitor and handle any exceptions that occur.
3. Consider Elastic logs: Note that the Elastic logs currently suffer from slow performance. This information is provided for awareness.
4. Optional - Access to production EKS: If necessary, grant the customer access to the production EKS (Elastic Kubernetes Service) using the kubectl config. The specific context to use is "arn:aws:eks:eu-west-1:564407622114:cluster/dev-kubiya-cluster". This access allows the customer to interact with the Kubernetes cluster.
5. Access customer logs: If required, access the customer's logs in their designated namespace within the Slack workspace. This can be done using the command "kubectl -n <customer namespace> logs -f".

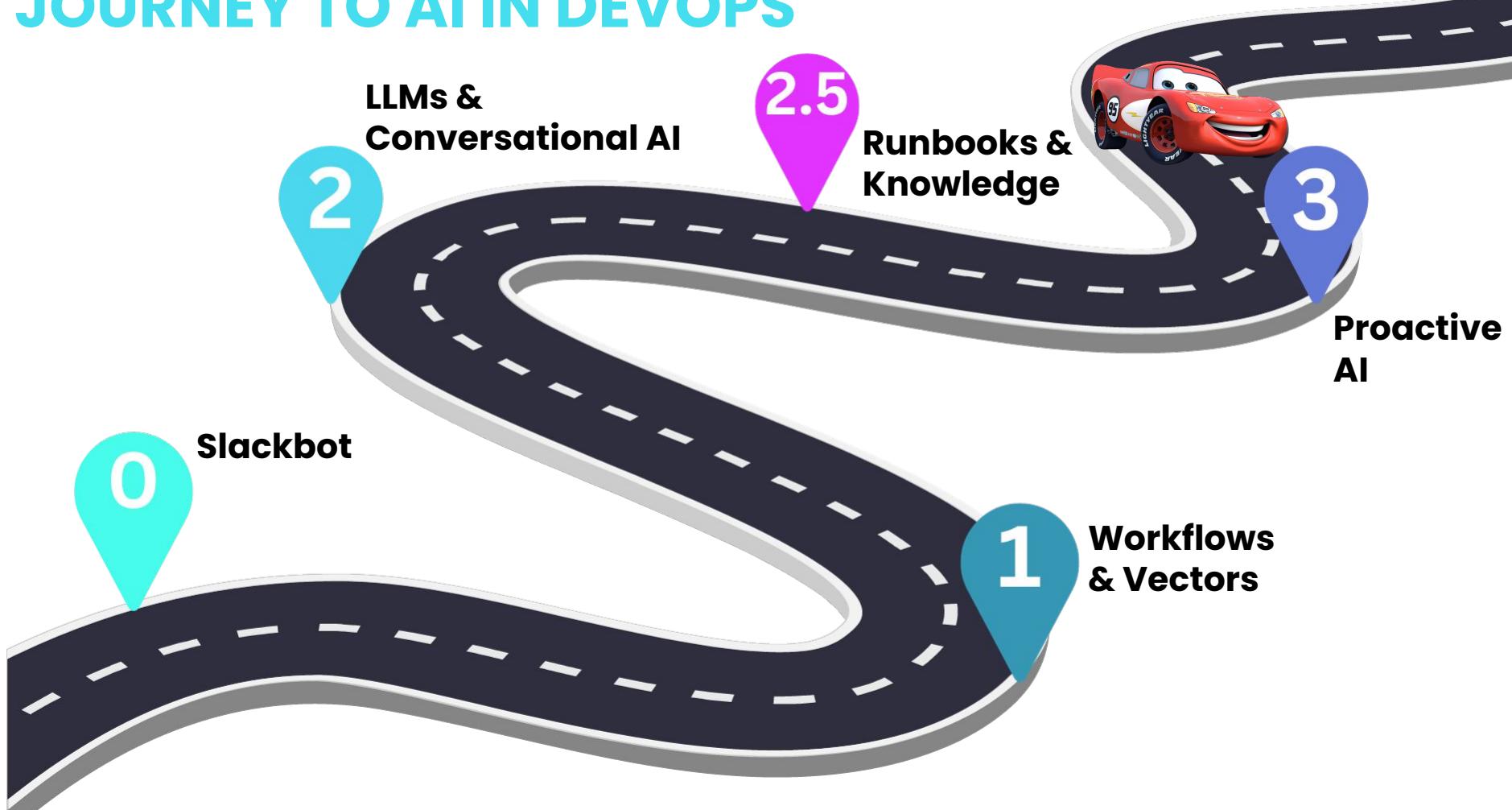
Please note that these steps are specific to the onboarding process and assume that the customer's environment has already been integrated, workflows have been created, and users have been invited. If you need information on these prior steps, please refer to the provided documents.

Sources:

- <https://kubiya.atlassian.net/wiki/spaces/SE/pages/39550977>
- <https://kubiya.atlassian.net/wiki/spaces/PLAT/pages/16809992>
- <https://kubiya.atlassian.net/wiki/spaces/PLAT/pages/2162689>
- <https://kubiya.atlassian.net/wiki/spaces/PLAT/pages/8192001>
- <https://kubiya.atlassian.net/wiki/spaces/SE/pages/48988161>

See less

# JOURNEY TO AI IN DEVOPS



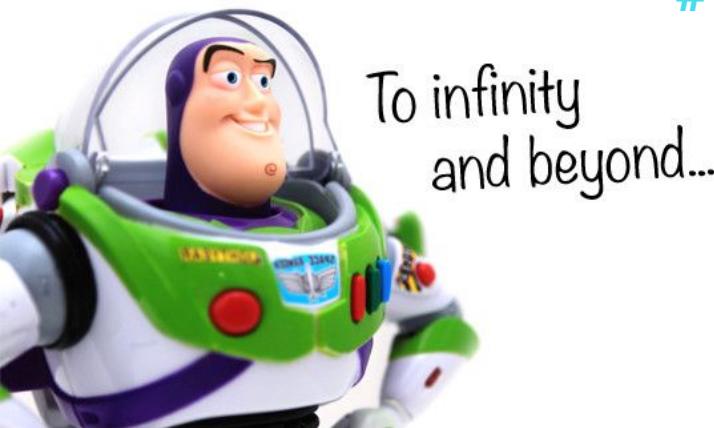
# GEN 3 – PROACTIVE ALERTS, TROUBLESHOOTING

## # Proactive AI

- “I noticed your pods are showing high utilization, would you like me to scale up?”
- “Your ec2 cost has increased over the last 7 days, would you like me to look into that?”

## # Can learn from prior conversations and ticketing systems

- Can suggest resolutions or guidance based on that.
- “In the past crashloop was resolved by restarting the worker. Would you like me to do that?”



**Double your DevOps capacity without the  
Doubling your headcount**

<https://kubiya.ai>



Shani Shoham  
shani@kubiya.ai



Omer Cohen  
omer.cohen@kubiya.ai

**Talk to me**



# HOW TO WORKAROUND

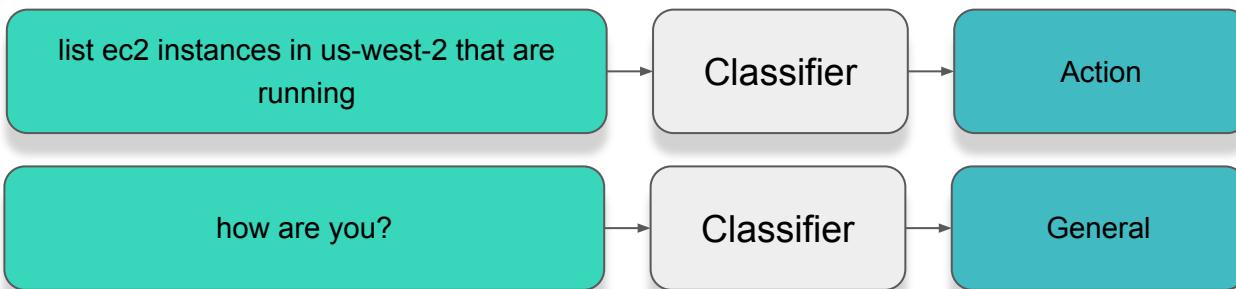
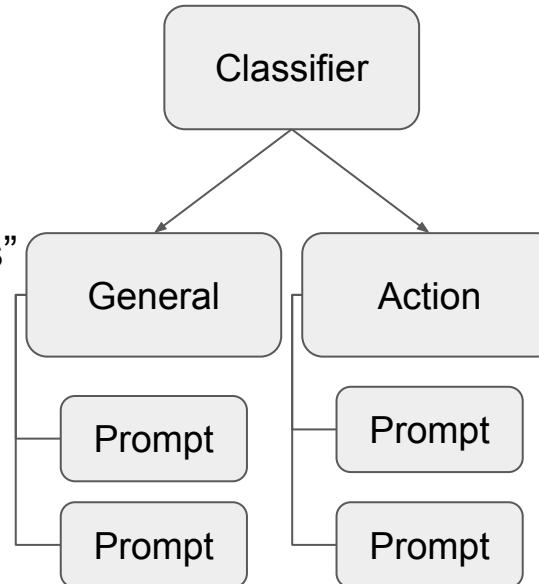
## # Split into multiple small prompts

- Finetune prompts (only one task, history ex..)
- Giving us a way to test changes and know their “Blast Radius”
- Ability to use different (best) models for specific tasks (finetune)

## # Orchestrator

- Deterministic programming for specific tasks

## # Classifier



```
from langchain.embeddings import CohereEmbeddings

embeddings = CohereEmbeddings(
    cohere_api_key=secrets.COHERE_API_KEY,
    model="embed-english-v2.0", # 4096 dimensions
) # type: ignore

embedded_vectors = embeddings.embed_documents(["Hello World!"])
print(embedded_vectors[0][:10]) # first 10 out of 4096 dimensions
```

```
[ 1.4453125,  
 0.3972168,  
 1.4052734,  
 -0.7338867,  
 -0.55810547,  
 0.15185547,  
 -0.7758789,  
 0.27026367,  
 1.1181641,  
 0.03744507  
 ]
```

Embeddings Model	Dimensions
Bge-small-en-v1.5 (BAAI)	384
Text-embedding-ada-002 (OpenAI)	1536
Embed-english-v2.0 (Cohere)	4096