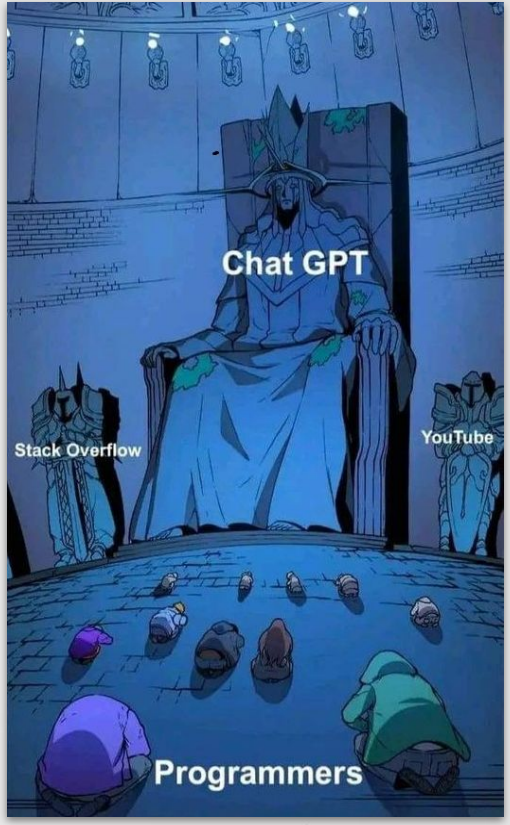
Balancing AI's Productivity Boost with Ethical Considerations in Cloud Native

Alex Jones,
AWS

Alex Jones
Principal Engineer
AWS
@AlexsJones

The rise of Artificial Intelligence

The reality...



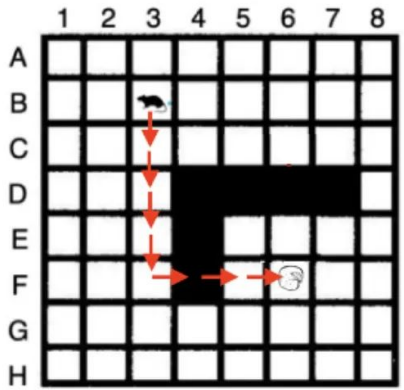
TU

suppose I have an 8x8 grid. The columns are labeled 1-8 from left to right, and the rows are labeled A-H from top to bottom. All cells are empty except for cell B-3 which has a rat, and cells F-6, which has cheese.

Suppose that the following cells are impassable, the rat cannot go through them: D4, D5, D6, D7, E4, F4.

What is the exact series of cells the rat can move through to get to the cheese as quickly as possible? Assume the rat can only move up-down-left-right to adjacent cells, and not diagonally.

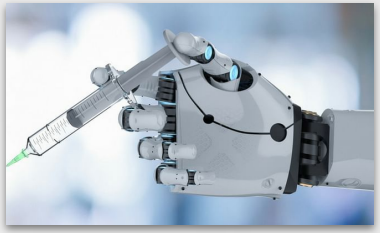
The rat can move through the following cells to get to the cheese as quickly as possible: B3, C3, D3, E3, F3, F4, F5, F6.



But, really…

Finance

Insurance



Healthcare

Security

Transportation

Virtual ——————————————————————————→ Physical


SPES
Iulio Bonasone ·f·

Agenda

Accretion of Knowledge

Esotericism Of Kubernetes

Introducing K8sGPT

Demo

An ethical dilemma

Brave new world

Accretion of Knowledge

Meta releases Torch
2002

IBM Watson showcased
2011

OpenAI launches
2015

Google announces BERT
2018

“Large-Scale Deep Unsupervised Learning Using Graphics Processors” published
2009

Deepmind, CNN
2013

Microsoft releases Turning Natural Language Generation
2017

ChatGPT released
2022

Generative AI excels at…

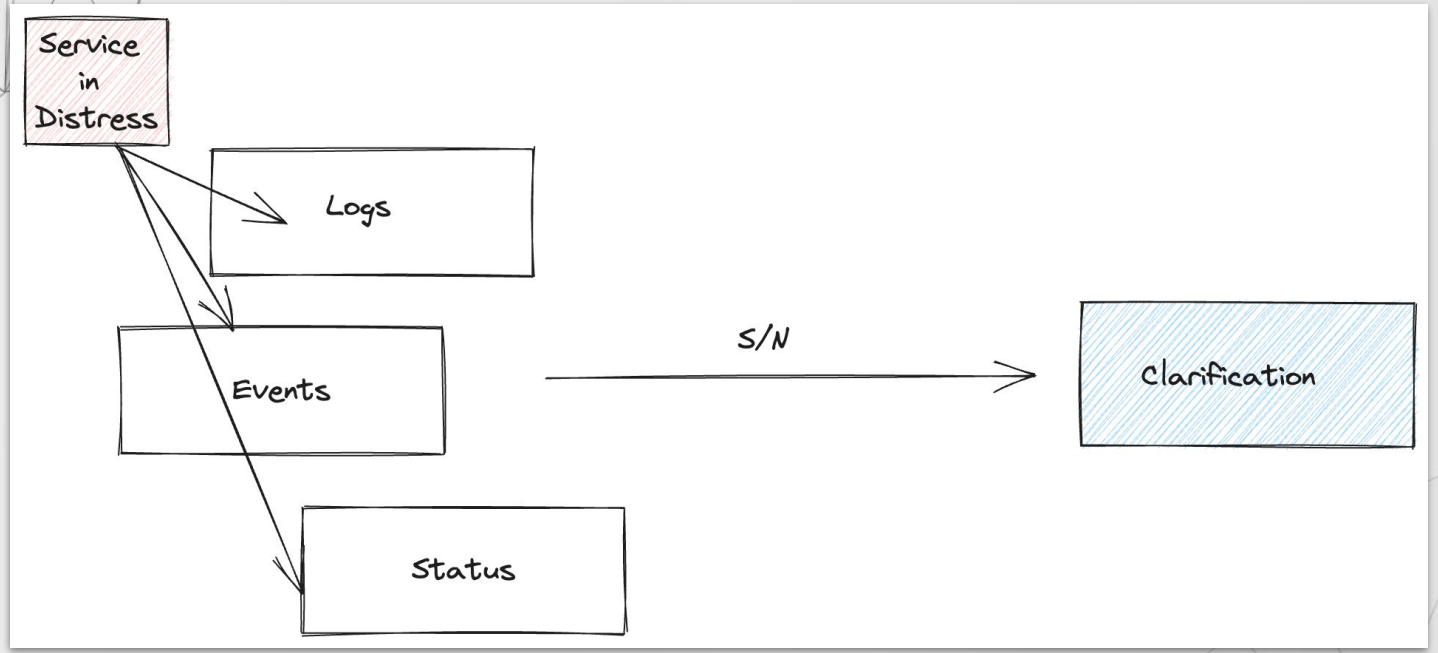
Pattern recognition
Translation
Prediction

Problems I have…

Pattern recognition
Translation
Prediction


Esotericism of
Kubernetes

Error messages,
Logs, Byzantine, Mystify,
Fixing’s rare for me


Service in Distress
Logs
Events
Status
S/N
Clarification

Many Kubernetes errors are really just Linux issues surfaced in a Byzantine way

Troubleshooting K8s is hard

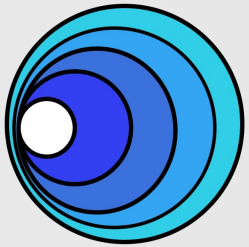
Amorphous/n-dimensional layers

Tacit knowledge formed

Signal/Noise ratio

Introducing K8sGPT

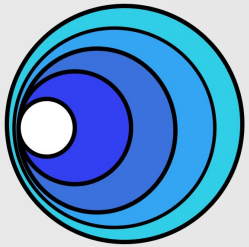
K8SGPT
KUBERNETES SUPERPOWERS



K8SGPT

KUBERNETES SUPERPOWERS

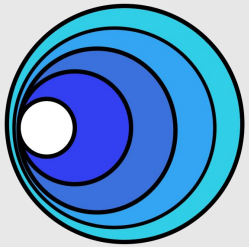
OpenAI



K8SGPT
KUBERNETES
SUPERPOWERS

LocalAI

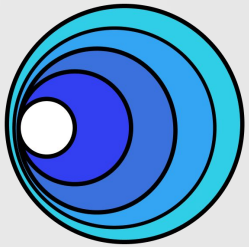
OpenAI



K8SGPT
KUBERNETES
SUPERPOWERS

LocalAI

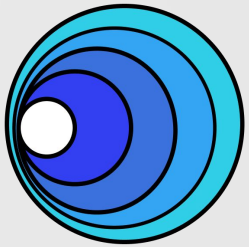
OpenAI



K8SGPT
KUBERNETES
SUPERPOWERS

LocalAI

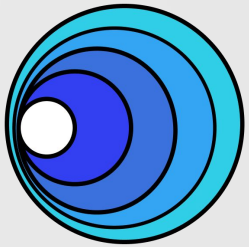
OpenAI



K8SGPT
KUBERNETES
SUPERPOWERS

LocalAI

OpenAI

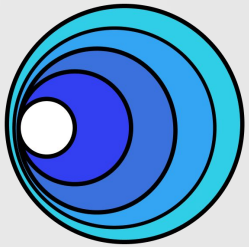


K8SGPT
KUBERNETES
SUPERPOWERS

cohere

LocalAI

OpenAI



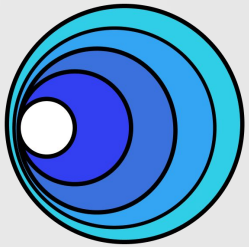
K8SGPT
KUBERNETES
SUPERPOWERS

Amazon Bedrock

cohere

LocalAI

OpenAI



K8SGPT
KUBERNETES
SUPERPOWERS

Amazon Bedrock

cohere

Amazon SageMaker

How does it actually work?

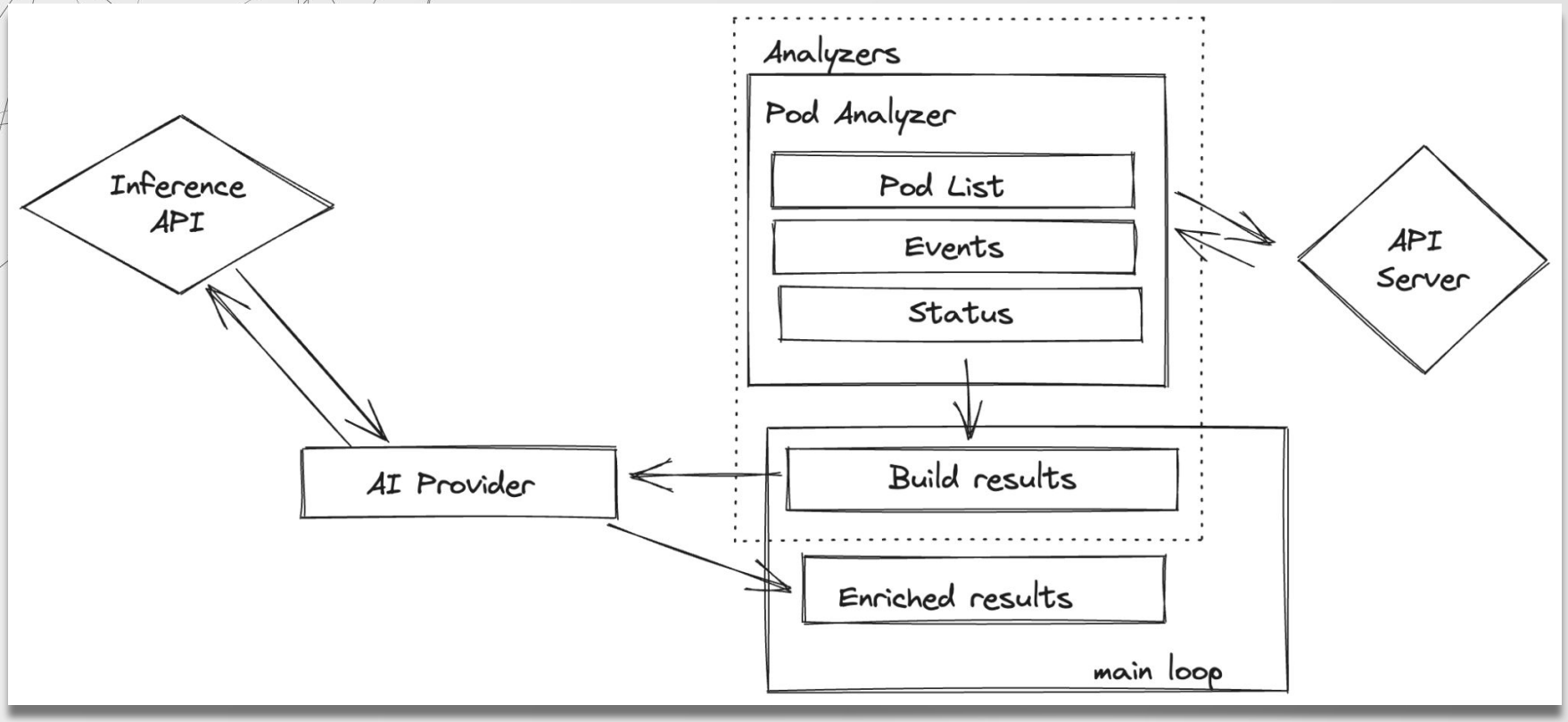
```
    var preAnalysis = map[string]common.PreAnalysis{}

    for _, pdb := range list.Items {
        var failures []common.Failure
        if pdb.Status.Conditions[0].Type == "DisruptionAllowed" && pdb.Status.Conditions[0].Status ==
"False" {
            var doc string
            if pdb.Spec.MaxUnavailable != nil {
                doc = apiDoc.GetApiDocV2("spec.maxUnavailable")
            }
            if pdb.Spec.MinAvailable != nil {
                doc = apiDoc.GetApiDocV2("spec.minAvailable")
            }
            if pdb.Spec.Selector != nil && pdb.Spec.Selector.MatchLabels != nil {
                for k, v := range pdb.Spec.Selector.MatchLabels {
                    failures = append(failures, common.Failure{
                        Text:          fmt.Sprintf("%s, expected pdb pod label %s=%s",
pdb.Status.Conditions[0].Reason, k, v),
                        KubernetesDoc: doc,
                        Sensitive: []common.Sensitive{
                            {
                                Unmasked: k,
                                Masked:   util.MaskString(k),
                            },
                            {
                                Unmasked: v,
                                Masked:   util.MaskString(v),
                            },
                        },
                    })
                }
            }
        }
```

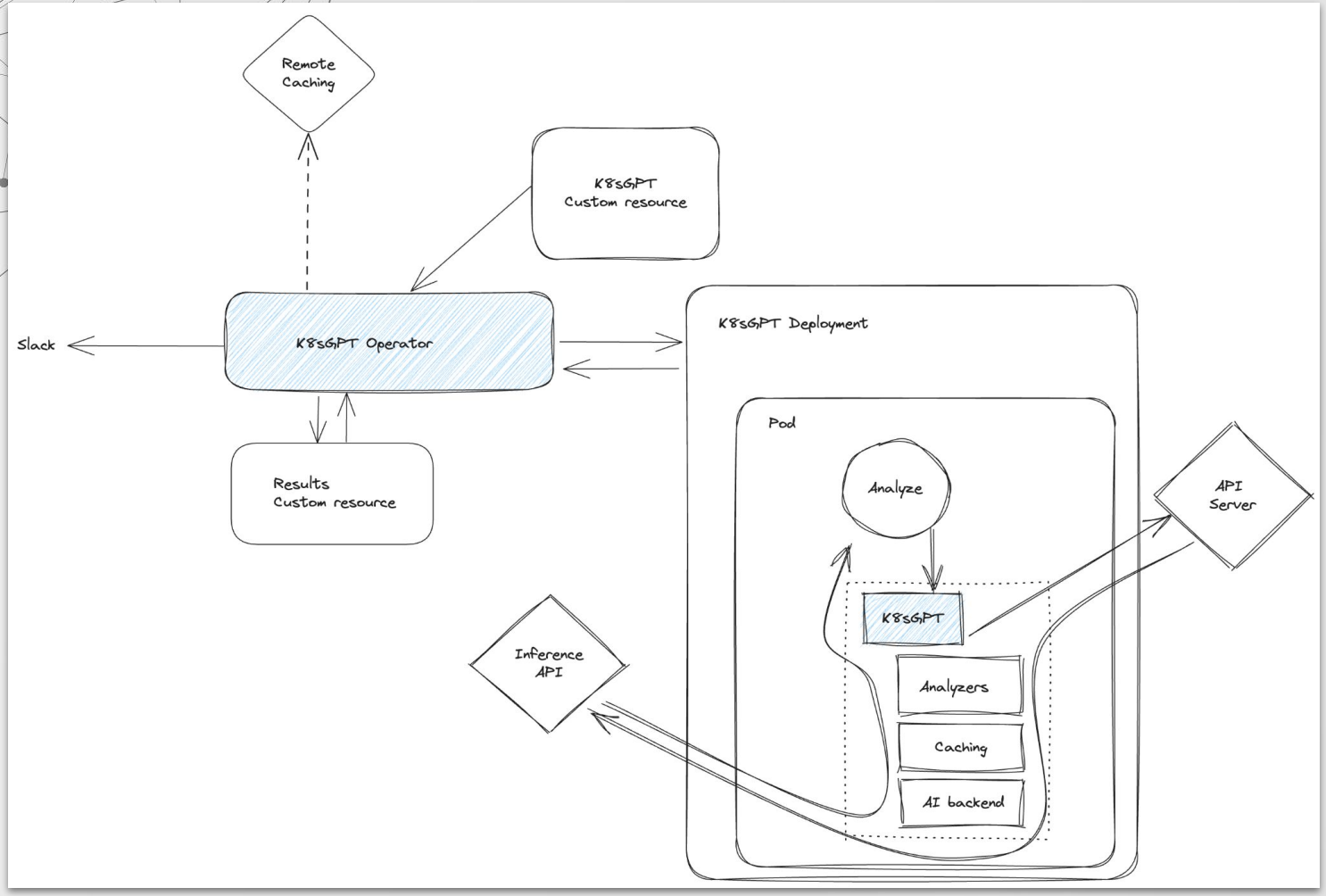
```
        // check if ingressclass exist
        if ingressClassName != nil {
            _, err := a.Client.GetClient().NetworkingV1().IngressClasses().Get(a.Context,
*ingressClassName, metav1.GetOptions{})
            if err != nil {
                doc := apiDoc.GetApiDocV2("spec.ingressClassName")

                failures = append(failures, common.Failure{
                    Text:          fmt.Sprintf("Ingress uses the ingress class %s which does not
exist.", *ingressClassName),
                    KubernetesDoc: doc,
                    Sensitive: []common.Sensitive{
                        {
                            Unmasked: *ingressClassName,
                            Masked:   util.MaskString(*ingressClassName),
                        },
                    },
                })
            }
        }
```

```go
func (c *CohereClient) GetCompletion(ctx context.Context, prompt, promptTmpl string) (string, error)
{   // Create a completion request
    if len(promptTmpl) == 0 {
        promptTmpl = PromptMap["default"]
    }
    resp, err := c.client.Generate(cohere.GenerateOptions{
        Model:             c.model,
        Prompt:            fmt.Sprintf(strings.TrimSpace(promptTmpl), c.language, prompt),
        MaxTokens:         cohere.Uint(2048),
        Temperature:       cohere.Float64(float64(c.temperature)),
        K:                 cohere.Int(0),
        StopSequences:     []string{},
        ReturnLikelihoods: "NONE",
    })
    if err != nil {
        return "", err
    }
    return resp.Generations[0].Text, nil
}
```

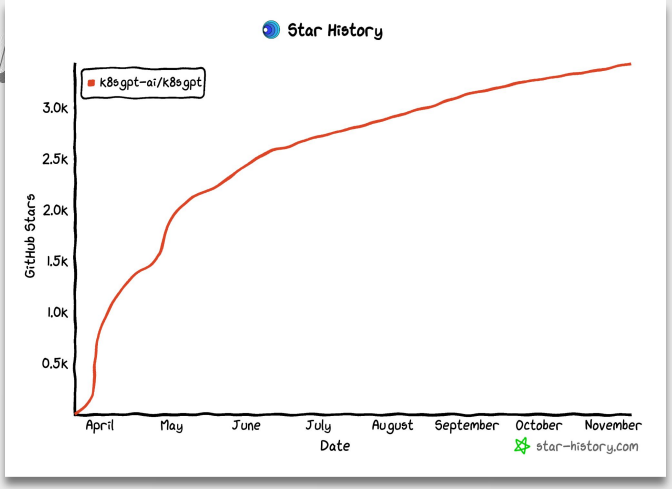


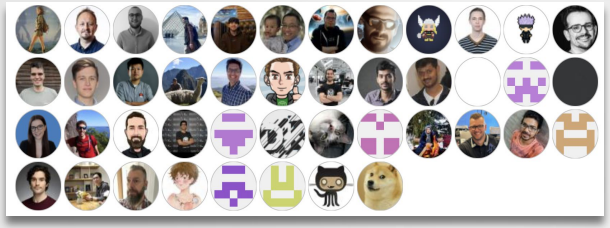
https://docs.k8sgpt.ai/



https://docs.k8sgpt.ai/

Founded, Kubecon Amsterdam 2023
30+ key contributors
3 Production usages
Made by the community, for the community
Builders from OpenAI, Cohere, AWS, Redhat, etc.





K8SGPT
KUBERNETES SUPERPOWERS

WOULD YOU LIKE TO KNOW MORE?

DEMO

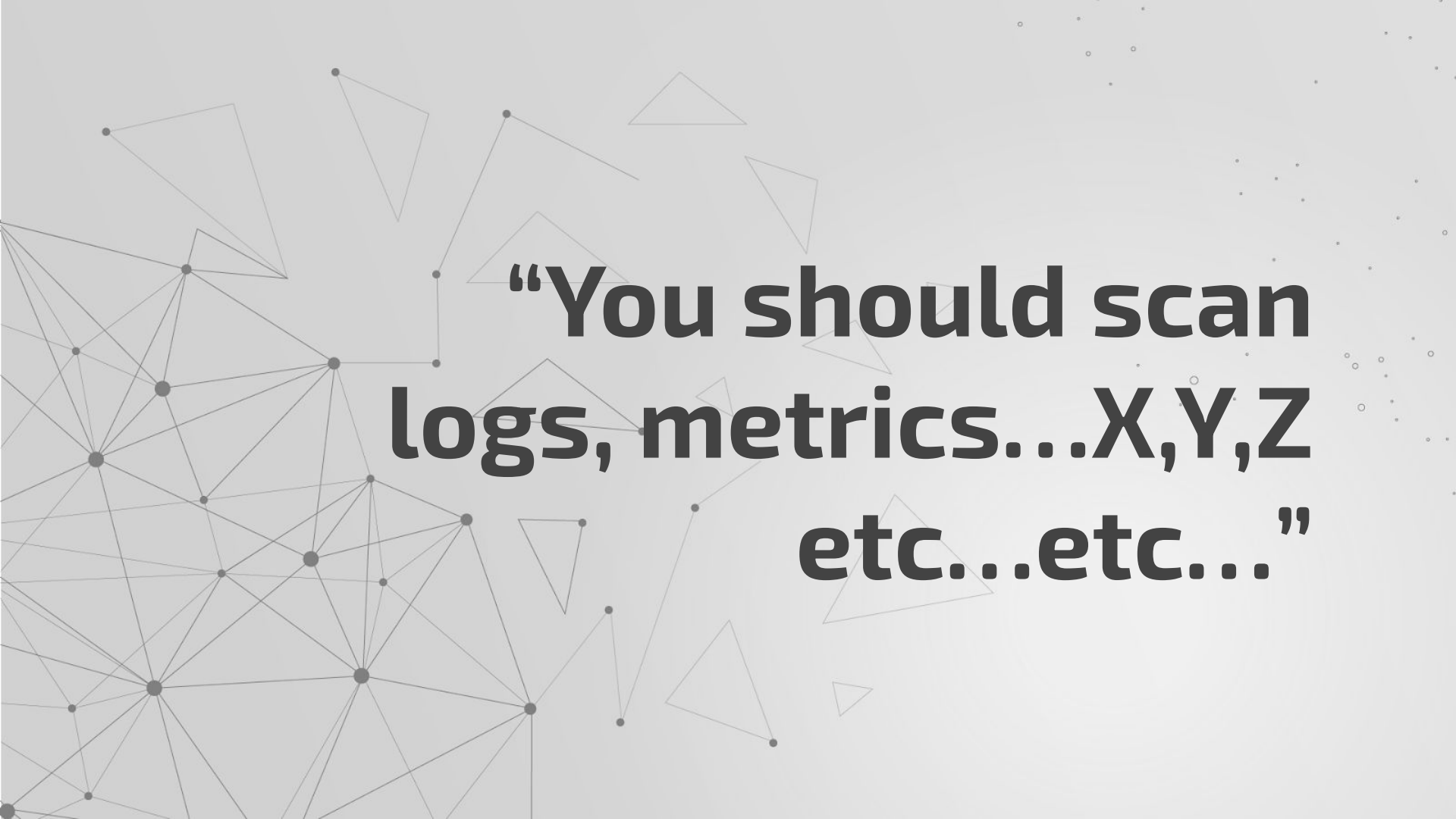
How has K8sGPT actually helped?

Faster debugging times

Codifying knowledge

Consistency of diagnostics

Lower the bar for Operators (human) to be effective


“You should scan
logs, metrics…X,Y,Z
etc…etc…”

An ethical dilemma

Who owns my **information**?
Who is **accountable** for decisions recommended by Ks8GPT?
What **due diligence** has been done to validate model training and data **provenance**?
Doesn't this detract from the function of SREs and diagnosticians?
How can we be sure advice isn't poisoned?

Compilation of GitHub issue questions

LocalAI

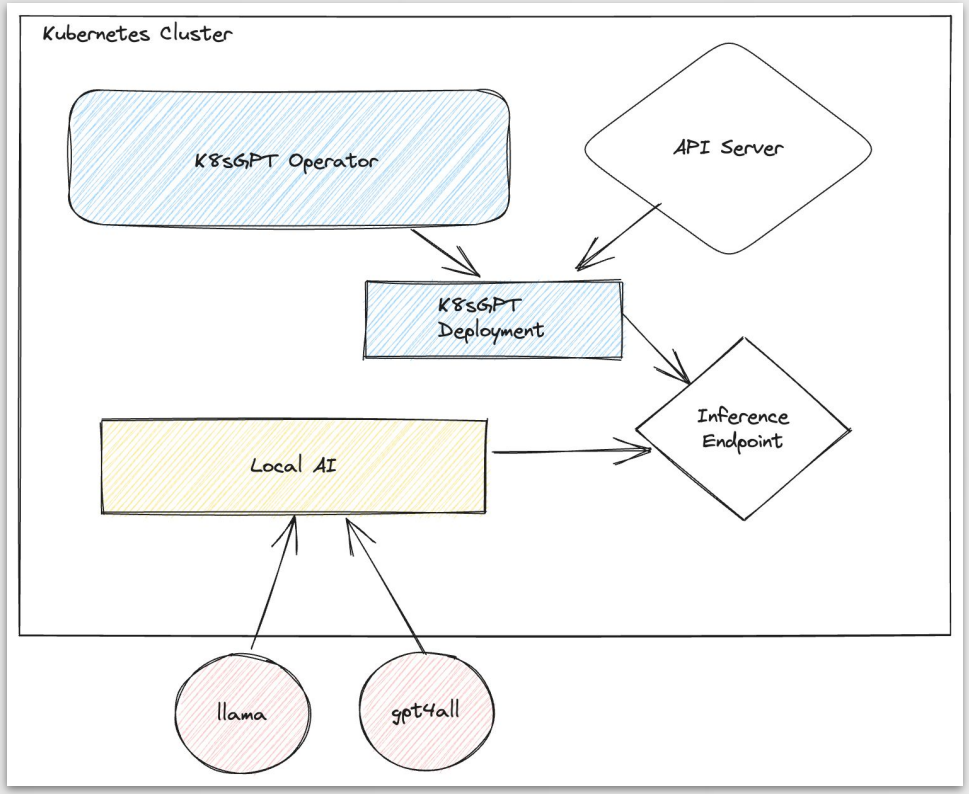
Local, OpenAI drop-in alternative REST API.

No GPU required

Supports multiple models

🏃 Once loaded the first time, it keep models loaded in memory for faster inference

⚡ Doesn't shell-out, but uses C++ bindings for a faster inference and better performance


Kubernetes Cluster
K8sGPT Operator
API Server
K8sGPT
Deployment
Inference
Endpoint
Local AI
llama
gpt4all

Amazon Titan

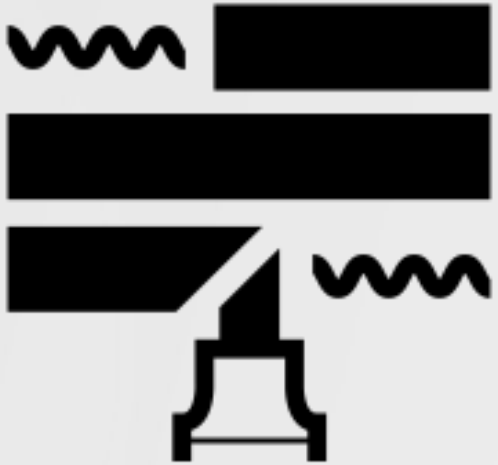
Foundational models

Built in support for “responsible AI”

Personalisation

Fine-tuning built in

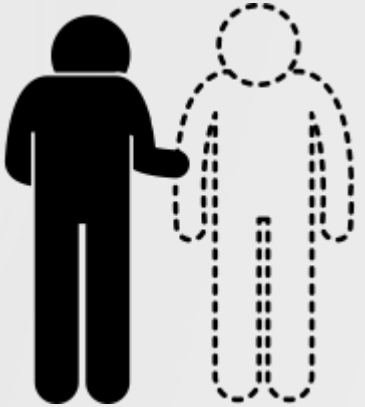
```
Sensitive: []common.Sensitive{
    {
        Unmasked: *ingressClassName,
        Masked:
util.MaskString(*ingressClassName),
},
```



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



Open project-level questions

How do we incrementally add value through new features such as logs without compromising safety, security and privacy?

Should we look towards Task Specific AI Models to avoid bias and misinformation inherent in the knowledge gaps and training of LLM?

Who is accountable for auto-remediated changes with K8sGPT?

Brave new world

“In no other field is the ethical compass more relevant than in artificial intelligence. These general-purpose technologies are re-shaping the way we work, interact, and live. The world is set to change at a pace not seen since the deployment of the printing press six centuries ago”.

Gabriela Ramos,
Assistant Director-General for Social and Human Sciences of UNESCO

We as a collective need to **study/educate/coordinate/clarify/leverage** the intersection of AI and cloud native

#wg-artificial-intelligence

AI HUB

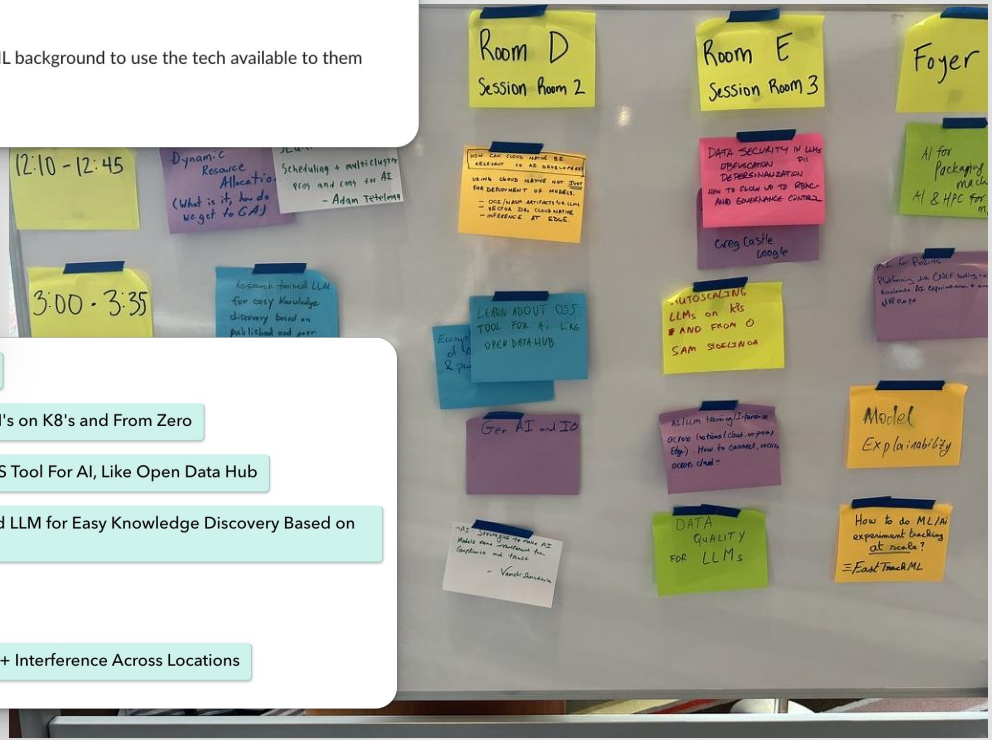
WEDNESDAY, NOVEMBER 8 | 11:00 AM-12:30 PM & 2:30-5:20 PM

Location: Hyatt Regency McCormick Place, Ballroom CDE

Phillip Carter 1:44 PM
👋 happy to collaborate here. I'd primarily love to advocate for ways for end-users without an ML background to use the tech available to them effectively.

Johnu George 1:48 AM
Thanks everyone. Great initiative and very much interested to collaborate on this.

- AI Hub: Unconference Session - AI For Products
- AI Hub: Unconference Session - Autoscaling LLM's on K8's and From Zero
- AI Hub: Unconference Session - Learn About OSS Tool For AI, Like Open Data Hub
- AI Hub: Unconference Session - Research Trained LLM for Easy Knowledge Discovery Based on Published and Peer Reviewed Data
- AI Hub: Unconference Session Block #2
- AI Hub: Unconference Session - AI/LLM Training + Interference Across Locations



The Future

Search or jump to... / Pull requests Issues Codespaces Marketplace Explore

cncf / sandbox Public

Watch 21 Fork 9 Star 77

Code Issues 22 Pull requests 1 Actions Projects 1 Security Insights

[Sandbox] K8sGPT #38

New issue

Open 2 tasks done AlexsJones opened this issue last week · 1 comment

AlexsJones commented last week • edited Tip

Application contact emails

alexsimonjones@gmail.com, thomas.schuetz@t-sc.eu

Project Summary

Kubernetes cluster analysis augmented with Artificial Intelligence

Project Description

The popularity of Kubernetes has skyrocketed, but it has also led to a significant amount of complex knowledge regarding the management of cluster workloads. As new innovations emerge, it becomes increasingly challenging to manage workloads and identify potential issues.

K8sGPT employs codified SRE techniques, utilizing Artificial Intelligence (either hosted or bring-your-own) to simplify the description of complex problems and provide easy-to-implement solutions. This tool is accessible through CLI or as an Operator and can be integrated with observability projects to facilitate continuous monitoring and straightforward triage.

The goal of K8sGPT is to act as a virtual engineer, reducing the number of personnel required on your team and eliminating two of the most significant obstacles to cloud-native adoption: cost and skill.

Assignees

No one assigned

Labels

New

K8SGPT

K8sGPT

k8sgpt-ai

Observability and Analysis · Monitoring

Giving Kubernetes Superpowers to everyone

Website	k8sgpt.ai
Repository	github.com/k8sgpt-ai/k8sgpt ★ 3,263
Crunchbase	crunchbase.com/organization/k8sgpt-ai
LinkedIn	linkedin.com/company/k8sgpt-ai
Twitter	@k8sgpt — Latest Tweet
First Commit	8 months ago — Latest Commit
Contributors	45 — Latest Release
Headquarters	London, United Kingdom — Headcount

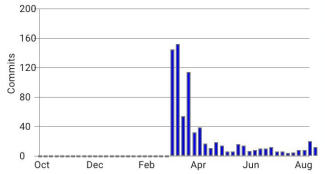
Open Source Software

License Apache License 2.0

OpenSSF Best Practices ✓

Tweet 1987

Go 97%
Makefile 2%
Smarty 1%
Dockerfile 1%



Tweets from @k8sgpt

k8sgpt @k8sgpt · 52m

AWS superpowers with the robustness of Rust and power of Bedrock

```
Restored session: Mon  6 Nov 2023 17:22:20 CST

isotope on   main is   v0.0.6 via   v1.72.0 on ☁   (eu-west-2)
❯ i
```

ISOTOPE

WOULD YOU LIKE TO KNOW MORE?

Thank you!

@AlexsJones