Vision Agents at Scale

Policy-aware image moderation using agentic reasoning

Bhavishya Pandit

Introduction

3.2 billion images are shared daily; manual review and fixed models lag; supervised retraining is slow and data-hungry, so new policy cases are often misclassified.

This post shows how vision agents, already deployed at Google and Meta for **zero/few-shot** screening, close the gap on scale and policy drift. Let’s dive in and see how it works for the use case of ad moderation.

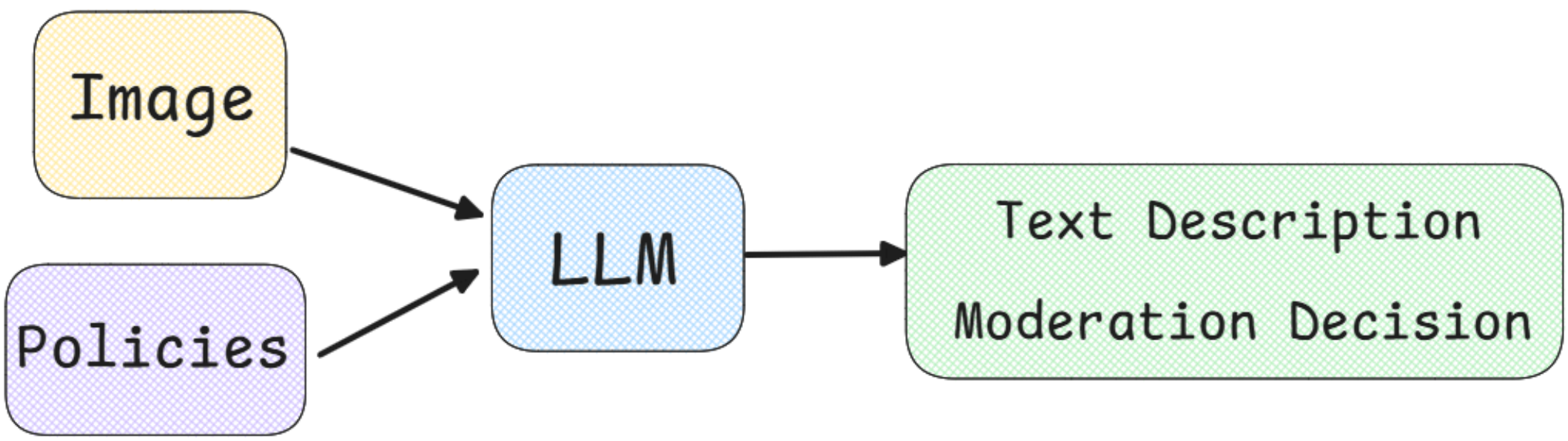
Bhavishya Pandit

Zero-Shot Image Moderation

First, lets see what is ad moderation?

> **Ad image moderation** is the process of reviewing and filtering visual content in advertisements to ensure compliance with platform policies, legal standards, and brand safety.
>
> To put simply, it determines whether an ad aligns with the platform's guidelines or violates them.

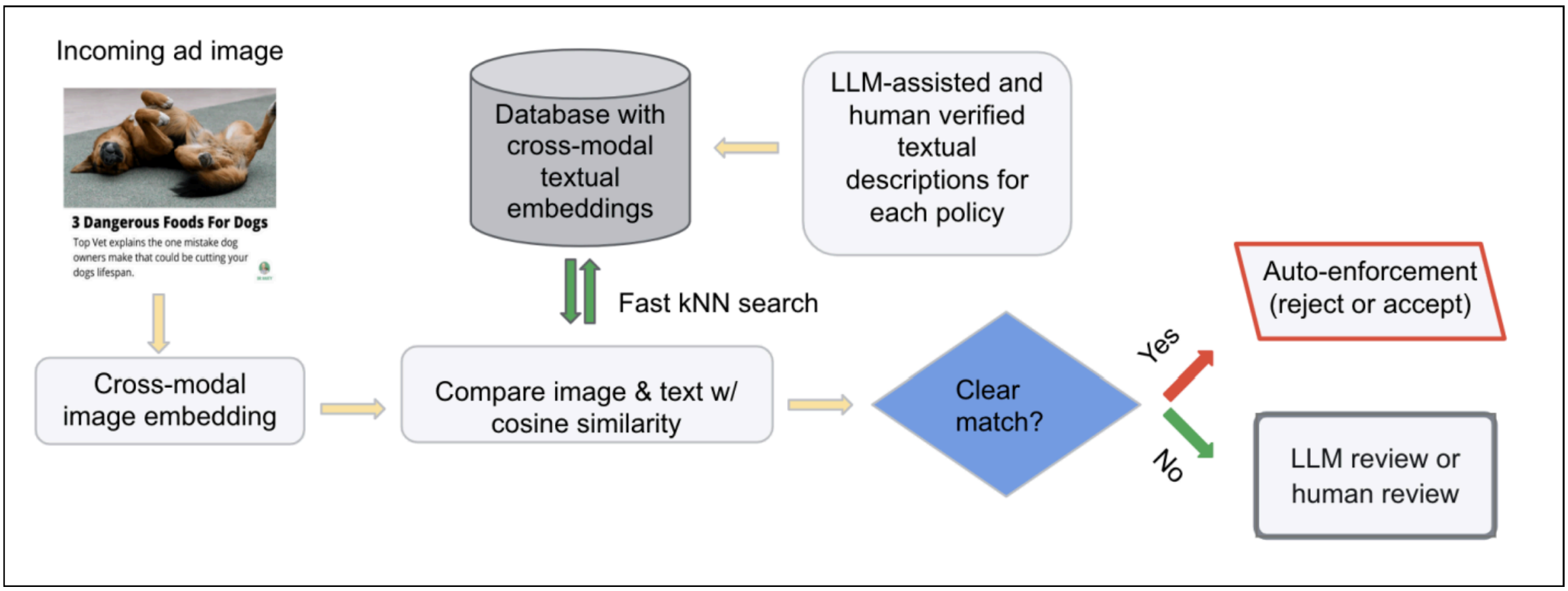
Zero-shot moderation eliminates the need of labelled dataset by leveraging LLMs to generate rich textual descriptions of images and cross-modal embeddings to evaluate content.



No pre-training on specific image categories required!

How Google Ads Does It?

So how do they do it without retraining a new model for every new type of violation?



Source

Steps:

- Image is encoded (e.g., CLIP) into a vector.
- Policies are textified, refined by LLMs (PaLM/Gemini), and human-reviewed.
- k-NN matches the image vector to policy vectors; if similarity ≥ threshold, auto-enforce; otherwise escalate to LLM or human.

What the Metrics Say

	Baseline Model	Google's Approach
Precision	89.1%	**90.8%**
Incremental Coverage Significance	57.3%	**107.3%**
Relative Recall	48.2%	**63.4%**

What this tells us:

- **Higher precision** means fewer mistakes when detecting policy-violating images, saving both time and trust
- It identifies **a broader range of policy violations**, even the subtle or previously unseen ones
- **Stronger recall** indicates, it catches more harmful content that traditional systems often miss

This proves that zero-shot moderation offers both scale and superior performance compared to conventional approaches.

Why a Game-Changer

This approach revolutionizes ad moderation with:

- **Minimal Training Data**

Focus on designing textual descriptions, not large-scale labeled datasets.

- **Speed & Efficient**

One scalable workflow for multiple policies and quick real-time moderation

- **Flexibility**

Handles diverse ad formats (banners, videos) and languages

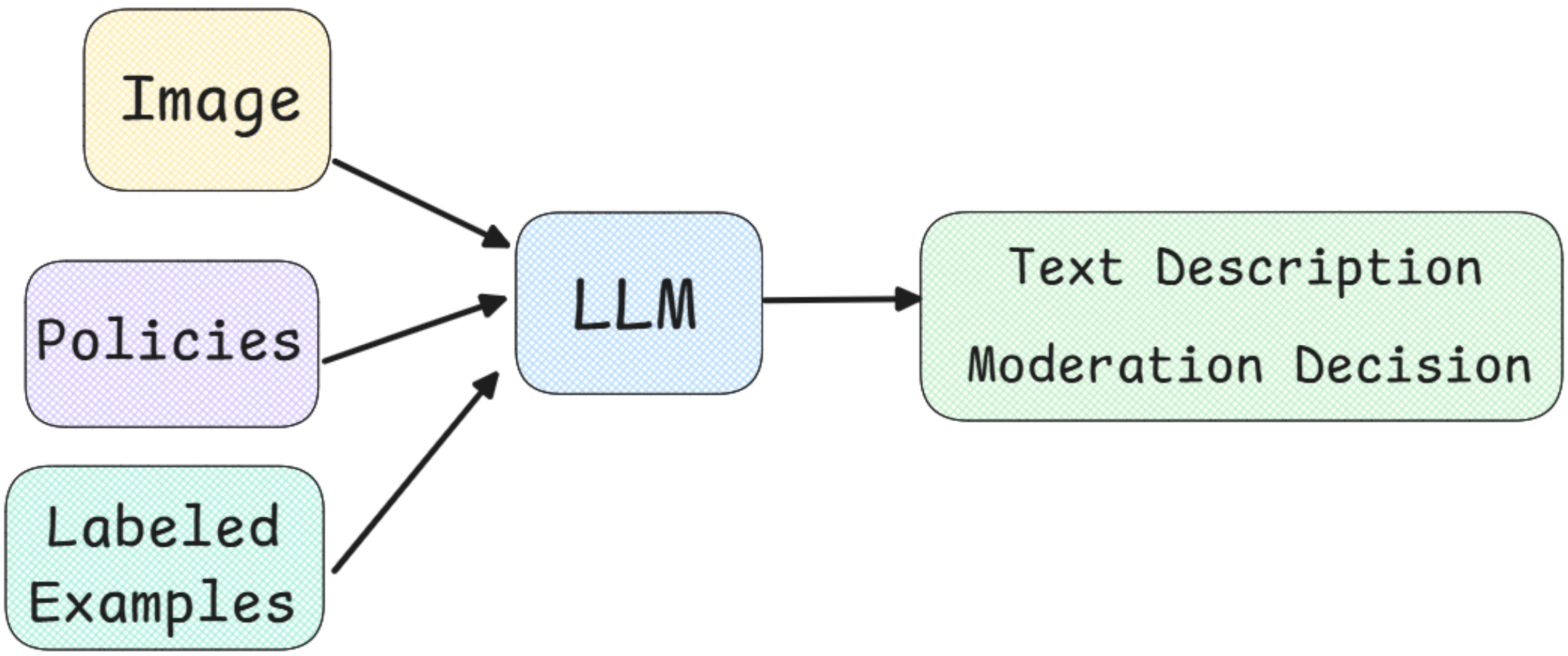
- **Accuracy**

Reduces false positives/negatives by understanding context

Few-Shot Image Moderation

Now, lets see what is few-shot moderation?

Few-shot moderation leverages a small set of labeled examples to prompt AI models, enabling them to learn new policy violations quickly and generalize to similar unseen content with minimal data and retraining.

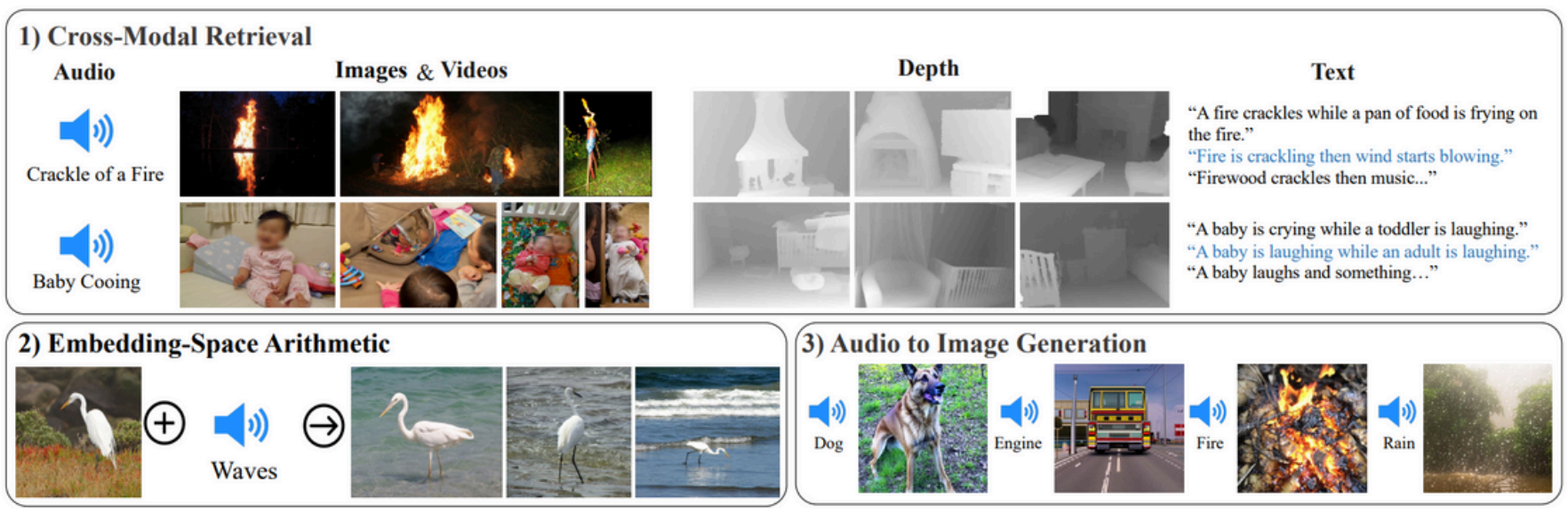


Meta’s Few-Shot Learner (FSL) enables rapid adaptation to new ad policy violations, reducing the response time from months to just weeks.

Let’s explore how Meta their approach to ad moderation today!

How Meta Does It?

Meta trains multimodal models like ImageBind to unify text, vision, and audio.



Source

Step 1 :
When a new ad policy is introduced, Meta's team l**abels a small set of violating examples**,typically fewer than 50. These examples could include images, ad copy, or both.

Step 2 :
These are then used to prompt or fine-tune a few-shot transformer model. It **supports zero-shot, few-shot, and low-shot configurations**.

Step 3 :
The trained system then begins flagging new content similar to the labeled examples. If confidence is high, Meta auto-enforces the decision, else it is escalated to human reviewers.

The results:

1. Fast turnaround
2. Multilingual moderation
3. Better performance in edge cases than traditional supervised models

What Meta Achieved?

Why This Matters for Ads :

1

Fast Moderation

Meta’s Few -Shot system catches new violations in less time.

2

Multi-Lingual

It works across 100+ languages and data types- both text and images

3

Accurate

Achieves up to 55% improvement over legacy few-shot models in accuracy on new violations

This makes FSL a powerful middle-ground more precise than zero-shot, but faster than full retraining.

Google & Meta are Not Alone

While Google leverages LLM-generated captions and cross modal co-embeddings, other platforms are too pushing their frontier.

OpenAI uses an **LLM-based Moderation API** to classify unsafe content.

Microsoft offers a real-time, enterprise-grade **moderation API through Azure** that handles text, images, and videos across multiple languages and platforms.

Reddit combines **community moderation with simple AI-assisted** tools for scale.

YouTube uses a combination of **ML models and human-in-the-loop** to moderate videos at scale, ensuring transparency through explainable AI systems.

TikTok integrates **Flamingo-style video-language models** to understand short videos.

Discord empowers communities with **automated moderation bots** and is now integrating **LLMs like Clyde AI** to better understand and manage user content in real time.

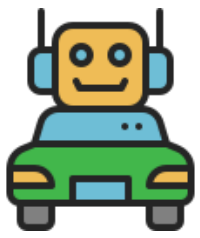
Applications Across Industries

Synthetic content detection

Deepfake and synthetic content detection using zero/few-shot vision-language agents that generalize across unseen forgery types.

Healthcare

Medical image anomaly detection and diagnostics where few or no labeled examples exist, speeding deployment in low-data clinical settings.



Autonomous Vehicles

Autonomous driving hazard and novel object detection via multi-agent vision-language reasoning to catch out-of-label dangers in real time.

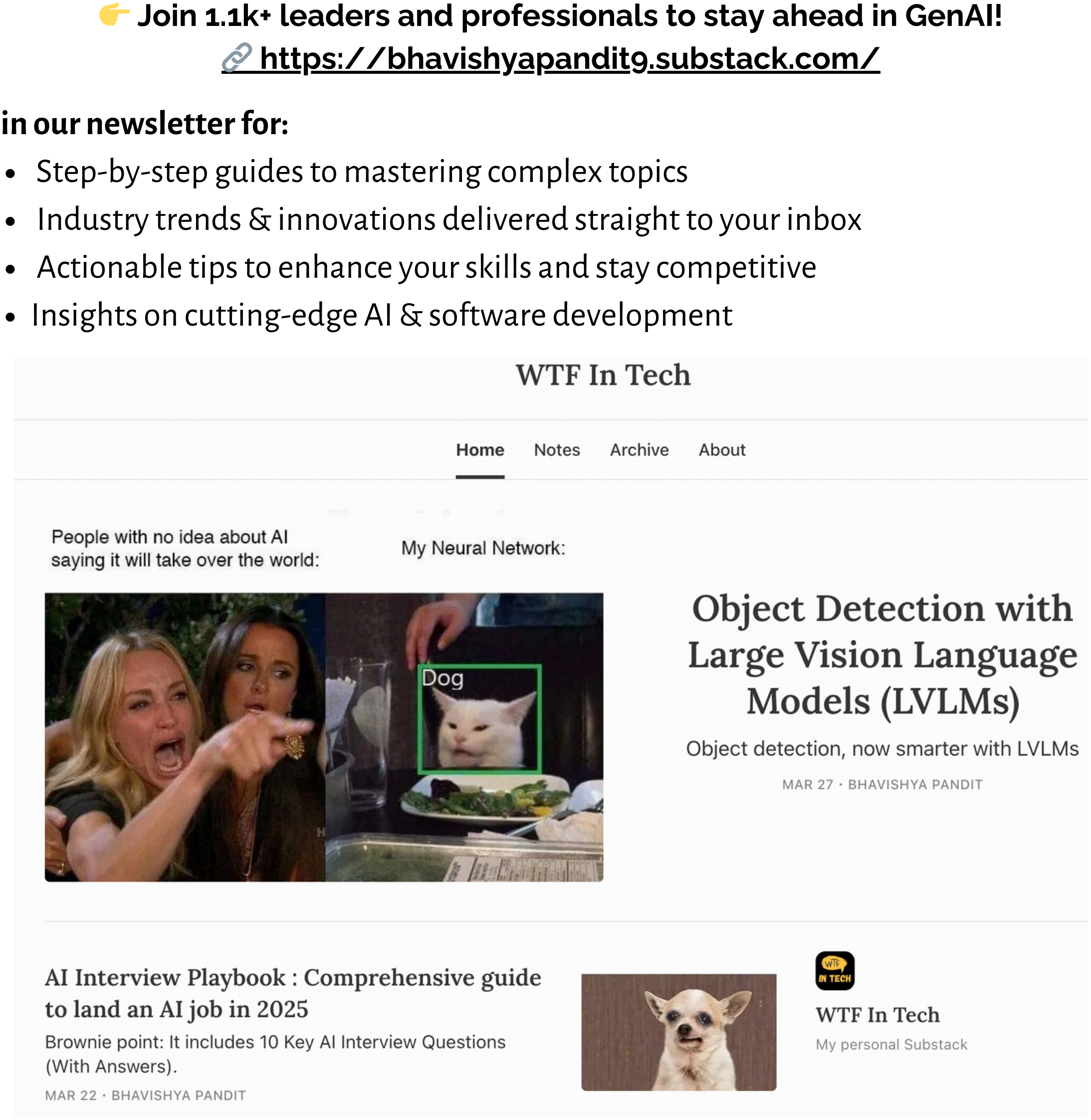
Stay Ahead with Our Tech Newsletter! 🚀

👉 **Join 1.1k+ leaders and professionals to stay ahead in GenAI!**

🔗 **https://bhavishyapandit9.substack.com/**

Join our newsletter for:

- Step-by-step guides to mastering complex topics
- Industry trends & innovations delivered straight to your inbox
- Actionable tips to enhance your skills and stay competitive
- Insights on cutting-edge AI & software development



💡 Whether you're a developer, researcher, or tech enthusiast, this newsletter is your shortcut to staying informed and ahead of the curve.

Bhavishya Pandit

Follow to stay updated on Generative AI

LIKE

COMMENT

REPOST