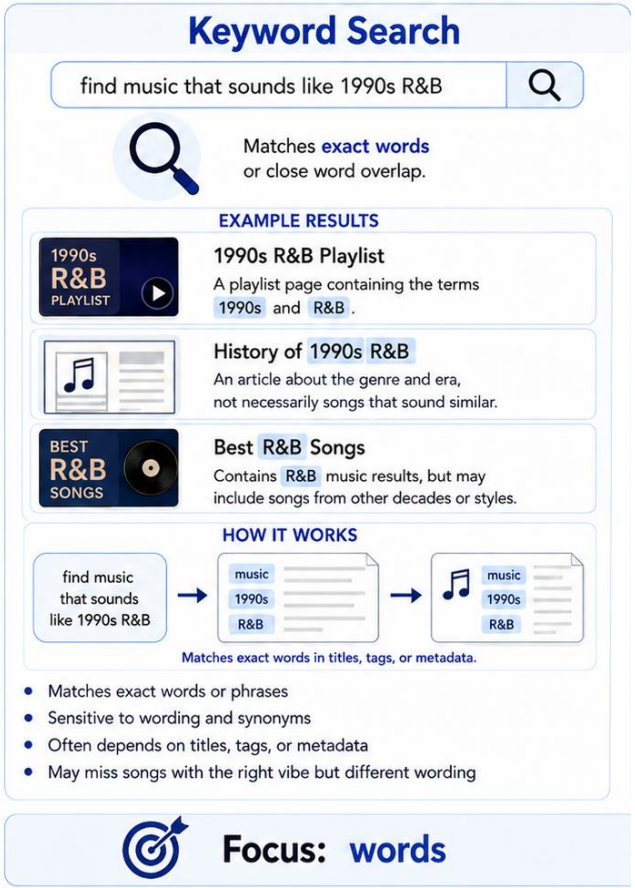
Adversarial Hubness in Multimodal Retrieval

Tingwei Zhang

Fnu Suya, Rishi Jha, Collin Zhang, Vitaly Shmatikov

Cornell Tech
University of Tennessee

Keyword Search vs. Semantic Search



Keyword Search vs. Semantic Search

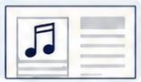
Keyword Search

find music that sounds like 1990s R&B

Matches **exact words** or close word overlap.

EXAMPLE RESULTS

1990s R&B Playlist
A playlist page containing the terms 1990s and R&B.



History of 1990s R&B
An article about the genre and era, not necessarily songs that sound similar.

Best R&B Songs
Contains R&B music results, but may include songs from other decades or styles.

HOW IT WORKS

Matches exact words in titles, tags, or metadata.

- Matches exact words or phrases
- Sensitive to wording and synonyms
- Often depends on titles, tags, or metadata
- May miss songs with the right vibe but different wording

Semantic Search

find music that sounds like 1990s R&B

Understands **meaning**, intent, and style behind the query.

Meaning-based

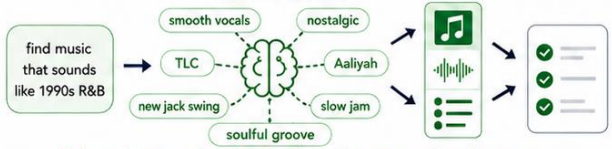
EXAMPLE RESULTS

Songs with a 1990s R&B Vibe
Finds tracks with **smooth grooves**, **soulful vocals**, and classic **90s R&B production**.

Artists Similar to TLC, Aaliyah, and Boyz II Men
Retrieves music related in **sound**, **mood**, and **era** even if the exact words are not in the title.

Playlists for Smooth Nostalgic R&B
Matches **mood**, **style**, and **listener intent** across tracks and playlists.

HOW IT WORKS



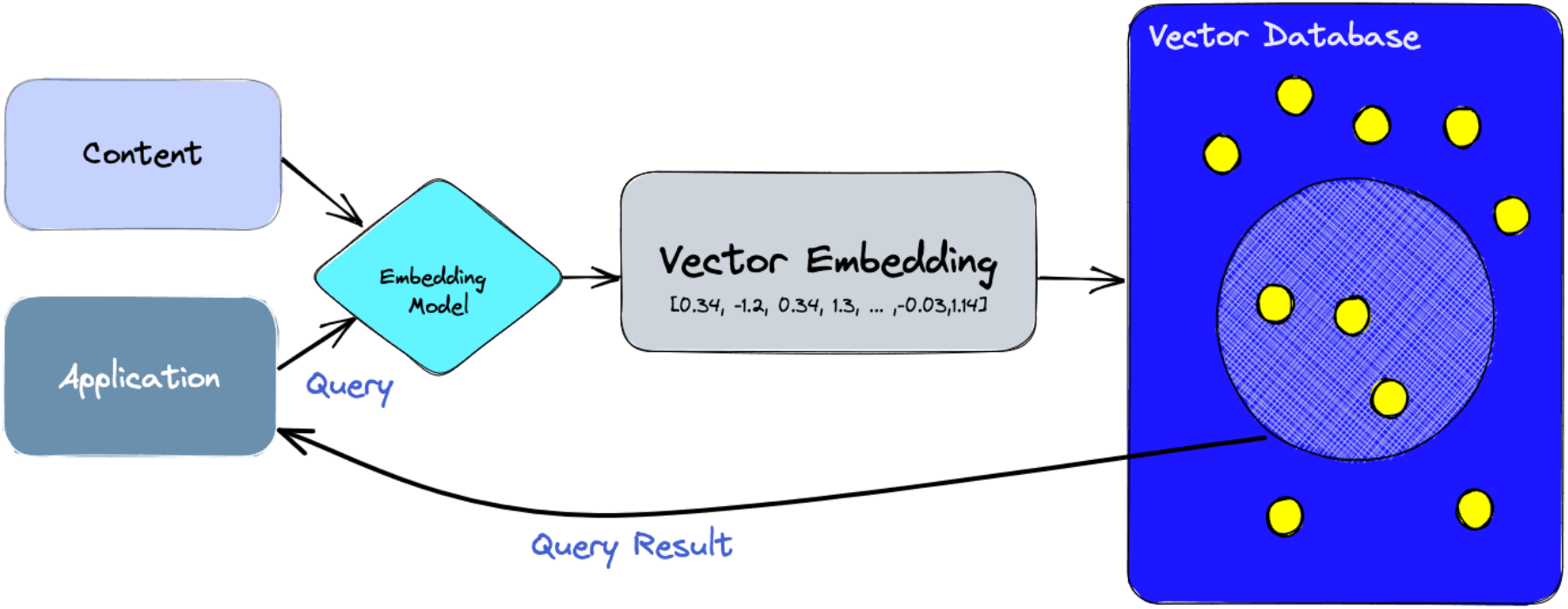
Understands intent and matches related sounds, artists, moods, and concepts.

- Understands meaning and user intent
- Finds similar sounds, moods, and genres
- Handles vibes, related artists, and natural language
- Better for nuanced or conversational queries

Focus: words

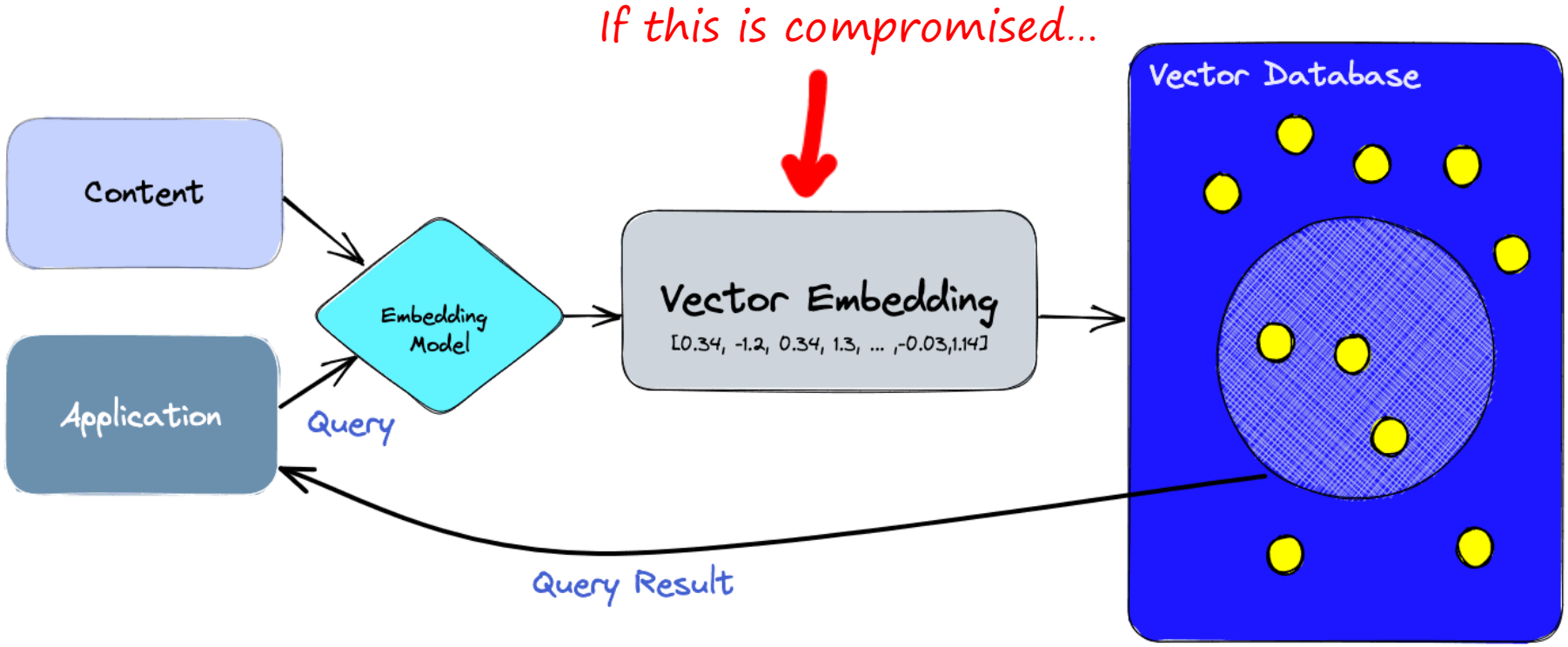
Focus: meaning

Semantic Retrieval

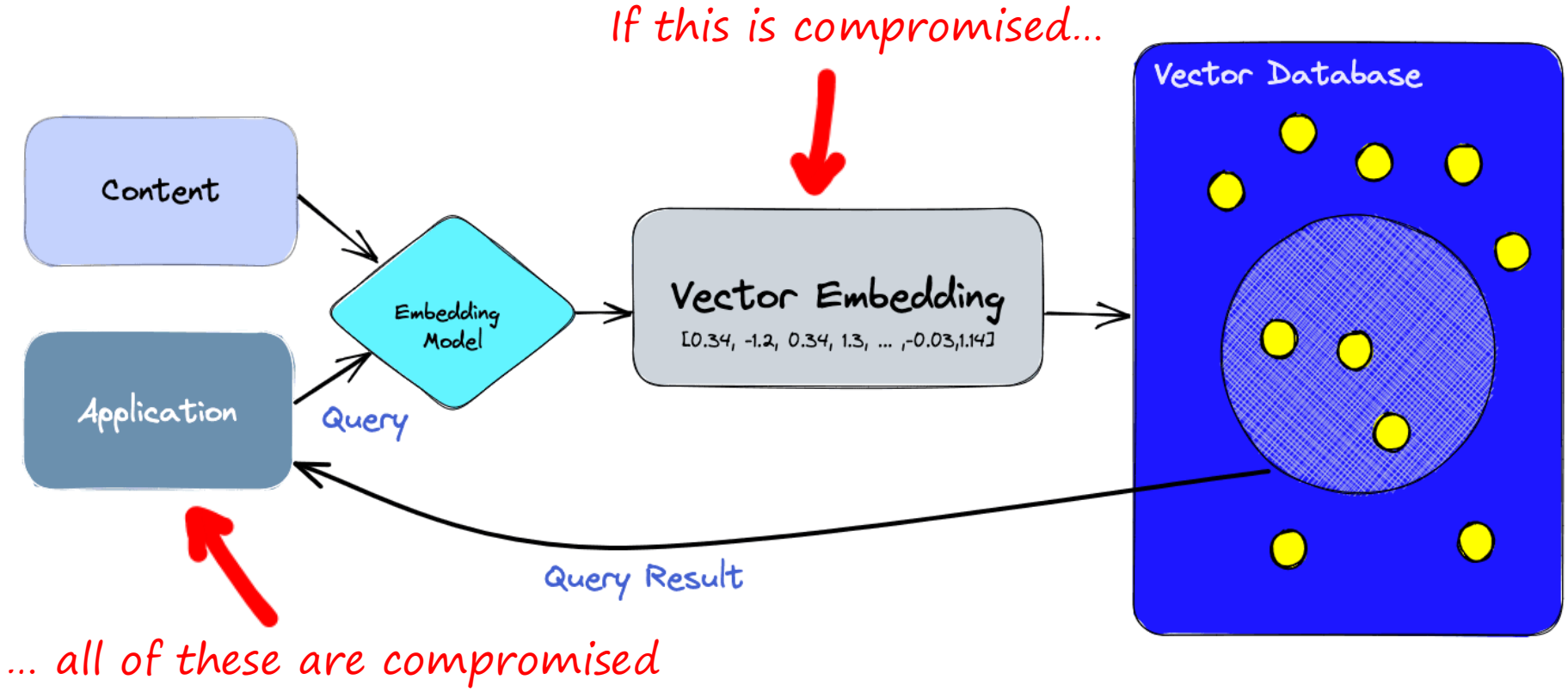


This is not theoretical!

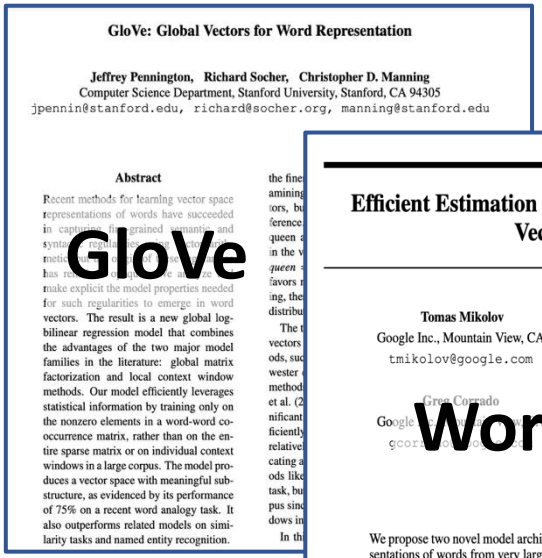
Semantic Retrieval



Semantic Retrieval



Word Embeddings



GloVe: Global Vectors for Word Representation

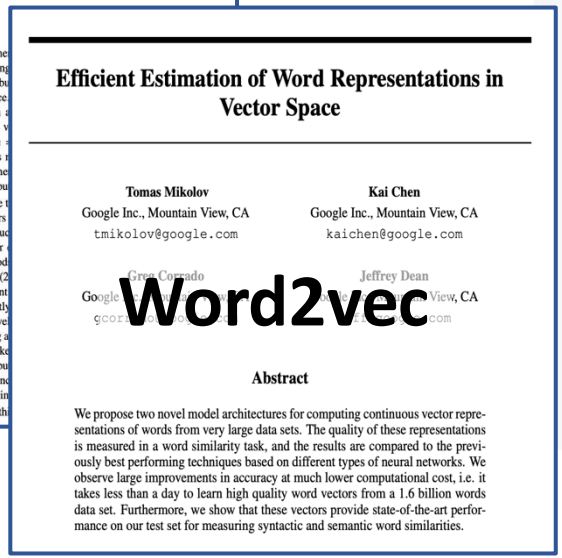
Jeffrey Pennington, Richard Socher, Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA 94305
jpennin@stanford.edu, richard@socher.org, manning@stanford.edu

Abstract

Recent methods for learning vector space representations of words have succeeded in capturing fine-grained semantic and syntactic regularities using vector arithmetic, but the origin of these regularities has remained opaque. We analyze and make explicit the model properties needed for such regularities to emerge in word vectors. The result is a new global log-bilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. Our model efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus. The model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a recent word analogy task. It also outperforms related models on similarity tasks and named entity recognition.

GloVe

50K citations



Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA

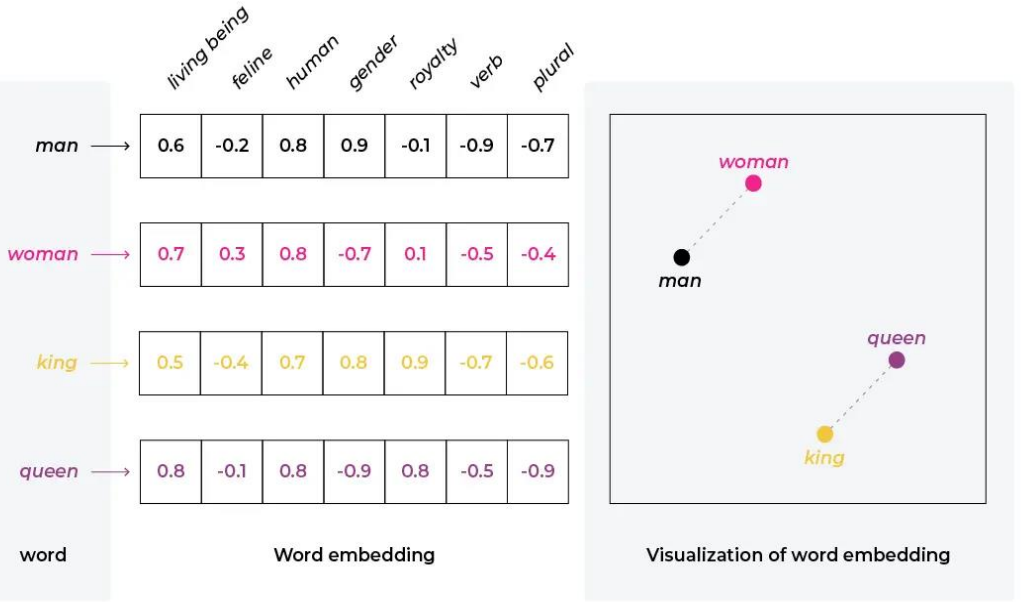
Jeffrey Dean
Google Inc., Mountain View, CA

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

Word2vec

53K citations



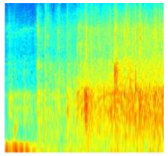
Words → Vector representations

Multi-modal Embeddings

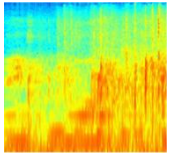
"Dog"

"Cat"

...



"Barking"



"Meowing"

Multi-modal encoders

ImageBind

Titan

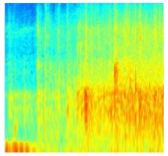
Vertex

Multi-modal Embeddings

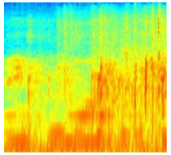
"Dog"

"Cat"

...



"Barking"



"Meowing"

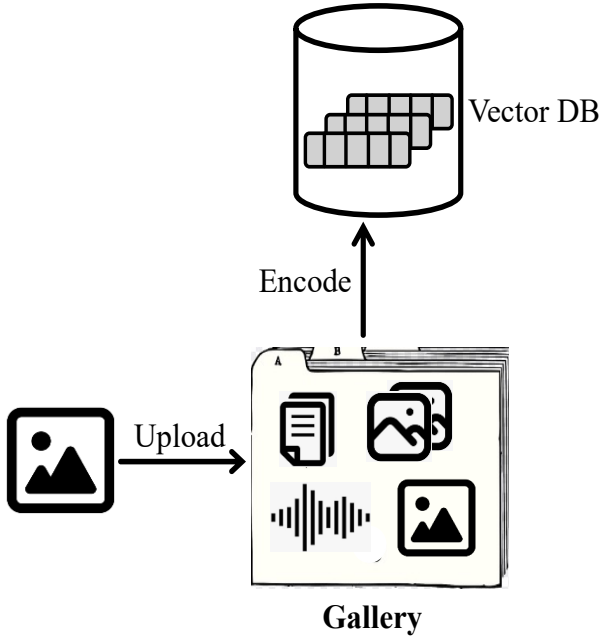
cosine similarity

ImageBind

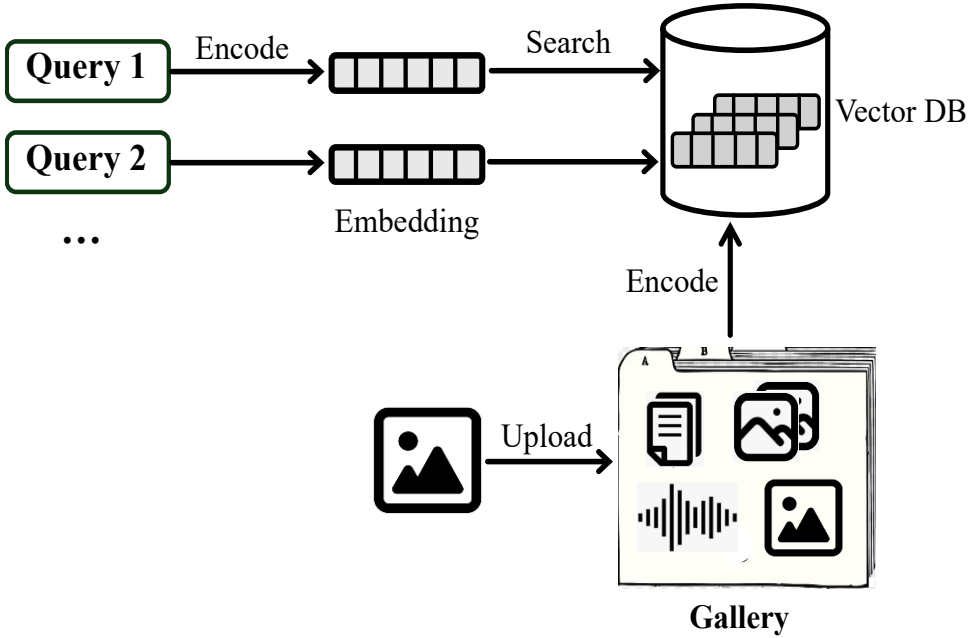
Titan

Vertex

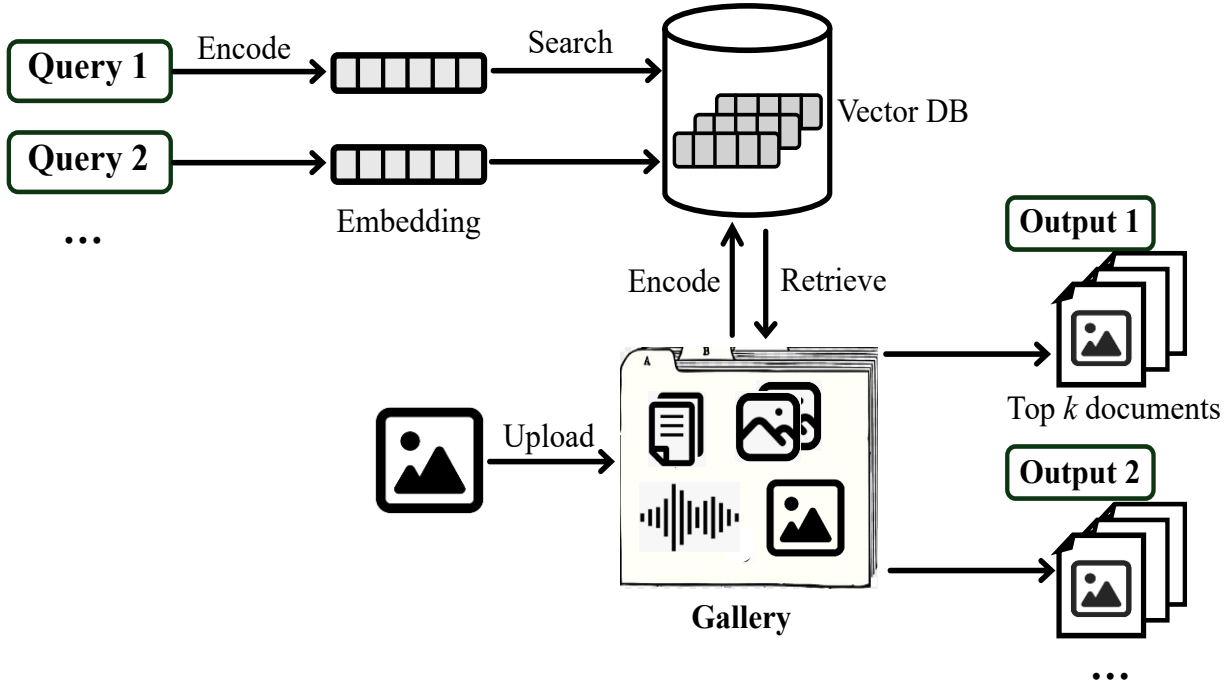
Modern Cross-Modal Retrieval



Modern Cross-Modal Retrieval



Modern Cross-Modal Retrieval

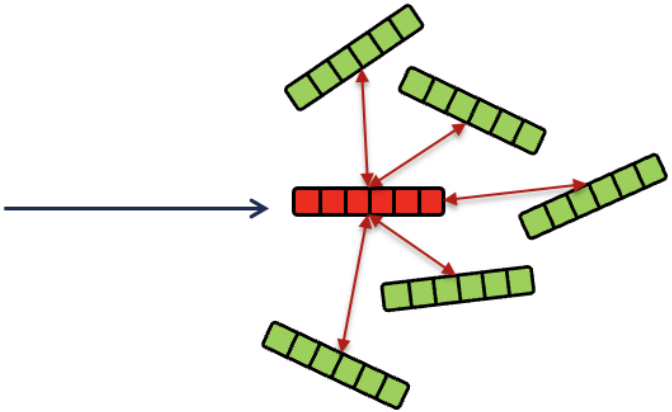


Threats to Information Retrieval

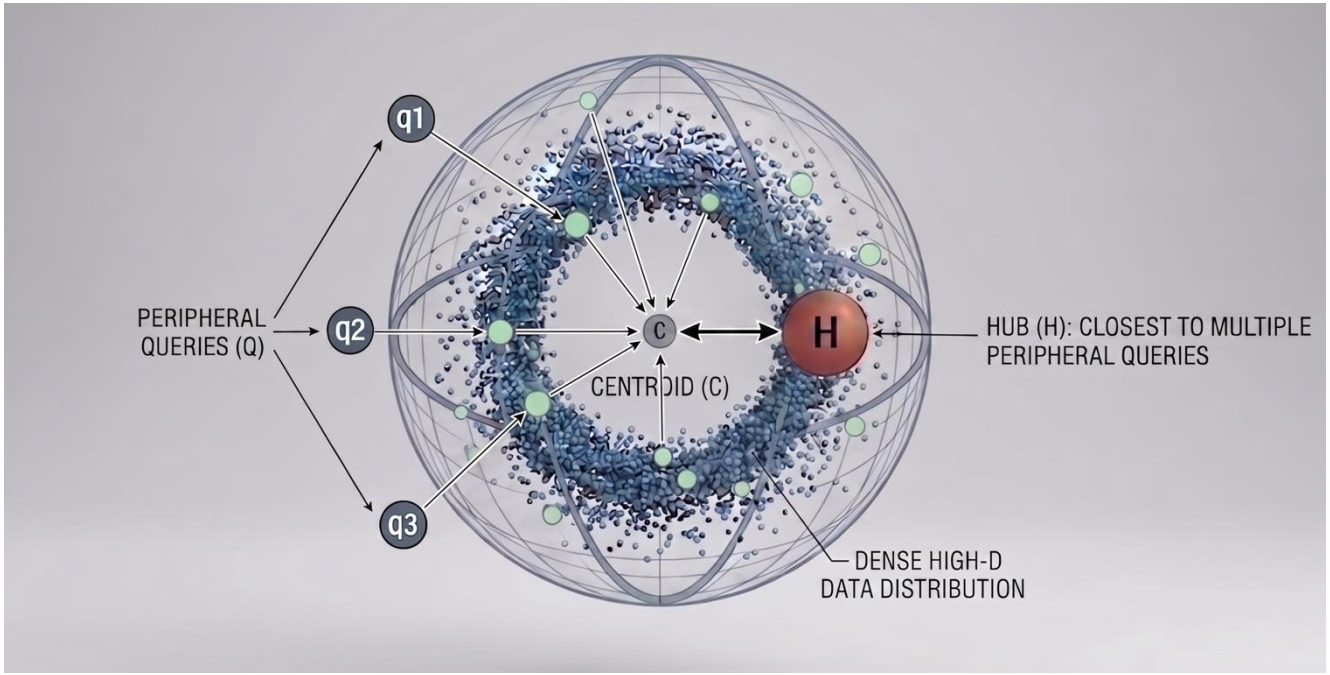
- Search/Generative Engine Optimization
- Spamming retrieval and search
- Spamming product listing
- Spreading misinformation

Embedding Search Makes SEO Easier

- Search/Generative Engine Optimization
- Spamming retrieval and search
- Spamming product listing
- Spreading misinformation



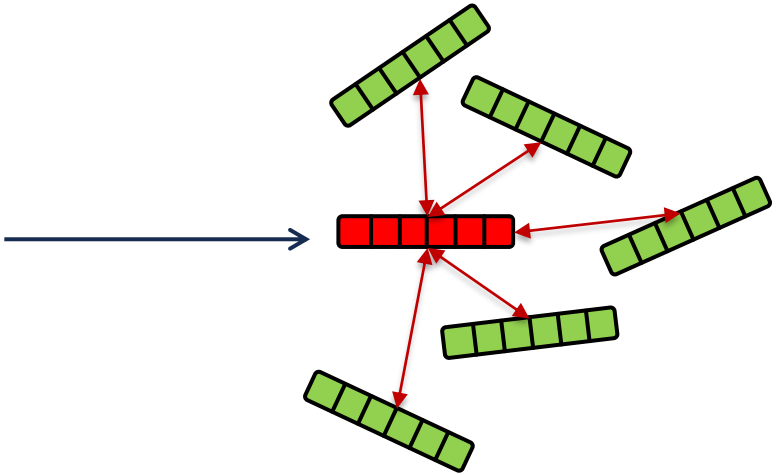
Natural Hubness



Natural Hubness: in high-dimensional vector spaces, a few points appear as neighbors of many other points. (e.g., Strongest natural hub in MS COCO (5 captions per image): one image retrieved by 102 of 25K captions queries with OpenCLIP.)

Question: Could adversaries exploit hubness?

Adversarial ML techniques can create hubs



Motivations

"People upload massive amounts of albums that are intended to be streaming fraud albums."

Another strategy is creating AI covers of popular songs and getting them onto popular playlists so normal people will stream them. Another involves bots "listening" to songs.

Earlier this year, a Danish man was sentenced to 18 months in prison for using bots to get about $300,000 in royalties. Another man, Michael Smith, was arrested and charged with defrauding streaming services out of $10 million over the course of seven years. Smith used AI tools to create hundreds of thousands of songs under the names of fake artists such as "Calm Baseball," "Calm Connected," and "Calm Knuckles." He then streamed the huge catalog using bots, billions of times, prosecutors allege. That diverted money that should have gone to real musicians that real people were really listening to. (In this case, Spotify paid only $60,000 to Smith, suggesting the company's protective measures worked to limit payments, Macowski said.)

Sifting Through A.I. Slop: What's Real, What's Not, and Why It Matters

Adam Nemeroff, The Conversation
Updated Tue, September 2, 2025 at 11:06 AM PDT
3 min read

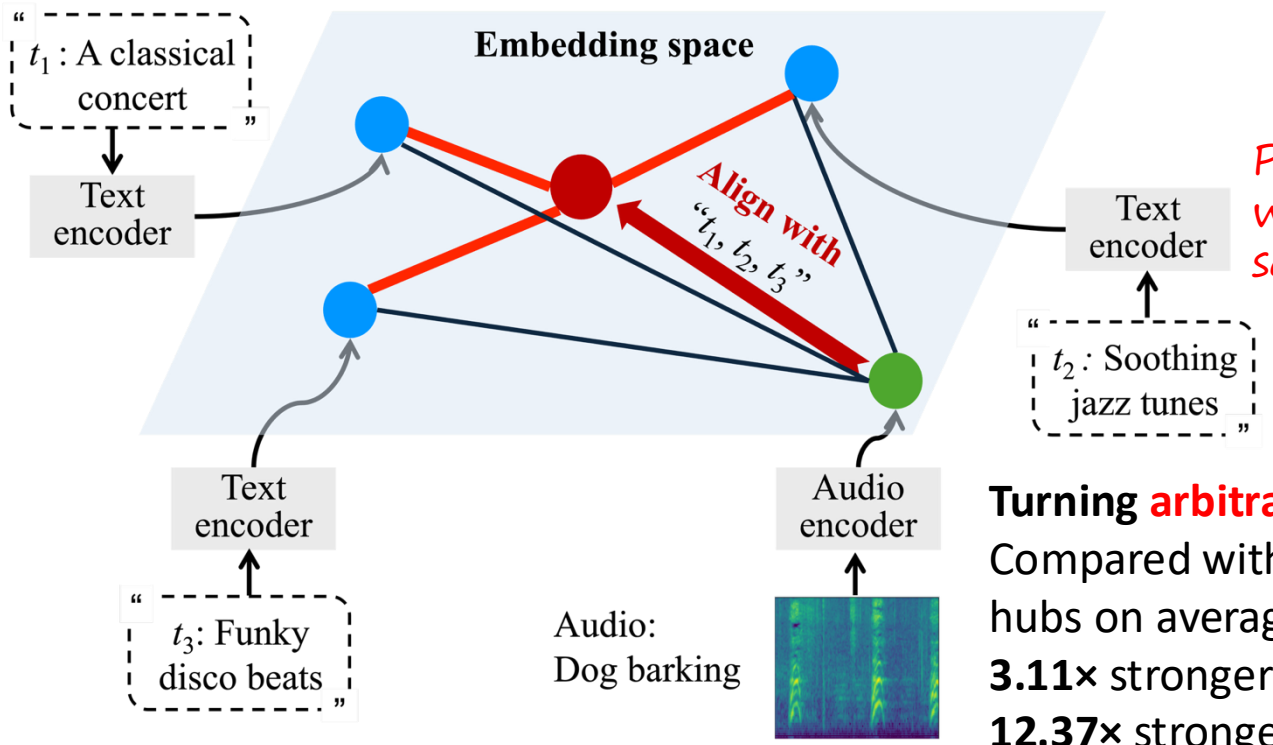
Add Yahoo on Google

Sifting Through A.I. Slop: What's Real, What's Not, and Why It Matters

You've probably encountered images in your social media feeds that look like a cross between photographs and computer-generated graphics. Some are fantastical — think Shrimp Jesus — and some are believable at a quick glance — remember the little girl clutching a puppy in a boat during a flood?

These are examples of A.I. slop: low- to mid-quality content — video, images, audio, text, or a mix — created with A.I. tools, often with little regard for accuracy. It's fast, easy, and inexpensive to make this content. A.I. slop producers typically place it on social media to exploit the economics of attention on the internet, displacing higher-quality material that could be more helpful.

Adversarial Hubness



Perturb input to align it with many targets at the same time.

Turning arbitrary inputs into powerful hubs

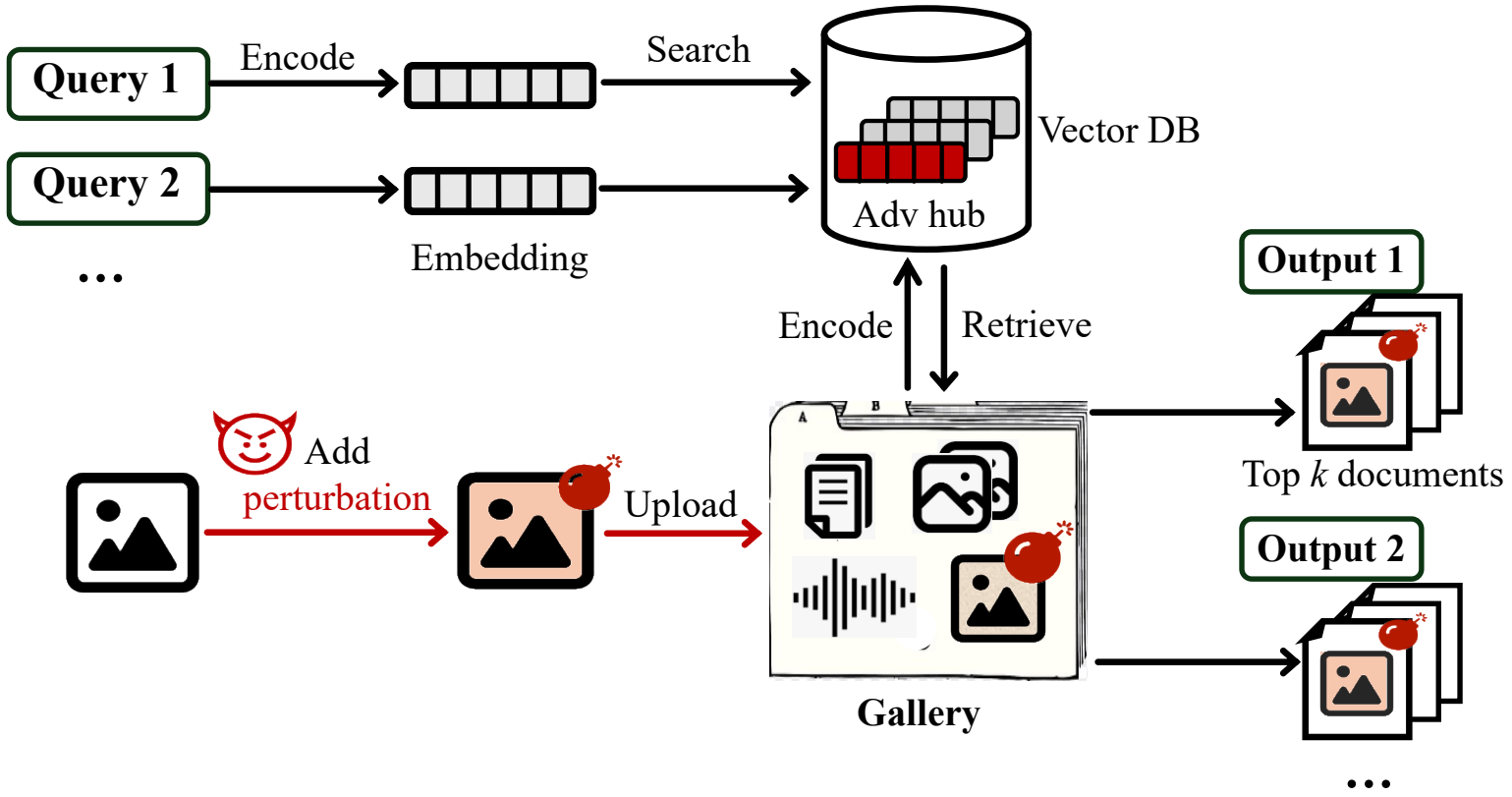
Compared with the strongest natural hub, our hubs on average are:

3.11× stronger for image-to-image

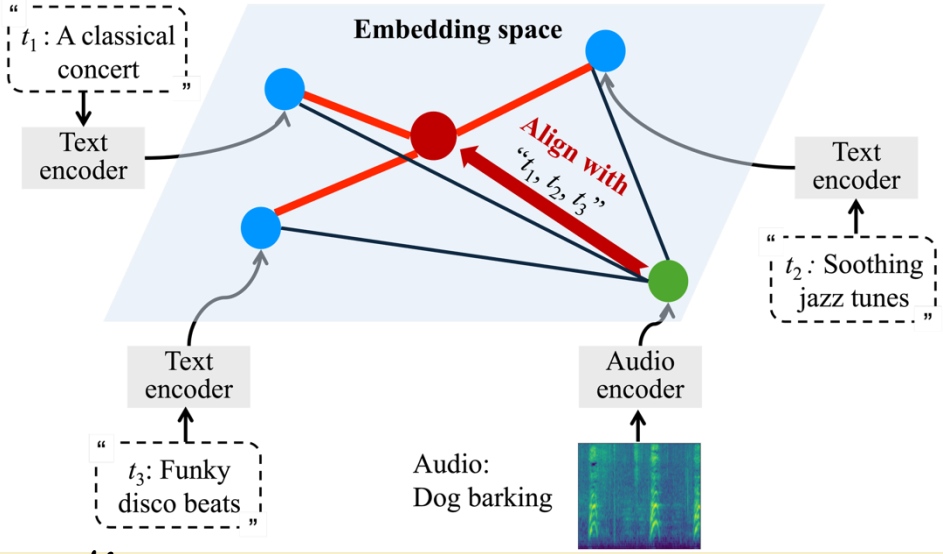
12.37× stronger for text-to-audio

155.96× stronger for text-to-image

Adversarial Hubness in Cross-Modal Retrieval



Multi-Modal Adversarial Hubs



Alignment

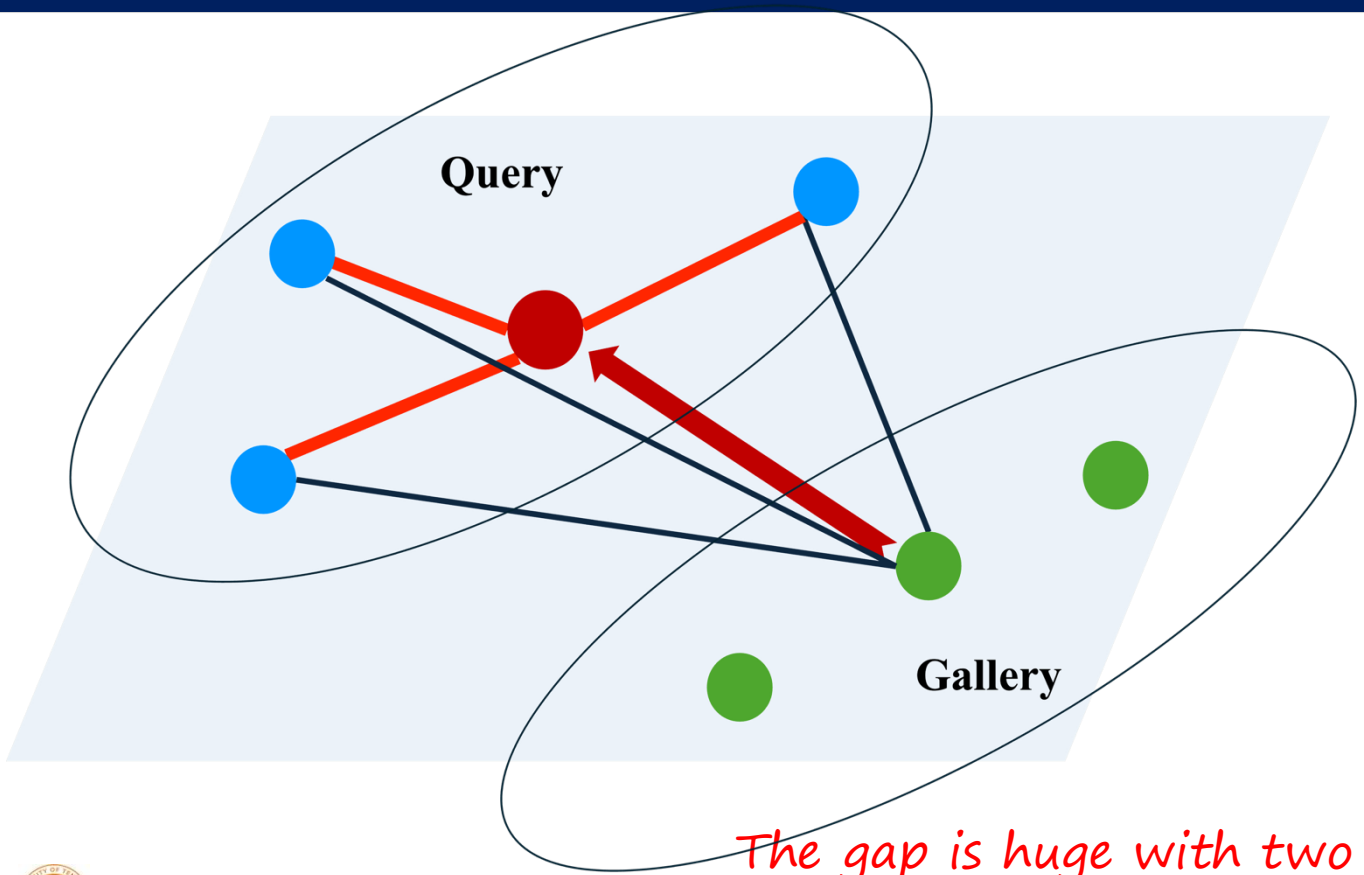
$$\theta^{\mathrm{m}}(x+\delta) \sim \mathrm{Centroid}(\theta^{\overline{m}}(\mathrm{Y_s})), \mathrm{Y_s} \subseteq Y$$

clean input

perturbation

randomly selected query set

Distribution Gap



The gap is huge with two different modalities

Adversarial Hub in Text-to-Image Retrieval

Query	Result			
“Golf cart perfect for courses”				
“Coolest looking sports car”				
“Small car that is easy to park in the city”				

Adversarial Hub in Image-to-Image Retrieval

Results

- **91.3%** top-1 success against text-to-image retrieval, aligning one image with 25k queries
- **58.0%** top-5 success against image-to-image retrieval, 45.7% success against **Pinecone image-search App (black-box)** Pinecone
- **75.3%** top-5 success against text-to-audio retrieval

Only used 100 randomly selected target query sample!

Takeaways

- Embedding search makes SEO easier: attackers only need one vector to be close to many queries.
- Hubness makes this feasible in high-dimensional spaces.
- Existing defenses for natural hubness do not stop concept-based adversarial hubs.

More details in the paper…

Companies are making moves

March 12, 2026 2 Comments

Artificial Intelligence - AI

Your Model's Memory Has Been Compromised: Adversarial Hubness in RAG Systems

3 min read

Idan Habler, Vineeth Sai Narajala

This blog is jointly written by Amy Chang, Idan Habler, and Vineeth Sai Narajala.

Prompt injections and jailbreaks remain a major concern for AI security, and for good reason: models remain susceptible to users tricking models into doing or saying things like bypassing guardrails or leaking system prompts. But AI deployments don't just process prompts at inference time (meaning when you are actively querying the model): they may also retrieve, rank, and synthesize external data in real time. Each of those steps is a potential adversarial entry point.

Retrieval-Augmented Generation (RAG) is now standard infrastructure for enterprise AI, allowing large language models (LLMs) to obtain external knowledge via vector similarity search. RAGs can connect LLMs to corporate knowledge repositories and customer support systems. But that grounding layer, known as the vector embedding space, introduces its own attack surface known as adversarial hubness, and most teams aren't looking for it yet.

But Cisco has you covered. We'd like to introduce our latest open source tool: Adversarial Hubness Detector.

Thank You!

https://arxiv.org/pdf/2412.14113

Our code is available!