~~#~~ BUILDING MULTIMODAL SEARCH AND RAG

- MULTIMEDIA CONTENT IS ALL AROUND US;

- DATA FROM DIFFERENTS SOURCES;

- TRAINING MULTIMODAL MODELS: START WITH SPECIALIST MODELS;

  (a) TEXT ENCODER; (c) AUDIO ENCODER;

  (b) IMAGE ENCODER; (d) VIDEO ENCODER;

- SIMILAR CONCEPTS = SIMILAR VECTORS;

  (DATA) (VECTORS REPRESENTATIONS)

  └▷ UNIFY THE SPECIALIST MODELS;

- UNIFY THE MODELS USING CONTRASTIVE LEARNING:

  └▷ PROCESS TO TRAIN ANY EMBEDDING MODEL;

  └▷ UNIFY MULTIPLE MODELS;

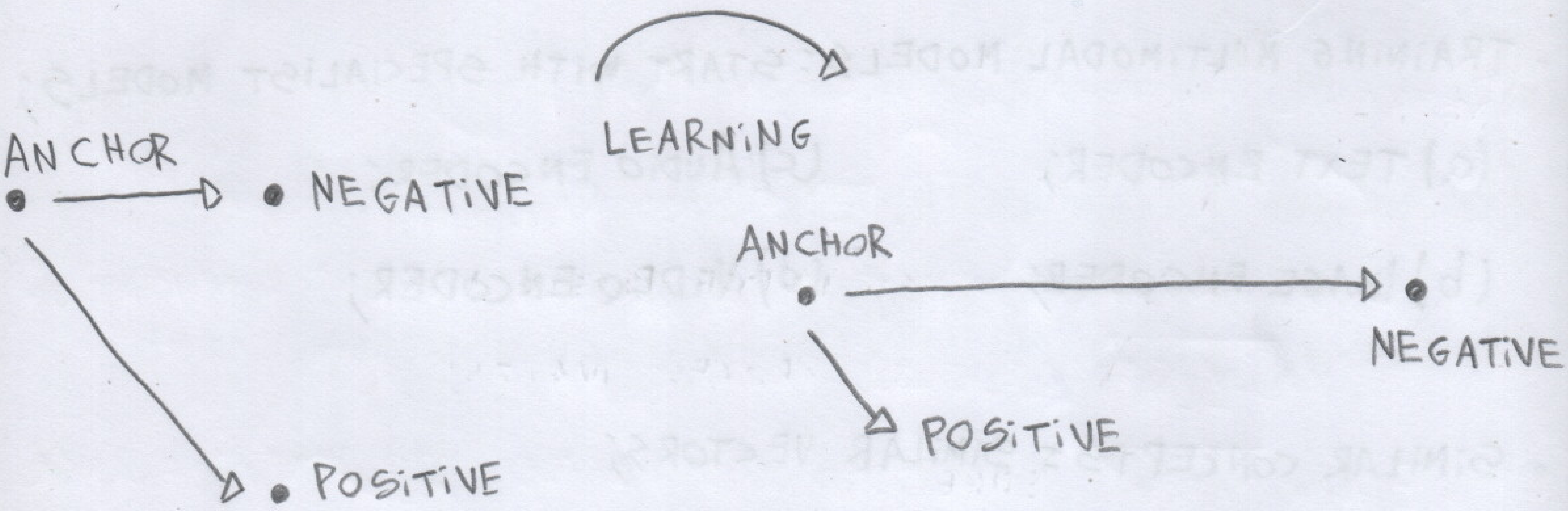
  └▷ CREATE ONE VECTOR SPACE;

  └▷ TUNE MODELS BY PROVIDING CONTRASTIVE EXAMPLES;

- E.g: ANCHOR → "HE COULD SMELL THE ROSES"

  POSITIVE EXAMPLE → "A FIELD OF FRAGRANT FLOWERS"

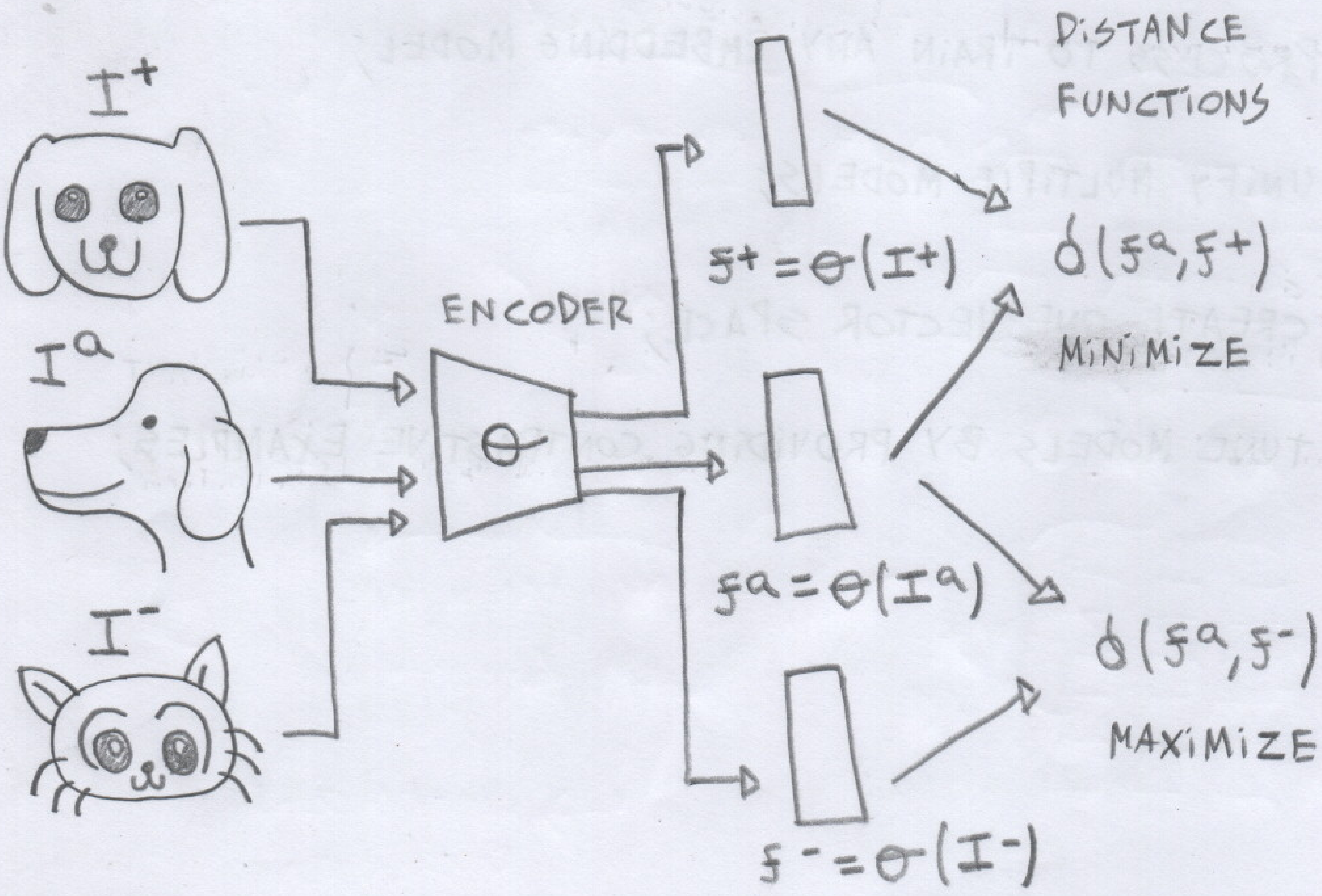
  NEGATIVE EXAMPLE → "THE LION ROARED MAJESTICALLY"



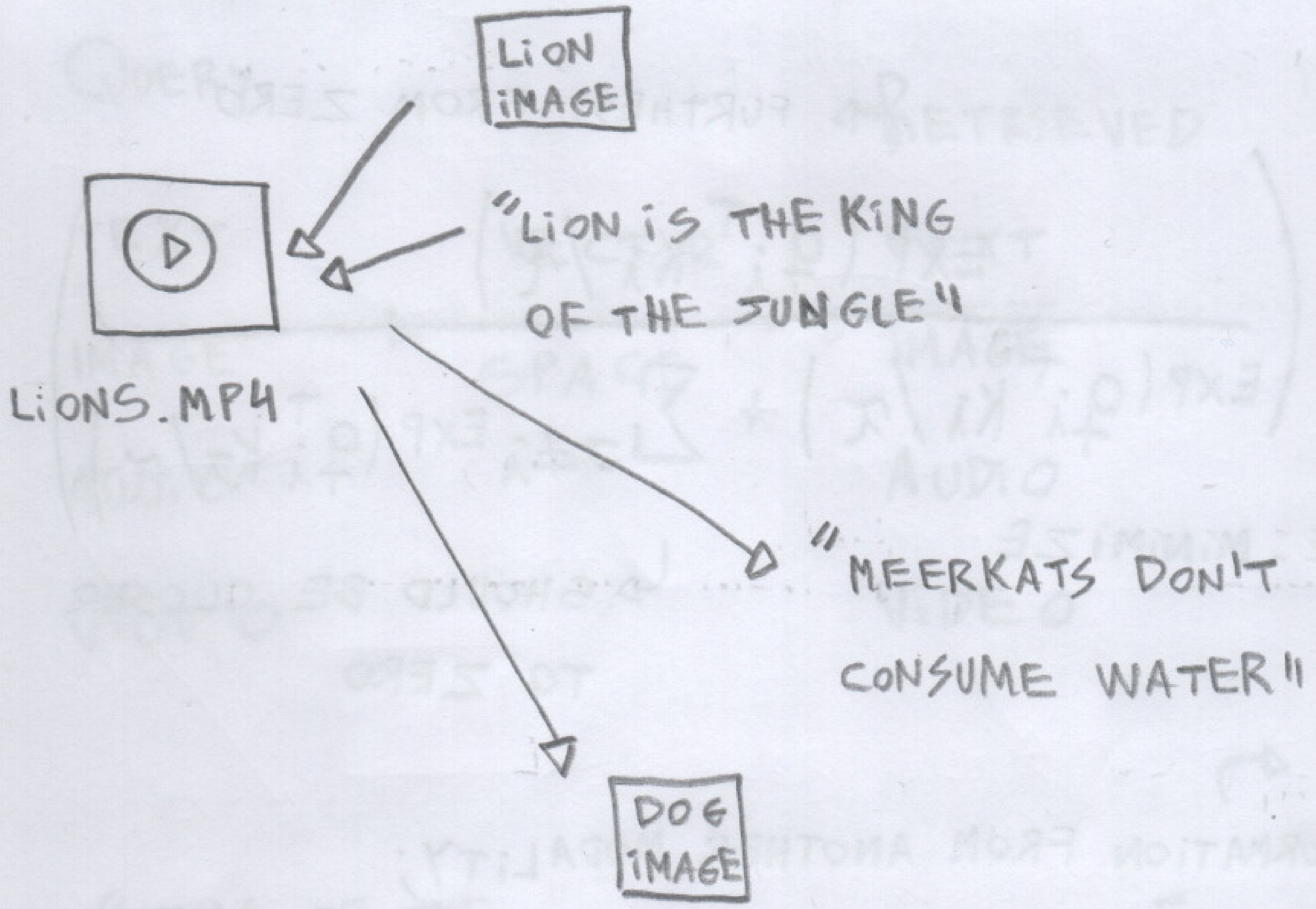
↳ SAME EXAMPLE WITH IMAGES, AUDIO AND VIDEO;

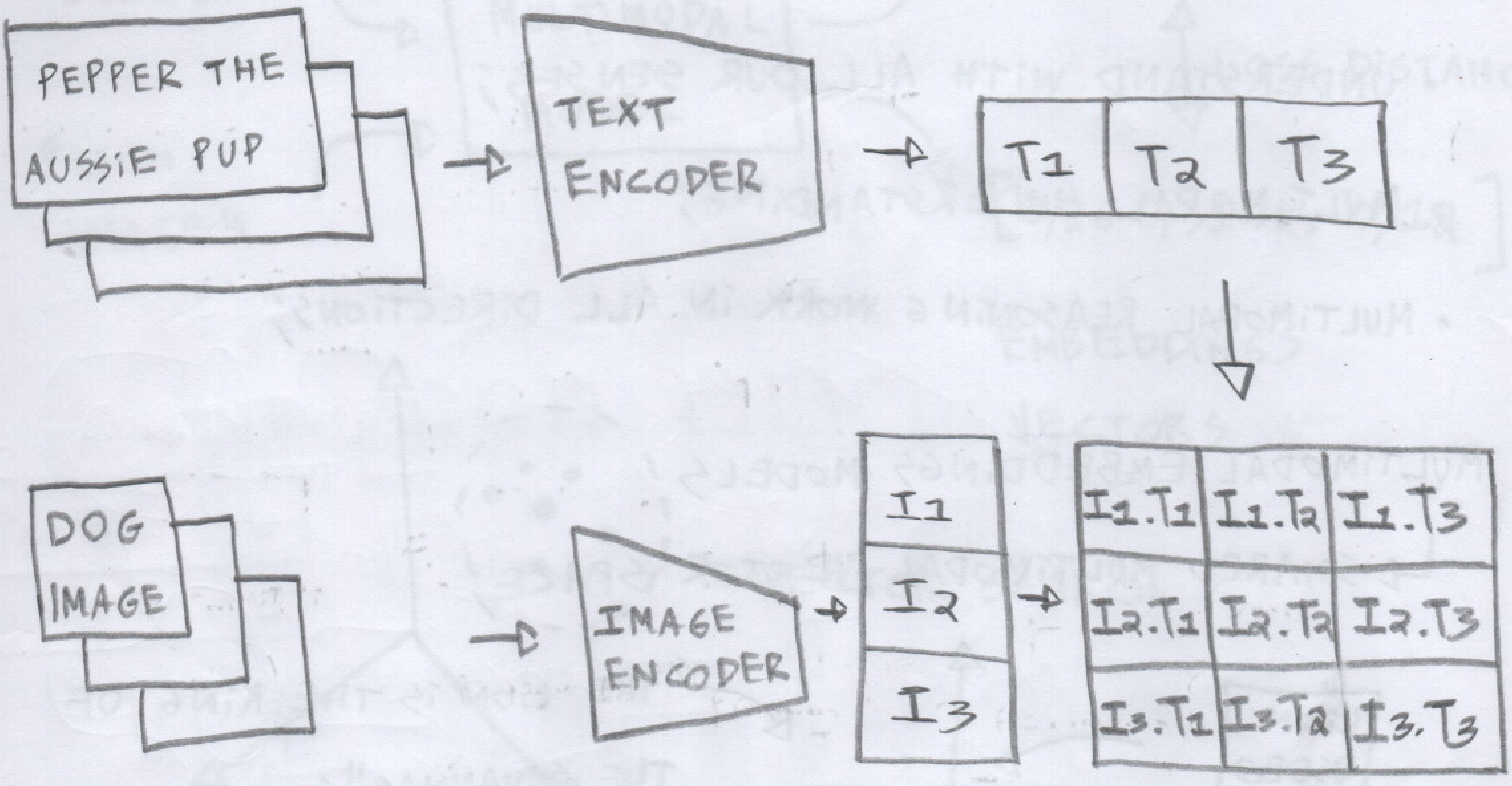
↳ PUSH NEGATIVE EXAMPLE; ↳ CONTRASTIVE LOSS FUNCTION

↳ PULL POSITIVE EXAMPLE;



E.g: APPLY CONTRASTIVE LOSS IN TEXT AND IMAGE MULTIMODAL DATA.


LION IMAGE
"LION IS THE KING OF THE JUNGLE"
LIONS.MP4
"MEERKATS DON'T CONSUME WATER"
DOG IMAGE



I1.T1	I1.T2	I1.T3
I2.T1	I2.T2	I2.T3
I3.T1	I3.T2	I3.T3

DIAGONAL IS POSITIVES EXAMPLES

$q_i = f(\text{"IMAGE OF LION"})$ AND $K_i = g(\text{"VIDEO OF LION"})$

↱ FURTHEST FROM ZERO

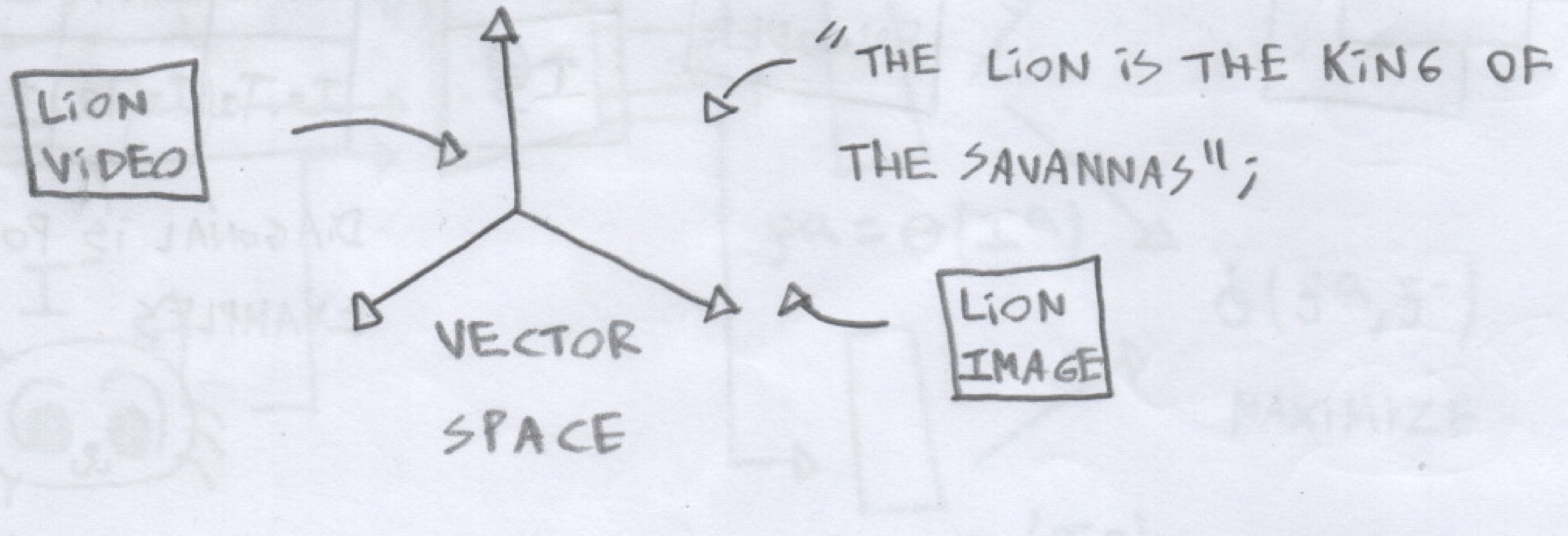
$$L_{I,M} = -\text{LOG}\left(\frac{\text{EXP}(q_i^T K_i/\tau)}{\text{EXP}(q_i^T K_i/\tau) + \sum_{j \neq i} \text{EXP}(q_i^T K_j/\tau)}\right)$$

↳ OBJECTIVE: MINIMIZE FUNCTION

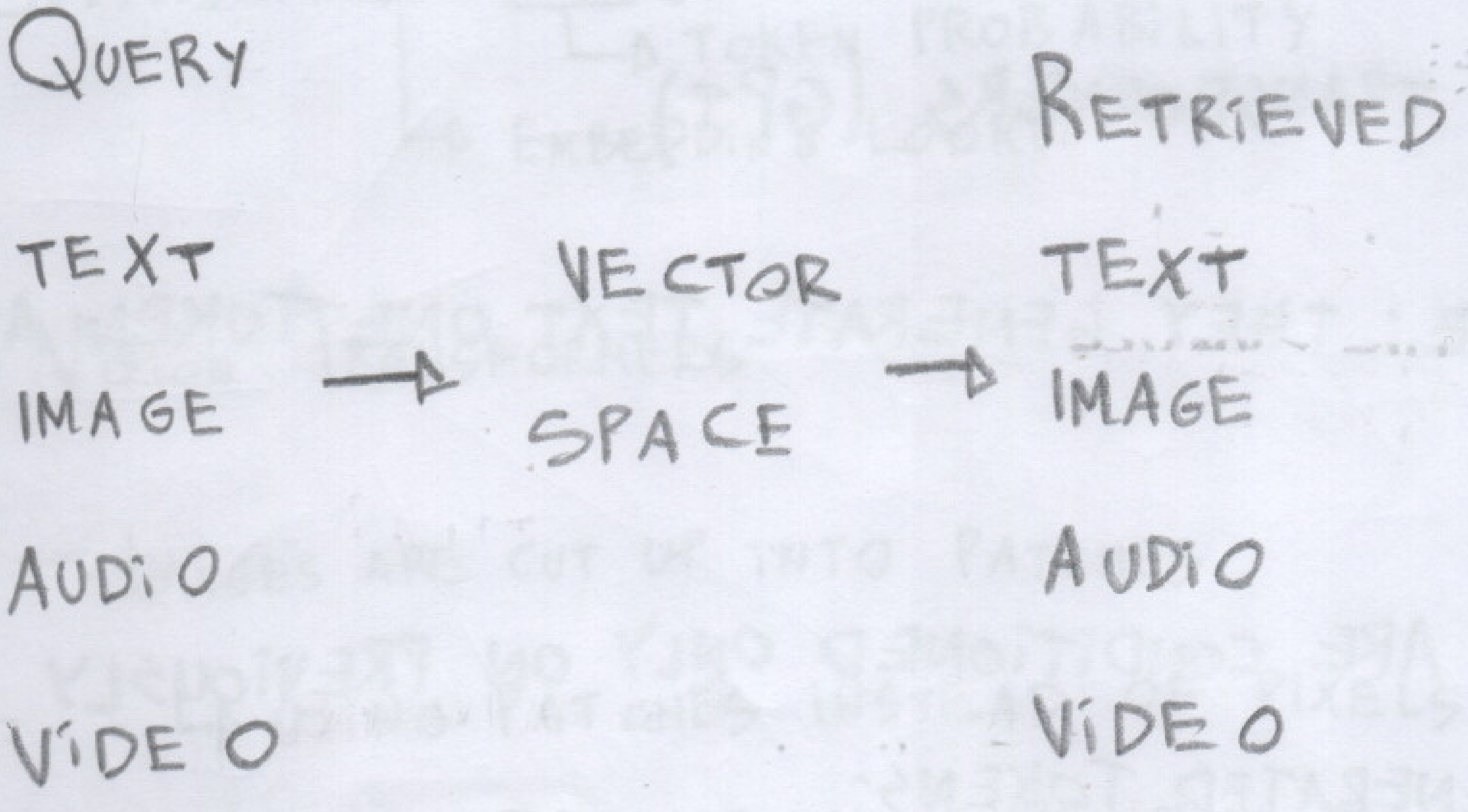
↳ SHOULD BE CLOSER TO ZERO

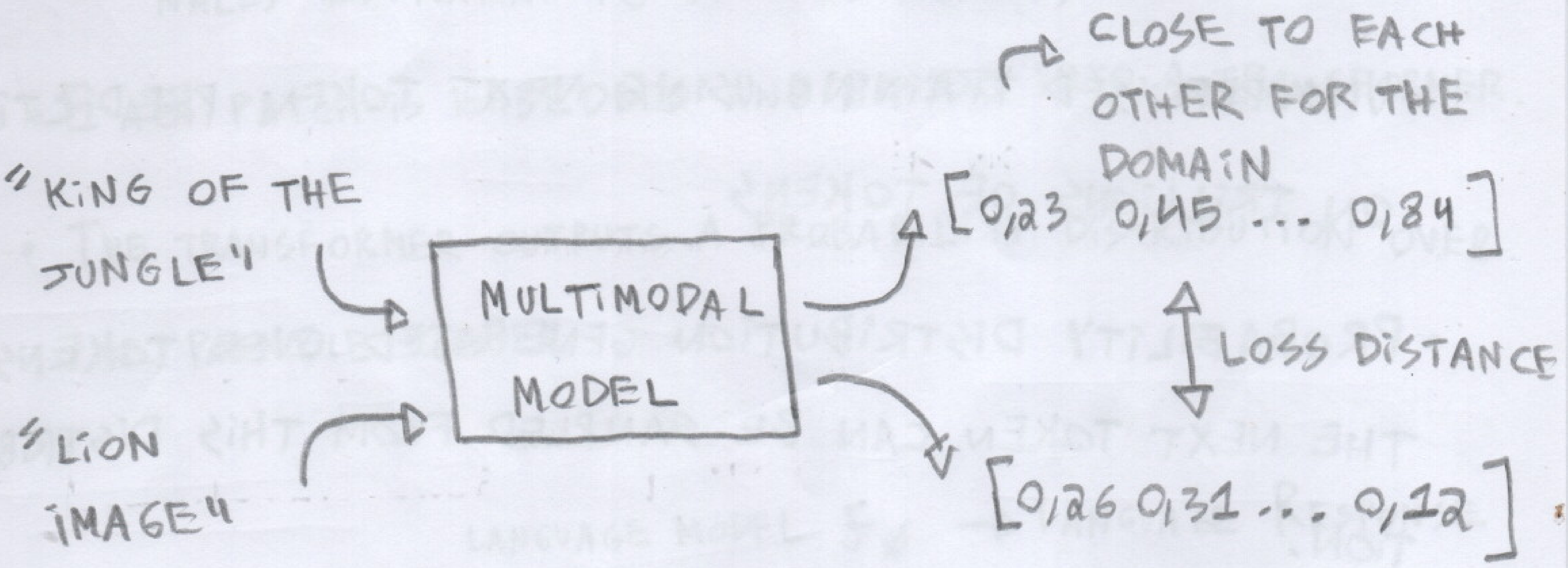
- WE CAN INFER INFORMATION FROM ANOTHER MODALITY;
- MULTIMODAL REASONING AND RETRIEVAL:
    - UNDERSTAND WITH ALL OUR SENSES;
    - MULTIMODAL UNDERSTANDING;
    - MULTIMODAL REASONING WORK IN ALL DIRECTIONS;

- MULTIMODAL EMBEDDINGS MODELS
    - ↳ SHARED MULTIMODAL VECTOR SPACE;

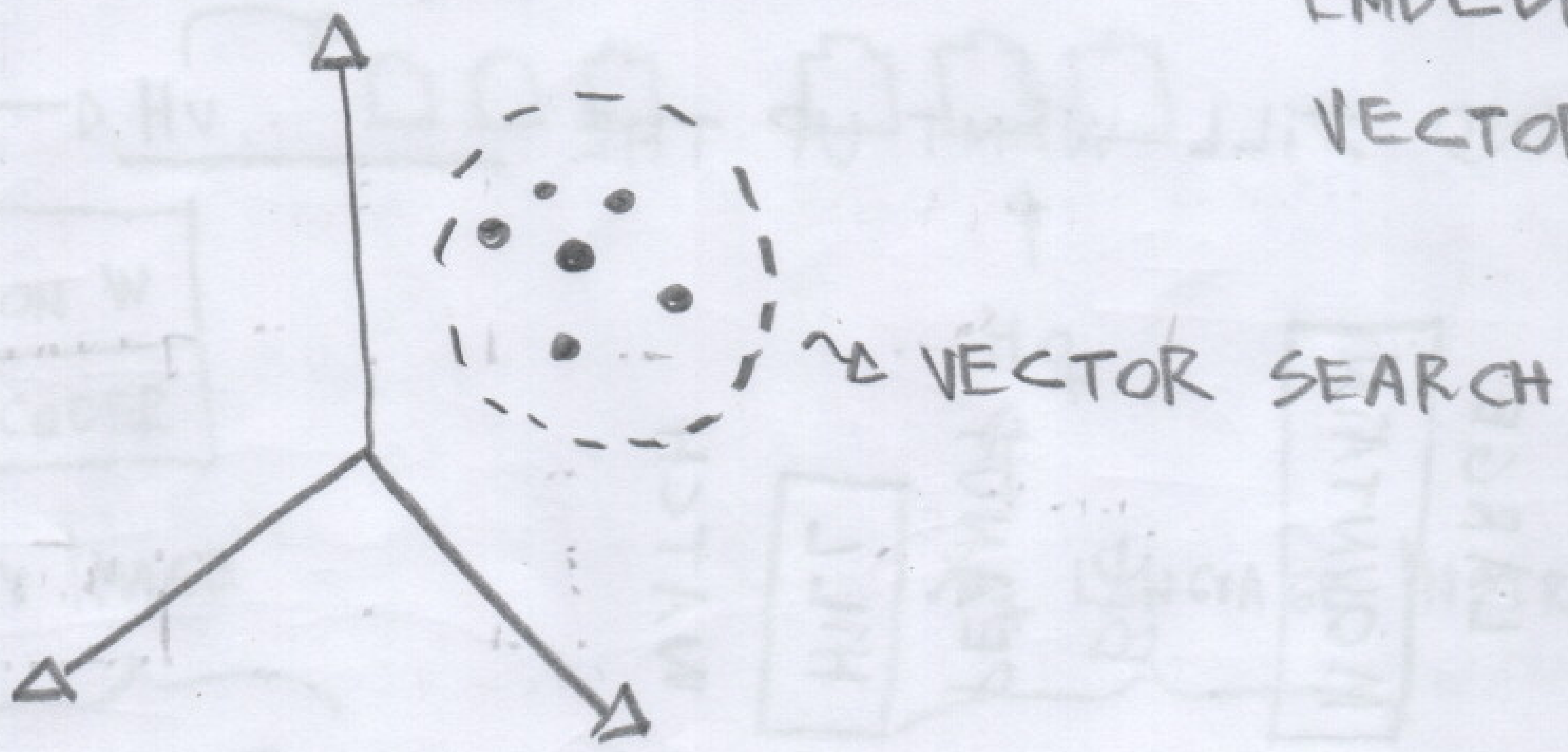


Any to Any Search





EMBEDDINGS VECTORS

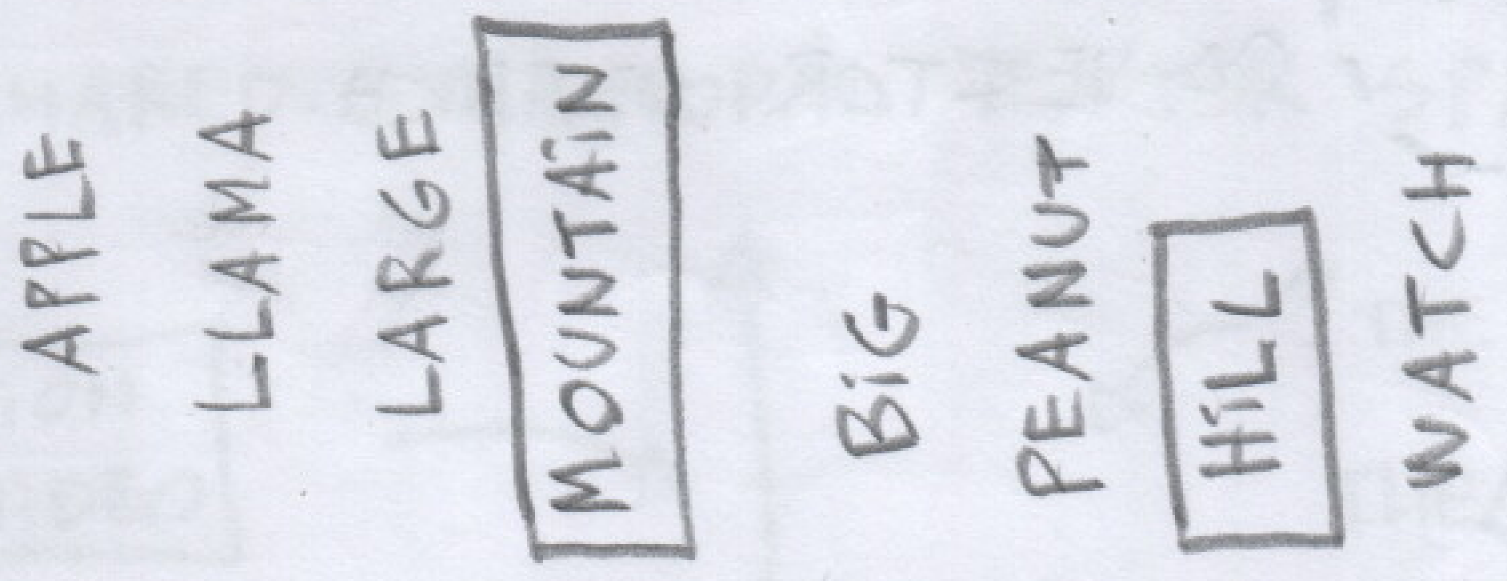


- HOW DO LARGE LANGUAGES MODELS WORK?

↳ THE MAJORITY OF CURRENT LLMs ARE GENERATIVE PRE-TRAINED TRANSFORMERS (GPT).

- AUTOREGRESSIVE: THEY GENERATE TEXT ONE TOKEN AT A TIME;
- FUTURE TOKENS ARE CONDITIONED ONLY ON PREVIOUSLY PROVIDED OR GENERATED TOKENS;
- UNSUPERVISED TRAINING USING NEXT TOKEN PREDICTION ON TRILLIONS OF TOKENS.
- PROBABILITY DISTRIBUTION GENERATED OVER TOKENS: THE NEXT TOKEN CAN BE SAMPLED FROM THIS DISTRIBUTION.

E.g: JACK AND JILL WENT UP THE ________



[1, 2, 3, 7, ..., 3, 1, 9, 1]

PREDICTION SCORE

- HOW DO GPT MODELS WORK?

PROMPT: THE ROCK

↳ TOKEN PROBABILITY

↳ EMBEDDING LOOKUP

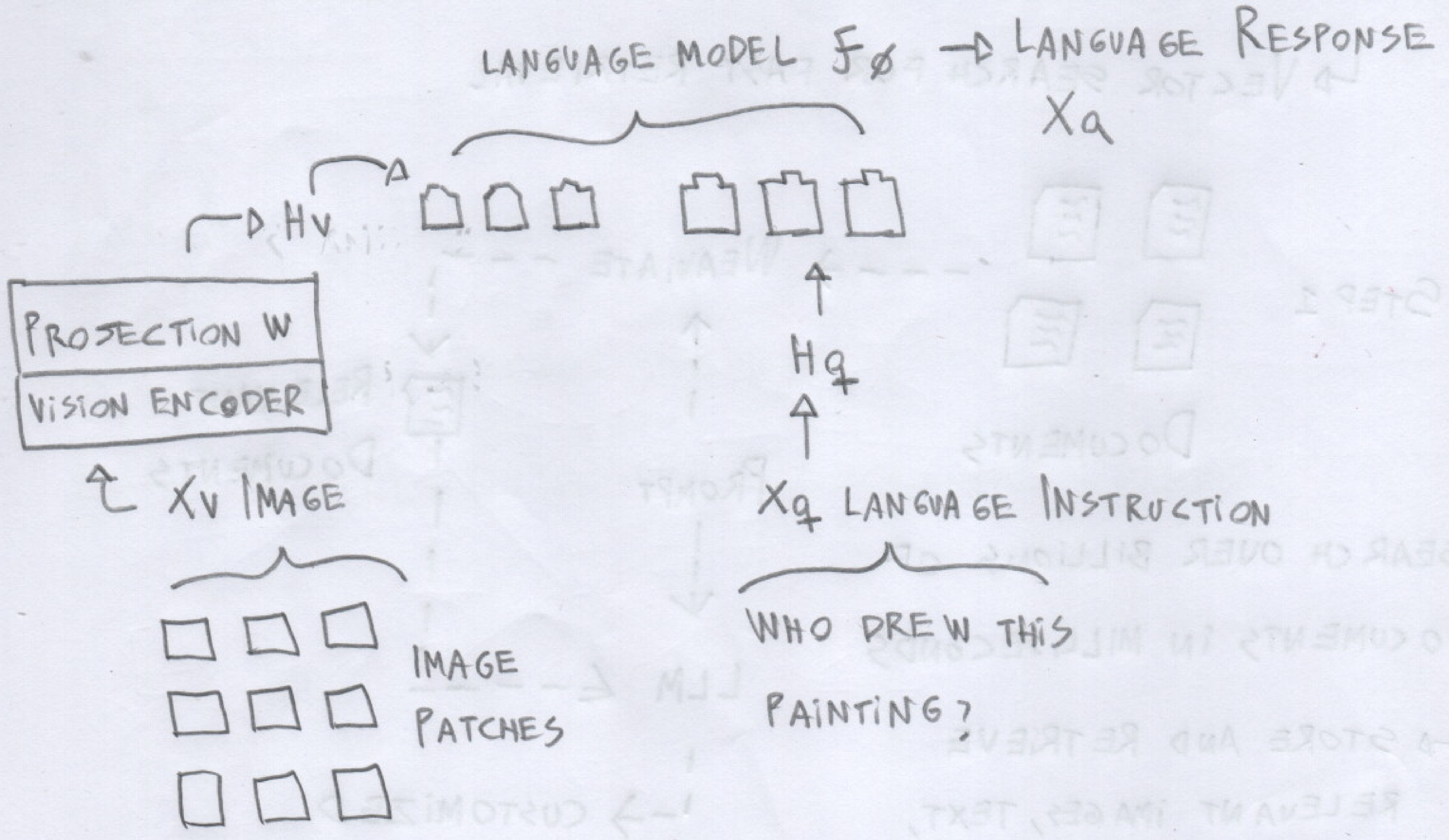
- VISION TRANSFORMERS

  - IMAGES ARE CUT UP INTO PATCHES.

    ↳ USING PATCHES INSTEAD OF PIXELS MAKE IT COMPUTATIONALLY EFFICIENT TO PROCESS IMAGES;

  - EACH PATCH IS EMBEDDED AND PASSED INTO A TRANSFORMER.

  - THE TRANSFORMER OUTPUTS A PROBABILITY DISTRIBUTION OVER THE POSSIBLE CLASSES.



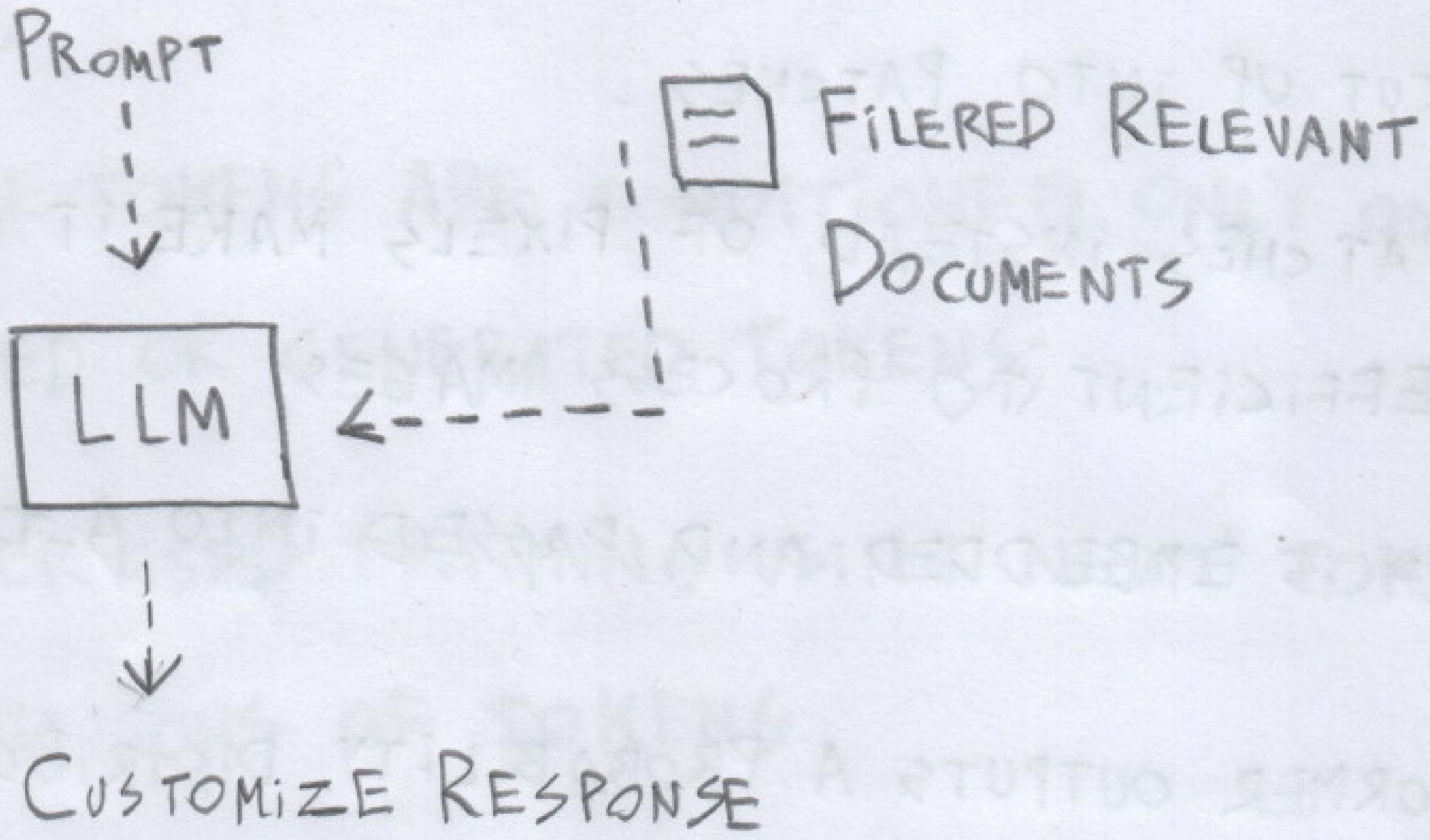
- THE PROBLEM WITH LLMs

"YOU DON'T KNOW WHAT YOU DON'T KNOW"

~SOCRATES

e.g: "WHO AT THE FAMILY PICNIC IS ALLERGIC TO NUTS?"

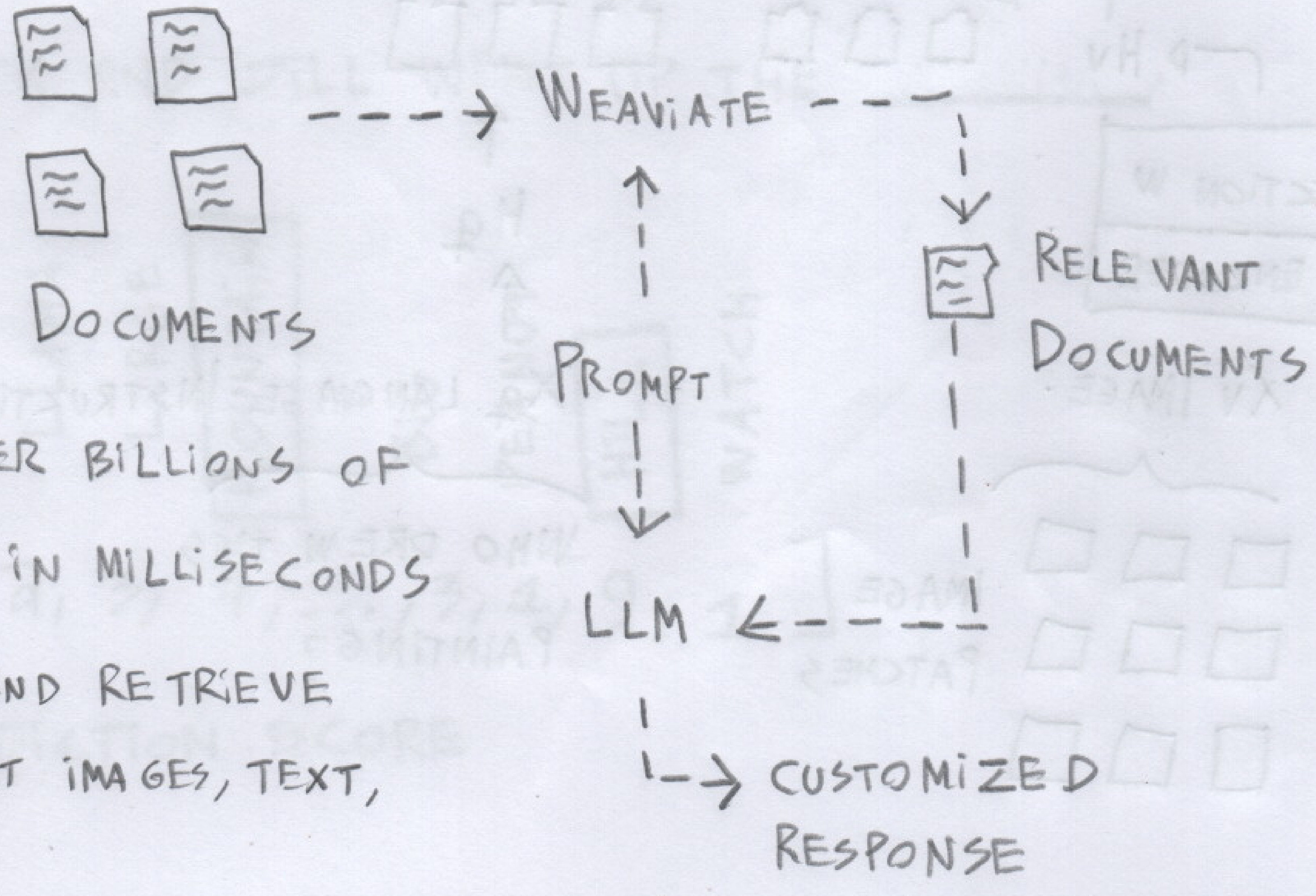
- RETRIEVAL AUGMENTED GENERATION (RAG)



- VECTOR SEARCH FOR FAST RETRIEVAL [WITH WEAVIATE (VECTOR DB)]

↳ VECTOR SEARCH FOR FAST RETRIEVAL

STEP 1



SEARCH OVER BILLIONS OF DOCUMENTS IN MILLISECONDS

↳ STORE AND RETRIEVE RELEVANT IMAGES, TEXT, VIDEO;

- APPLICATIONS OF MULTIMODALITY IN INDUSTRY

[I] STRUCTURED DATA GENERATION

[II] TABLE CREATION

[III] UNDERSTAND LOGIC FLOWCHARTS

- MULTIMODAL RECOMMENDER SYSTEM

SEARCH IS OBJECTIVE AND RECOMMENDATION IS SUBJECTIVE.