AI: The Big Picture

> Mapping AI concepts and mechanisms to human behavior

The Intern

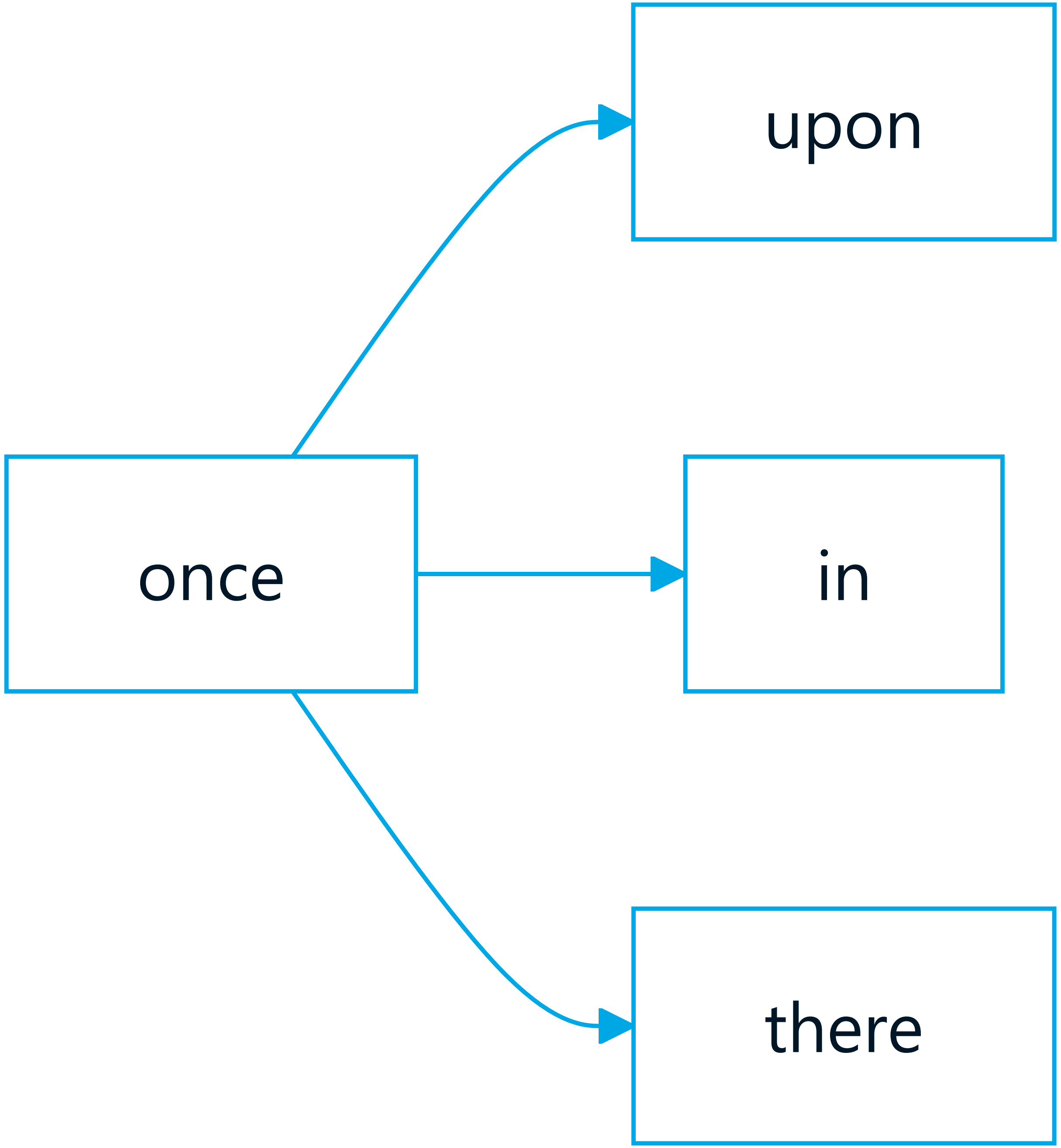
1. Tireless
2. Read **all** the books, knows **all** the words
3. Doesn't **remember** sentences
4. Doesn't inherently **understand.**

Assignment A

> Tell me a story starting with the word `Once` ...

What should the intern do?

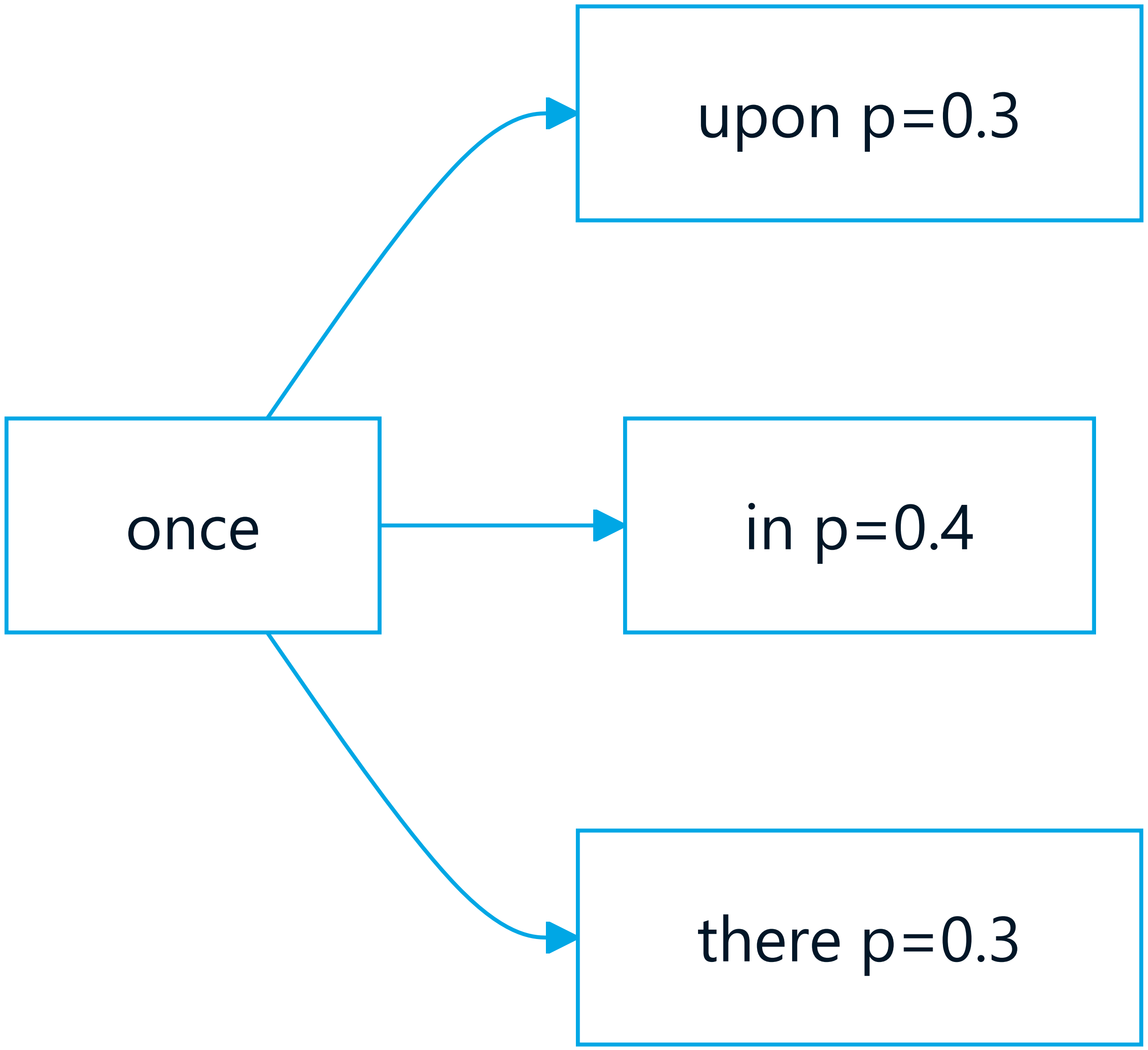
1. Write down the word `once`
2. Write words after that
3. Repeat until done.



Yes, And?

> What is the next word in the sentence?

- Given the word "once"
- What is the next word?

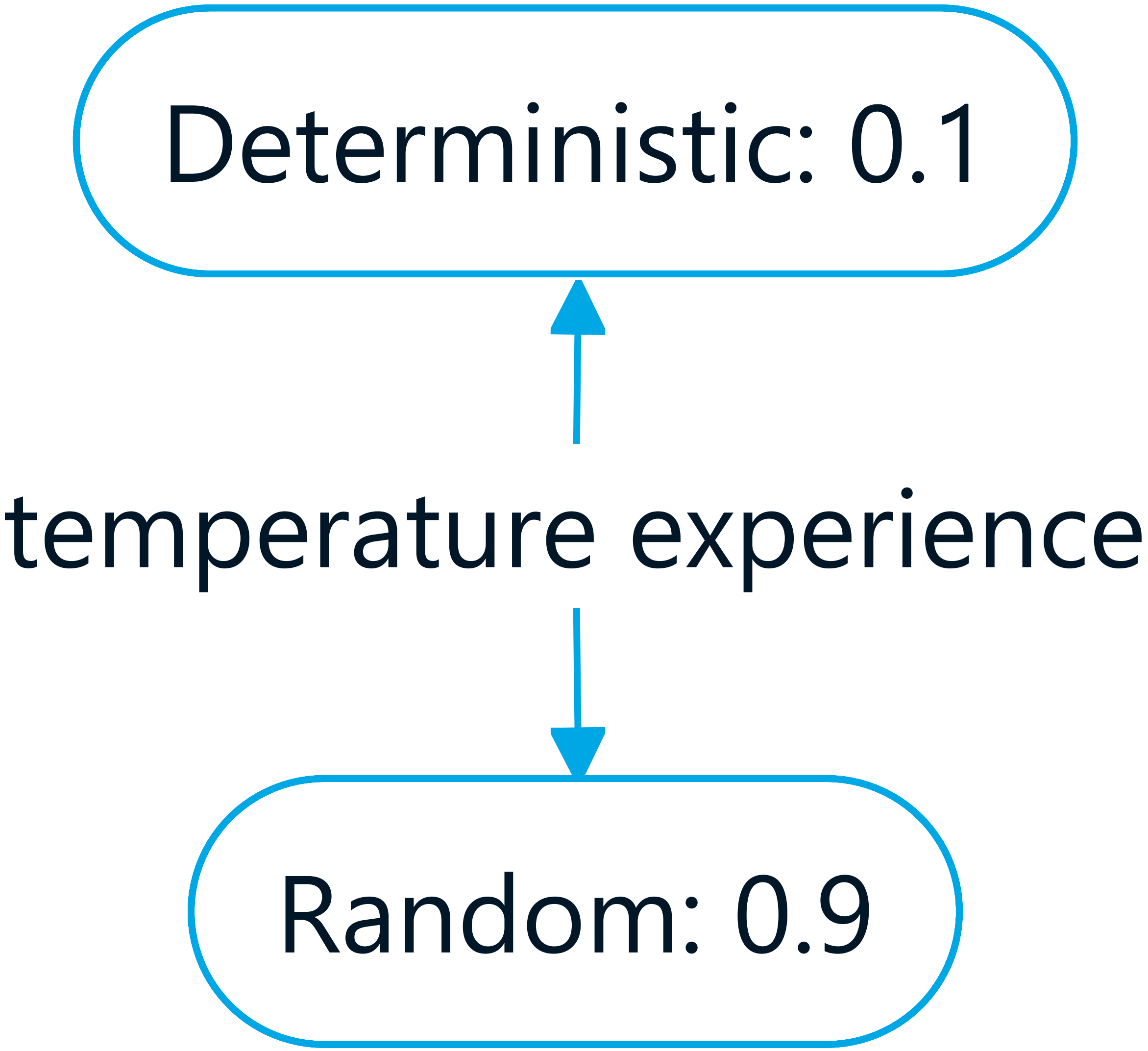


Organization Will Set You Free

Word + probability makes deciding easy (?)

What Would Oscar Wilde do?

Logic	Human Impression	Fit for
Pick #1 (Greedy)	Robotic / Rigid, loop-prone	Coding, Math, Facts
pick from Top *K*	Consistent	General Chat
pick from Top *K* of mass(p)	Dynamic / Human	Creative Writing, Storytelling



Deterministic: 0.1

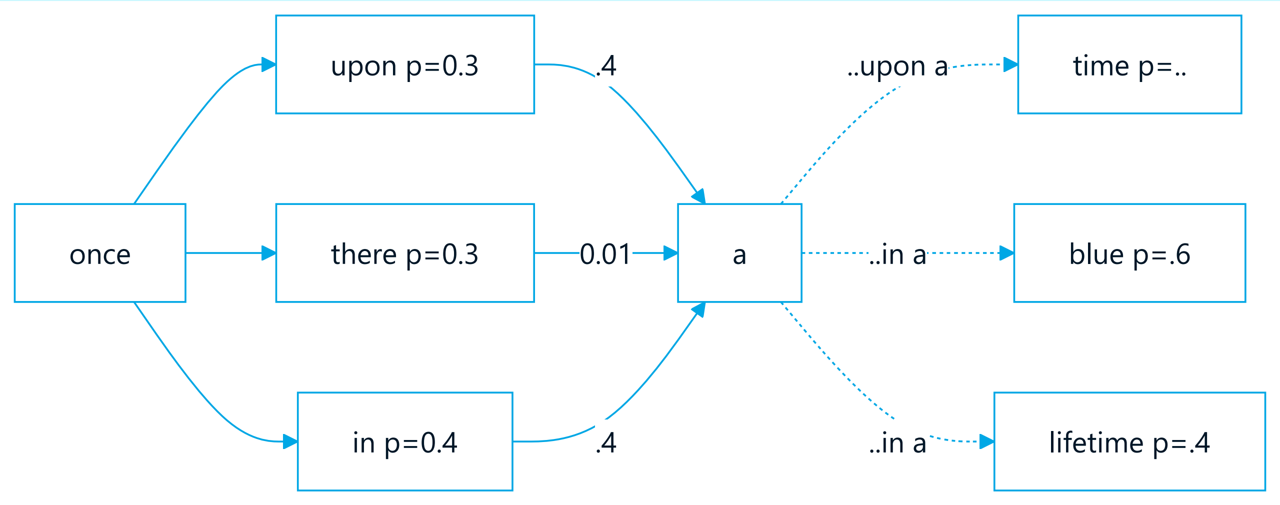
temperature experience

Random: 0.9

What's the Weather Like

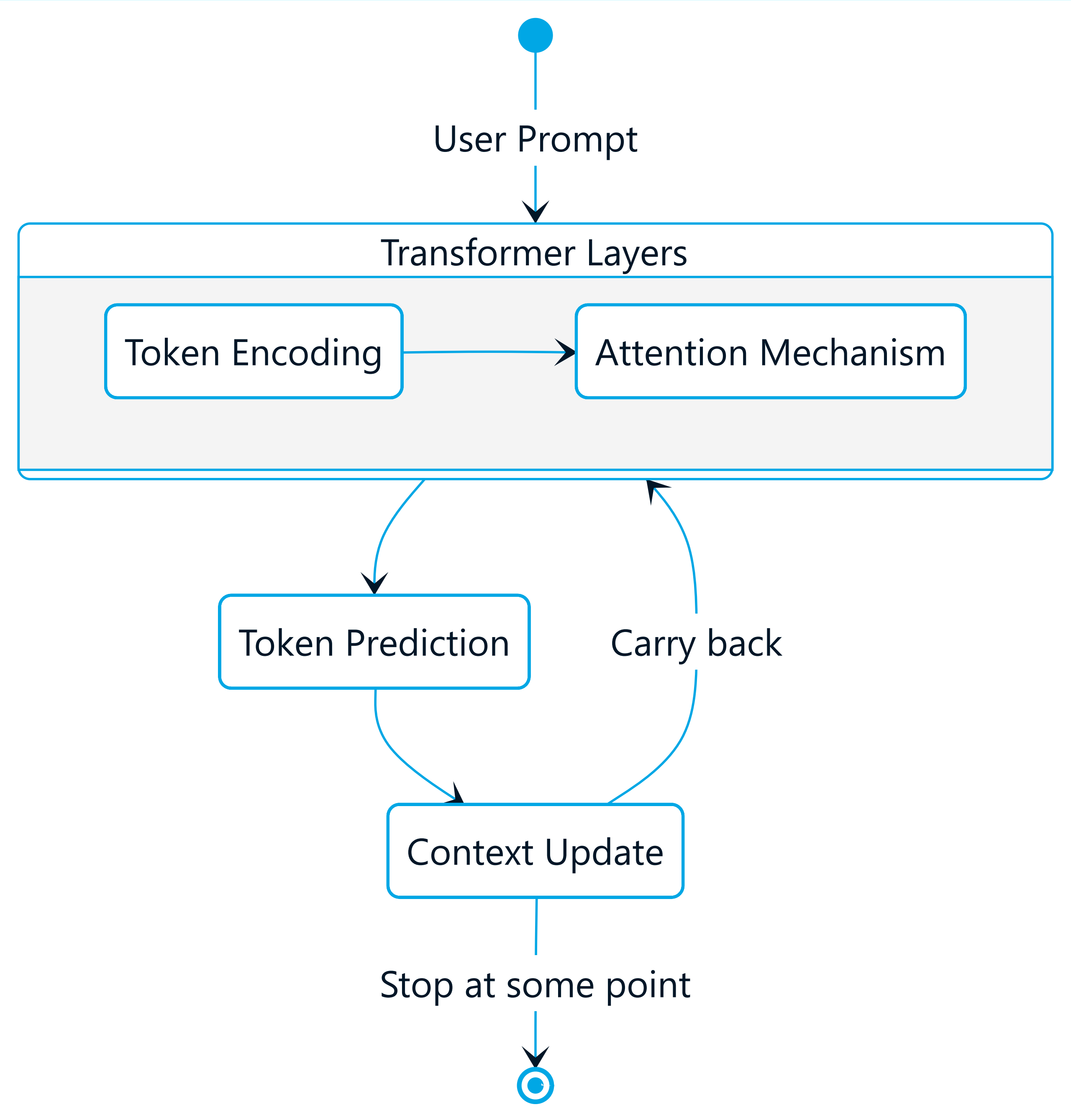
> **Temperature** is a *divisor* on individual token/logit *raw score*

Note: Value LLM version specific.



History Lesson?

> *Just* the last word isn't enough...
> The *next* word wants "context"



Context

- Better prediction
- **Lots** of work

Model - What is it?

1. **Parameter File**
   - Containing (billions) of numerical weights
2. **Transformer Architecture**
   - Organizes how data flows between *layers*
3. **Vocabulary**
   - Enables translating words -> numerical ids and vice-versa.

Transformer Parts

🔦 Attention

> The "Flashlight": Helps find likely/relevant previous tokens
>
> *What was "important" thus far?*

👀 Feed Forward

> The "Supplies": Semantics / Knowledge for the token.
>
> *What do we "know" about it?*

Attention- Intern Focuse on Relevant Parts of Conversation

Model contains immutable **Weight Matrices** for `Q`, `K`, `V`

Conceptual	Actual
Latest prediction added to list	Context is a matrix that gets a new row each token.
List scanned to influence possible next token	Matrix computed with W(q),W(k),W(v), emitting possible values (v)to next layer

Mitigating Token Limits

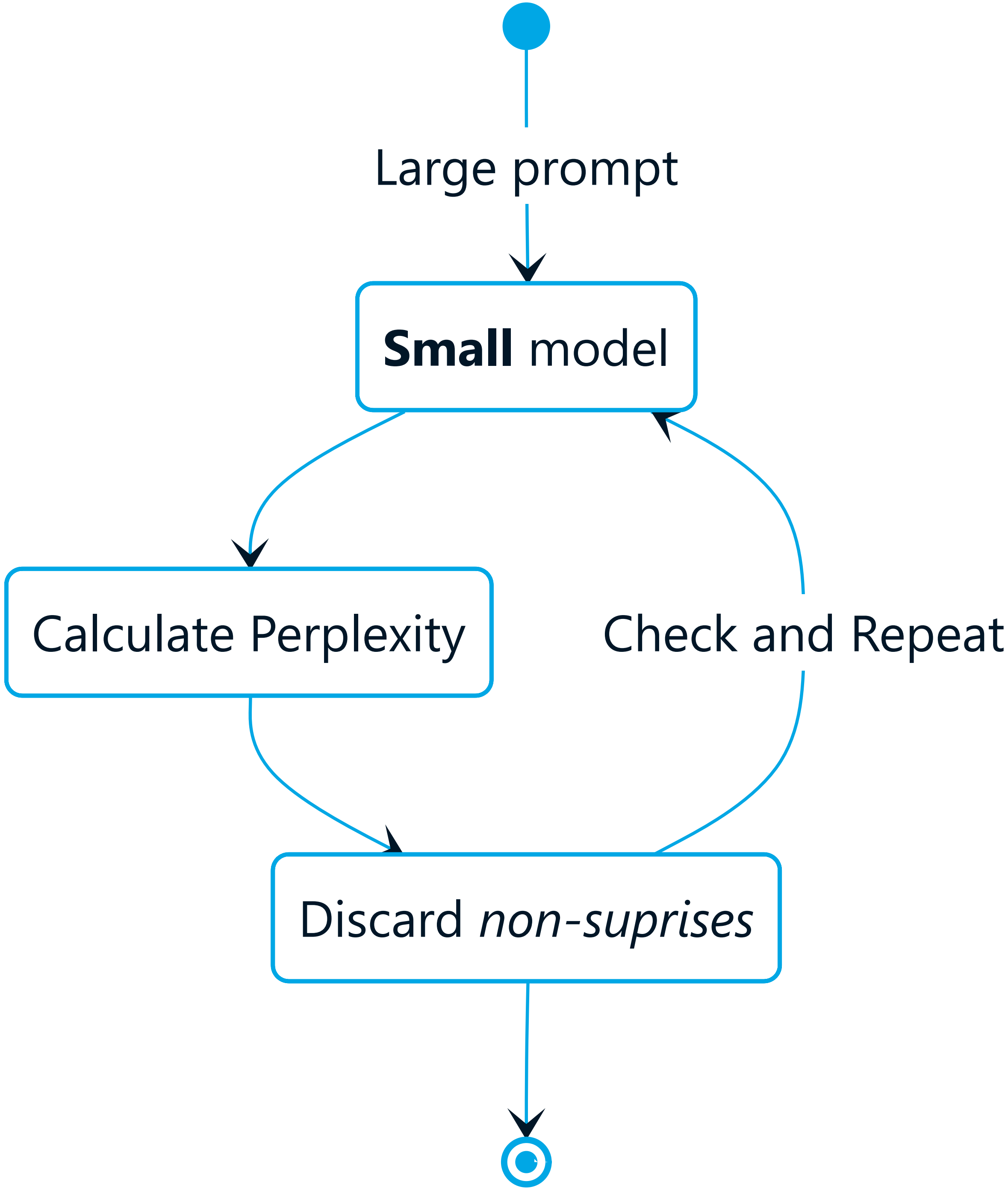
> How can I use less tokens
> But still get good results

Compression Methods: Text

Compression	Primary Benefit	Trade-off
Summarization	Highly readable	Loss of detail/nuance
LLMLingua	Massive token savings	Extra model "pass"

Compression Methods: Latent

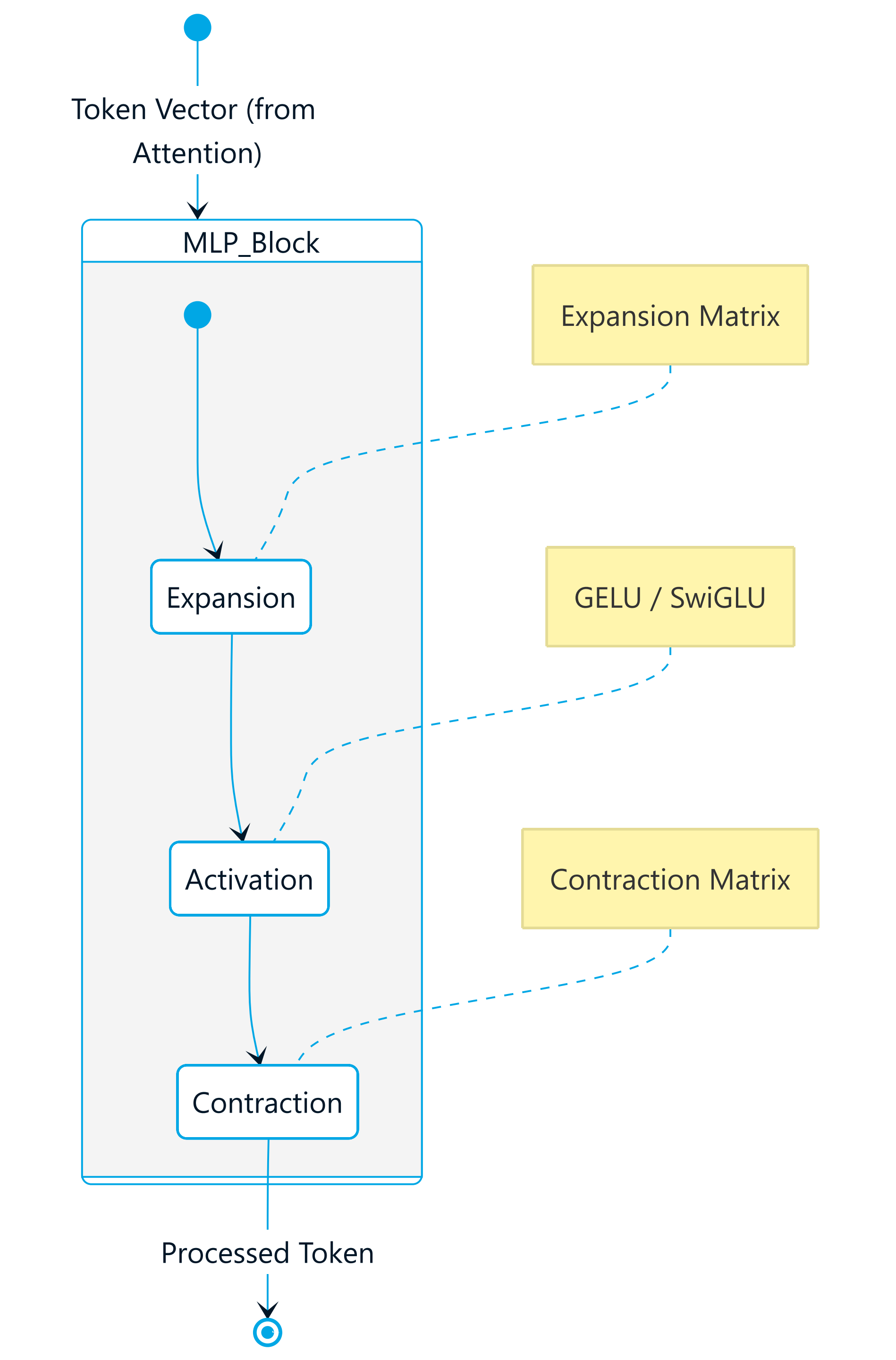
Compression	Primary Benefit	Trade-off
KV Cache Pruning	Memory efficiency	Can appear to forget early facts
Gist Tokens	Extreme compression	Hard to interpret/debug



The Path of Most Surprise

LLMLingua

1. Read prompt
2. Calculate `Perplexity` of each word
3. Discard low-perplexity and re-compute



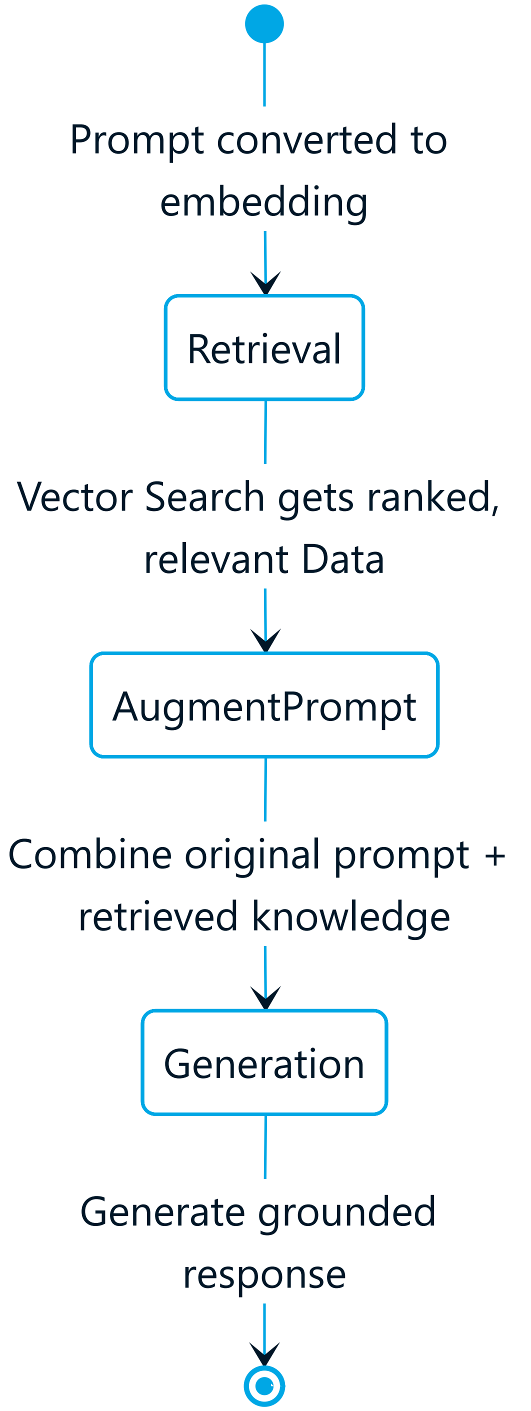
Feed-Forward (MLP)

> "Multi-Layer Perceptron"

You OK, GPT?

Hallucenations mapped to human excuses. The tell-tale signs of coverups:

Excuse	Impression
TL;DR:	Ungrounded, fabricated
It just made sense	Fabricated, no actual basis in training
Oh yeah, forgot...	Instruction loss, ungrounded
I got distracted	Low probability token chase
I confused the terms	Early context mixed with later one


Prompt converted to embedding
Retrieval
Vector Search gets ranked, relevant Data
AugmentPrompt
Combine original prompt + retrieved knowledge
Generation
Generate grounded response

RAG: Give Intern Fresh Relevant Info

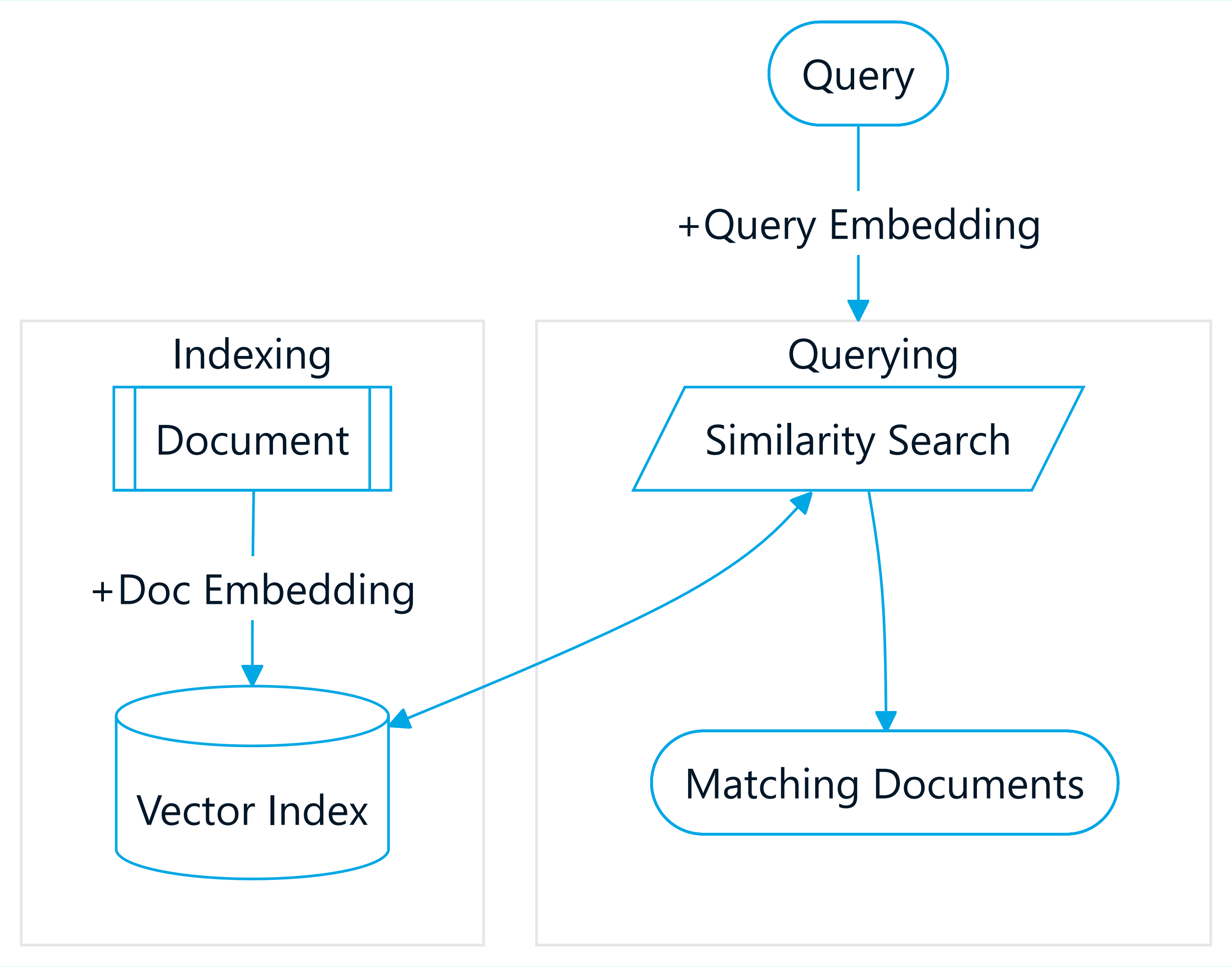
Retreival Augmented Generation overcomes context problems:

1. Intern didn't read **that** book
2. Intern read many books, and it's a blur

Vector Search

An *Embedding model* encodes prompt as *Vector*.

In *Vector Space* **proximity** represents **similarity**.



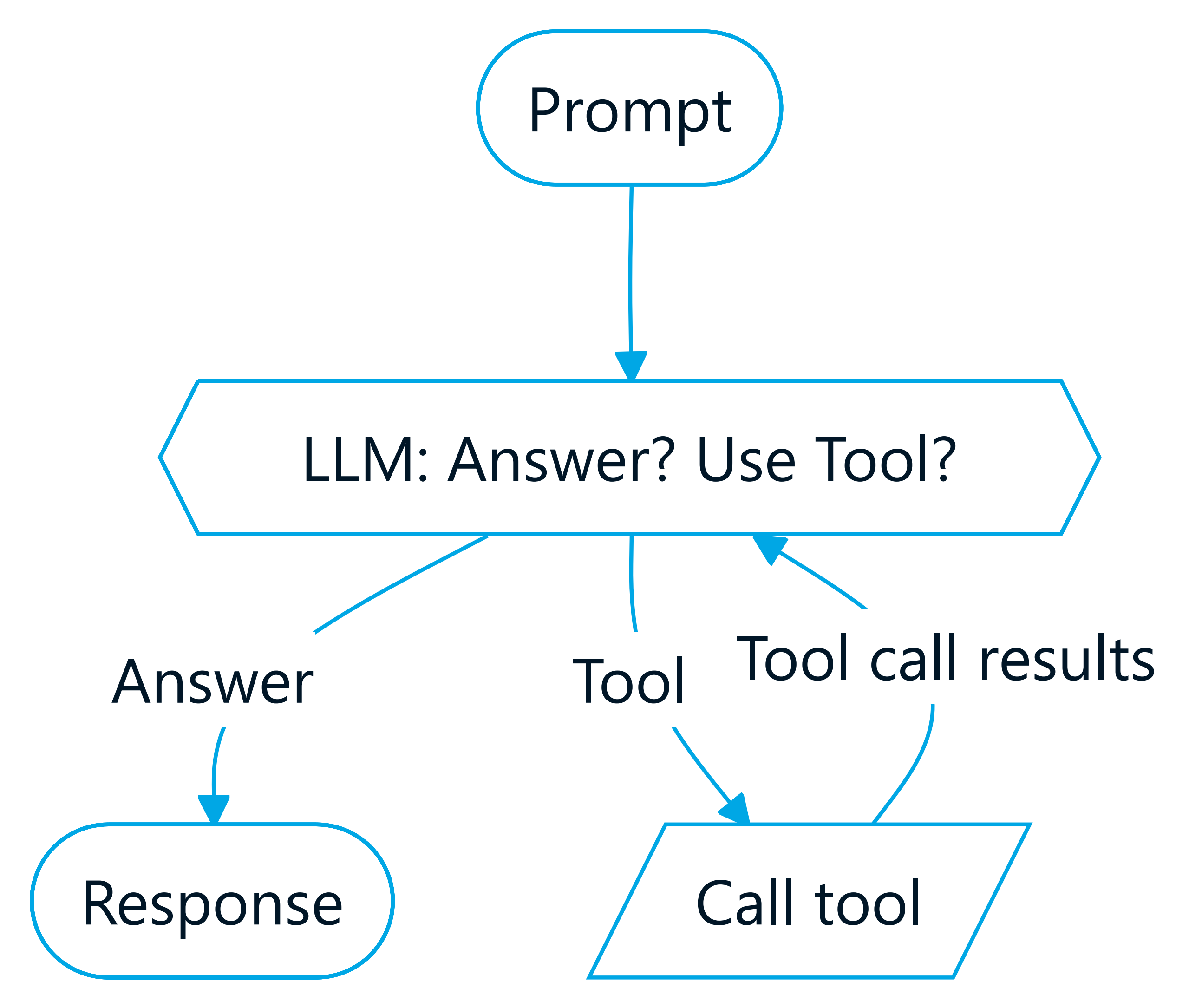
Agents - The Intern as a Task Master

The **Divide & Conquer** approach

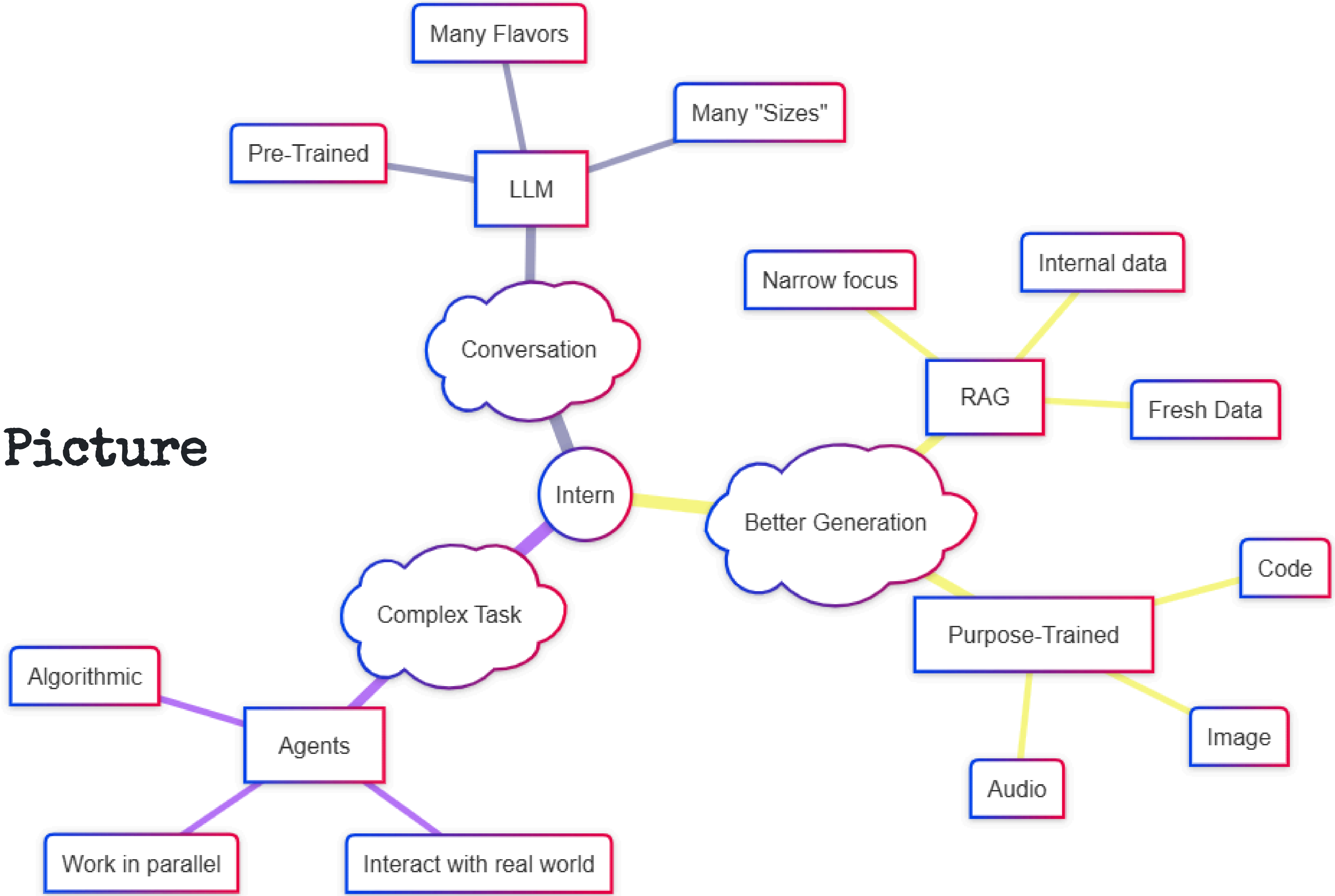
1. LLM as a *supervisor*
2. *Tools/Skills* as workers, offload
   1. Tedium
   2. Algorithmic tasks
   3. Focused or separate extra context

Agentic Internals

An **AI Agent** is an *application* built around a "conversation", prompting in a *loop*.



Big Picture


Many Flavors
Pre-Trained
LLM
Many "Sizes"
Conversation
Intern
Narrow focus
Internal data
RAG
Fresh Data
Better Generation
Complex Task
Code
Purpose-Trained
Algorithmic
Agents
Image
Audio
Work in parallel
Interact with real world

Parting Thoughts

- 🫶Treat your Interns well!
- 🔀AI concepts map to certain human ways of doing things
- 🌱 LLM architecture, math, evolving rapidly

Thank You ❤️