

COMPLEMENTARY THINKING

Language Models,
Ontologies, and
Knowledge Graphs

By Kurt Cagle





KURT CAGLE BIO

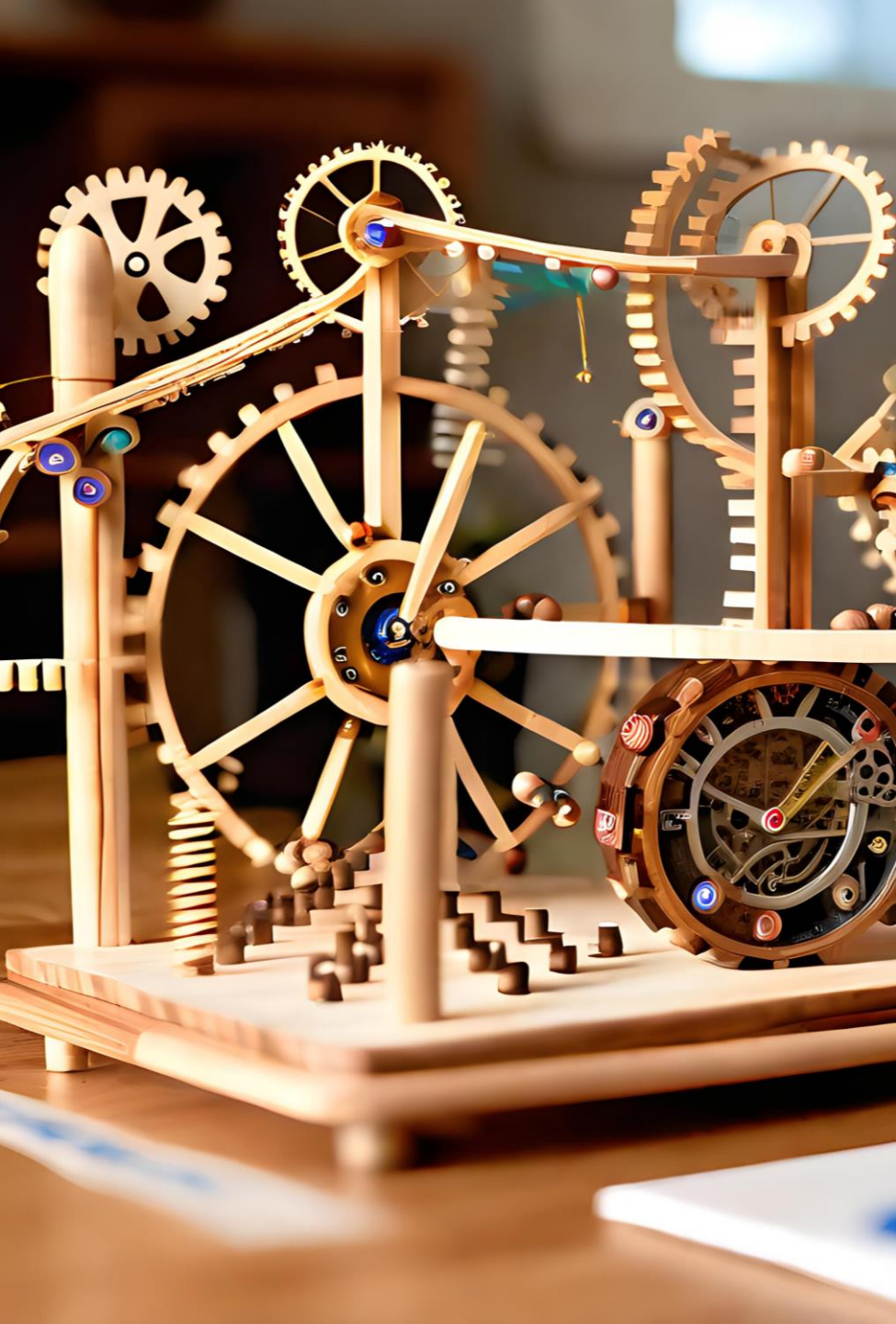
Kurt Cagle is the managing editor of [The Cagle Report](#) and editor for [The Ontologist](#).

He is a writer, speaker, futurist, and system architect who has worked with several Fortune 500 Companies and US, Canadian and EU Agencies over the last thirty years, and has written 25 books. He is currently focusing on knowledge graphs and AI systems, and their underlying ontological dilemma.

He lives in Bellevue, WA and is looking for new work or consulting opportunities. He can be reached at kurt.cagle@gmail.com, or you can set up a free office hour with him through his Calendly account at <https://calendly.com/semantical>.

KNOWLEDGE GRAPH

Ontologies



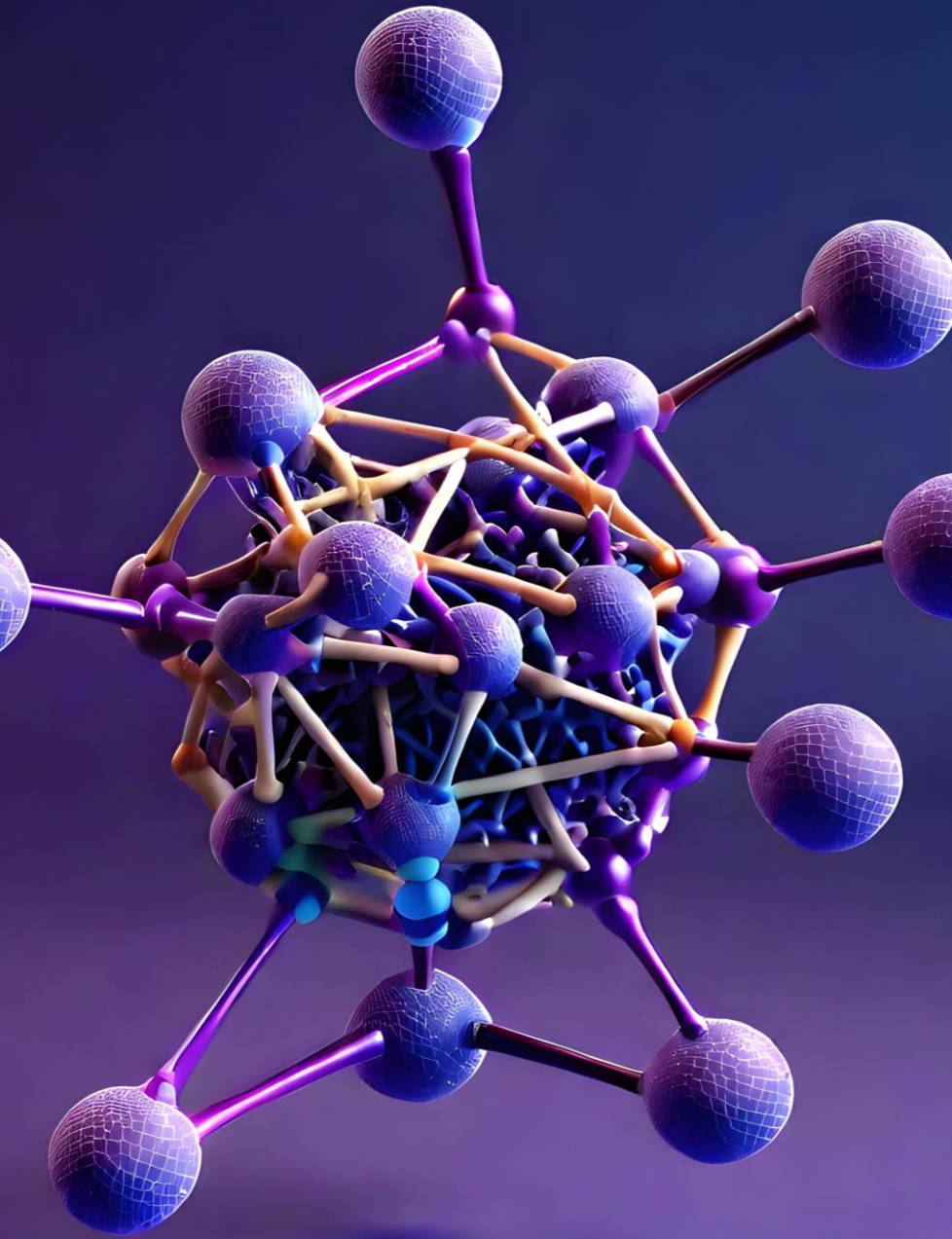
ONTOLOGIES

- A schema that specifies classes, properties and constraints of a particular domain of concern.
- It can also be thought of as the collection of all *data shapes* that can describe objects in a graph.
- Ontologies do not have to be expressed as RDF, but RDF is a language for ontological expression.
- Ontologies can be identified via URIs, and can be composed and nested

ASSERTIONS

- An assertion is a simple statement that is "asserted" or claimed to be valid.
- The subject of the assertion identifies the focus of the statement. (Think row ID)
- The predicate of the assertion indicates a relationship. (Think column name)
- The object of the assertion identifies an object or value. (Think cell value or foreign key reference).





GRAPHS

- A graph consists of a set of assertions, with the object of one assertion (potentially) forming the subject of another assertion.
- Graphs can be disconnected.
- Most data structures can be modeled as graphs.
- Graphs can be named and referenced.
- Use set operations to select and aggregate.
- The structure of graphs are described by shapes or schemas.



KNOWLEDGE GRAPH

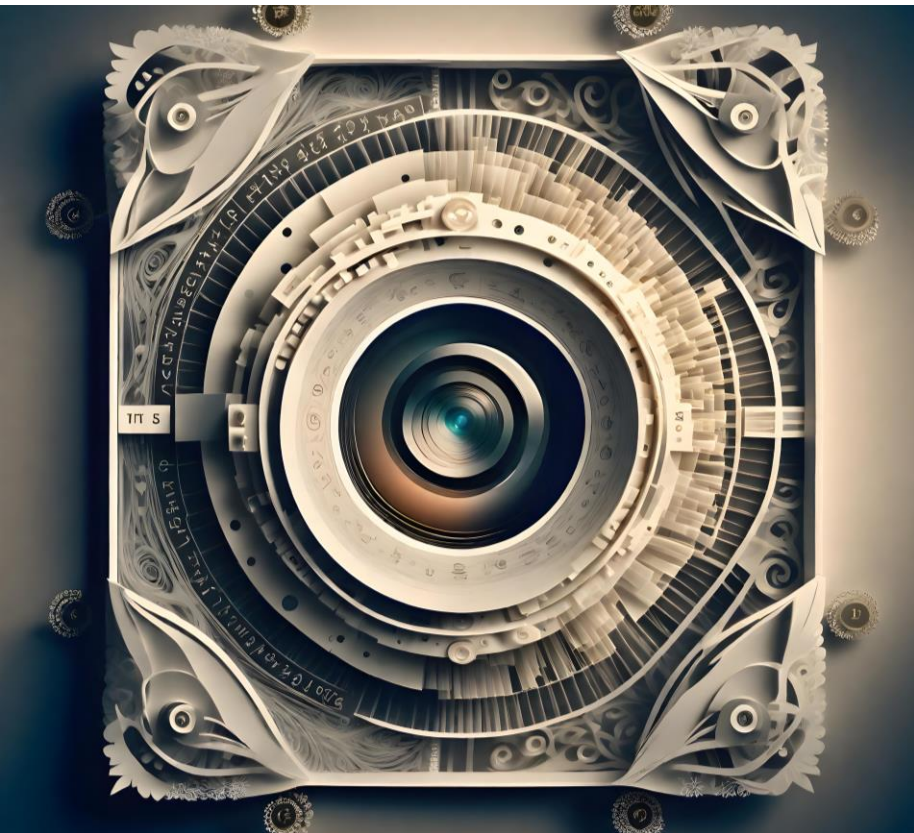
A knowledge graph (KG) is a domain specific collection of graphs that form an information mesh.

KGs may incorporate multiple ontologies, and span multiple knowledge portals.

Knowledge Graph and Knowledge Portal are sometimes used synonymously, though the KG usually refers to the graph available through the portal.

KGs are a form of data mesh.

DATA SHAPES



- A data shape identifies a set of node relationships based on constraints, and is analogous to a complex type in XML.
- Data shapes can be used to validate a given set of nodes as having a given structure.
- Data shapes also contain metadata that determine cardinality and constraints

KNOWLEDGE PORTAL/SERVER

- A knowledge portal or server is a set of related endpoints for knowledge graph servers.
- A knowledge graph server is a database server that is optimized for storing, retrieving and manipulating graphs.
- Knowledge portals need to have the ability to query and manipulated metadata, and most are RDF-adjacent.
- Older KG servers are usually known as triple stores.

Kurt Cagle is the managing editor of The Cagle Report and editor for The Ontologist. He is a writer, speaker, futurist, and system architect who has worked with several Fortune 500 Companies and US, Canadian and EU Agencies over the last thirty years, and has written 25 books. He is currently focusing on



LARGE LANGUAGE MODEL

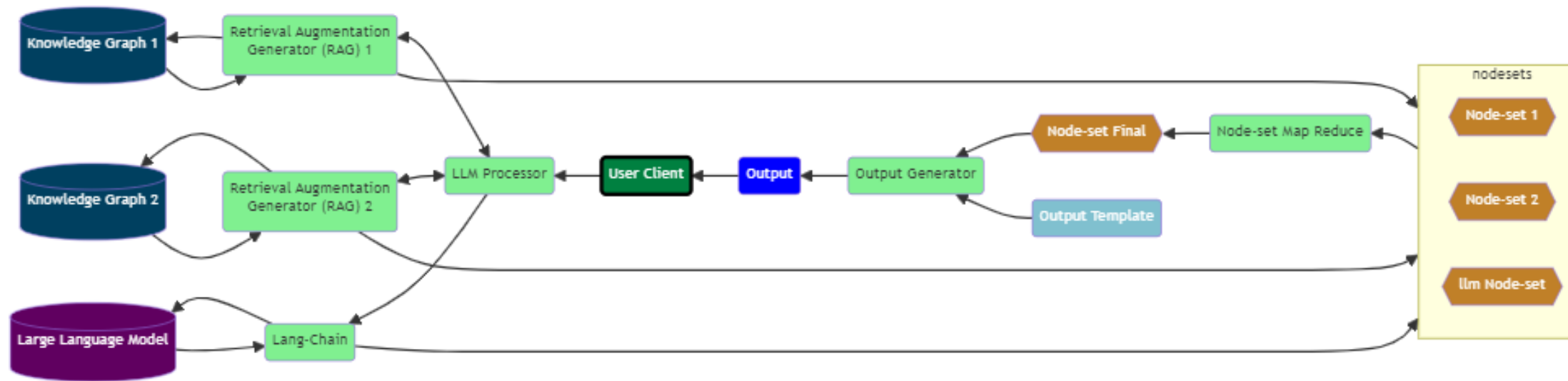
Ontologies

LARGE LANGUAGE MODELS



- Deep learning neural network systems that generate responses to prompts, often abbreviated as LLMs.
- Models are large (often billions of features) and require extensive corpuses
- LLMs tokenize prompts, using these to retrieve sets of concept nodes, then transforms these nodes into resulting text.
- LLMs are considered generative AI.
- LLMs have some reasoning capability, primarily through embedded graphs.

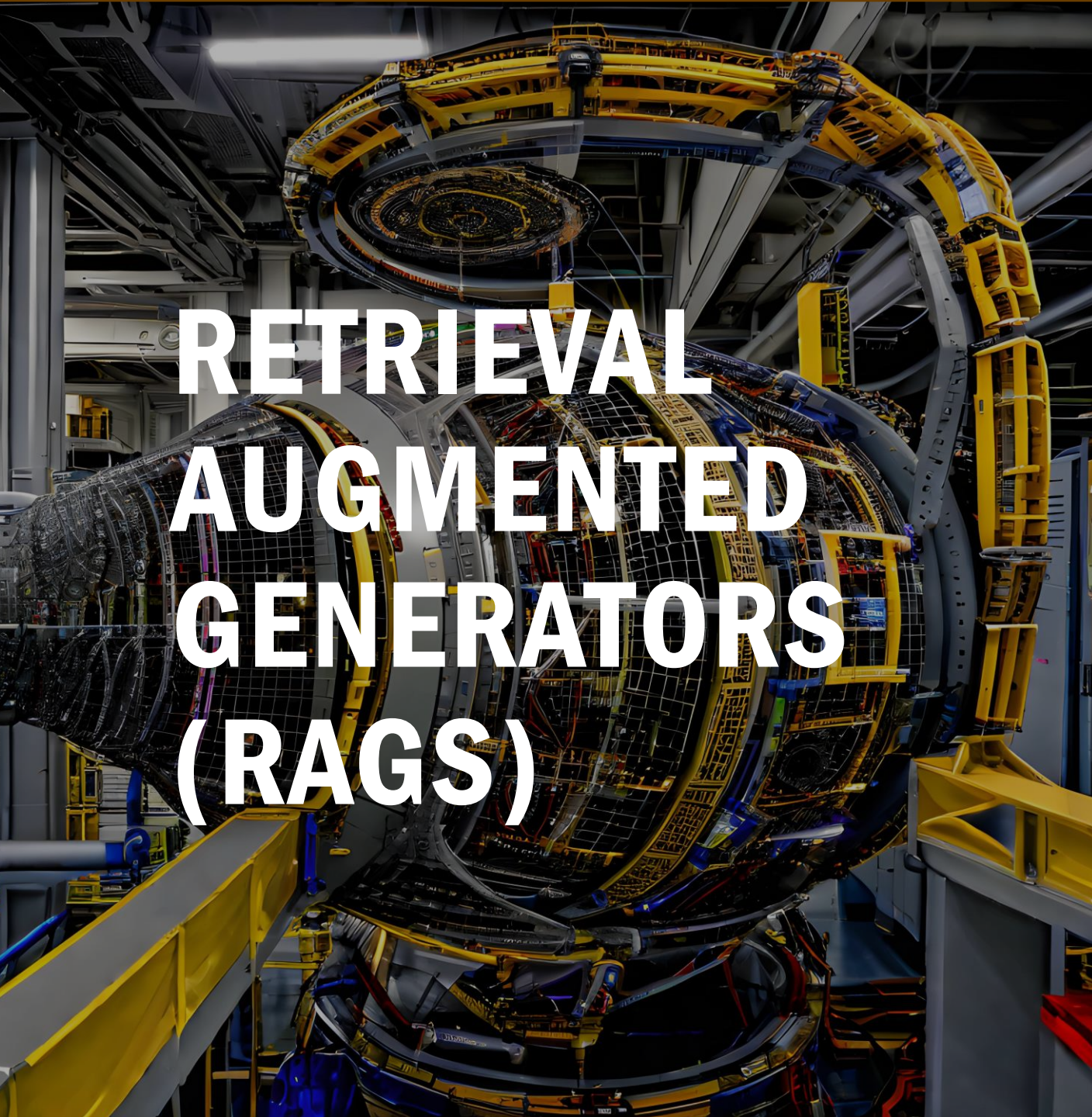
THE LLM PIPELINE



LANG-CHAINS

- Lang-chains are processor modules that sit between the user client and the language model
- A langchain parses the prompt tokens, and uses these to retrieve a response from the LLM as concept node sets.
- Lang-chains may be called recursively, building more detailed information as necessary
- Lang-chains typically are associated with a single LLM source.





RETRIEVAL AUGMENTED GENERATORS (RAGS)

These are similar to lang chains, but are intended primarily to communicate with non-LLM based data sources

They are typically implemented as client plugins, rather than middleware components.

They add to the resulting node-sets (and can change the weight of such sets).

They can also control the output transformations.

It is possible to have a GPT just using RAGs.

BRIDGING THE TWO

Ontologies

ONTOLOGIES (KG VS LLM)

KNOWLEDGE GRAPHS

- Concepts have globally unique identifiers.
- They are schematically defined.
- They are precise.
- They are composable.
- They are explicitly reifiable.
- Relationships have well defined rules and constraints
- Context can be easily switched.

LARGE LANGUAGE MODEL



Concepts are contextually defined.



They lack explicit global identifiers and as such are fuzzy.



No distinction between class and instance.



Reasoning is possible but limited to what was specifically in the corpus.



Relationships are soft and may vary from object to object, even of the same "class".



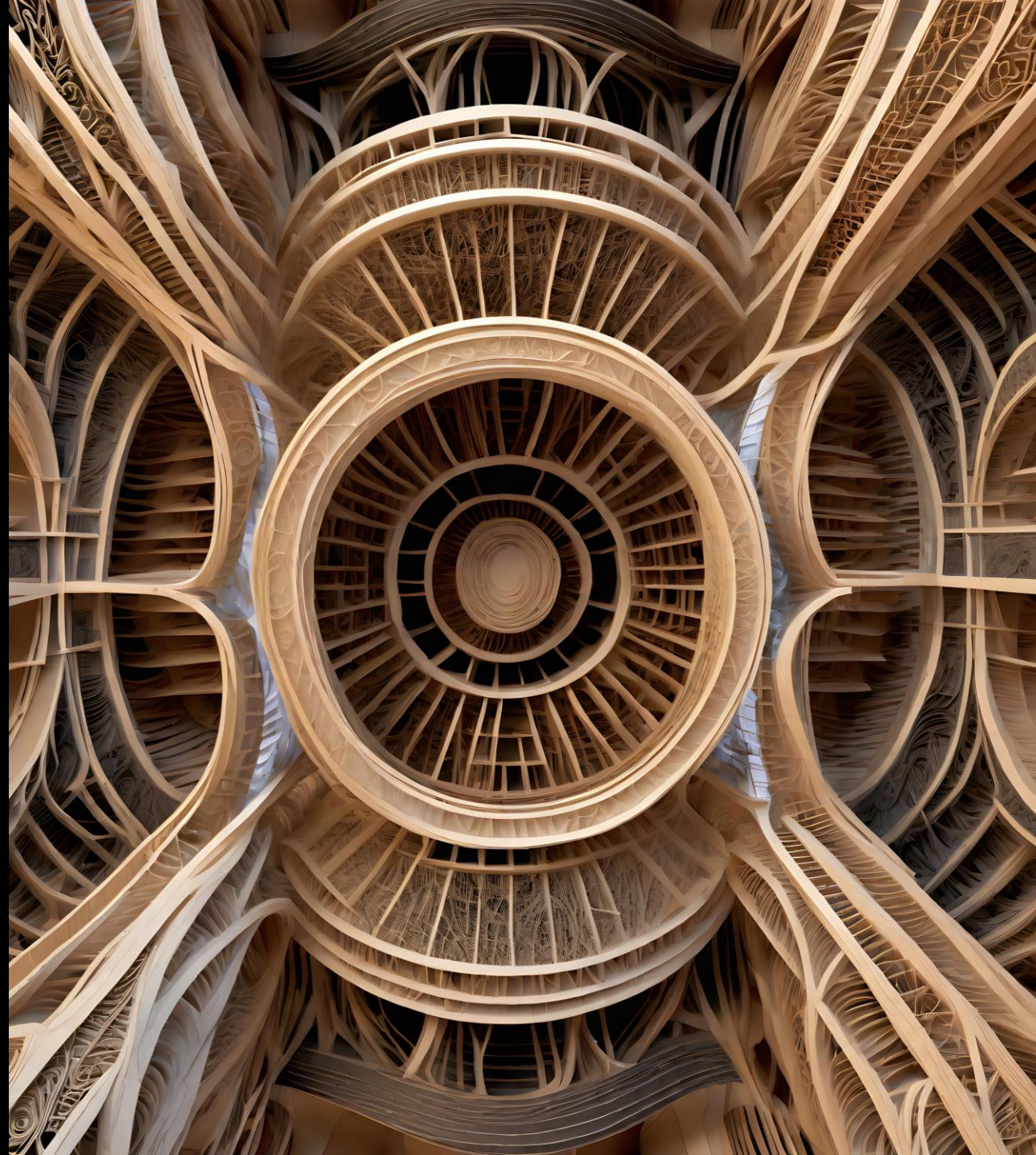
Limited reifiability, though with *this* and similar constructs.



Context is baked in.

MAPPING LLM CONCEPTS TO KNOWLEDGE GRAPHS

- ChatGPT is surprisingly good at matching LLM concepts to knowledge graph instances over specific classes.
- Materialized ontologies, such as Schema.org or NIEM can make a good target. Upper ontologies (such as GIST) are more problematic.
- Defining a schema in OWL or SHACL, then referencing that via a RAG plugin (such as GPT4's Link Reader) can also map data to that schema.
- ChatGPT in particular can read and understand Turtle, RDF/XML and JSON-LD.





MAPPING KG CONCEPTS TO LLMS

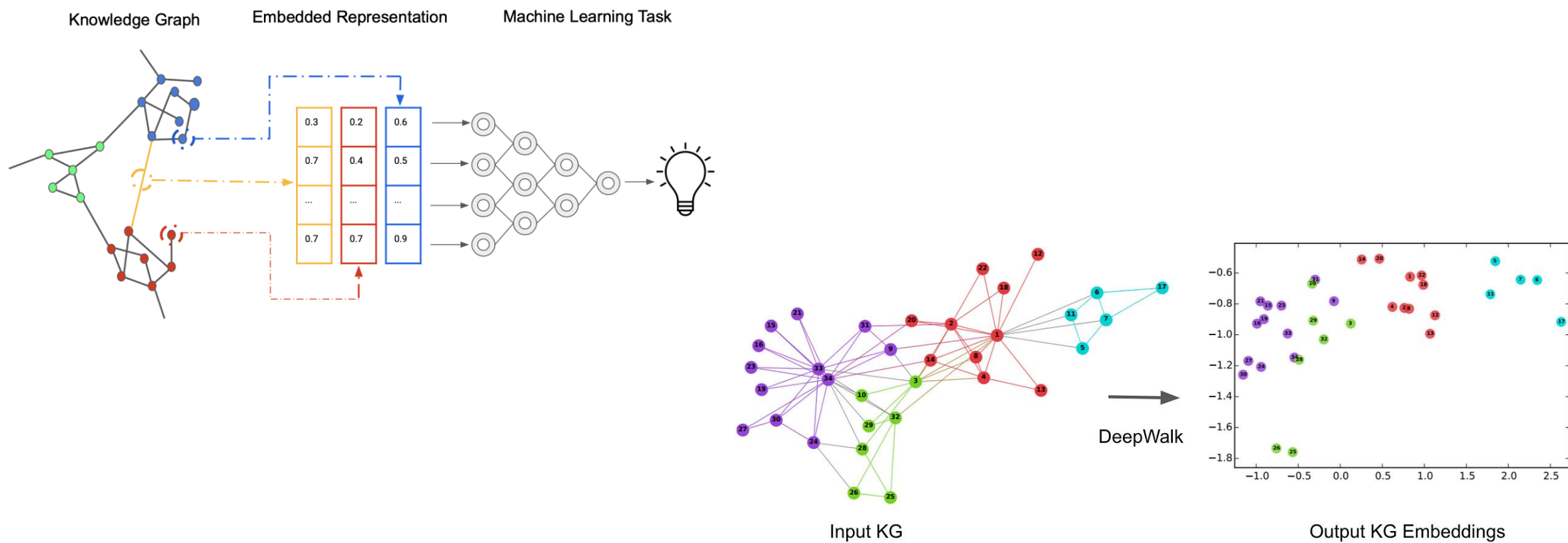
- LLMs use a {source,question,answer} format for node assertions.
- When generating answers, include node URIs as properties for resources and classes.
 - (Jane Doe has URI identifier <https://www.example.com/ns/Person/JaneDoe>)
- When documents are being selected for the corpus, where possible incorporate RDFa or similar microformats.

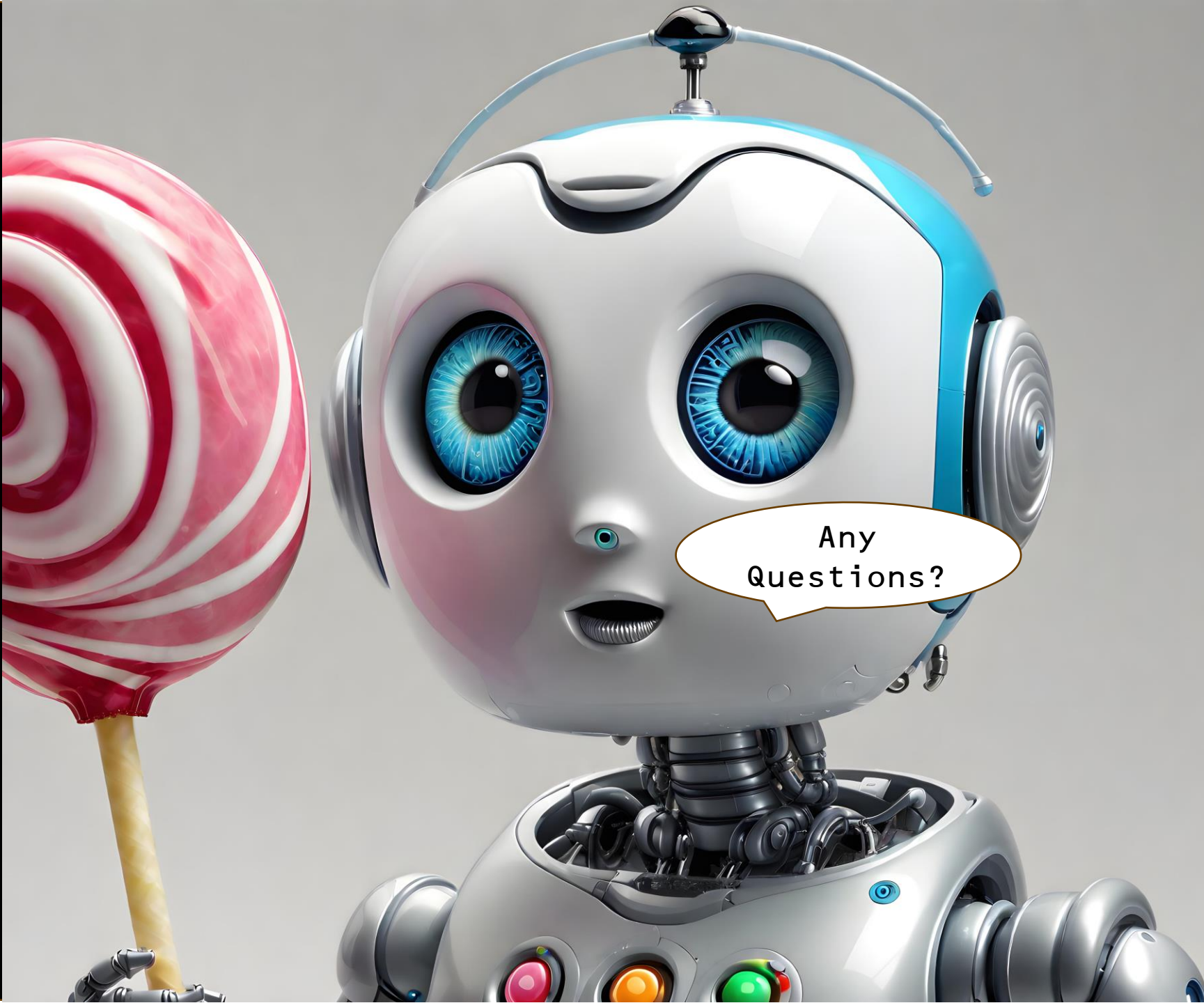
KNOWLEDGE GRAPH EMBEDDINGS

- Knowledge graphs can also be converted into vector space representations as *knowledge graph embeddings (KGE)*.
- These are specialized documents added to the corpus that incorporate KG logic directly into the LLM itself.
- Unlike RAGs, KGEs are part of the LLM and greatly enhance the ability to reason on that LLM through the LangChains.
- KGEs are incorporated when the model is compiled, which provides performance at the cost of novelty.



UNDERSTANDING KGES





Any
Questions?