Triplex — SOTA LLM for Knowledge Graph Construction

An open-source model for 10x cheaper knowledge graph construction

Owen C., Nolan T., Shreyas P. / 07/09/2024

4 min read

Triplex is an innovative new model that allows you to convert large amounts of unstructured data into structured data. Triplex exceeds the performance of gpt-4o at knowledge graph construction for less than one-tenth the cost.

Triplex is open source and available on **HuggingFace here** and **ollama here**.
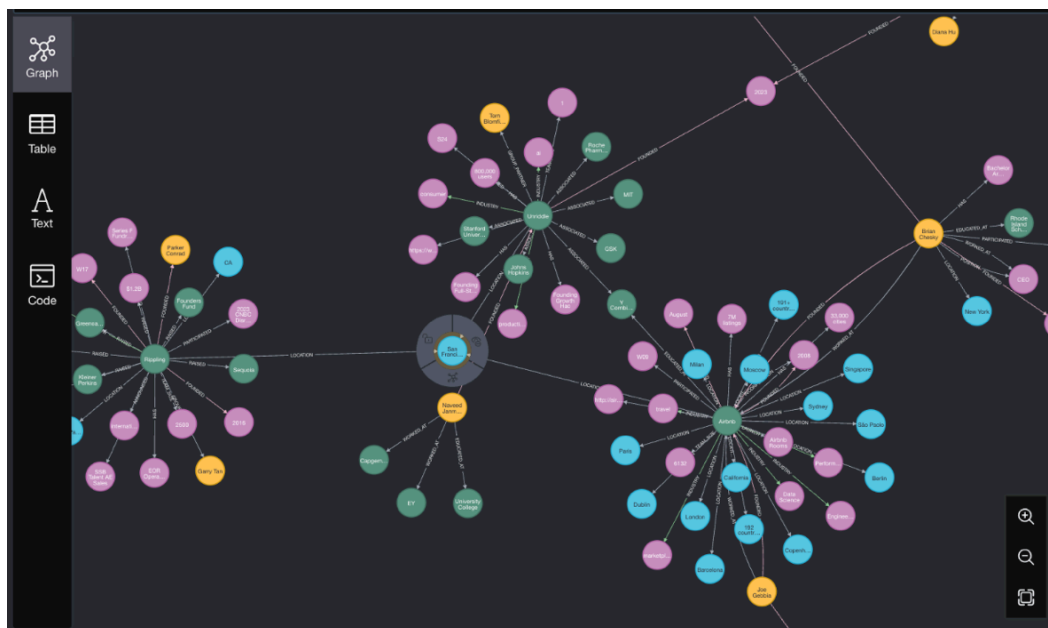

Figure 1: Illustration of a knowledge graph structure

Knowledge graphs excel at answering queries that traditional search methods often struggle with, particularly population-level relational queries. For instance, "Provide a list of AI employees who attended technology schools." Interest in knowledge graphs has intensified after Microsoft's recent GraphRAG paper [1].

However, the process of knowledge graph construction has traditionally been complex and resource-intensive [2], limiting its widespread adoption. Recent

Ask AI

estimates suggest that Microsoft's GraphRAG procedure is particularly costly, requiring at least one generated output token for every ingested input token. This high cost makes it impractical for most applications.

Triplex aims to radically disrupt this paradigm by reducing the generation cost of knowledge graphs tenfold. This cost reduction is achieved through Triplex's efficient method of converting unstructured text into "semantic triples" - the building blocks of knowledge graphs. To demonstrate how Triplex creates these graphs, let's observe how it processes simple sentences:

```
# Inputs
# Entity Types <- CITY, COUNTRY
# Relationships <- CAPITAL_OF, LOCATED_IN
# Text <- Paris is the capital of France

# Output -> (subject > predicate > object)
CITY: Paris > CAPITAL_OF > COUNTRY: France
CITY: Paris > LOCATED_IN > COUNTRY: France
```
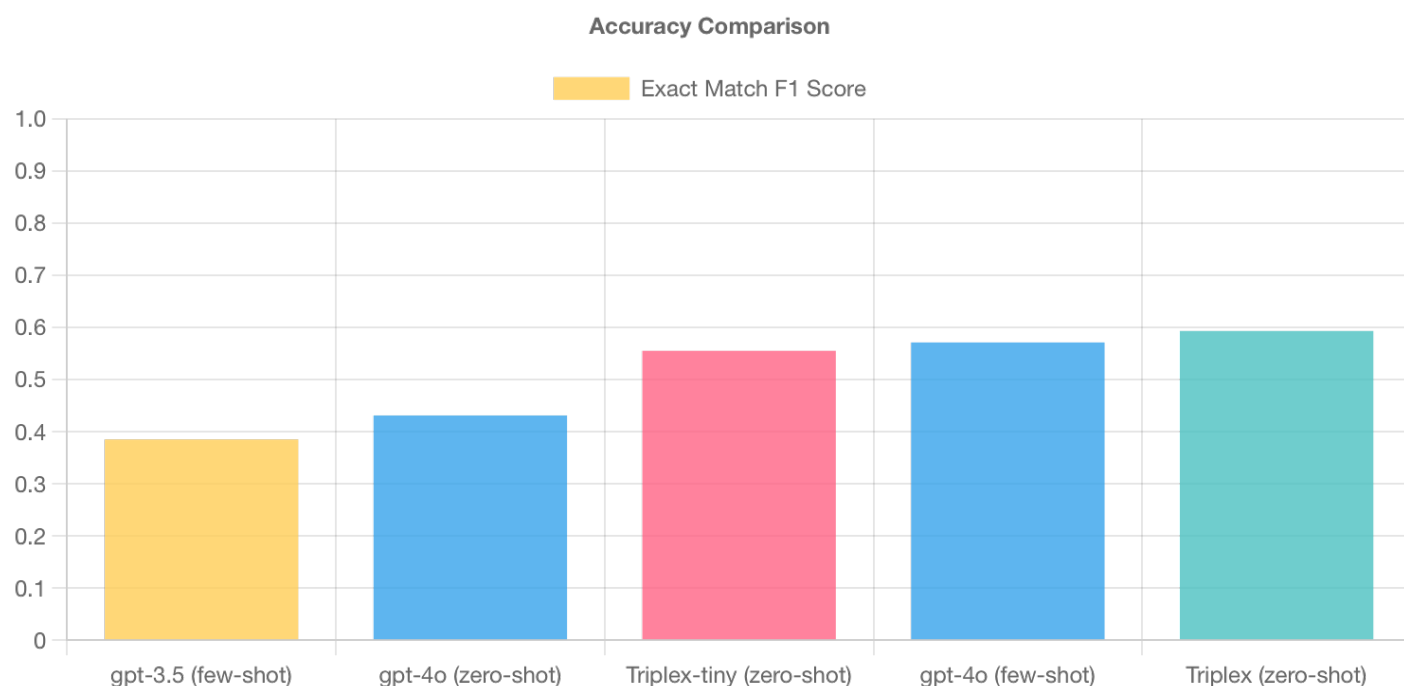
And on a more complex input:

```
# Inputs
# Entity Types <- ARTIST, ARTWORK, ART_MOVEMENT
# Relationships <- CREATED_BY, BELONGS_TO_MOVEMENT
# Text <- Vincent van Gogh, a post-impressionist painter, created "The Starry Night" i

# Output -> (subject > predicate > object)
ARTIST:Vincent van Gogh > BELONGS_TO_MOVEMENT > ART_MOVEMENT:post-impressionist
ARTWORK:The Starry Night > CREATED_BY > ARTIST:Vincent van Gogh
ARTIST:Vincent van Gogh > BELONGS_TO_MOVEMENT > ART_MOVEMENT:Expressionism
ARTIST:Vincent van Gogh > BELONGS_TO_MOVEMENT > ART_MOVEMENT:Fauvism
```
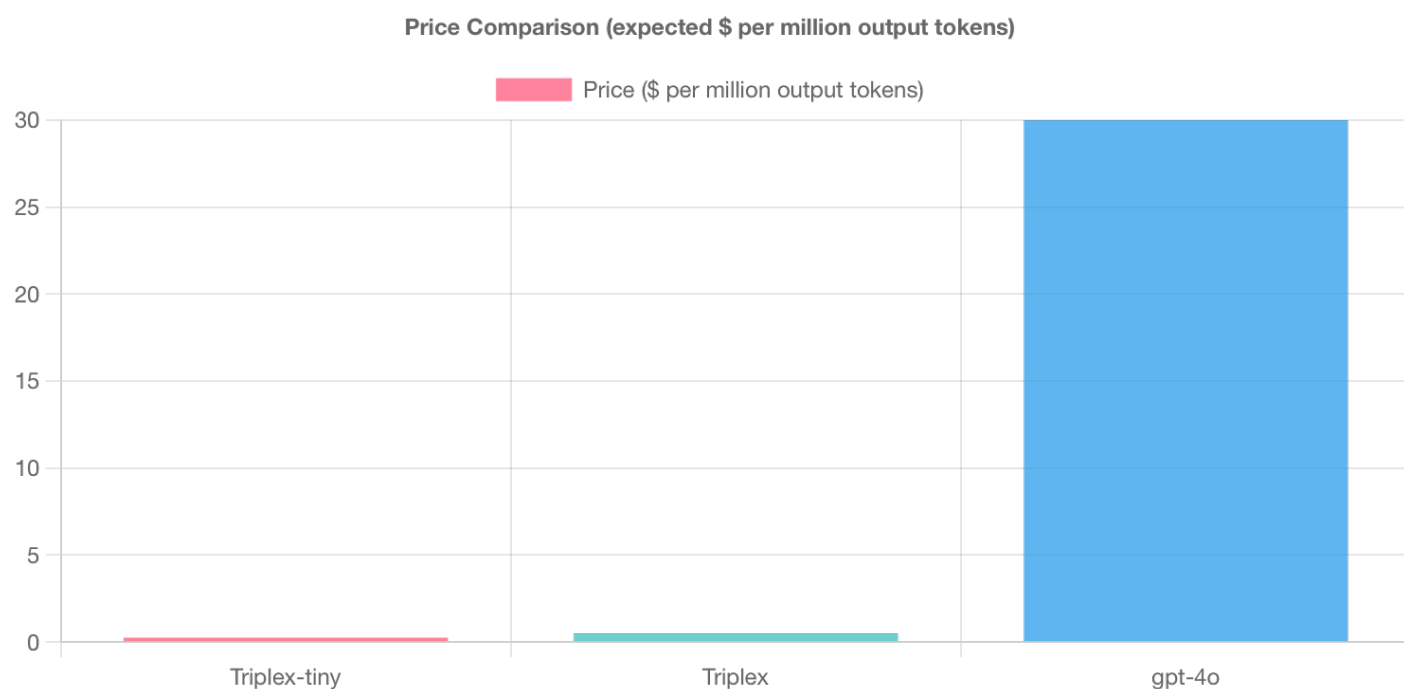
Our performance measurements revealed that Triplex significantly outperformed gpt-4o in cost and performance.

Accuracy Comparison

Accuracy Comparison

Exact Match F1 Score

| | gpt-3.5 (few-shot) | gpt-4o (zero-shot) | Triplex-tiny (zero-shot) | gpt-4o (few-shot) | Triplex (zero-shot) |

Price Comparison

Price Comparison (expected $ per million output tokens)

Price ($ per million output tokens)

The triple extraction model achieves results comparable to GPT-4, but at a fraction of the cost. This significant cost reduction is made possible by Triplex's smaller model size and its ability to operate without the need for few-shot context.

Building upon the SFT model, we generated additional preference-based dataset using majority voting and topological sorting to further train Triplex using DPO and KTO. These additional training steps yielded substantial improvements in

model performance. To accurately assess these nuanced enhancements, we conducted a rigorous evaluation using Claude-3.5 Sonnet. Our evaluation involved head-to-head comparisons between three models: triplex-base, triplex-kto, and GPT-4o. The results are presented in the table below:

Model 1	Model 2	Model 1 Win	Model 2 Win	Tie
triplex-base	gpt-4o	54%	43%	3%
triplex-kto	triplex-base	66%	26%	8%
triplex-kto	gpt-4o	56%	40%	4%

The exceptional performance of Triplex results from extensive training on a diverse and comprehensive dataset. Our model leverages proprietary datasets generated from authoritative sources such as DBPedia and Wikidata, as well as web-based text sources and synthetically generated datasets. This broad foundation ensures Triplex's versatility and robustness across a wide range of applications.

Usage

We have designed the **R2R RAG engine**, together with **Neo4J**, to immediately leverage Triplex for local knowledge graph construction, a use case that has only now become more viable due to this work. Please read more about how to get started in the **documentation here**, or try out **triplex directly here**.

For additional information or to explore how Triplex can transform your data insights, please reach out to us at **founders@sciphi.ai**. Join us in shaping the future of knowledge representation and extraction!

References

[1] Zhong, L., Wu, J., Li, Q., Peng, H. and Wu, X., 2023. A comprehensive survey on automatic knowledge graph construction. ACM Computing Surveys, 56(4), pp.1-62. [Link]

[2] Edge, Darren, et al. "From local to global: A graph rag approach to query-focused summarization." arXiv preprint arXiv:2404.16130 (2024). [Link]

Was this helpful?

😃   😐   😞