

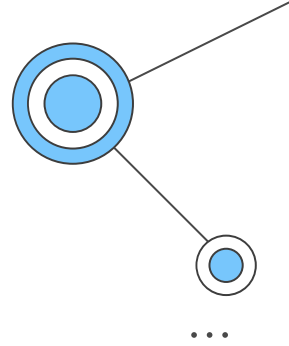
AskWiki

Capstone Spring 2023
Annie Cui, John Stilb, Max
Ziff, & Shrinivas Joshi



Value Proposition and Vision

Retrieving information from large datasets has never felt more natural!



Eliminate time spent searching for the needle in a haystack



Lower the learning curve for querying languages



Assist analytical roles that require large amounts of info gathering

Our Vision

Is to Enable enterprises to capitalize on the bigger value proposition of using large language models to **motivate a safe and secure knowledge-driven question-answering eco system which is free from training data constraints and is accessible to all** enterprises irrespective of their size, strength and technical prowess.

Real-Life Application

Question

What was the average wait period for emergency patient check-ins this week?



Answer

The average wait period for emergency patient check-ins this week was 3.45 hours

Market and Industry

Enterprise Knowledge Management Software
Market size growing at a
CAGR of 12.5% from 2022 to 2030.



2023

April 20th

April 13th

April 12th

March 28th

March 14th



Amazon Bedrock





Data



Data: Natural Language to SPARQL



LC-QuAD 2.0 Dataset:
30k pairs of questions and
corresponding sparql queries

Question (Features)

How many children did J.S. Bach
have?

SPARQL (Labels)

```
SELECT (COUNT(?children) as  
?count) WHERE { wd:Q1339  
wdt:P40 ?children . }
```



Data: Summarization



WebNLG 2020 Dataset: 13,211 triples with 35,426 target texts across 372 properties

Triples: Subject | Property | Object (Features)

- Abner_W._Sibal | battle | World_War_II
- World_War_II | commander | Chiang_Kai-shek
- Abner_W._Sibal | militaryBranch | United_States_Army

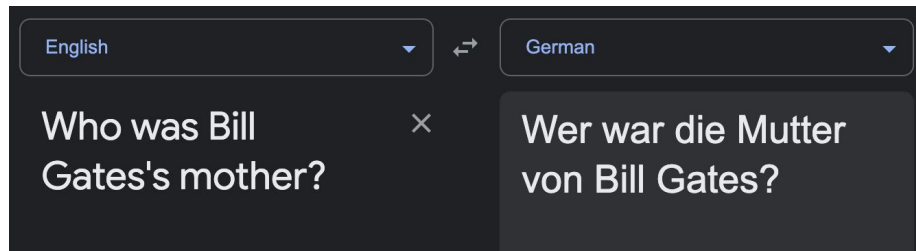
Target Texts (Labels)

- World War II had Chiang Kai-shek as a commander and United States Army soldier Abner W. Sibal.
- Abner W. Sibal served in the United States Army during the Second World War and during that war Chiang Kai-shek was one of the commanders.
- Abner W. Sibal, served in the United States Army and fought in World War II, one of the commanders of which, was Chiang Kai-shek.

NL to SPARQL



Experimentation



T5 is very good at NL -> NL translation

But Sparql is a formal language
A misplaced curly brace is a complete fail!
T5 translations almost all failed.



Fine-tuning GPT-3

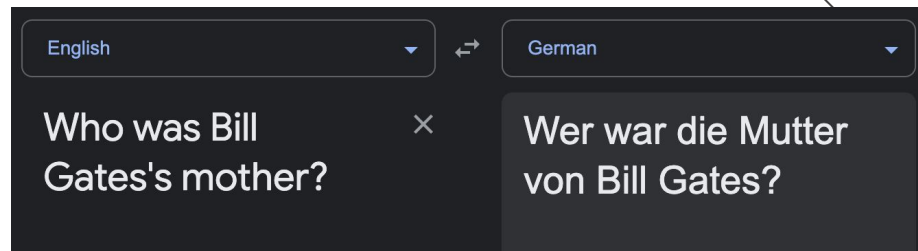
OpenAi offers hosted fine-tuning - for GPT-3 only

Models: ada, babbage, curie and davinci

All models know very little about Sparql initially, but learn the syntax very quickly – less than 1000 examples

Even the basic models generate syntactically correct sparql

But what about vocabulary?



No need to translate “Bill Gates”!



Did you know that the wikibase word for “Bill Gates” is Q5284? How would our model know?



Beyond translation to reasoning: Langchain, chatGPT, and gpt-4



Teaching LLMs to use tools is a very hot research area:
ToolFormer, MRKL, ReAct

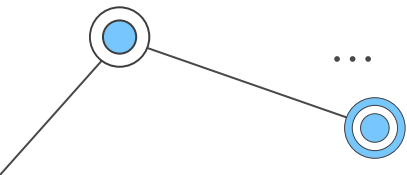
Implementation frameworks are moving very fast:
langchain, gpt-plugins

langchain abstractions:

- Prompts
- Tools
- Agents



LangChain



The LLM
does its
own query
planning!

> Entering new Agent Executor chain...

Thought: I need to find the q-number for J.S. Bach and the p-number for the property "number of children".

Action: ItemLookup

Action Input: J.S. Bach

Observation: **Q1339** I need to find the p-number for the property "number of children".

Action: Property Lookup

Action Input: number of children

Observation: **P1971** I need to run a sparql query to get the answer.

Action: Sparql Query Runner

Action Input:

```
SELECT ?children WHERE {  
  wd:Q1339 p:P1971 ?statement .  
  ?statement ps:P1971 ?children .  
}
```

Observation: [{"children": {"datatype": "http://www.w3.org/2001/XMLSchema#decimal", "type": "literal", "value": "20"}}] I now know the final answer.

Final Answer: J.S. Bach had 20 children.

> Finished chain.



A decorative network diagram with blue nodes and lines. The nodes are represented by concentric circles, with some having a solid blue center and others being hollow. They are connected by thin black lines. There are three main paths: one in the top right, one in the bottom left, and one in the bottom right. Each path starts with an ellipsis (...).

Summarization



Objective

Natural Language Answer generation from Knowledge Graph

Capitalize on the information and structure of the knowledge graph to generate natural language [english] response utilizing pre trained large language models.

Trigger

Fine Tune

Establish

**Control and
Summarize**

Achieve

Feature & Label Generation

Wikidata object represents language specific labels, multiple property value pairs, related qualifiers and references



item: *Earth*



property: *highest point*



value: *Everest*



Sparql Results

Data for Mountains		
Mountain	Property	Value
Mount Everest	height	8,848 m
K2	hauteur	8,611 m
Kanchenjunga	height	8,586 m
Lhotse	height	27940 ft

Triples

Earth (Q2) (item) → highest point (P610) (property) → Mount Everest (Q513) (value)

Mount Everest (Q513) (item) → instance of (P31) (property) → mountain (Q8502) (value)

Askwiki will

- **Execute** the sparql generated in earlier pipeline on wikidata and gather results
- **Extract** embedded wikidata objects in the result
- **Parse** the wikidata objects properties
- **Generate TRIPLES** from the above information indicating positional embeddings of subject predicate and object from wikidata
- **Aggregate TRIPLES** in a certain way so that it can be fed into the NLG & Summarization model
- **Present** the answer back to the user

Models and Experimentation

Model	Training	Observation
T5 large	Trained WebNLG dataset ,on 70% training split for about 24K+ examples,for 4 epochs	top_p ,top_k sampling generation produced 35 words per response as compared to 33 words per response for beam search,across 100 sample inputs [top sampling reading score =69.34,beam search reading score=68.77]
Open AI da-vinci	Trained with WebNLG dataset with 246 random examples	Few shot trained model produced 17 words per response [more effective in linking input triples] for 100 sample inputs without compromising readability [reading score =68.10]

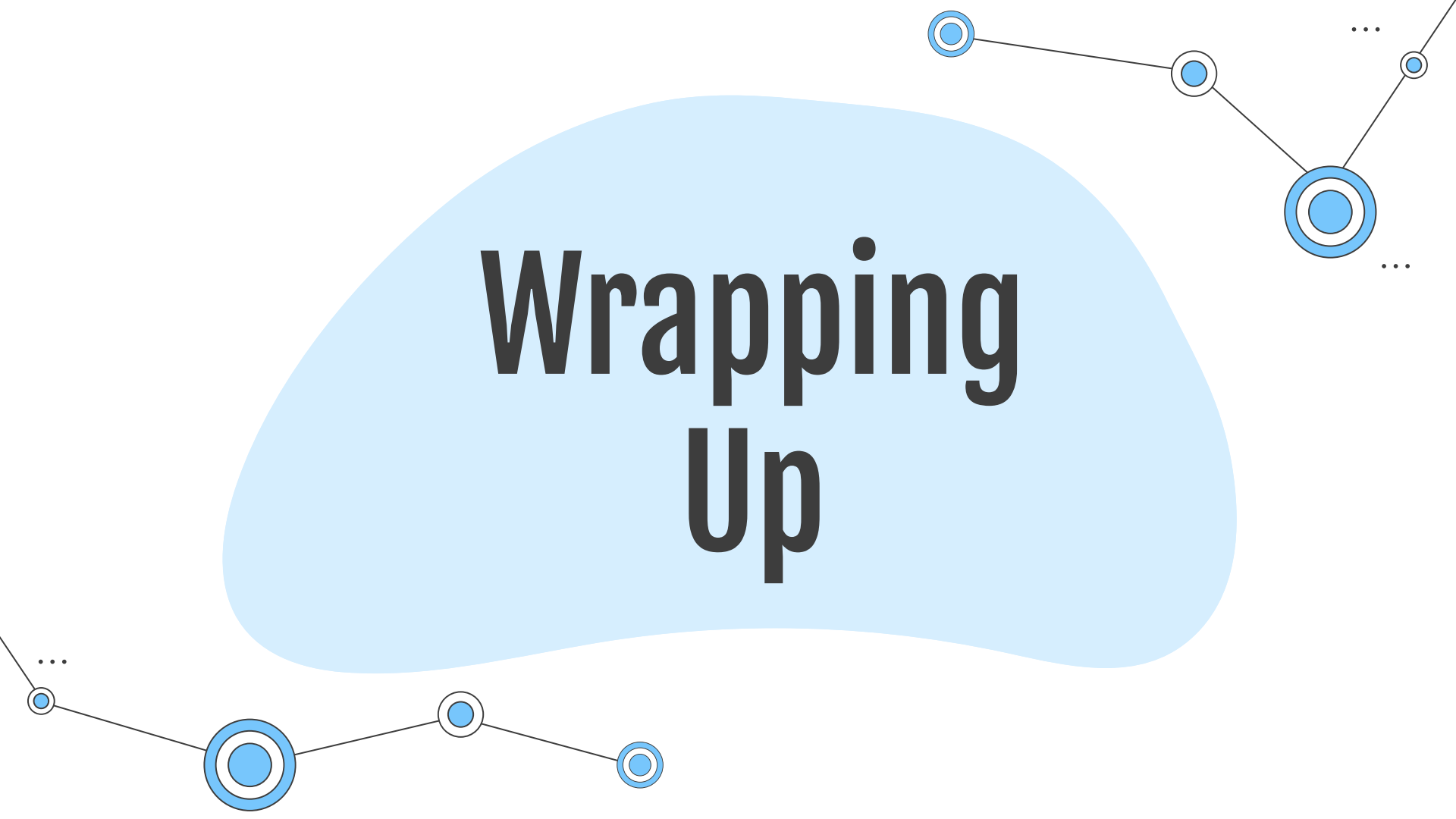


Demo



The diagram features a central light blue blob containing the word 'Demo'. Surrounding this are several blue circular nodes connected by black lines. One node is significantly larger than the others. Ellipses (...) are placed at the ends of the lines to suggest a larger, interconnected network.

Wrapping Up



Enterprise Adoption

Domain Specific Data

Askwiki did not utilize **domain specific** curated dataset, **enterprise datasets** can be curated & converted into knowledge graphs and would provide better feature and label generation.

Query Accuracy

Wikidata is a *live data source*, posing challenges on **accuracy of the sparql**, **enterprise metadata** can be effectively managed to generate more accurate queries.

Factual Accuracy

Quality of data on wikidata can be questionable, **enterprise data quality** can be improved resulting in more accurate answers.



Value Proposition

AskWiki

1. Establish the **path for effective adoption and exploration** for real life use cases where **LLMs can be applied in enterprise settings.**
2. Help **educate** enterprises to understand **role their “data” will plays.**
3. Separate the **“hype”** from **“reality”** and get organizations to actually think on how to use the technology.
4. Make LLM capabilities **accessible to all** without [*] trading their data and abstract the complexities of model evaluations, tuning, prompt experimentation, etc.
5. **Fast track and automate** routine information gathering and knowledge management tasks.

*we anticipate that pretty soon offerings will enable consumers to use LLM services without trading their proprietary training data

Conclusion

Ask a question and let AskWiki do the heavy lifting

