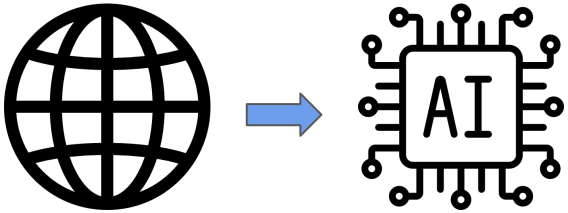
Can Foundation Models Wrangle Your Data?

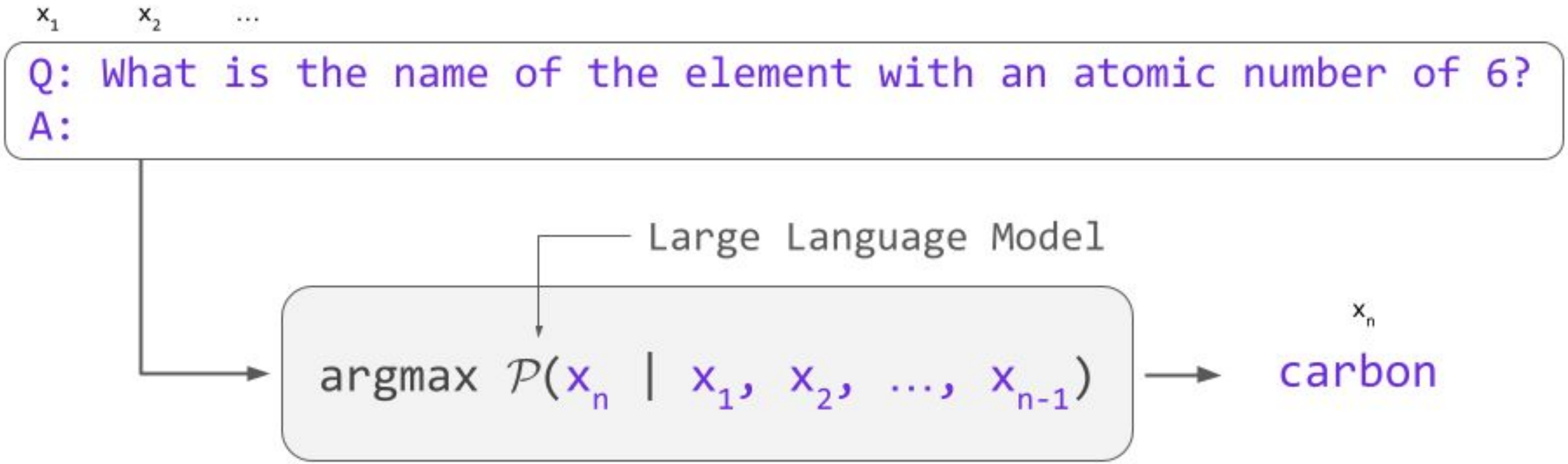
Avanika Narayan, Ines Chami, Laurel Orr, Simran Arora, Christopher Ré

Sahil Bhatia

Large Language Models



Large Language Models



Large Language Models

Cause & Effect

Prompt

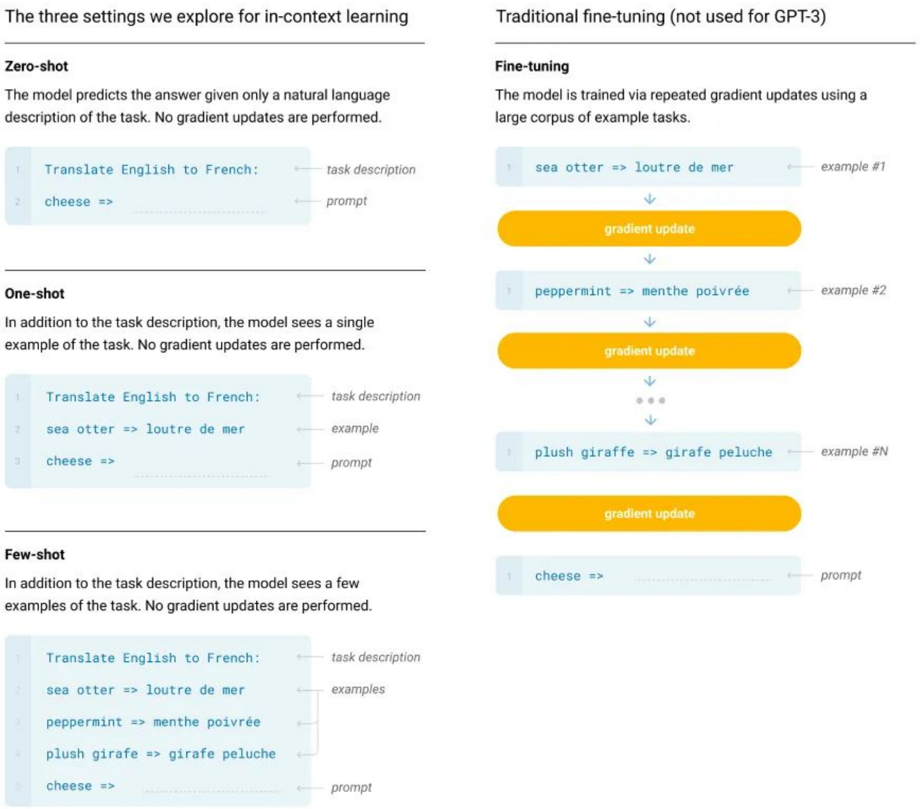
Wh

Large Language Models

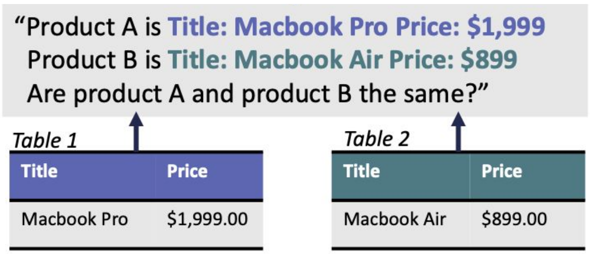
Prompt	Model Response

```
// Translate from C to Python
int add_one ( int x ){
 int m = 1;
 while ( x & m ) {
   x = x ^ m;
   m <<= 1;
 }
 x = x ^ m;
 return x; }
```

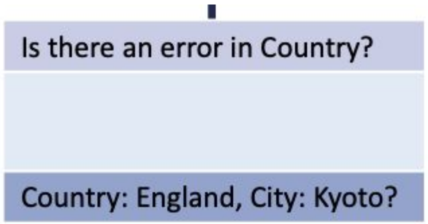
LLM are Few Shot Learners



Data Wrangling



Entity Matching



Error Detection

	Height	Weight	Country	Place	Number of days	Some column
0	12.0	35.0	India	Bengaluru	1.0	NaN
1	NaN	36.0	US	New York	2.0	NaN
2	13.0	32.0	UK	London	NaN	NaN
3	15.0	NaN	France	Paris	4.0	NaN
4	16.0	39.0	US	California	5.0	12.0
5	NaN	NaN	NaN	Mumbai	NaN	NaN
6	NaN	NaN	NaN	NaN	6.0	NaN

Data Imputation

ML vs LLM for Data Wrangling

Property	**Traditional ML**	**LLM**
Task Specific Architecture	Architectural changes and Fine-tuning	Natural language Interface
Hard-Coded Knowledge	Domain knowledge and commonsense reasoning with human engineered rules	Trained on generic data (inherit)
Labeled Data	Massive amounts of labelled data for training	Little to no data (zero-shot, few-shot)

Problem Statement

Can advances in **LLM help** in these hard **data tasks**?

LLMs for Data Task - Serialization

1. Convert structured data to text
   a. Given a table with columns $attr_1$, …, $attr_m$ and entry (e) with values val_1, …, val_m

$$\mathtt{serialize}(e) \coloneqq \mathtt{attr}_1 : \mathtt{val}_1 \ldots \mathtt{attr}_m : \mathtt{val}_m$$

LLMs for Data Task - Naturalization

1. Convert structured data to text
   a. Given a table with columns $attr_1$, …, $attr_m$ and entry (e) with values val_1, …, val_m

$$\texttt{serialize}(e) := \texttt{attr}_1 : \texttt{val}_1 \ldots \texttt{attr}_m : \texttt{val}_m$$

2. Convert data tasks to natural language tasks (prompts)

Product A is serialize(e). Product B is serialize(e').
Are Product A and Product B the same?

Entity Matching

given an entry e and attribute j to infer, we use
$attr_1$:val_1...$attr_j$?

Data Imputation

given an entry e and attribute j to classify as erroneous, we use
Is there an error in $attr_j$:val_j?

Error Detection

LLMs for Data Task - Demonstration

1. Random - sample random examples from labelled dataset
2. Manual - carefully select examples on basis of performance on validation set

Experimental Setup

1. Model - GPT3-175B parameter model

Task	Dataset	Baseline	Eval Metric
Entity Matching	Magellan	Ditto (BERT)	F1 score
Data Transformation	TDE	Search-based Solution	Accuracy
Schema Matching	Synthea	SMAT (attention bi-LSTM)	F1 score
Imputation	Restaurant and Buy	IMP (finetunes RoBERTa)	Accuracy
Error Detection	Hospital and Adult	HoloClean and HoloDetect (ML)	F1 score

Entity Matching

Dataset	Magellan	Ditto	GPT3-175B (k=0)	GPT3-175B (k=10)
Fodors-Zagats	100	100	87.2	**100**
Beer	78.8	94.37	78.6	**100**
iTunes-Amazon	91.2	97.06	65.9	**98.2**
Walmart-Amazon	71.9	86.76	60.6	**87.0**
DBLP-ACM	98.4	**98.99**	93.5	96.6
DBLP-Google	92.3	**95.60**	64.6	83.8
Amazon-Google	49.1	**75.58**	54.3	63.5

Entity Matching

1. LLMs struggle with data domains that contain jargons
   a. Amazon-Google has product specific identifier in description

name: pcanywhere 11.0 host only cd-rom xp 98 nt w2k me. manufacturer: symantec. price: NULL” and “name: symantec pcanywhere 11.0 windows. manufacturer: NULL. price: 19.99.”

Imputation and Error Detection

Task	Imputation		Error Detection	
Dataset	Restaurant	Buy	Hospital	Adult
HoloClean	33.1	16.2	51.4	54.5
IMP	77.2	96.5	-	-
HoloDetect	-	-	94.4	99.1
GPT3-175B (k=0)	70.9	84.6	6.9	0.0
GPT3-6.7B (k=10)	80.2	86.2	2.1	99.1
GPT3-175B (k=10)	**88.4**	**98.5**	**97.8**	**99.1**

Imputation and Error Detection

1. LLMs understand how to complete task
2. Have encoded knowledge (dependencies between zip codes and address)

Transformation and Schema Matching

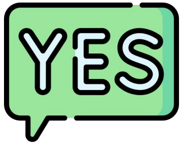
Task	Data Transformation		Schema Matching
Dataset	StackOverflow	Bing-QueryLogs	Synthea
Previous SoTA	63.0	32.0	38.5
GPT3-175B (k=0)	32.7	24.0	0.5
GPT3-175B (k=3)	**65.3**	**54.0**	**45.2**

Ablation Study

Prompt Format	Beer	iTunes-Amazon	Walmart-Amazon
Prompt 1 (w. Attr. & Example Select.)	**100** ± 0.00	**98.2** ± 0.00	88.9 ± 0.00
Prompt 1 (w/o Example Select.)	91.1 ± 0.05	86.6 ± 0.02	65.2 ± 0.04
Prompt 1 (w/o Attr. Select.)	76.9 ± 0.00	94.1 ± 0.00	75.0 ± 0.00
Prompt 1 (w. Attr. & w/o Attr. names)	80.0 ± 0.00	94.5± 0.00	84.2 ± 0.00
Prompt 2 (w. Attr. & Example Select.)	96.3 ± 0.00	84.7 ± 0.00	**100** ± 0.00

Prompt 1: “`Are Product A and Product B the same?`”
Prompt 2: “`Are Product A and Product B equivalent?`”

Problem Statement



Can advances in **LLM help** in these hard **data tasks**?

Future Opportunities

1. Natural Interactions
    a. Systems would be more accessible to non-machine learning experts
2. Unstructured Data to Structured
    a. Organizations with data from streams can organize using LLMs
3. Integration in Data Management
    a. Everything is not text. Lot of actions using GUIs
    b. Multimodal models are going to be helpful
4. Integrate with existing system
    a. Systematically incorporate with existing data management systems

Challenges

1. Domain Specificity
   a. Highly specialized data
   b. Fine tuning
2. Privacy
   a. Open-source LLMs which can be fine tuned easily
3. Prompt Engineering
   a. Retrieval augmented
   b. Automatic prompt engineering