Accepted by the Tabular Representation Learning workshop at NeurIPS 2022
Link to the paper:
https://github.com/jiaweitang0202/generic.entity.resolution/blob/main/ger.pdf

Generic Entity Resolution Models

Jiawei Tang
American School of Doha, Qatar

Yifei Zuo
USTC, China

Lei Cao, Samuel Madden
CSAIL, MIT, USA

The Entity Resolution Problem

Given two entities **a** and **b**, decide if (**a**, **b**) match or not?
Answer: 1 (match)

a

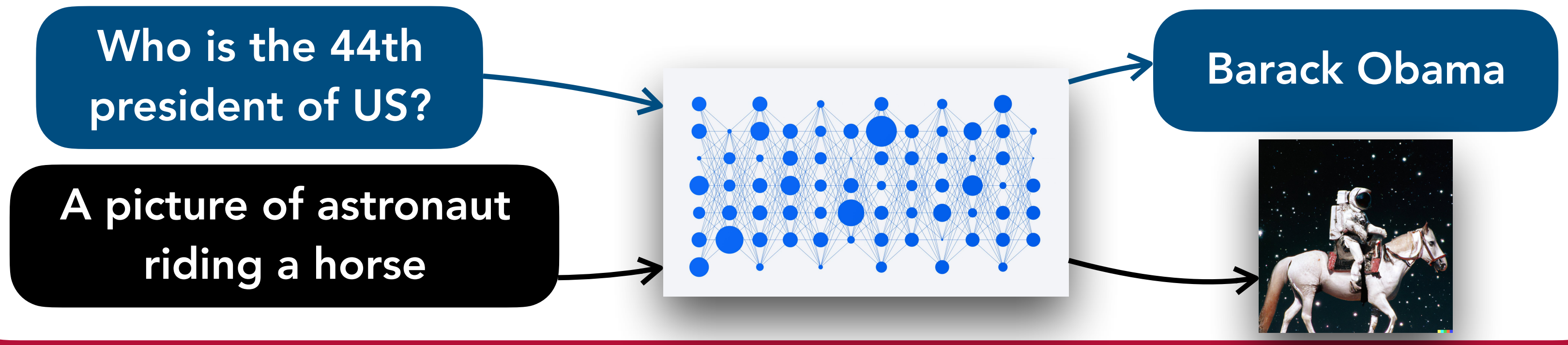
Company	Address	City
P Sherman Orthodontics	32 Wallaby Way	Sydney, NSW

b

Company	Address	City
Sherman Orthodontics Ltd.	Sky Tower, 32 Wallaby Way	Sydney

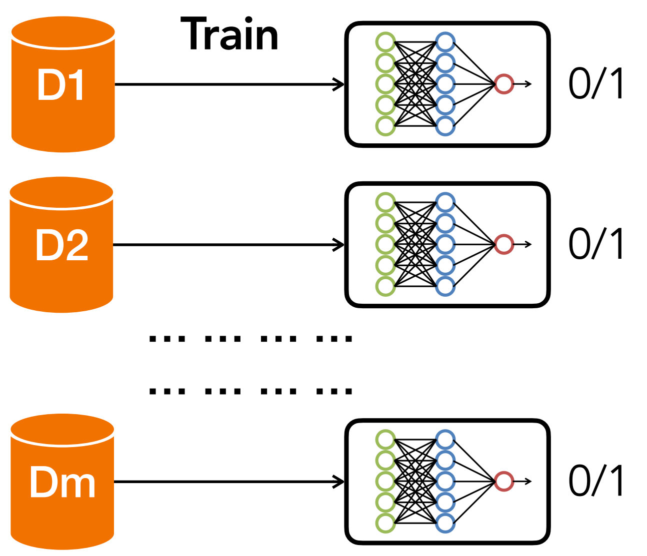
Foundation Models

Foundation models are giant artificial intelligence models trained on a large corpus of data and can perform many tasks

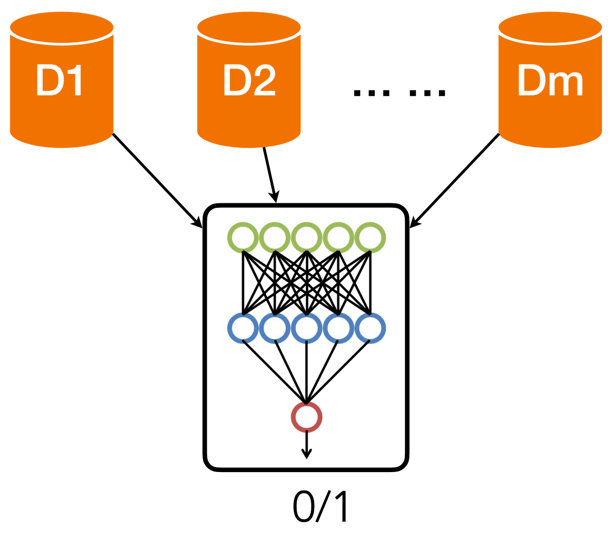


Specific Models vs. Generic Solutions

Specific Solutions
- traditional wisdom -



Generic Solutions
- goal of this project -



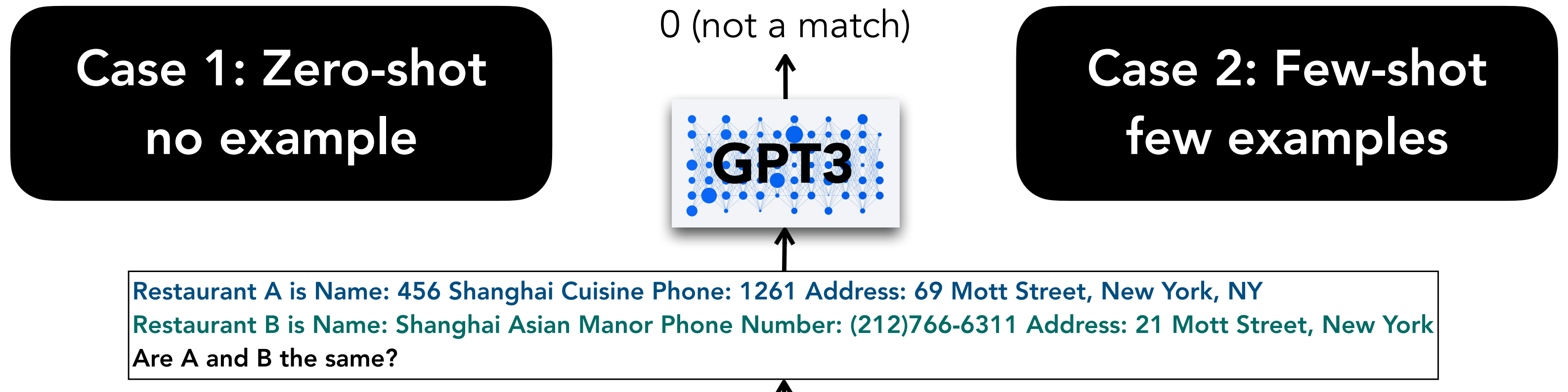
Limitations

1. Need a lot of train data for each dataset
2. Lack of generalizability
3. Large sizes

Opportunities

1. Foundation models
2. A trained generic model

Foundation Models for Entity Resolution



Name	Phone	Address
456 Shanghai Cuisine	1261	69 Mott Street, New York, NY

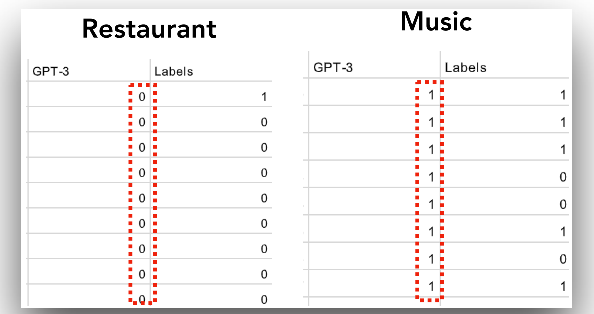
Name	Phone	Address
Shanghai Asian Manor	(212)766-6311	21 Mott St, New York

Experiments with Foundation Models

Zero-shot results

Dataset	Precision	Recall	F-messure
Restaurant	0.991	0.919	0.954
Bike	1	0.1	0.182
Movie	0.714	0.368	0.486
Book	0.632	0.598	0.615

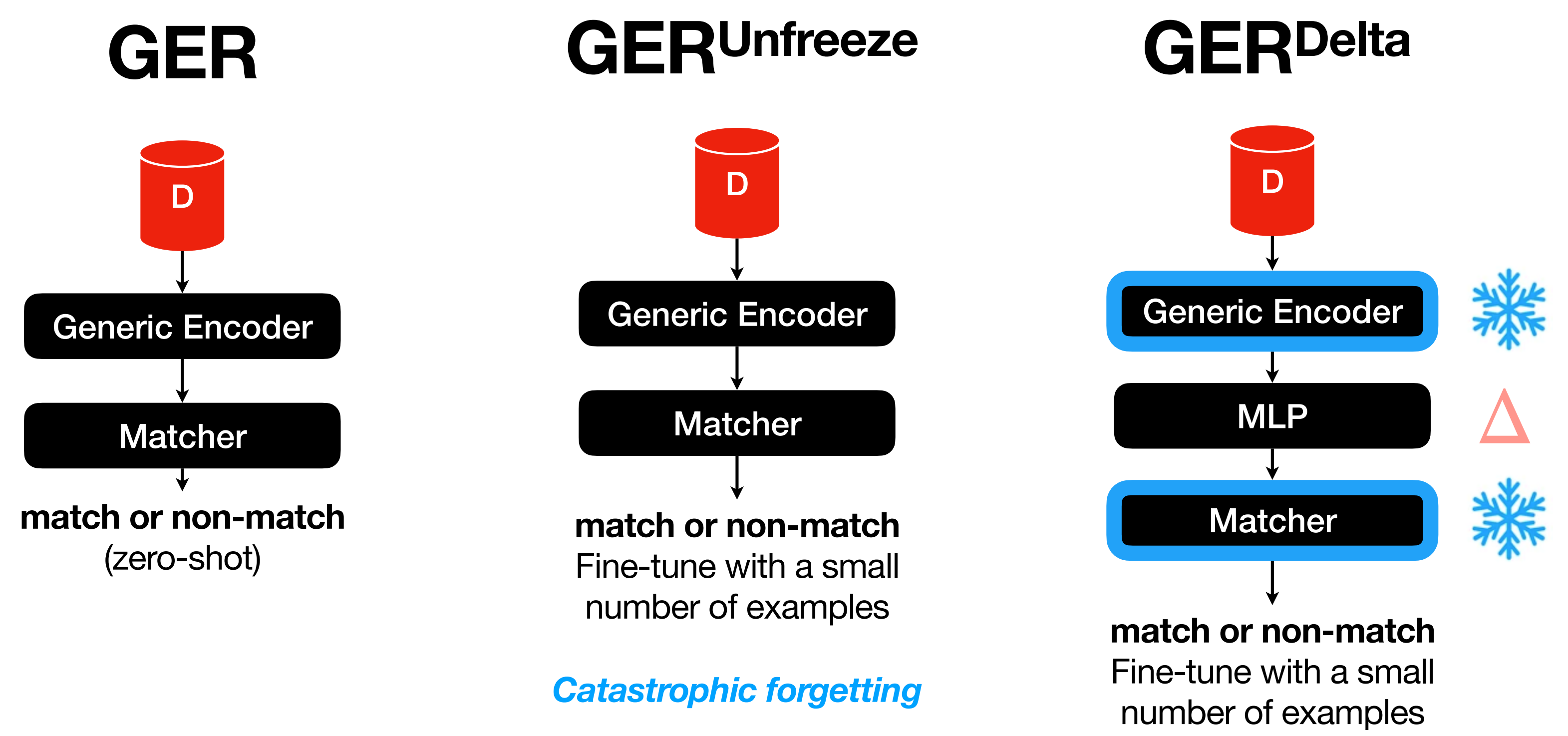
Few-shot results



1. Under zero-shot, GPT3 works, but underperforms specific solutions
2. Few-shot gives very biases results (either all 0's or all 1's)
3. **Conclusion**: Zero-shot should be used on GPT3 for entity resolution

A Trained Generic Model

Designed three solutions for building generic models



Experiments with Generic Models

	Specific Model			GPT-3			GER		
Dataset	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Restaurant	0.941	0.941	0.941	0.991	0.919	0.954	0.978	0.936	0.956
Bike	0.773	0.548	0.642	1	0.1	0.182	0.405	0.81	0.54
Movie	0.978	0.957	0.968	0.714	0.368	0.486	0.873	0.984	0.925
Book	1	1	1	0.632	0.598	0.615	0.837	1	0.911

	$\text{GER}^{\text{Unfreeze}}$			$\text{GER}^{\text{Delta}}$		
Dataset	**P**	**R**	**F1**	**P**	**R**	**F1**
Restaurant	0.979	0.957	**0.968**	0.939	0.979	0.958
Bike	0.554	0.857	0.673	0.56	0.881	**0.685**
Movie	1	0.984	0.992	1	0.984	**0.992**
Book	0.973	1	**0.986**	0.935	1	0.966

1. Both $\text{GER}^{\text{Unfreeze}}$ and $\text{GER}^{\text{Delta}}$ can be adapted to a new task with a few train examples
2. If one still wants the entity resolution model to perform well on previously trained ER datasets, then $\text{GER}^{\text{Delta}}$ should be used