Pool-Search-Demonstrate:

Improving Foundation Models in Data Wrangling Tasks

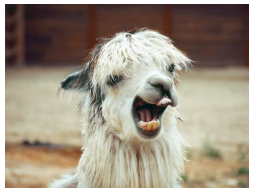
Joon Suk Huh, Changho Shin, Elina Choi

Data wrangling – Data cleaning for downstream tasks

- **Data Imputation –** Filling in missing values.
- **Entity Matching –** Identifying records referring to the same real-world entity.
- **Error Detection –** Detect erroneous values.

Foundation Models (FM)

- **Large** ML models trained on massive datasets.
- **Adaptable** to a wide range of downstream tasks without fine-tuning.
- **Examples** are Large Language Models (LLM).
  - **ChatGPT**
  - **LLaMA**
  - **Vicuna**
  - **T5**
  - ...



Foundation Models for Data Wrangling

Task demonstrations

"Address: 1720 university blvd State: AL ZipCode? 32533

Address: 26025 pacific coast hwy Phone number: 310/456-5733 City? Malibu"

Task description

"Address: 804 north point st Phone number: 415-775-7036 City?"

Main Question

Q: What demonstrations are good demonstrations?

Main Question

Q: What demonstrations are good demonstrations?

Manual > Random sampling [**Narayan *et.al.* (2022)**]

Main Question

Q: What demonstrations are good demonstrations?

Manual > Random sampling [**Narayan *et.al.* (2022)**]

Refined Q:
How the **quantity**, **relevancy** and **diversity** of demonstrations affect the data wrangling performance?

Proposed method: Pool–Search–Demonstrate

Proposed method: Pool–Search–Demonstrate

"Address: 1720 university blvd
State: AL ZipCode? 32533"

"Address: 26025 pacific coast hwy
Phone number: 310/456-5733
City? Malibu"

"Address: 1632 north point st
number: 510/653-3394 City? San
Francisco"

⋮

Example set

Proposed method: Pool–Search–Demonstrate

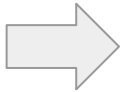
"Address: 1720 university blvd State: AL ZipCode? 32533"

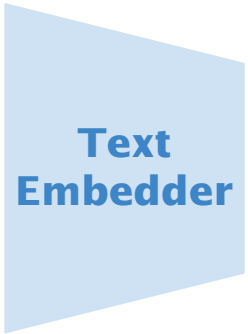
"Address: 26025 pacific coast hwy Phone number: 310/456-5733 City? Malibu"

"Address: 1632 north point st number: 510/653-3394 City? San Francisco"

⋮

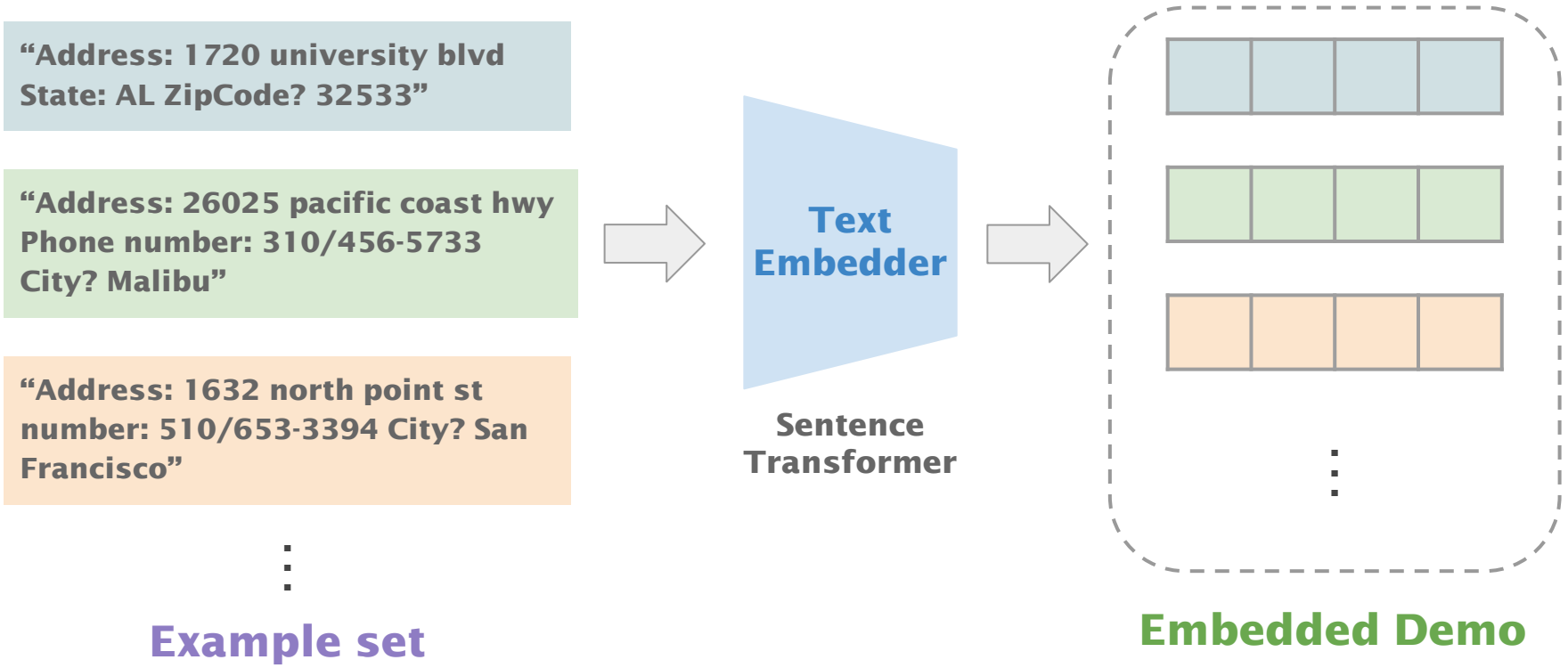
Example set



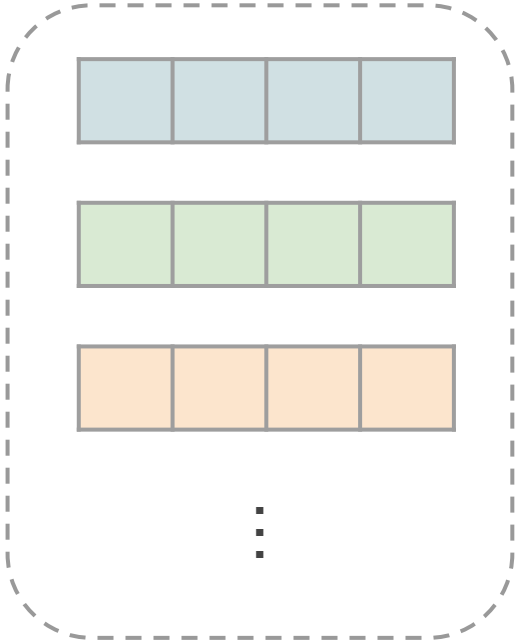


Sentence Transformer

Proposed method: Pool–Search–Demonstrate

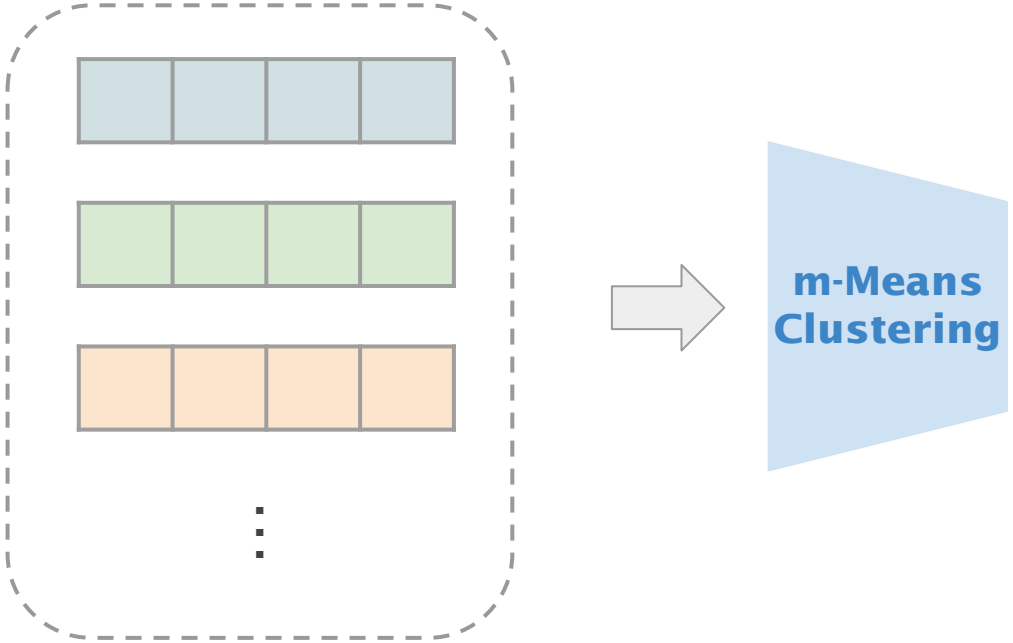


Proposed method: Pool–Search–Demonstrate



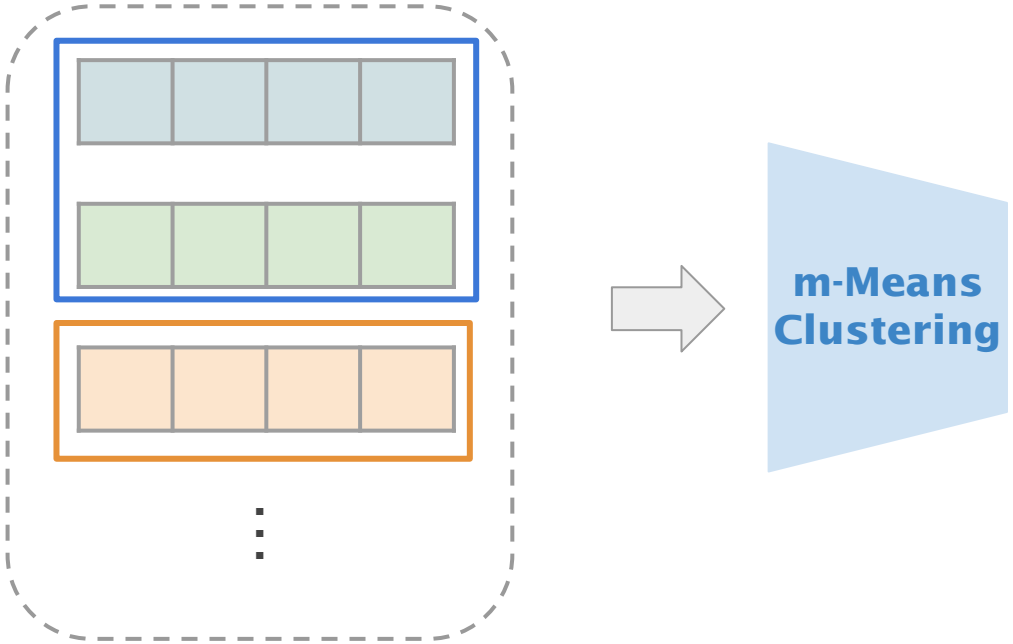
Embedded Demo

Proposed method: Pool–Search–Demonstrate



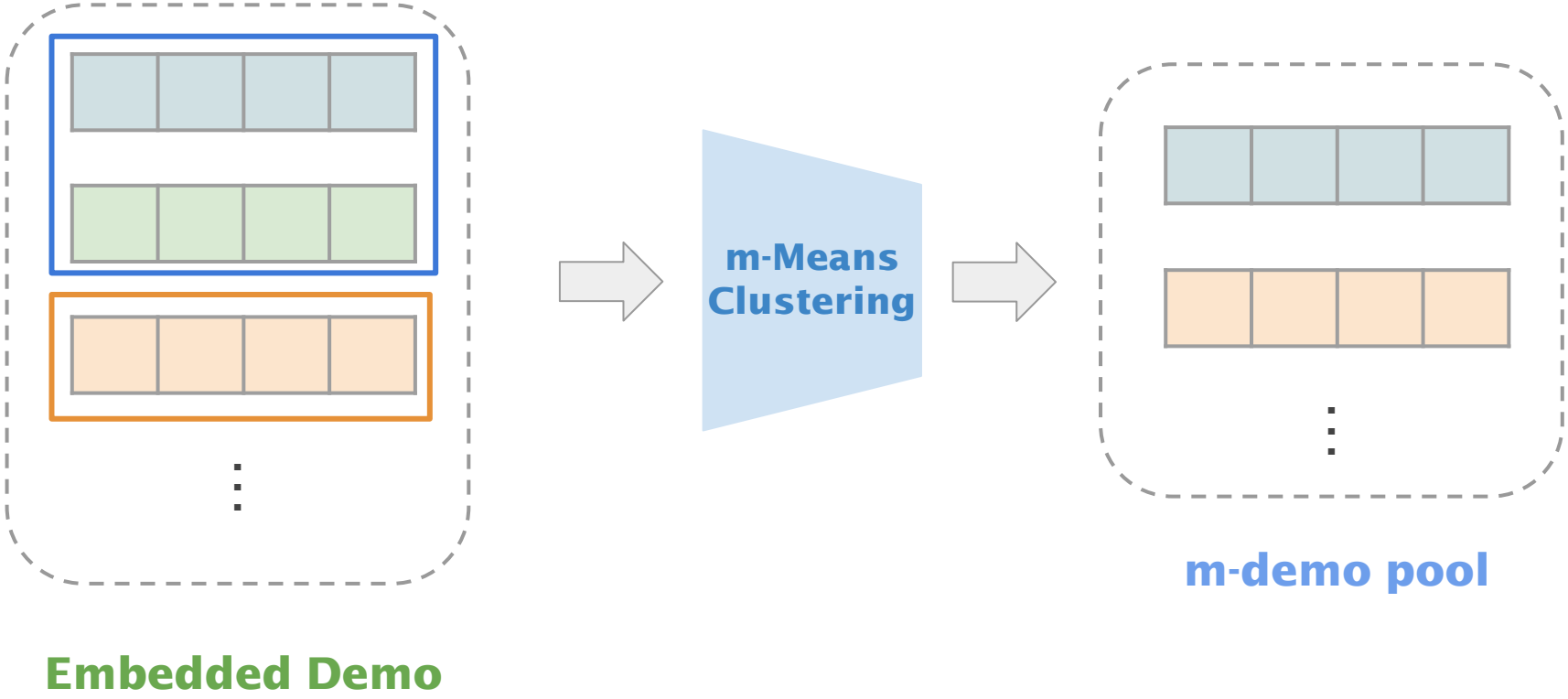
Embedded Demo

Proposed method: Pool–Search–Demonstrate

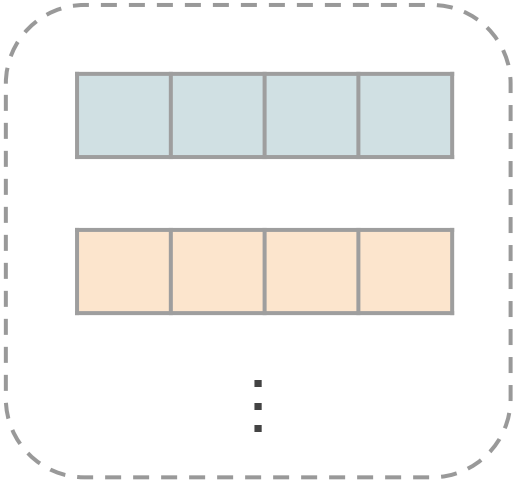


Embedded Demo

Proposed method: Pool–Search–Demonstrate



Proposed method: Pool–Search–Demonstrate

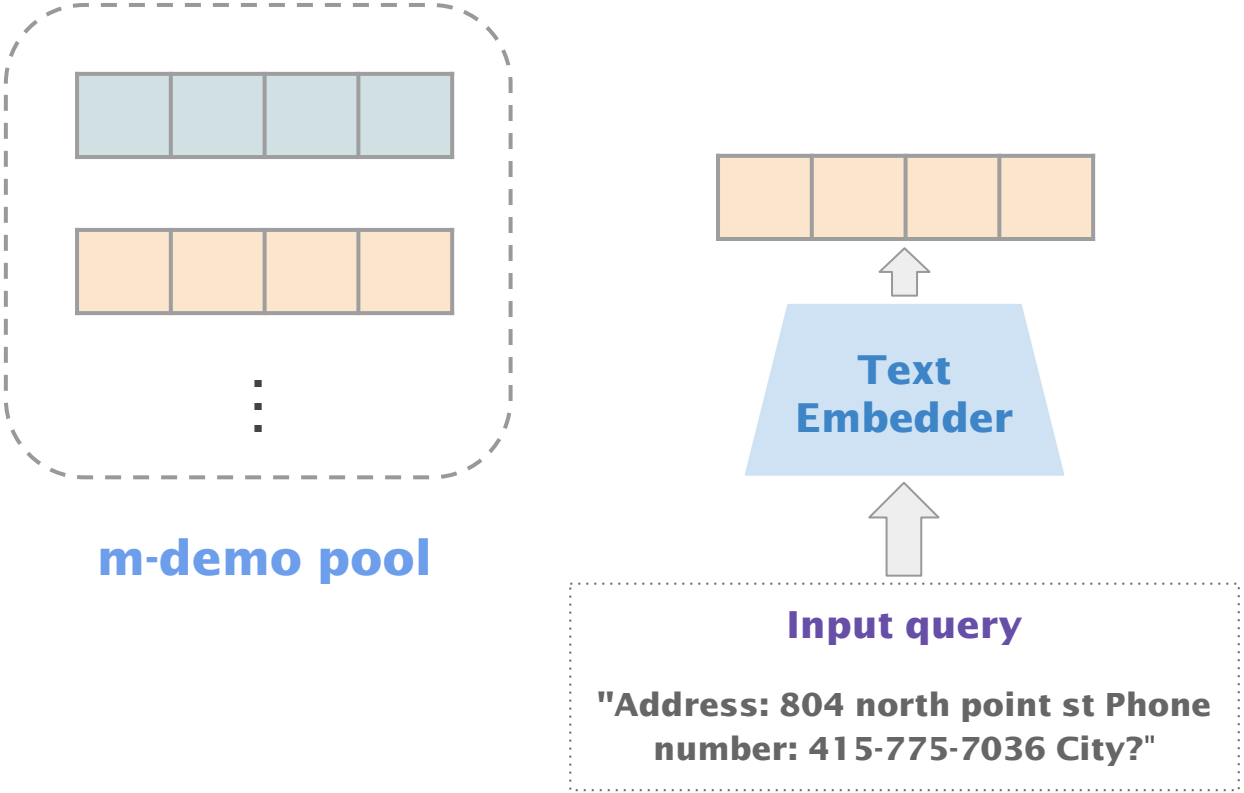


m-demo pool

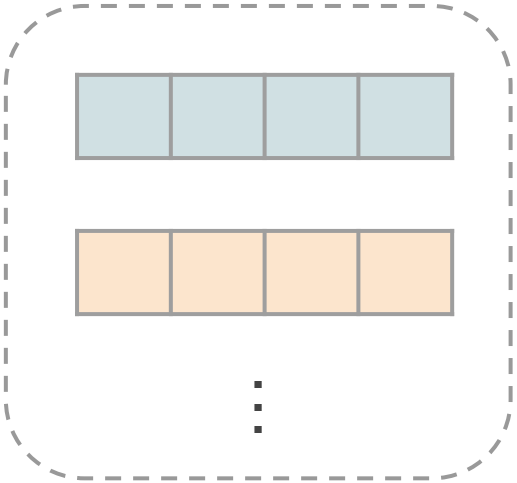
Input query

"Address: 804 north point st Phone number: 415-775-7036 City?"

Proposed method: Pool–Search–Demonstrate

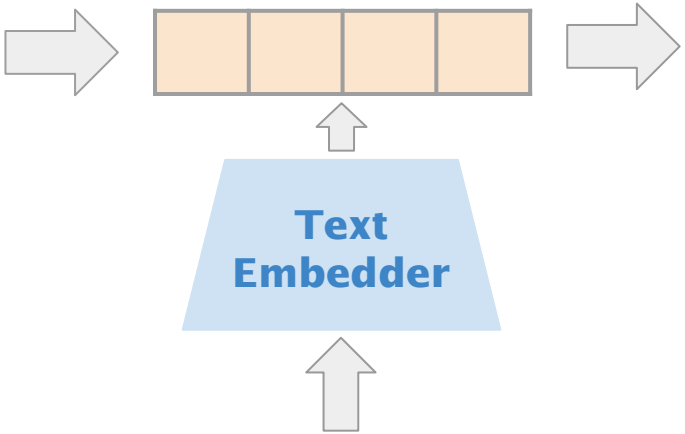


Proposed method: Pool–Search–Demonstrate



m-demo pool

Compute distances to the input query.
Pick k nearest demos.



Input query

"Address: 804 north point st Phone number: 415-775-7036 City?"

"Address: 1632 north point st number: 510/653-3394 City? San Francisco"

"Address: 586 hayes st number: 415/549-7445 City? San Francisco"

(m,k)–demo set

Proposed method: Pool–Search–Demonstrate

Prompt =
(m,k)-demo set
+ Input query

“Address: 1632 north point st number: 510/653-3394 City? San Francisco\n

Address: 586 hayes st number: 415/549-7445 City? San Francisco\n
…

Address: 804 north point st Phone number: 415-775-7036 City?”

Proposed method: Pool–Search–Demonstrate

Prompt =
(m,k)-demo set
+ Input query

"Address: 1632 north point st number: 510/653-3394 City? San Francisco\n

Address: 586 hayes st number: 415/549-7445 City? San Francisco\n
…

Address: 804 north point st Phone number: 415-775-7036 City?"

Large Language Model

Proposed method: Pool–Search–Demonstrate

Prompt =
(m,k)-demo set
+ Input query

"Address: 1632 north point st number: 510/653-3394 City? San Francisco\n

Address: 586 hayes st number: 415/549-7445 City? San Francisco\n
…

Address: 804 north point st Phone number: 415-775-7036 City?"

Large Language Model

"San Francisco"

Experiment setup

Tasks

- **Data Imputation** – Buy, Restaurant
- **Entity Matching** – DBLP-ACM, Walmart
- **Error Detection** – Hospital, Adult

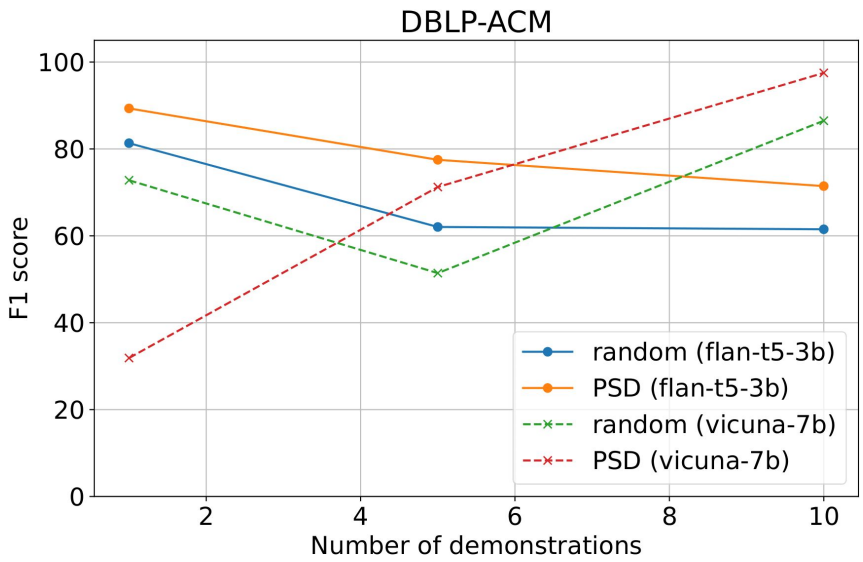
Models

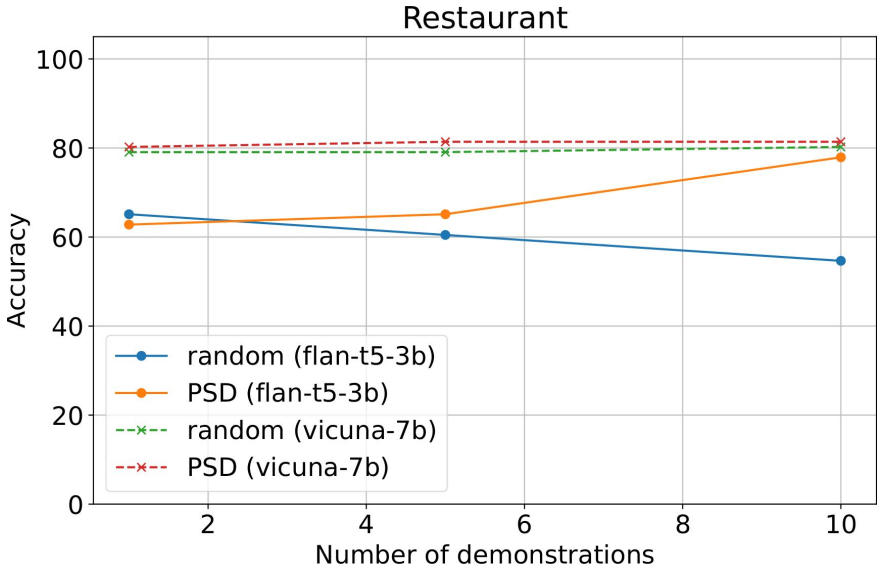
- **Flan-T5-3B**
- **Vicuna-7B**

Demonstration Methods

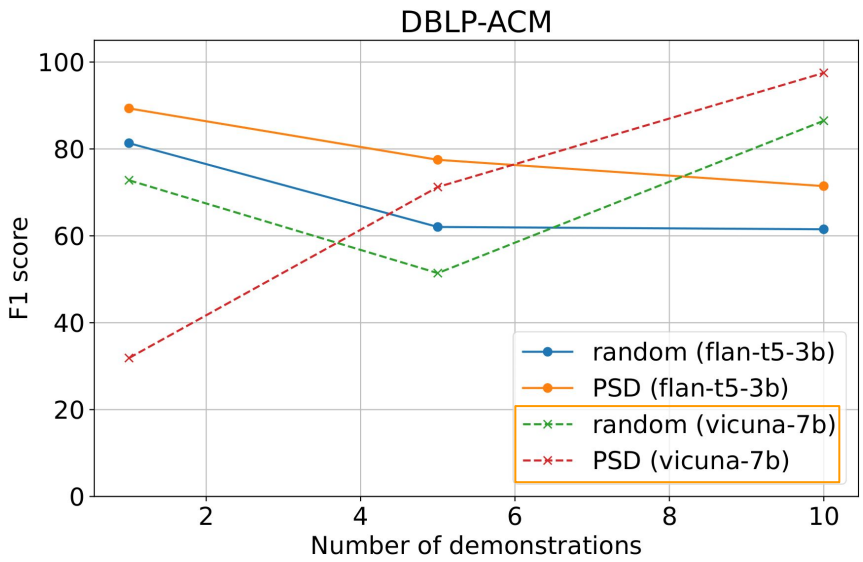
- **Random** – k=1, 5, 10
- **Manual (Narayan *et.al.*)** – k=10
- **Pool-Search-Demonstration (Ours)** – k=1, 5, 10

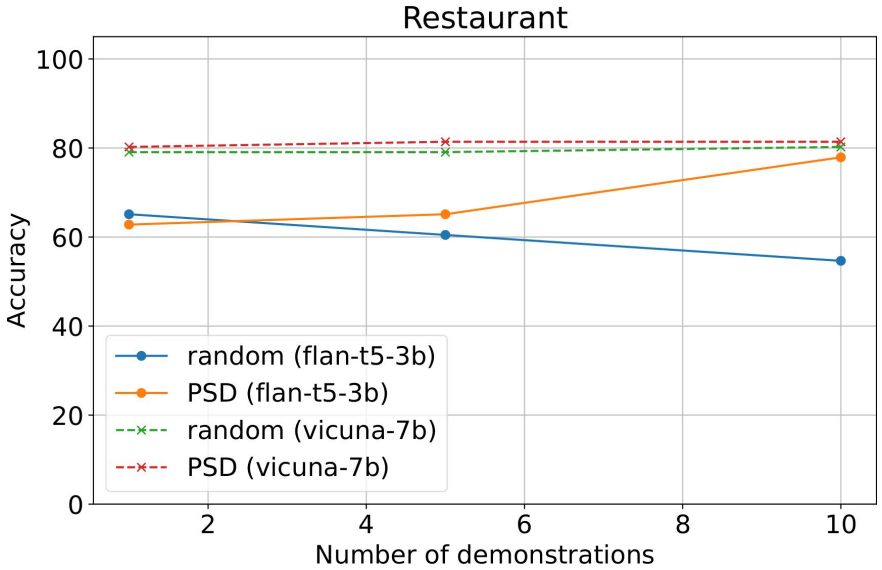
Effects of increasing number of demos (k)





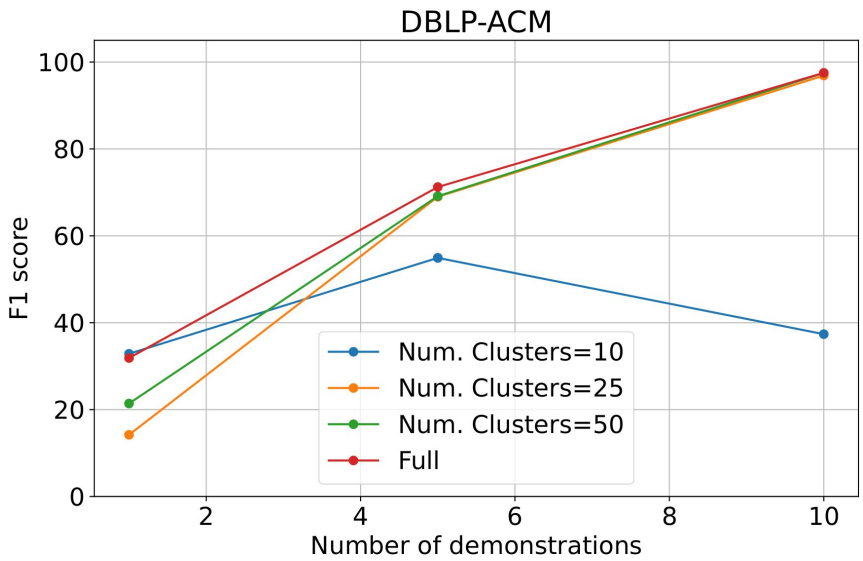
Effects of increasing number of demos (k)



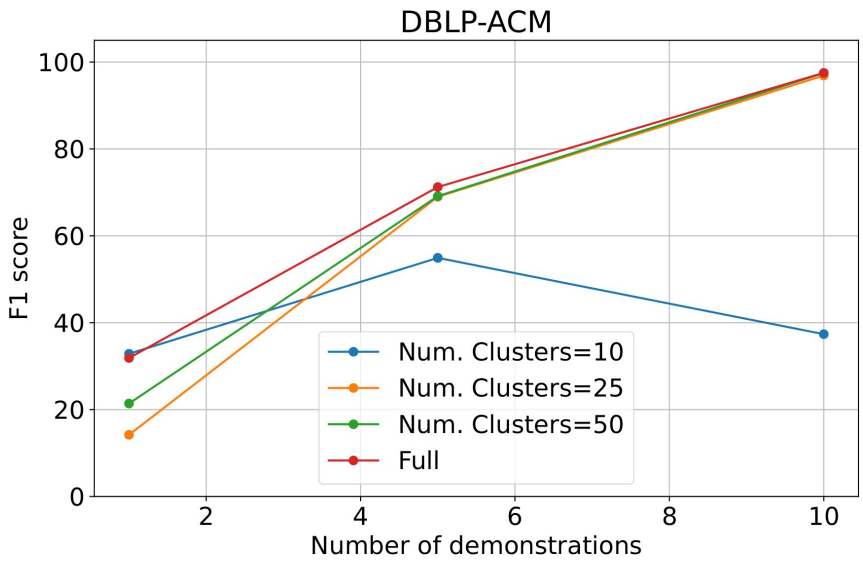


More is Better for Larger.

Effects of diversity (m=# of clusters)



Effects of diversity (m=# of clusters)



Diversity helps.

Results (Best hyperparameters)

		Data imputation (Acc.)		Entity matching (F1)		Error detection (F1)	
Model	Demo. method	Buy	Restaurant	DBLP-ACM	Walmart	Hospital	Adult
Prev.ML Best	X	96.5	77.2	99.0	86.8	94.4	99.1
FLAN-T5-3B	random	98.5	65.1	81.3	**78.2**	10.9	**58.5**
	manual	98.5	**88.4**	78.0	75.9	15.5	0
	PSD	**100.0**	77.9	**89.3**	72.7	**67.4**	34.6
Vicuna-7B	random	98.5	80.2	86.5	68.7	4.6	58.0
	manual	**100.0**	**87.2**	73.1	59.0	30.7	38.0
	PSD	**100.0**	81.4	**97.5**	**82.5**	**89.4**	**87.3**

Summary

Foundation models can perform data wrangling tasks ***without re-training or fine-tuning***.

Performance depends on the ***quality*** of demonstrations given in prompt.

Our method shows ***more, diverse and relevant*** demos help foundation models to perform better in data wrangling tasks.

Thank you!