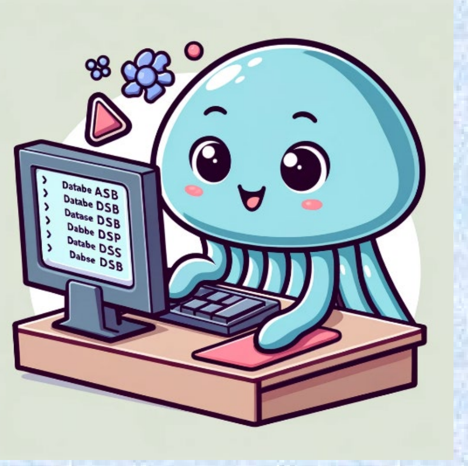


Jellyfish: Instruction-Tuning Local Large Language Models for Data Preprocessing

Haochen Zhang[1], Yuyang Dong[2],

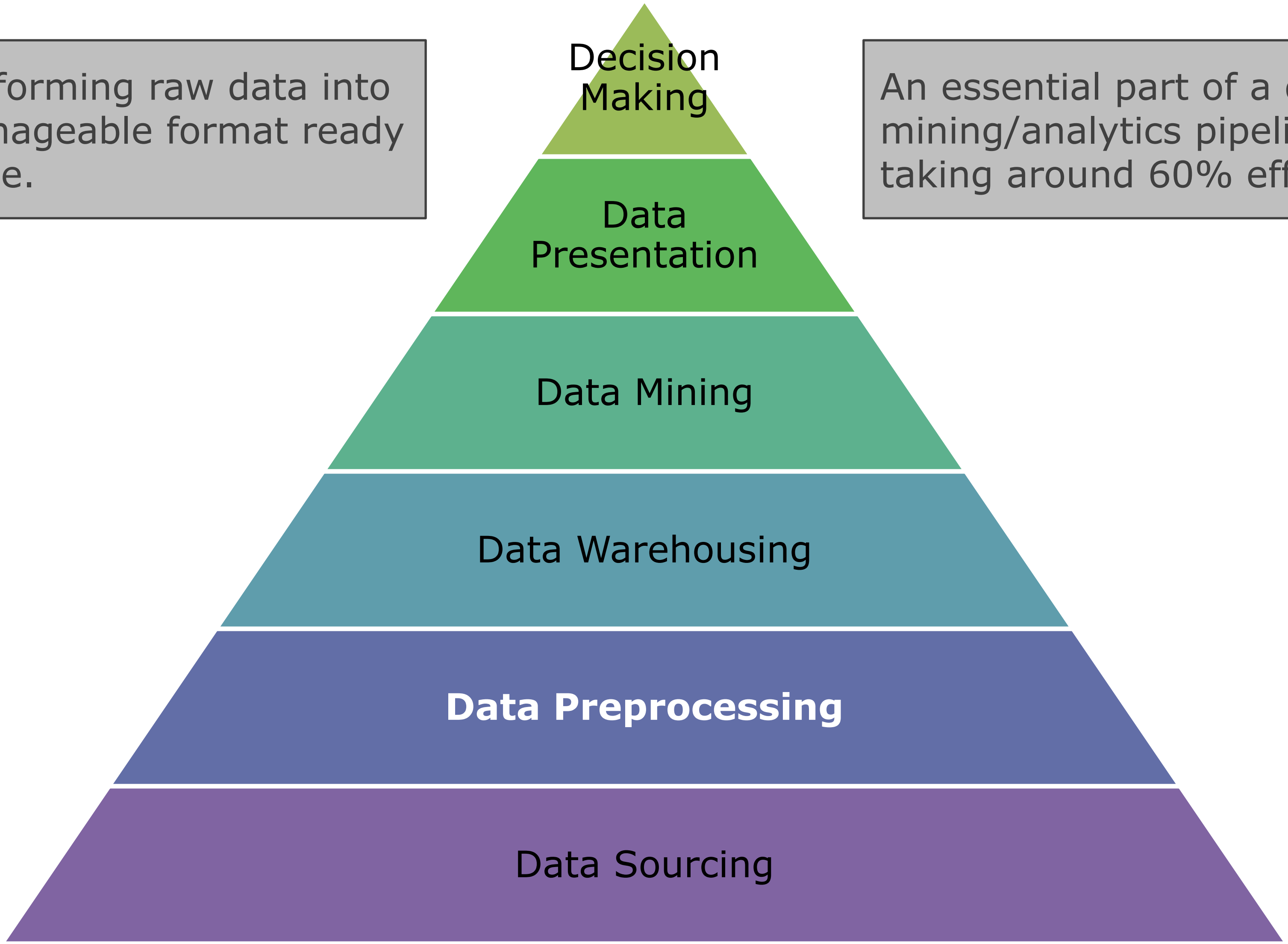
Chuan Xiao[1,3], and Masafumi Oyamada[2]

[1]Osaka University, [2]NEC Corporation, [3]Nagoya University

https://huggingface.co/NECOUDBFM/Jellyfish

Data Preprocessing (DP)

Transforming raw data into a manageable format ready for use.



An essential part of a data mining/analytics pipeline, taking around 60% effort.

Why Is DP Important?

Data in real world are often:

- Dirty: missing values, noise, duplicates …
- Heterogeneous: multiple sources, different formats
- Raw
- Complex
- Space-consuming

No quality data, no quality results!

- Duplicate or missing data may cause incorrect or even misleading statistics.
- Prediction models need to be built upon quality data.

Typical Procedures in DP

Data Cleaning

Detect and repair errors.

missing *inconsistency*

ID	Name	Date of Birth	Prefecture	Postal Code	Height
1	Yuka	2003/02/26	Hokkaido	540-8570	165
2	Nana		Aichi	464-0804	157
3	Nana	2003/03/30	Aichi	464-0804	157
4	Miho	2001/06/25	Kangawa	2208799	1.60

duplicate *typo* *format* *outlier*

ID	Name	Date of Birth	Prefecture	Postal Code	Height
1	Yuka	2003/02/26	Osaka	540-8570	165
2	Nana	2003/03/30	Aichi	464-0804	157
4	Miho	2001/06/25	Kanagawa	220-8799	160

https://huggingface.co/NECOUDBFM/Jellyfish

Data Integration

Merge data from multiple sources to address heterogeneity.

id	name	loc	# of employees
1	Apple	CA	154,000
2	IBM	NY	282,000
3	...	...	...

id	name	addr	rev
1	IBM Corp	NY	$57B
2	Apple Inc	CA	$366B
3	GE	MA	$74B

name	loc	# of employees	rev
Apple Inc	CA	154,000	$366B
IBM Corp	NY	282,000	$57B
GE	MA	205,000	$74B

Data Transformation

Convert raw data to a unified format for analysis.

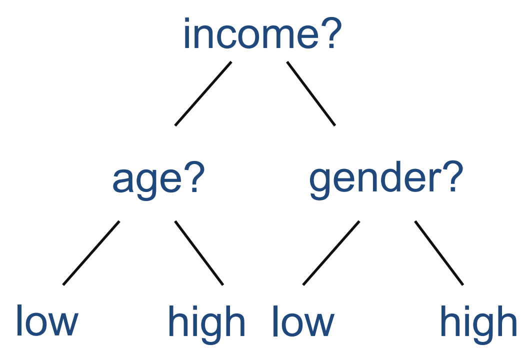
Name	Phone	Birth Date	State
Smith, John	445-881-4478	August 12, 1989	Maine
Jennifer Tal	+1-189-456-4513	11/12/1965	Tx
Gates, Bill	(876)546-8165	June 15, 72	Kansas
Alan Fitch	5493156648	2-6-1985	Oh
Jacob Alan	(205)1564896	1986 January 3	Alabama

Name	Phone	Birth Date	State
John Smith	445-881-4478	1989-08-12	Maine
Jennifer Tal	189-456-4513	1965-11-12	Texas
Bill Gates	876-546-8165	1972-06-15	Kansas
Alan Fitch	549-315-6648	1985-02-06	Ohio
Jacob Alan	205-156-4896	1986-01-03	Alabama

Data Reduction

Obtain reduced representation in volume but produce same/similar analytical results.

name	gender	age	date of birth	income	state
John Doe	male	38	02/27/1985	$120k	CA
Jane Sato	female	26	10/06/1997	$80k	NC



selected feature = {income, age, gender}

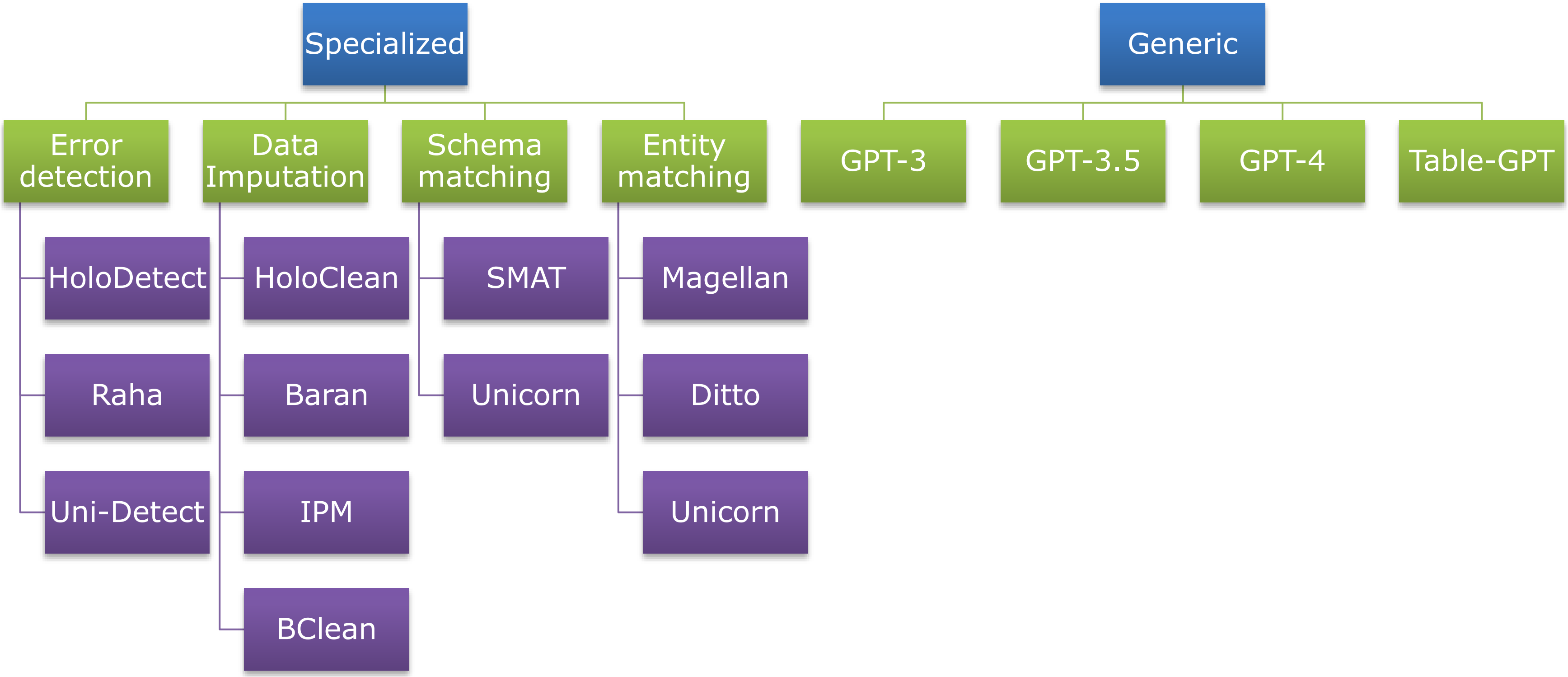
Typical Tasks in DP

Data cleaning	Error detection	Data repairing	Data imputation
Data integration	Schema matching	Entity matching	Data fusion
Data transformation	Text wrangling	Normalization	Smoothing
Data reduction	Feature selection	Dimensionality reduction	Compression

https://huggingface.co/NECOUDBFM/Jellyfish

Existing DP Approaches

Prior to the prevalence of large language models (LLMs), most solutions are tailored to one or two tasks.



https://huggingface.co/NECOUDBFM/Jellyfish

Pros & Cons in LLM Solutions to DP

Strengths

- Interaction in natural language
- Built-in knowledge
- Reasoning/interpretation ability
- Generalizability
- Adaptability
- Model conditioning with prompt engineering

Weaknesses

- Data security issue
- Difficulty in customization
- Large computation overhead
- Token limitation
- Lack of memory
- Hallucination

https://huggingface.co/NECOUDBFM/Jellyfish

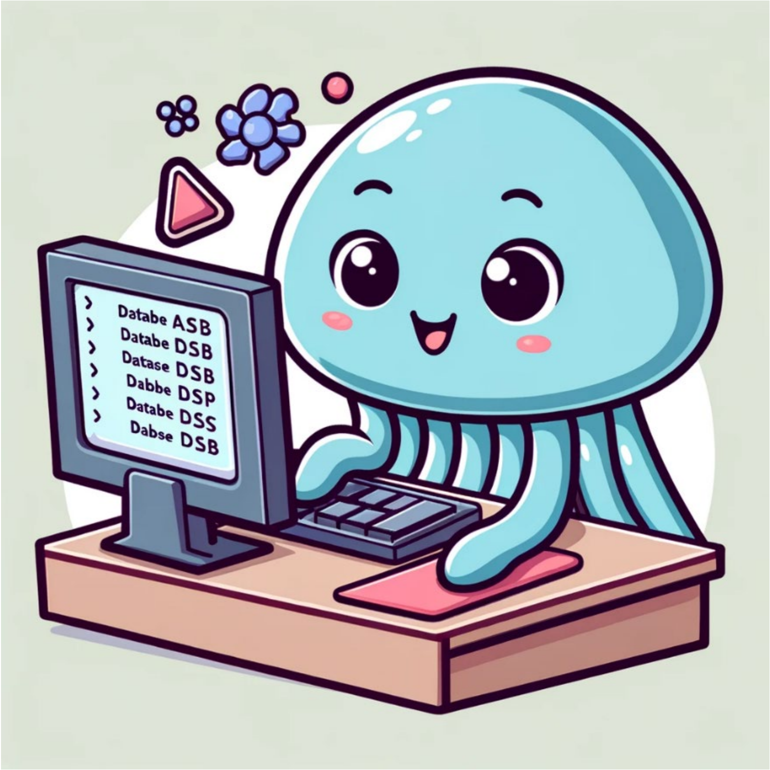
Jellyfish

Universal solution to DP

Instruction-tuned for 4 tasks, generalizing to unseen tasks.

Based on Mistral-7B or Llama 2-13B, running on a local, single, low-priced GPU.

Ensuring data security and allowing further customization.



Supporting handcrafted natural language instructions.

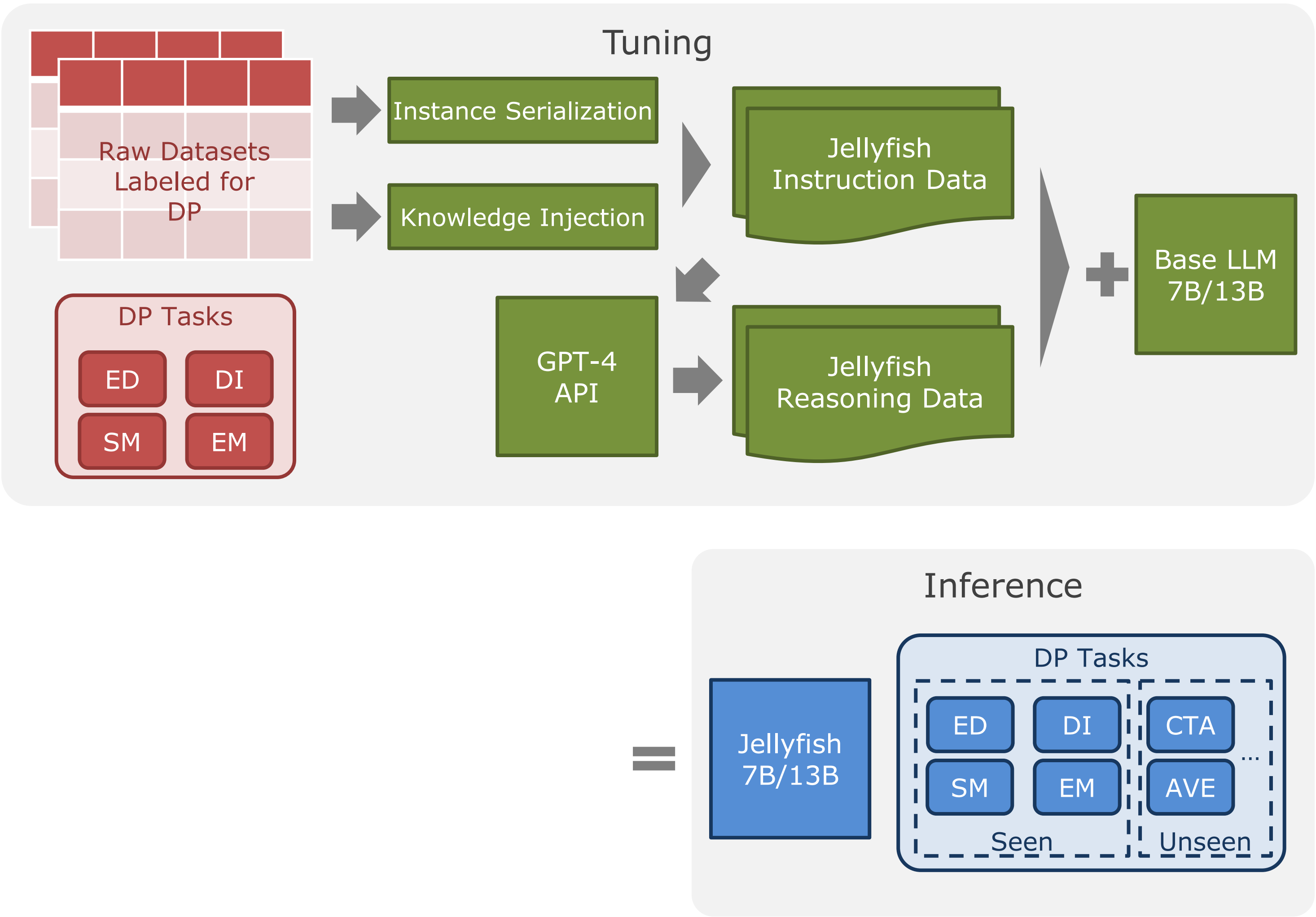
Built-in domain knowledge and optional knowledge specification.

Model interpreter for explaining its outputs.

Performance on a par with GPT-3.5/4.

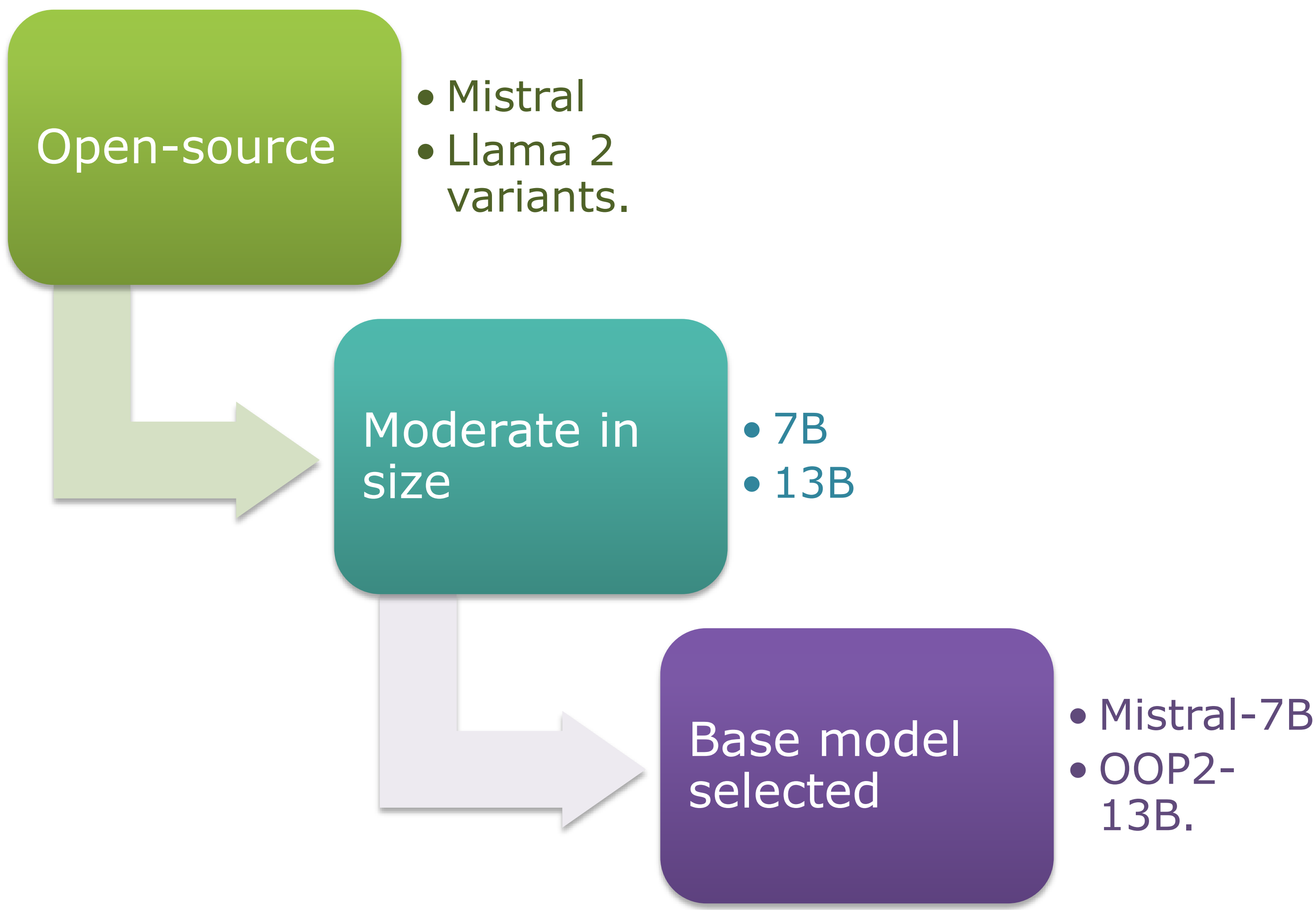
https://huggingface.co/NECOUDBFM/Jellyfish

Jellyfish Framework

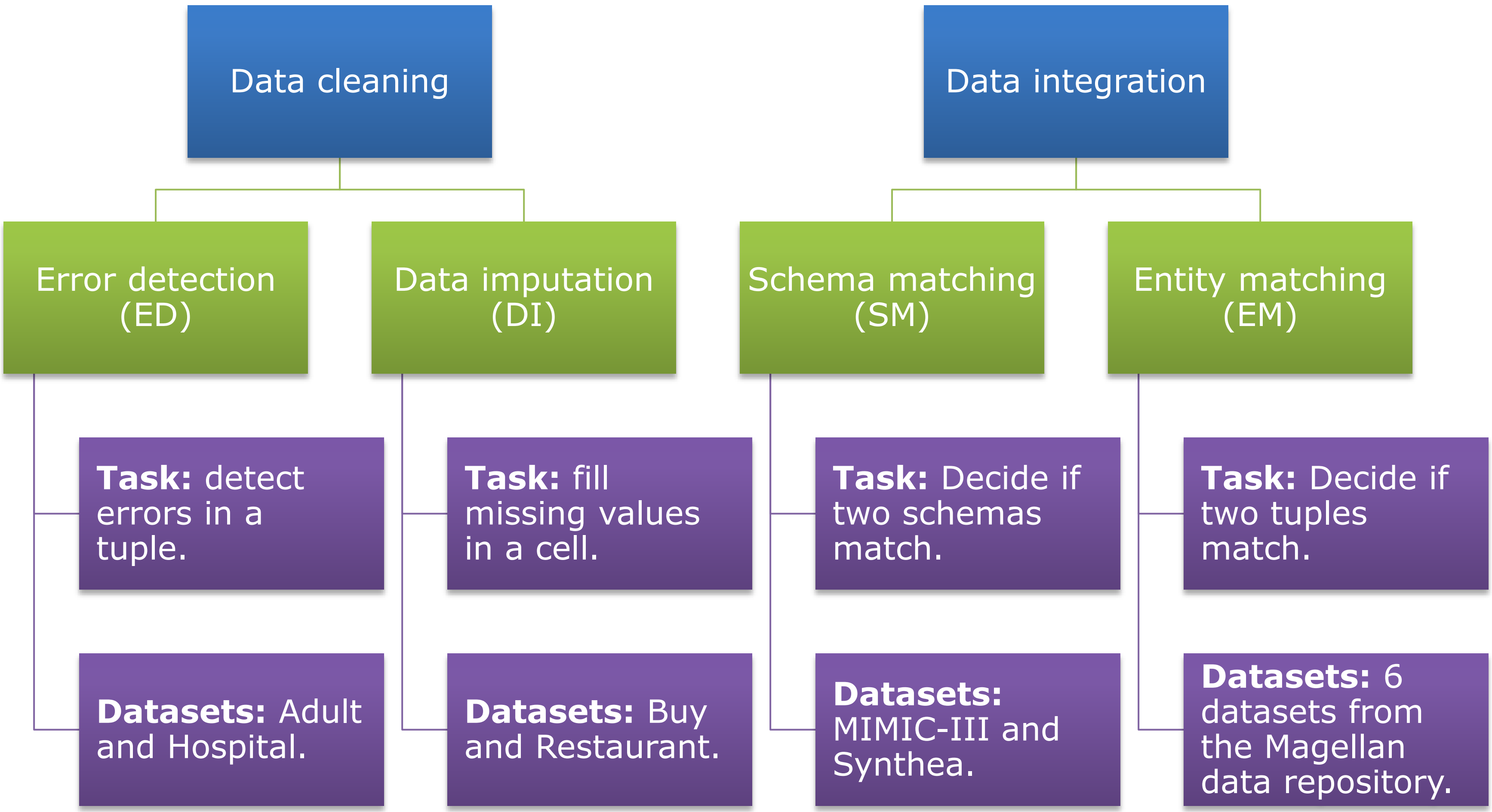


https://huggingface.co/NECOUDBFM/Jellyfish

Base Model Selection



Instruction Tuning



https://huggingface.co/NECOUDBFM/Jellyfish

Instruction Data Construction

Tuning the DP task solver

Instance serializer: transform each instance in the dataset (data element in a DP task) to a text sequence (namely the "**prompt**").

name	loc	# of employees	rev
NEC	JP	114,714	$26B

name: "NEC", loc: "JP", # of employees: 114714, rev: $26B

Knowledge injector: inject knowledge as prompt.

General knowledge: general DP rules (e.g., ignoring missing values).

Specific knowledge: knowledge specific to the task or the dataset (e.g., constraints and important features).

Knowledge can be generalized to <u>unseen datasets</u> within the same domain.

https://huggingface.co/NECOUDBFM/Jellyfish

Instruction Data Example

Beer dataset for entity matching
(comments in boldface)

(system message) You are an AI assistant that follows instruction extremely well. User will give you a question. Your task is to answer as faithfully as you can.
(task description) You are tasked with determining whether two Products listed below are the same based on the information provided. Carefully compare all the attributes before making your decision.
(injected knowledge) Note that missing values (N/A or "nan") should not be used as a basis for your decision.
(instance content) Product A: [name: "Sequoia American Amber Ale", factory: "Wig And Pen"]
Product B: [name: "Aarhus Cains Triple A American Amber Ale", factory: "Aarhus Bryghus"]
(question) Are Product A and Product B the same Product?
(output format) Choose your answer from: [Yes, No]

https://huggingface.co/NECOUDBFM/Jellyfish

Reasoning Data Construction

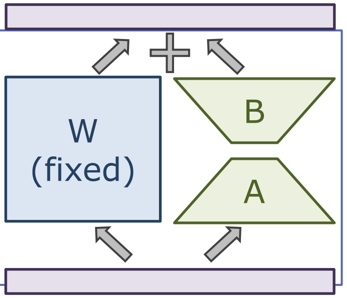
Tuning the model interpreter

Data collection

- Resort to GPT-4 to obtain reasonable answers.
- Jellyfish interpreter can be regarded as distilled from GPT-4's knowledge of DP tasks.

Model switch

- We tune models with low-rank adaptation (LoRA).
- Task solver X: $W + B_1A_1$.
- Interpreter Y: $W + B_2A_2 = X - B_1A_1 + B_2A_2$.



One model as both task solver and interpreter

- Mistral-7B is fine.
- Difficult to tune OOP2-13B to adequately perform both DP task-solving and interpretation.

https://huggingface.co/NECOUDBFM/Jellyfish

Reasoning Data Example

Beer dataset for entity matching

Reasoning Data

(system message) [same as Instruction Data] While answering, provide detailed explanation and justify your answer.
(task description - question) [same as Instruction Data]
(output format) After your reasoning, finish your response in a separate line with and ONLY with your final answer. Choose your final answer from [Yes, No].

Reasoning Ground Truth Collection

(system message - output format) [same as Reasoning Data]
(injected knowledge) Note that different factories can belong to the same parent company. The company name of Product B may occur in its product name.
(answer hint) You can use the "Hint" below, but your response cannot contain any information from it.
Hint: the final answer is "No"

https://huggingface.co/NECOUDBFM/Jellyfish

DP with Jellyfish Models

Instance Serialization

- Same as the prompts used for instruction tuning.

Feature Engineering

- Select beforehand.
- Specified in the prompt.

Prompt Engineering

- Few-shot prompting (optional, default = on)
- Knowledge injection (optional, default = off)

https://huggingface.co/NECOUDBFM/Jellyfish

Few-Shot Prompting

Condition the model to learn from a small selection of examples drawn from the dataset.

3 examples recommended, containing

Positives

Negatives

Few-shot examples can be handcrafted or automatically generated.

ED & DI: Randomly injected errors (missing values, typographical/formatting errors, and swapping values).

SM & EM: Rule-based methods for quickly finding matching instances.

Few-Shot Prompting Example

Beer dataset for entity matching

```
(system message - injected knowledge) [same as Instruction
Data]
(1st example's instance content) ### Instruction: Product
A: [name: "Shirt Tail Amber", factory: "Iron Hill Brewery &
Restaurant"]
Product B: [name: "Iron Hill Shirt Tail Amber", factory:
"Iron Hill Maple Shade"]
(1st example's question) Are Product A and Product B the
same Product?
(1st example's output format) Choose your answer from:
[Yes, No]
(1st example's answer) ### Response: Yes
(other examples) ...
(instance content - output format) [same as Instruction
Data] ### Response:
```

Experiments

Models

- Jellyfish-7B = Mistral-7B + instruction.
- Jellyfish-7B-I = Mistral-7B + instruction + reasoning.
- Jellyfish-13B = OOP2-13B + instruction.

Environment

- Tuning: NVIDIA RTX A6000 GPU 48GB x 2
- Inference: NVIDIA RTX A6000 GPU 48GB x 1

Seen tasks

Unseen tasks

Result interpretation

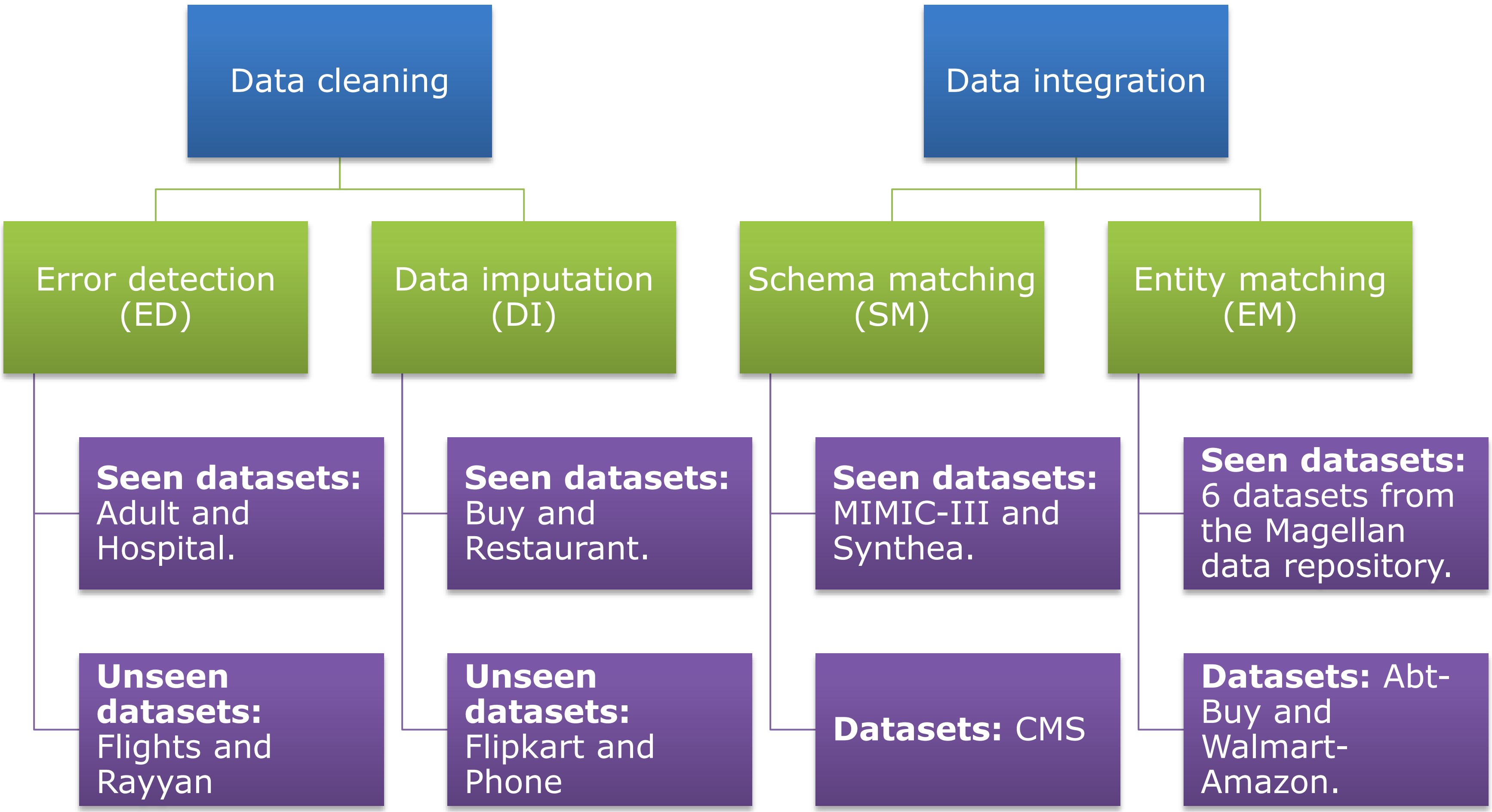
NLP performance

Ablation studies

Efficiency

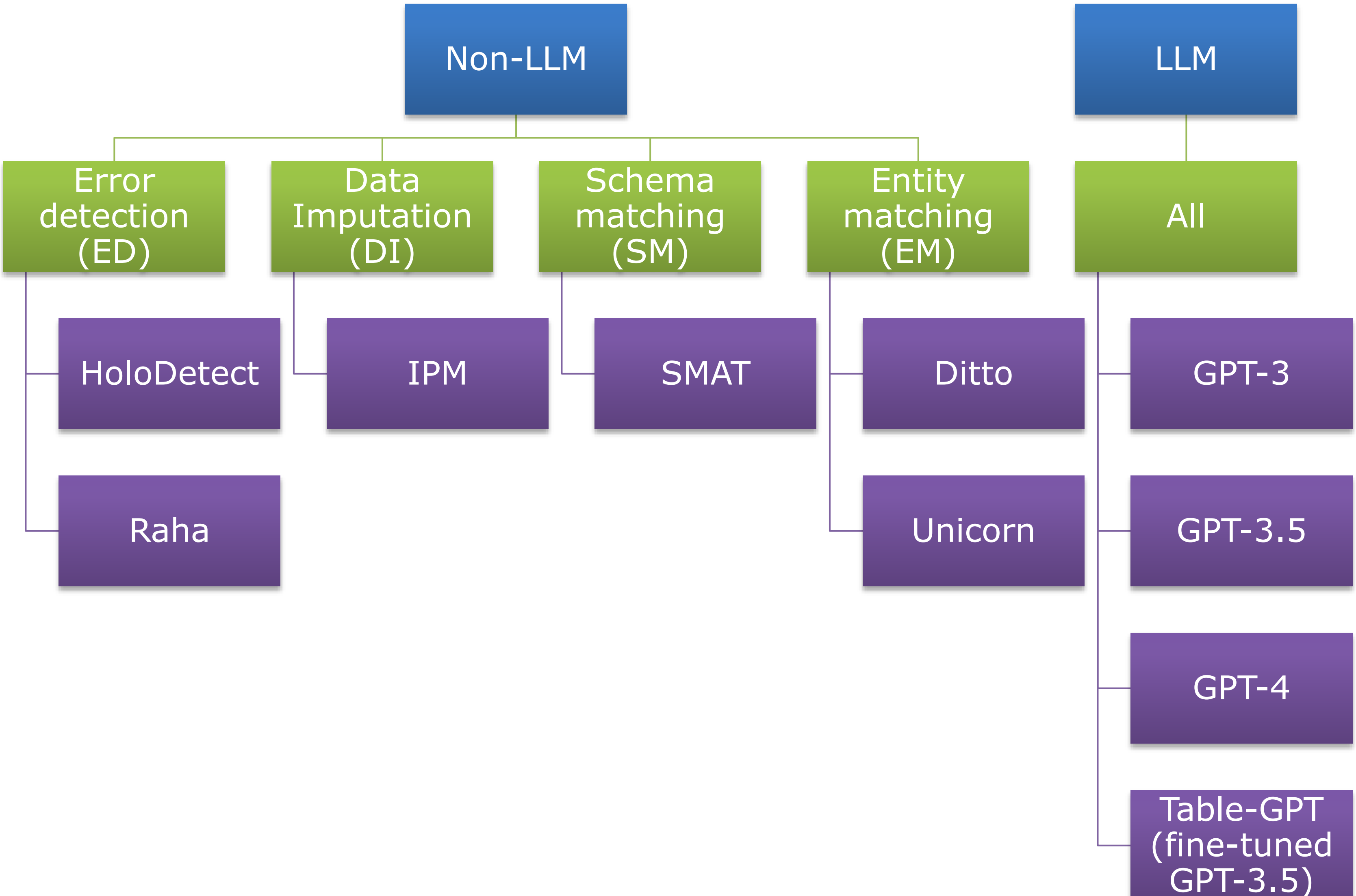
https://huggingface.co/NECOUDBFM/Jellyfish

Seen Tasks – Datasets



https://huggingface.co/NECOUDBFM/Jellyfish

Seen Tasks – Baseline Methods



https://huggingface.co/NECOUDBFM/Jellyfish

Seen Tasks – Performance

Task	Type	Dataset	Non-LLM (all seen)	GPT-3	GPT-3.5	Table-GPT	GPT-4	Jellyfish-7B	Jellyfish-7B-I	Jellyfish-13B
ED (F1)	Seen	Adult	<u>99.10</u>	<u>99.10</u>	92.01	–	92.01	94.70	91.96	**99.33**
		Hospital	94.40	**97.80**	90.74	–	90.74	95.09	<u>96.27</u>	95.59
	Unseen	Flights	81.00	–	–	–	**83.48**	65.30	66.92	<u>82.52</u>
		Rayyan	79.00	–	–	–	<u>81.95</u>	73.81	69.82	**90.65**
DI (Acc)	Seen	Buy	96.50	98.50	98.46	–	**100**	98.46	96.92	**100**
		Restaurant	77.20	88.40	<u>94.19</u>	–	**97.67**	86.05	88.37	89.53
	Unseen	Flipkart	68.00	–	–	–	**89.94**	<u>81.87</u>	79.44	81.68
		Phone	86.70	–	–	–	**90.79**	83.67	85.00	<u>87.21</u>
SM (F1)	Seen	MIMIC-III	20.00	–	–	–	<u>40.00</u>	**43.14**	<u>40.00</u>	<u>40.00</u>
		Synthea	38.50	45.20	<u>57.14</u>	–	**66.67**	55.55	44.44	56.00
	Unseen	CMS	<u>50.00</u>	–	–	–	19.35	20.00	13.79	**59.29**
EM (F1)	Seen	AMZN-Google	75.58	63.50	66.50	70.10	74.21	<u>81.29</u>	80.83	**81.34**
		Beer	94.37	**100**	96.30	96.30	**100**	96.30	96.55	96.77
		DBLP-ACM	**98.99**	96.60	96.99	93.80	97.44	98.54	98.88	<u>98.98</u>
		DBLP-GS	<u>95.70</u>	83.80	76.12	92.40	91.87	94.89	95.16	**98.51**
		Fodors-Zagats	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
		iTunes-Amazon	97.06	<u>98.20</u>	96.40	94.30	**100**	96.30	96.30	98.11
	Unseen	Abt-Buy	89.33	–	–	–	**92.77**	79.78	82.38	<u>89.58</u>
		WMT-AMZN	86.89	87.00	86.17	82.40	**90.27**	78.22	85.64	<u>89.42</u>

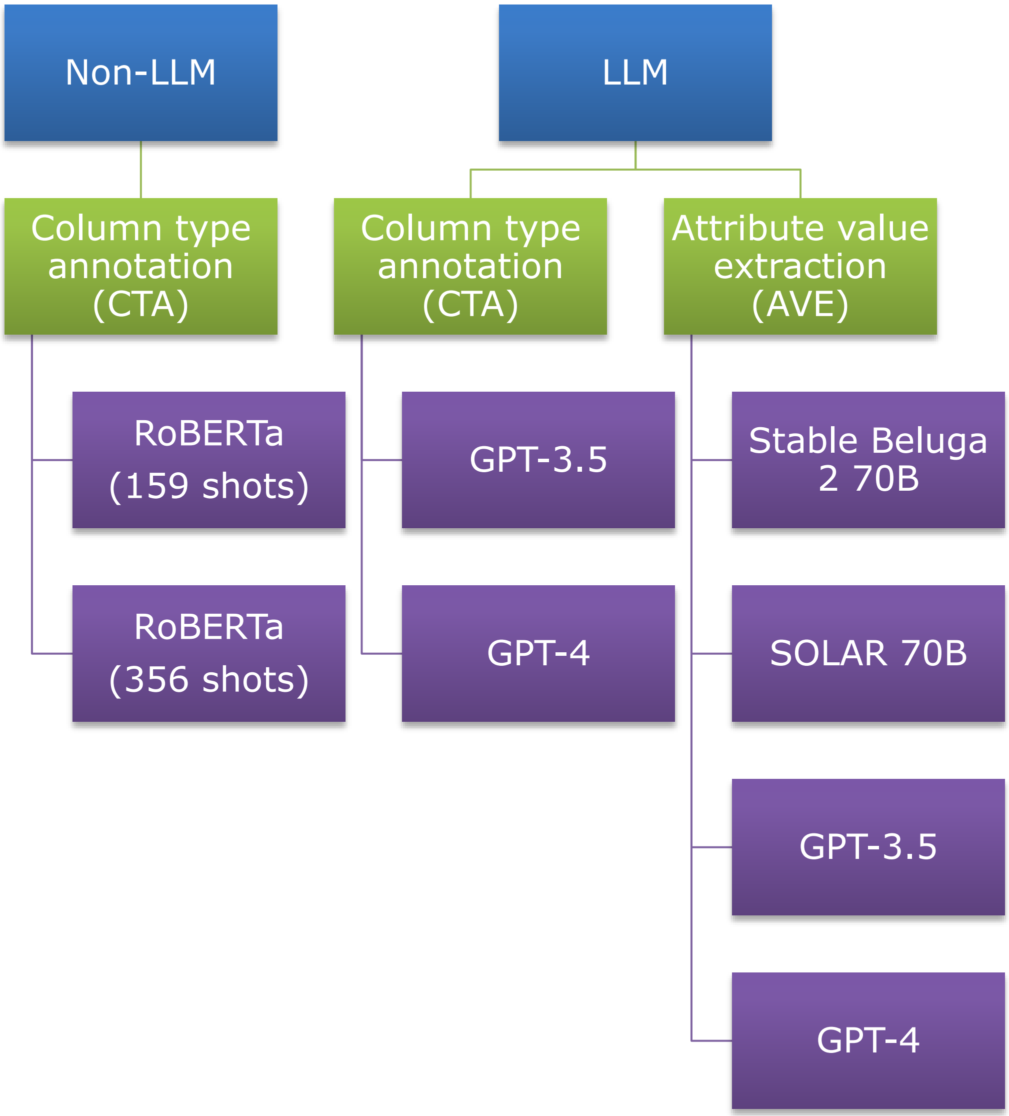
https://huggingface.co/NECOUDBFM/Jellyfish

Unseen Tasks & Datasets

- Table understanding
  - Column type annotation (CTA)
    - **Task:** infer the type of each column
    - **Datasets:** SOTAB
    - **Prompts:** Follow a study on using GPT-3.5 & GPT-4
- Information extraction
  - Attribute value extraction (AVE)
    - **Task:** extract attribute values from a text description
    - **Datasets:** AE-110k and OA-Mine
    - **Prompts:** Follow a study on using GPT-3.5 (tuned) & GPT-4.

https://huggingface.co/NECOUDBFM/Jellyfish

Unseen Tasks – Baseline Methods



https://huggingface.co/NECOUDBFM/Jellyfish

Unseen Tasks – Performance

Column Type Annotation (micro-F1)

Dataset	RoBERTa 159 shots	RoBERTa 356 shots	GPT-3.5	GPT-4	Jellyfish-7B	Jellyfish-7B-I	Jellyfish-13B
SOTAB	79.20	<u>89.73</u>	89.47	**91.55**	83.54	80.89	82.00

Attribute Value Extraction (F1)

Dataset	Stable Beluga 2 70B	SOLAR 70B	GPT-3.5	GPT-4	Jellyfish-7B	Jellyfish-7B-I	Jellyfish-13B
AE-110k	52.10	49.20	61.30	55.50	<u>74.17</u>	**76.85**	58.12
OA-Mine	50.80	55.20	62.70	68.90	<u>75.35</u>	**76.04**	55.96

https://huggingface.co/NECOUDBFM/Jellyfish

Result Interpretation

Head-to-head comparison with GPT-3.5, judged by GPT-4.

Task	Dataset	GPT-3.5	Jellyfish-7B-I
ED	Adult	**11**	9
	Hospital	9	**11**
DI	Buy	0	**20**
	Restaurant	**10**	**10**
SM	Synthea	8	**12**
EM	Amazon-Google	8	**12**
	Beer	7	**13**
	DBLP-ACM	8	**12**
	DBLP-GoogleScholar	4	**16**
	Fodors-Zagats	**12**	8
	iTunes-Amazon	**19**	1
Total		96	**124**
Winning Rate		43.64%	**56.36%**

NLP Performance

Model	MMLU	Wino Grande	ARC	Truthful QA	GSM8K	Hella Swag	Average
OOP2-13B	54.49	74.03	62.63	52.56	25.32	83.24	58.71
Jellyfish-13B (OOP2-13B + Jellyfish)	53.04 (-1.45)	74.19 (+0.16)	62.88 (+0.25)	52.56 (+0.00)	24.26 (-1.06)	83.16 (-0.08)	58.35 (-0.36)
Mistral-7B	62.91	73.88	63.48	66.91	41.32	84.79	65.55
Jellyfish-7B (Mistral-7B + Jellyfish)	62.08 (-0.83)	72.69 (-1.19)	63.48 (+0.00)	64.76 (-2.15)	37.91 (-3.41)	84.48 (-0.31)	64.23 (-1.32)

https://huggingface.co/NECOUDBFM/Jellyfish

Impact of Instruction Tuning

Task	Dataset	Llama 2 13B	Llama 2 13B + Orca	Llama 2 13B + Platypus	OOP2 13B	Llama 2 13B + DP-Preview	Llama 2 13B + Orca + DP-Preview	Llama 2 13B + Platypus + DP-Preview	OOP2 13B + DP-Preview
ED (F1)	Adult	5.92	33.67	7.73	**42.77**	93.62	93.49	93.49	**96.62**
	Hospital	8.78	64.05	6.29	**63.24**	81.55	89.67	90.58	**92.01**
DI (Acc)	Buy	**95.38**	75.38	41.54	89.23	92.31	90.77	87.69	**100**
	Restaurant	**90.70**	88.37	86.05	81.40	89.53	**90.70**	88.37	89.53
SM (F1)	Synthea	0.97	0.00	0.68	**22.22**	22.22	22.22	28.57	**36.36**
EM (F1)	Amazon-Google	14.58	25.62	25.64	**36.70**	40.00	**49.77**	42.35	48.20
	Beer	39.13	81.48	11.76	**85.71**	95.55	93.33	93.33	**96.55**
	DBLP-ACM	45.95	78.84	0.00	**78.86**	97.45	**97.66**	97.35	97.35
	DBLP-GoogleScholar	35.71	56.07	40.73	**59.48**	92.27	92.22	**92.87**	92.83
	Fodors-Zagats	42.86	84.21	39.56	**92.68**	97.67	**100**	**100**	**100**
	iTunes-Amazon	30.43	63.53	0.00	**57.45**	96.15	96.15	96.15	**96.30**

https://huggingface.co/NECOUDBFM/Jellyfish

Impact of Knowledge Injection

Entity Matching (F1)

Type	Dataset	OOP2-13B	OOP2-13B + EM-Preview w/o Knowledge	OOP2-13B + EM-Preview w/ Knowledge
Seen	Amazon-Google	36.70	47.54	**50.53**
	Beer	85.71	85.71	**92.86**
	DBLP-ACM	78.86	85.33	**90.26**
	DBLP-GoogleScholar	59.48	90.46	**91.54**
	Fodors-Zagats	92.68	**100**	**100**
	iTunes-Amazon	57.45	98.11	**98.18**
Unseen	Abt-Buy	61.78	83.35	**84.44**
	Walmart-Amazon	67.29	71.71	**73.18**

https://huggingface.co/NECOUDBFM/Jellyfish

Efficiency

Fine-tuning

- Jellyfish-13B: 5 hours
- Jellyfish-7B: 2.5 hours
- Jellyfish-7B-I: 3.5 hours

Inference

- Jellyfish-13B: 0.08 – 0.15 sec/instance.
- GPT-4: 1 – 8 sec/instance.

https://huggingface.co/NECOUDBFM/Jellyfish

Conclusions

- We developed Jellyfish, a dataset for instruction-tuning LLMs for DP.
- Models are tuned for 4 DP tasks and generalize to unseen tasks as a universal DP task solver.
- With 7B/13B parameters that can be further tuned, the models can operate on a local GPU without compromising data security.
- The tuned models understand natural language instructions, enabling users to craft prompts.
- The tuned models acquire domain knowledge during its tuning process and employ optional knowledge injection during inference.
- The dataset features reasoning data used for tuning a model interpreter that provides explanations for the model’s outputs.
- Our evaluation demonstrated the competitiveness, generalizability, and reasoning abilities of the tuned models, and justified the usefulness of the techniques in building the models.

https://huggingface.co/NECOUDBFM/Jellyfish

Future Works

- Data from a wider range of domains.
- More DP tasks (e.g., data repairing and text wrangling).
- Quantized or distilled model for faster speed.
- Multi-agent system for code-free, adaptable, conversational DP pipeline.

Model Released at Hugging Face

https://huggingface.co/NECOUDBFM/Jellyfish