

# Seminar: Table Representation Learning (TRL)

## Introduction

14.10.2024

Francesco Pugnaroni, Lukas Laskowski,  
Christoph Hönes

**Design IT.  
Create Knowledge.**

[www.hpi.de](http://www.hpi.de)



# Agenda



- Goals of the seminar
- Introduction to Table Representation Learning
- Presentation of topics
- Organizational stuff

# Goals of the seminar



- Learning goals:
  - Reading and writing research papers
  - Conducting scientific experiments and presenting their results
- Team activities:
  - Identifying a research question given a topic and relevant papers
  - Running experiments to answer the research question
  - Writing a report
- Deliverables for each team:
  - Paper-style technical report
  - Code, models, and datasets produced
  - Midterm and final presentations

# Introduction - Representation learning

- Representation Learning (RL) aims to discover meaningful representations of objects in different modalities to make them easier to process or understand.
- A representation is a vector with a variable number of dimensions, i.e., an embedding.
- Possible modalities are:
  - Images
  - Text
  - Structured data, i.e., tables
  - ...

The modern **jaguar**'s ancestors probably entered the Americas from Eurasia during the Early Pleistocene via the land bridge that once spanned the Bering Strait.



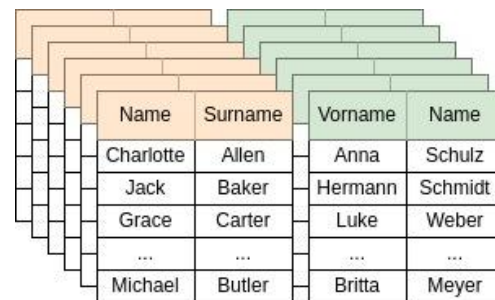
The **Jaguar** 420 (pronounced "four-twenty") and its Daimler Sovereign equivalent were introduced at the October 1966 London Motor Show

AIR Index ratings for JLR cars (Euro 6 allows up to 80 mg/km NOx emission)

Model	Year	Fuel	NOx measured (mg/km)	AIR Index rating
<a href="#">Jaguar E-Pace HSE 2.0i 180 hp</a>	2019	Diesel	14	A
<a href="#">Land Rover Range Rover Evoque TD4 2.0i 180 hp</a>	2019	Diesel	17	A
<a href="#">Land Rover Discovery 3.0 TD6 HSE</a>	2018	Diesel	33	A
<a href="#">Land Rover Discovery Sport 2.0i 180 hp</a>	2019	Diesel	34	A

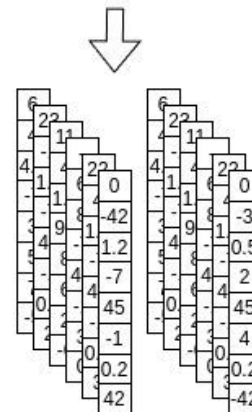
# Introduction - Table representation learning

- Table Representation Learning (TRL) aims to discovery meaningful representations of tabular data.



Name	Surname	Vorname	Name
Charlotte	Allen	Anna	Schulz
Jack	Baker	Hermann	Schmidt
Grace	Carter	Luke	Weber
...	...	...	...
Michael	Butler	Britta	Meyer

- Can be exploited for solving most of the standard table-related tasks.



- Table retrieval
- Duplicate detection
- Schema matching
- Entity resolution
- Table-to-text summarization
- Retrieval Augmented Generation
- Table matching
- ...

Tabular data can be embedded at different granularities:

- Cells
- Entities
- Rows
- Columns
- Captions
- Metadata
- Tables
- Databases
- Rows inside .csv files
- ...

Caption

The table below summarizes some of the most important relational database terms and the corresponding SQL term:

SQL term	Relational database term	Description
<i>Row</i>	<i>Tuple</i> or <i>record</i>	A data set representing a single item
<i>Column</i>	<i>Attribute</i> or <i>field</i>	A labeled element of a tuple, e.g. "Address" or "Date of birth"
<i>Table</i>	<i>Relation</i> or <i>Base relvar</i>	A set of tuples sharing the same attributes; a set of columns and rows
<i>View</i> or <i>result set</i>	<i>Derived relvar</i>	Any set of tuples; a data report from the RDBMS in response to a <i>query</i>

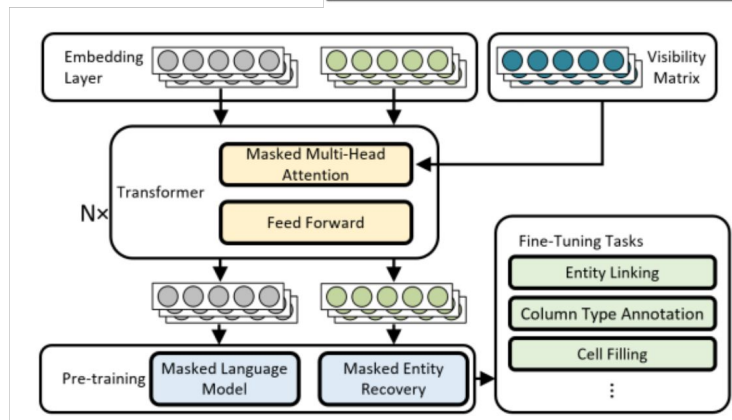
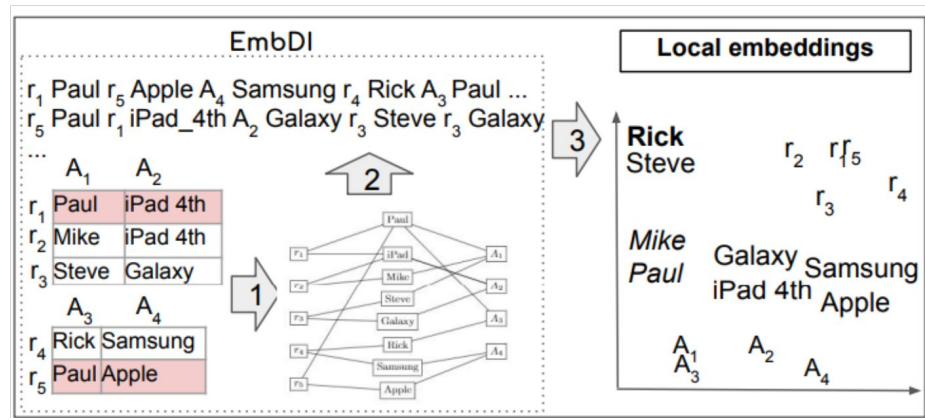
Column      Cell      Table      Entities

# Introduction - Table representation learning - Technologies



And it can be done using different technologies:

- Transformers
- Graphs
- Word embeddings
- ...



- Transformers for Tabular Data Representation: A Survey of Models and Applications [\[1\]](#)
- Observatory: Characterizing Embeddings of Relational Tables [\[2\]](#)
- Revisiting Deep Learning Models for Tabular Data [\[3\]](#)

[1] Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. 2023. Transformers for Tabular Data Representation: A Survey of Models and Applications. Trans. Assoc. Comput. Linguistics 11, (2023), 227–249.

[2] Tianji Cong, Madelon Hulsebos, Zhenjie Sun, Paul Groth, and H. V. Jagadish. 2023. Observatory: Characterizing Embeddings of Relational Tables. Proc. VLDB Endow. 17, 4 (2023), 849–862.

[3] Yury Gorishniy, Ivan Rubachev, Valentin Khulkov, and Artem Babenko. 2021. Revisiting Deep Learning Models for Tabular Data. In NeurIPS, 18932–18943.



# Seminar research topics

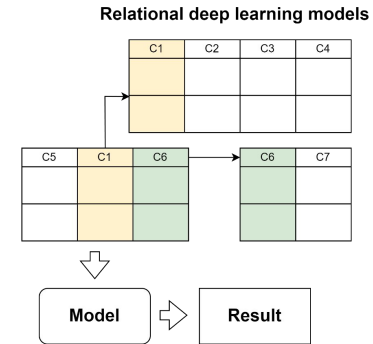
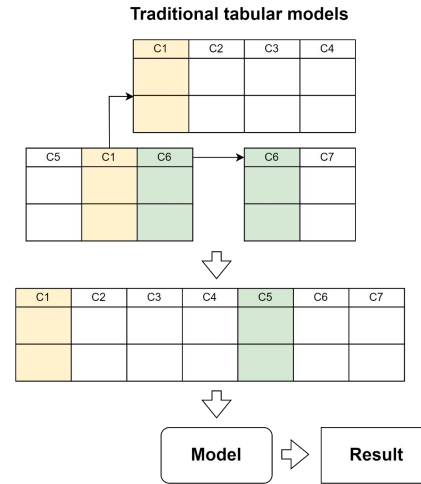


Each group of two people will work on one of these topics:

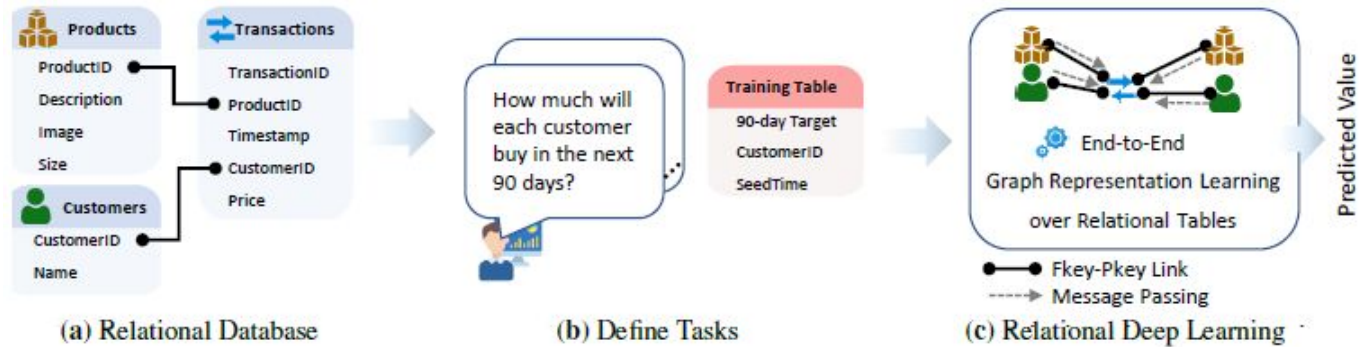
- Relational deep learning
- Data cleaning
- Natural language table management through LLMs

# Relational Deep Learning (RDL) - Intuition

- Problem: most of tabular learning approaches can process at most a table at once
- Consequence: when working with relational database it is necessary to (1) join the tables into a single one and to (2) train a model on that lose signals such as foreign case, relative size of the tables, and belonging of rows and columns to different tables.
- RDL is *“A blueprint for an end-to-end deep learning paradigm for relational tables”*



# Relational Deep Learning (RDL) - Pipeline



- Given a database and a task, to obtain a working RDL framework:
  - Generate training samples using historical data
  - Build a graph that preserve the structure of the database
  - Train a model

# Relational Deep Learning (RDL) - Papers



## RDL:

- Position: Relational Deep Learning - Graph Representation Learning on Relational Databases [\[1\]](#)
- RELBENCH: A Benchmark for Deep Learning on Relational Databases [\[2\]](#)

## Graph embeddings:

- GraphSAGE: Inductive Representation Learning on Large Graphs [\[3\]](#)
- node2vec: Scalable Feature Learning for Networks [\[4\]](#)

## Tabular learning:

- PyTorch Frame: A Modular Framework for Multi-Modal Tabular Learning [\[5\]](#)

[1]Matthias Fey, Weihua Hu, Kexin Huang, Jan Eric Lenssen, Rishabh Ranjan, Joshua Robinson, Rex Ying, Jiaxuan You, and Jure Leskovec. 2024. Position: Relational Deep Learning - Graph Representation Learning on Relational Databases. In ICML, OpenReview.net.

[2]Joshua Robinson, Rishabh Ranjan, Weihua Hu, Kexin Huang, Jiaqi Han, Alejandro Dobles, Matthias Fey, Jan Eric Lenssen, Yiwen Yuan, Zecheng Zhang, Xinwei He, and Jure Leskovec. 2024. RelBench: A Benchmark for Deep Learning on Relational Databases. CoRR abs/2407.20060, (2024).

[3]William L. Hamilton, Zitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In NIPS, 1024–1034.

[4]Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In KDD, ACM, 855–864.

[5]Weihua Hu, Yiwen Yuan, Zecheng Zhang, Akihiro Nitta, Kaidi Cao, Vid Kocijan, Jure Leskovec, and Matthias Fey. 2024. PyTorch Frame: A Modular Framework for Multi-Modal Tabular Learning. CoRR abs/2404.00776, (2024).

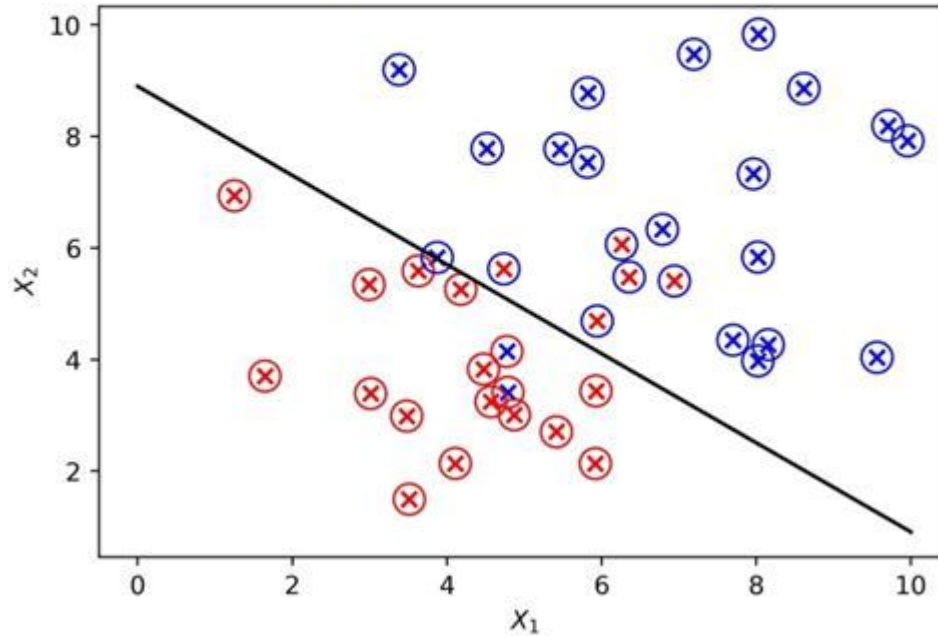
**Process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in data to improve its quality and usability.**

- Duplicate Detection
- Schema Matching
- Error Detection
- Data Imputation

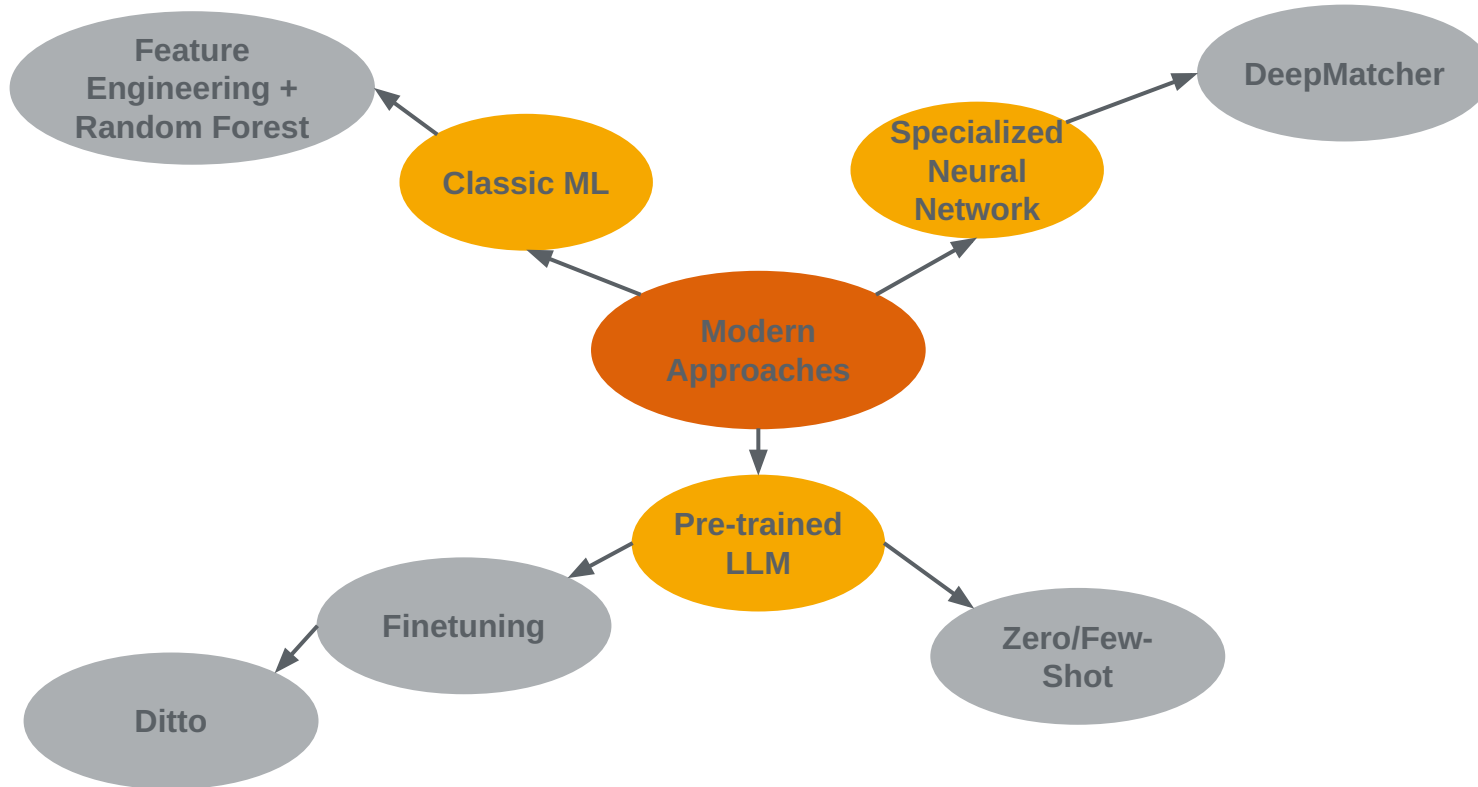
Modern solutions rely on Large Language Models (finetuned or zero-shot)

# Duplicate Detection

Treat Duplicate Detection as an imbalance binary classification task

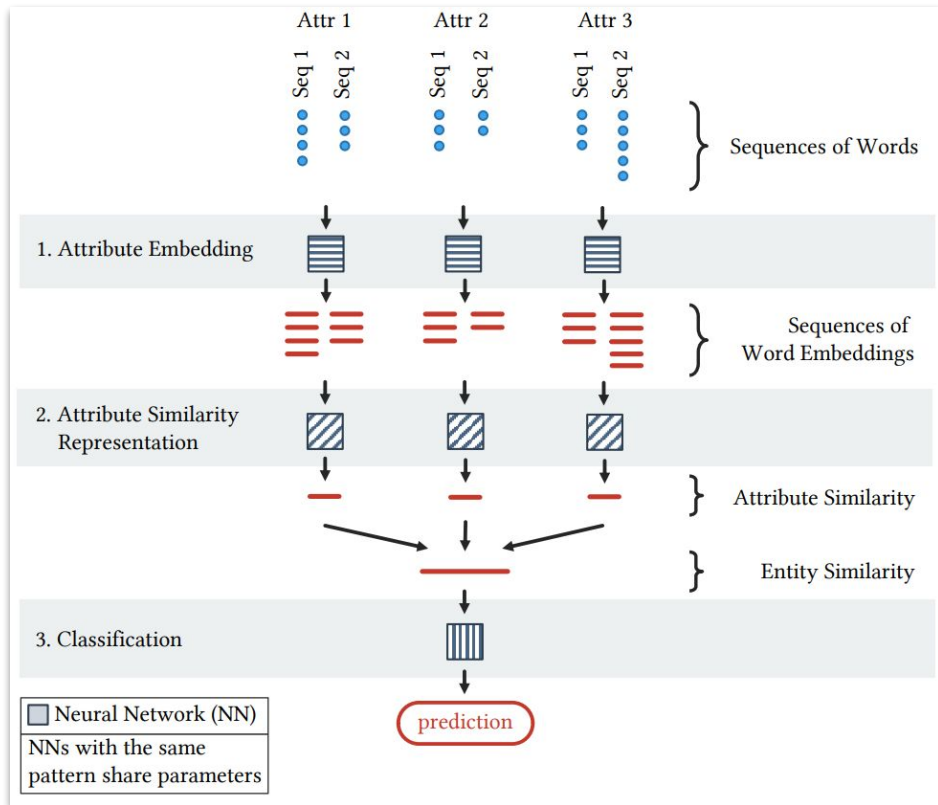


# Duplicate Detection

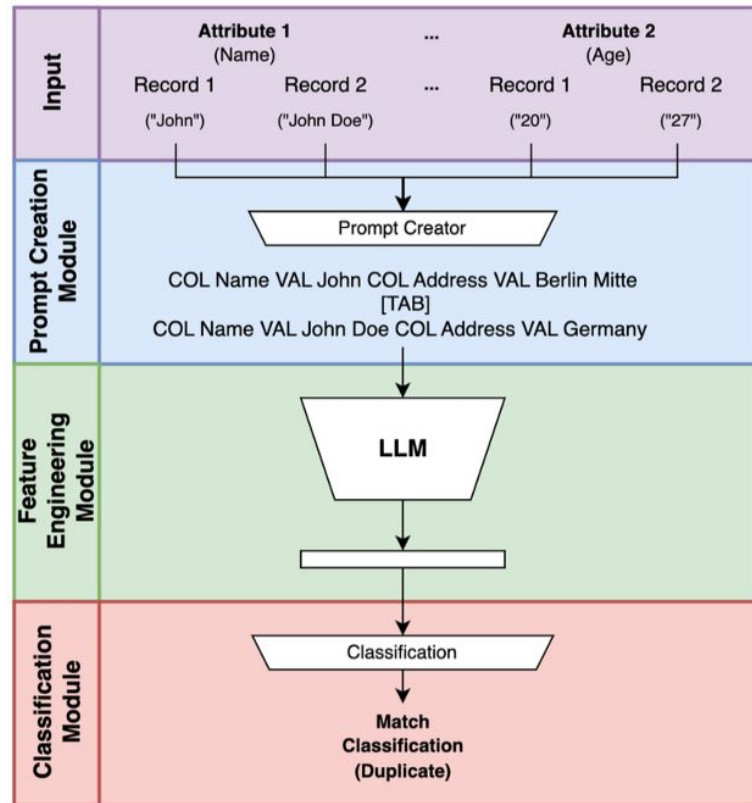


# Duplicate Detection

## Deep Matcher



## Ditto



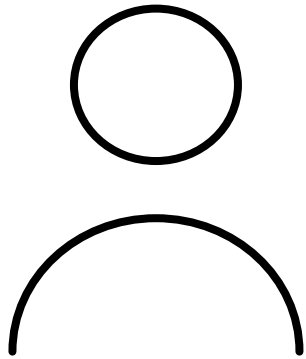


- EmbDI: <https://dl.acm.org/doi/pdf/10.1145/3318464.3389742>
- Ditto: <https://www.vldb.org/pvldb/vol14/p50-li.pdf>
- EM with contrastive loss: <https://arxiv.org/abs/2202.02098>
- Analysis of pre-trained embeddings: <https://www.vldb.org/pvldb/vol16/p2225-skoutas.pdf>
- Data Wrangling with LLMs: <https://www.vldb.org/pvldb/vol16/p738-narayan.pdf>

# (L)LMs as Natural Language Interface for Tables & DBs

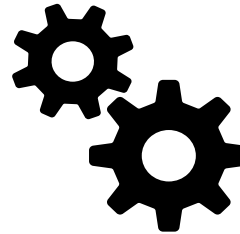
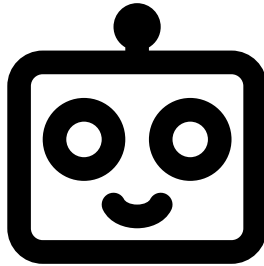


(Non-Tech)  
User



Conversation  
in Natural  
Language

LLM



SQL  
Python  
End2End

External Data



# Use Cases



- Fact Checking / Knowledge Grounding
- Table Question Answering
- Automated Data Analysis / Data Engineering
- Data Explanation (Data2Text, Data2Viz)

# Approaches



- 1. End-to-End Table Language Models**
2. Text-to-SQL translation (semantic parsing)
3. Agentic LLM Pipelines

# TAPAS (not quite End-to-End)

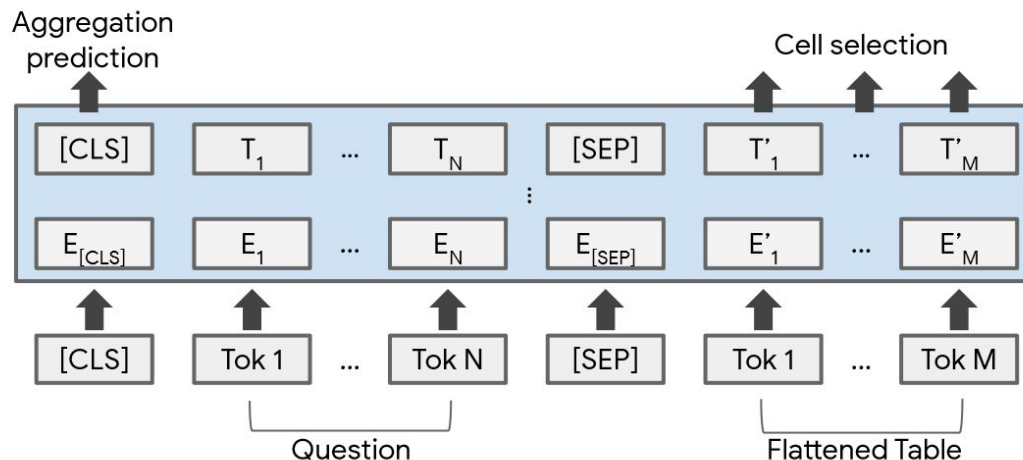


- BERT-style encoder-only model
- Linearize table into a text string
- Operator Classification
- Binary cell selection
- Limited operator set and context length

op	$P_a(op)$	compute(op, $P_s, T$ )
NONE	0	-
COUNT	0.1	.9 + .9 + .2 = 2
SUM	0.8	.9×37 + .9×31 + .2×15 = 64.2
AVG	0.1	64.2 ÷ 2 = 32.1

$$s_{pred} = .1 \times 2 + .8 \times 64.2 + .1 \times 32.1 = 54.8$$

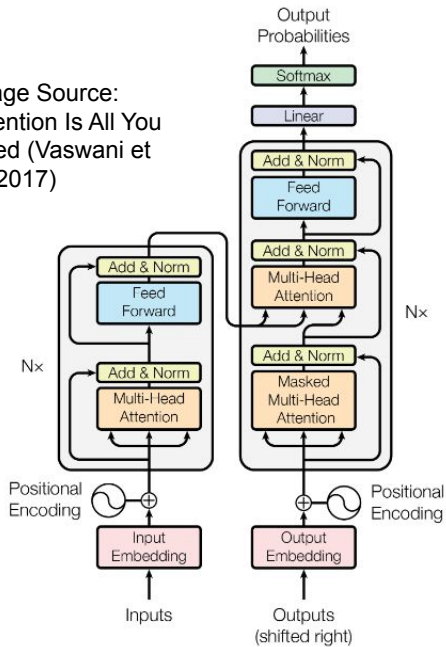
Rank	...	Days	$P_s$
1	...	37	0.9
2	...	31	0.9
3	...	17	0
4	...	15	0.2
...	...	...	0



# Representing Tables as Text: Example TAPEX



Image Source:  
Attention Is All You Need (Vaswani et al. 2017)



player	year	round	result	opponent
1 raymond van barneveld	2009	quarter-final	won	jelle klasen
2 reymond van barneveld	20120	2nd round	won	brendan dolan
3 adrian lewis	2011	final	won	gary anderson

Who are the only players listed that played in 2011 ? [HEAD] player | year | round | result | opponent [ROW] 1 ray mond van bar ne ve ld | 2009 | quarter - final | won | j elle k la as en [ROW] 2 ray mond van bar ne ve ld | 2010 | 2 nd round | won | bre nd an d olan [ROW] 3 ad rian le w is | 2011 | final | won | g ary and erson

## Encoder-Decoder Transformer (BART)

Figure 4: The visualization results of attention weights from other tokens to the cell “adrian lewis”. Intuitively, the darker the color, the more closely the word is associated with “adrian lewis”.

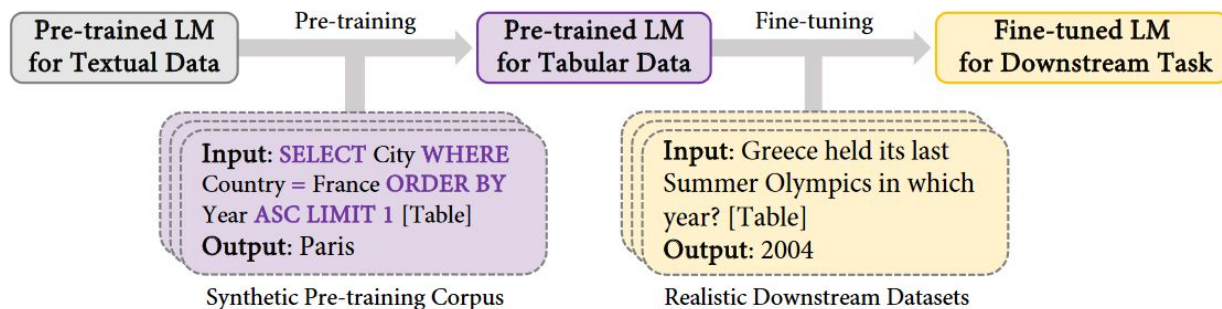


Figure 1: The schematic overview of our method. For the sake of brevity, the table content in the input is simplified with the symbol [Table].

- Start from pre-trained (text) LM (BART)
- Pre-train further on simulating a SQL execution engine
- Fine-tune on NL Question Answering (implicitly transfer SQL understanding to NL queries )

# Approaches



1. End-to-End Table Language Models
- 2. Text-to-SQL translation (semantic parsing)**
3. Agentic LLM Pipelines

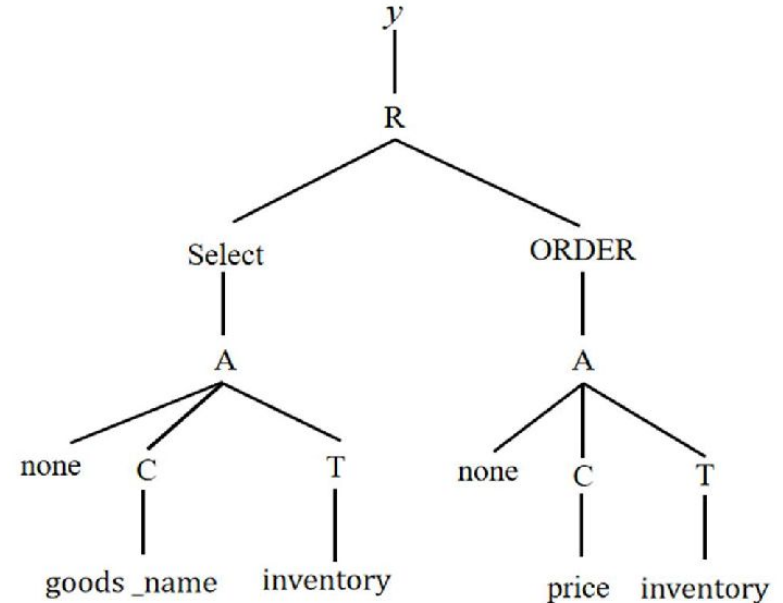


# Types of Text-to-SQL techniques



- Rule-based
- Template-based
- Considering only schema only vs. table-text attention
- Constrained decoding (e.g. grammar guided)
- Chain of Thought (CoT) LLM prompting for complex multi-turn questions

Question: List the names of goods in inventory sorted by price.  
SQL: `SELECT goods _name FROM inventory ORDER BY price`



- Exploit semantic column names
- Use Filtering & Ranking of schema items
- SQL skeleton as syntactic prior
- Fill in the slots of in the skeleton
- Execution-guided beam search decoding

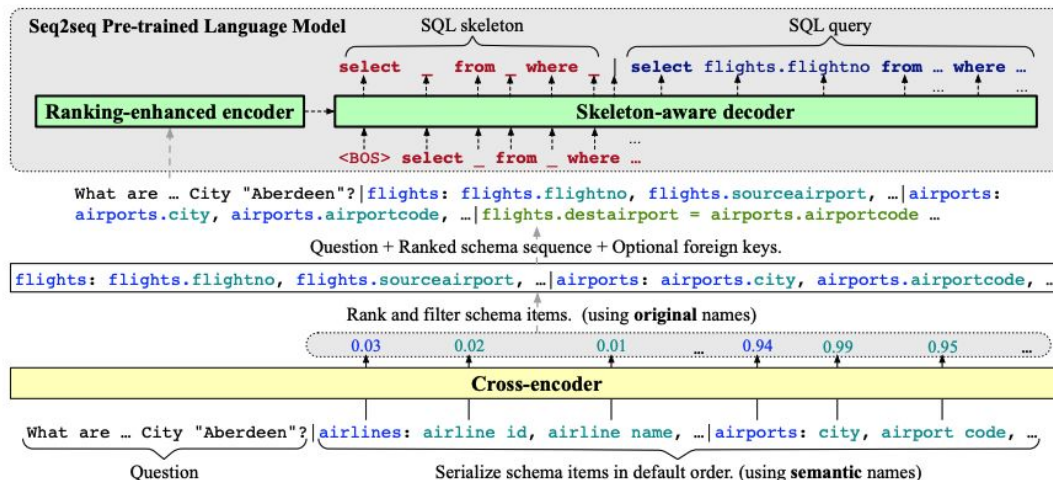


Figure 2: An overview of the ranking-enhanced encoding and skeleton-aware decoding framework. We train a cross-encoder for classifying the schema items. Then we take the question, the ranked schema sequence, and optional foreign keys as the input of the ranking-enhanced encoder. The skeleton-aware decoder first decodes the SQL skeleton and then the actual SQL query.

# Constrained Decoding: Example PICARD

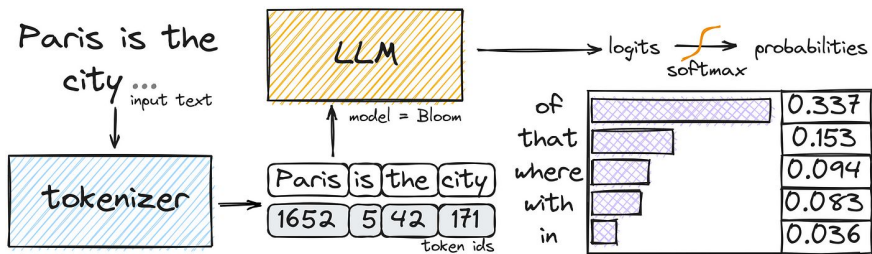


Image Source:

<https://pub.towardsai.net/how-does-an-llm-generate-text-fd9c57781217>

- Works well with beam search decoding
- Constrained decoding with different checks (lexical, syntactic parsing, semantic guards)
- Tradeoff between different levels of constraint complexity and execution time
- Only applied at inference, no additional training required, architecture agnostic

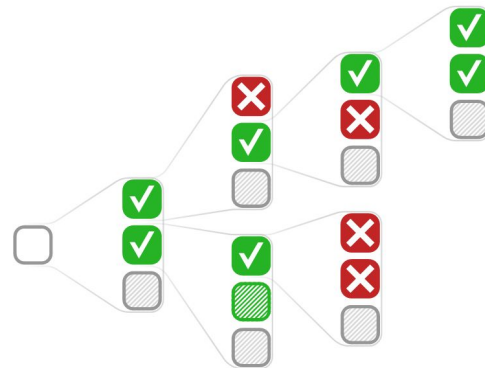


Figure 2: Illustration of constrained beam search with beam size 2 and PICARD. Each vertical column represents three token predictions for a hypothesis from top to bottom in descending order by probability. In this example, PICARD is configured to only check the top-2 highest ones. The rest is automatically dismissed by setting their score to  $-\infty$ . Tokens rejected by PICARD (red,  $\times$ ) are also assigned a score of  $-\infty$ . Accepted tokens (green,  $\checkmark$ ) keep their original score.

# Approaches



1. End-to-End Table Language Models
2. Text-to-SQL translation (semantic parsing)
- 3. Agentic LLM Pipelines**

# What are LLM Agents?

- Multi-step LLM prompting with some additional assets
- Complex sequential reasoning
- Handling and execution of external resources (similarities to RAG)

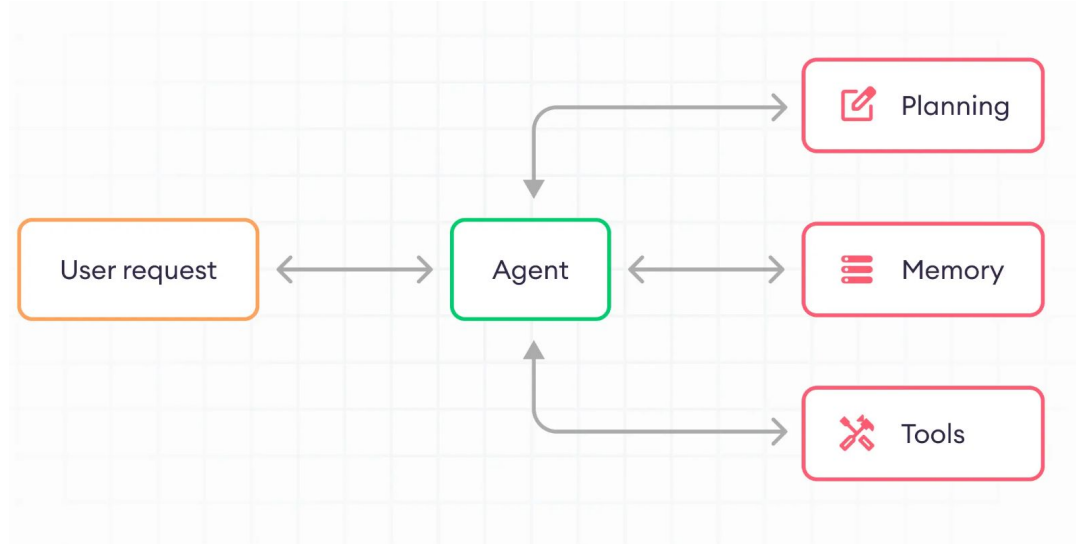
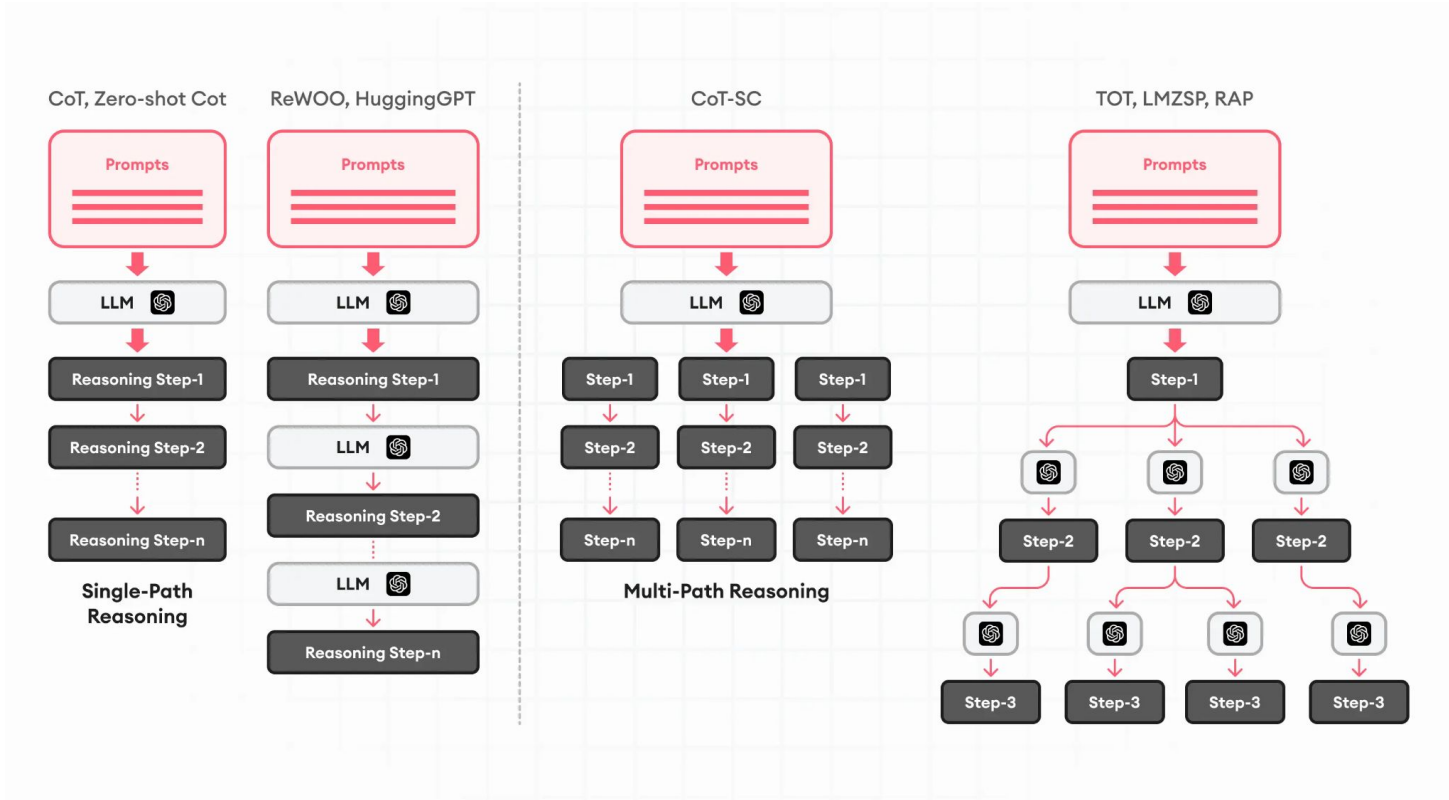


Image source: <https://www.superannotate.com/blog/llm-agents>

# Different reasoning strategies



# Example: TaPERA

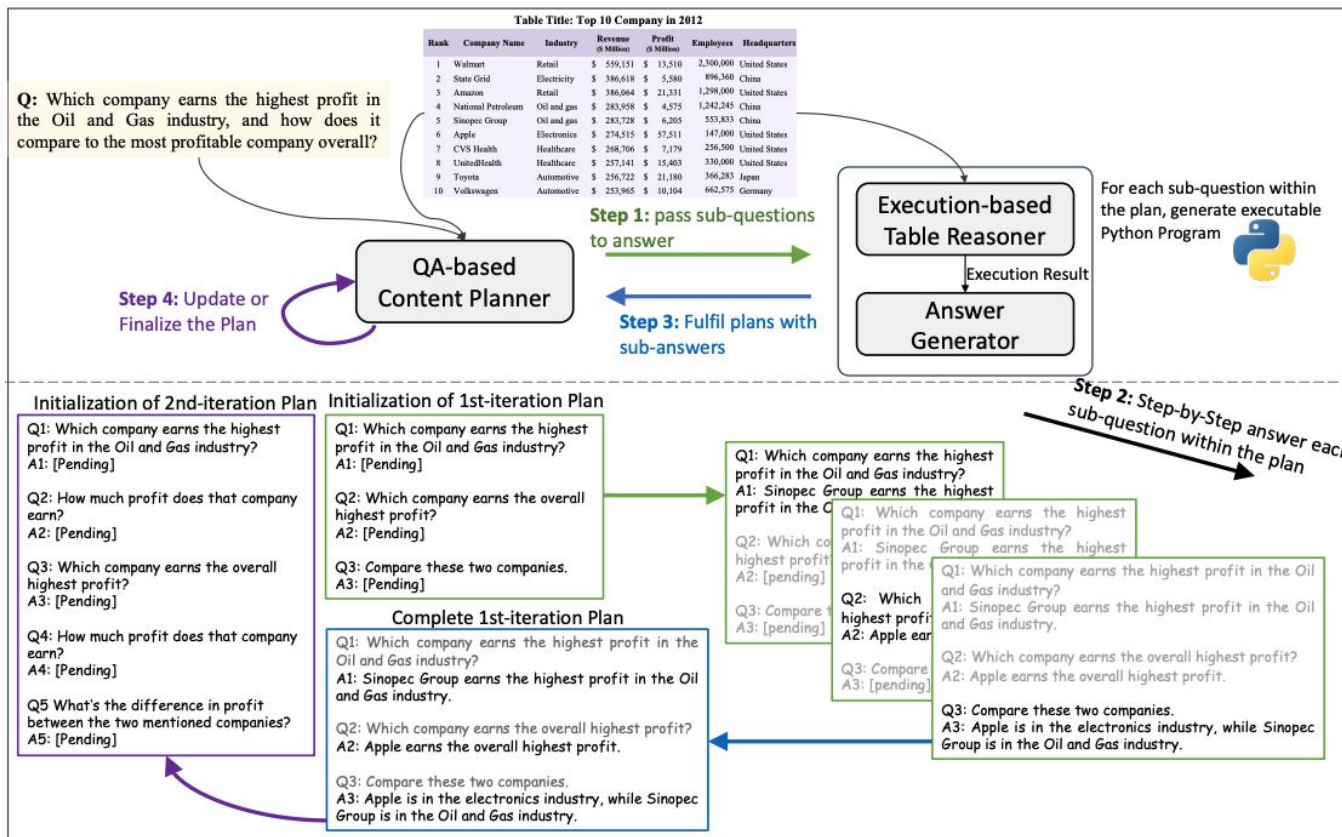


Figure 2: An illustration of TAPER. Top: The workflow of our modular framework. The QA-based content planner first generates a plan with QA-pairs, then Step 1-3 fills in the plan with sub-answers, and Step 4 finalizes or refines the plan. Bottom: A running example for the first iteration of the QA-based plan generation and refinement.

## End-to-End Table QA:

- TaPas: Weakly Supervised Table Parsing via Pre-training (<https://aclanthology.org/2020.acl-main.398>) (Herzig et al., ACL 2020)
- TAPEX: Table Pre-training via Learning a Neural SQL Executor (<https://arxiv.org/abs/2107.07653>) (Liu et al., ICLR 2022)

## Text-to-SQL:

- RESDSQL: Decoupling Schema Linking and Skeleton Parsing for Text-to-SQL (<https://arxiv.org/abs/2302.05965>) (Li et al., AAAI 2023)
- PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models (<https://arxiv.org/abs/2109.05093>) (Scholak et al., EMNLP 2021)

## Agentic LLM Pipeline:

- TAPERA: Enhancing Faithfulness and Interpretability in Long-Form Table QA by Content Planning and Execution-based Reasoning (<https://aclanthology.org/2024.acl-long.692.pdf>) (Zhao et al., ACL 2024)



# Grading



- Participation & Approach development (35%)
- Written report (35%)
- Midterm presentation (10%)
- Final presentation (20%)

- -How to read a paper- lesson jointly with “DQ4AI: Data Quality Assessment” and “Advanced Data Profiling” seminars:  
22nd of October, 15:15-16:45, Room F-E.06, replaces the normal lesson
- Teaching end: beginning of February
- Midterm presentation: 10th of December
- Final presentation: 11th of February
- Final deadline report: 18th of February
  
- Tuesdays: 17 - 18:30

# Schedule - Next Steps



- Introduction (15th of October)
  - Send a mail with registration and prior experiences to us until 19th of October, we will then notify you and create a slack
  - Add a preferred partner (student) and topic (supervisor) to the mail if you would like to
- How to read a paper (22nd of October)
- Paper presentations by groups and discussions (29th of October)
- Finalization of Research Ideas and Kick-off (5th of November)

# Code of Conduct - Overview



At DEF/HPI, we are committed to fostering a high-quality, inclusive environment where students and staff can thrive both scientifically and personally. Our Code of Conduct is key to ensuring this, emphasizing respect, safety, and positive collaboration across our community.

**Respect for All:** Treat everyone with dignity, regardless of ethnicity, religion, gender, sexual orientation, or other personal traits.

**Inclusivity:** Collaborate with others, valuing the diversity that enriches our academic community.

**Safety:** Maintain a safe environment, free from threats, harassment, or any form of assault.

**Appropriate Behavior:** Use language and behavior that is respectful and suitable for everyone.

**Positive Interaction:** Engage positively with others' work and experiences, avoiding negative tones like sarcasm or irony.

**Constructive Feedback:** Give and receive feedback thoughtfully, mindful of its impact.

**Support and Stand Up:** Advocate against any violations of this Code and support others in fostering a respectful community.

# Code of Conduct - How you can get help and support



Violations of this code are taken seriously. If you witness or experience any inappropriate behavior, report it to a lecturer or any DEF/HPI contact point. All reports will be handled confidentially and with care.

Please be aware of **further contact points and support** structures at DEF/HPI, including:

- Diversity Manager (Oliwia Gust)
- Equal Opportunities Officer (Charlotte Weiss)
- Ombudsman for good scientific practice (Prof. Tilmann Rabl)
- Student Trusted Advisors (Hanna, Joscha and Zero)
- Psychological counseling hotline (0800 7777015)
- Incident Response System – coming October 2024
- as well as the respective offers of the University of Potsdam (Mental Health Counseling Service, Psychosocial Counseling of Studentenwerk, Nightline)