PluRel: Synthetic Data unlocks Scaling Laws for Relational Foundation Models

Vignesh Kothapalli

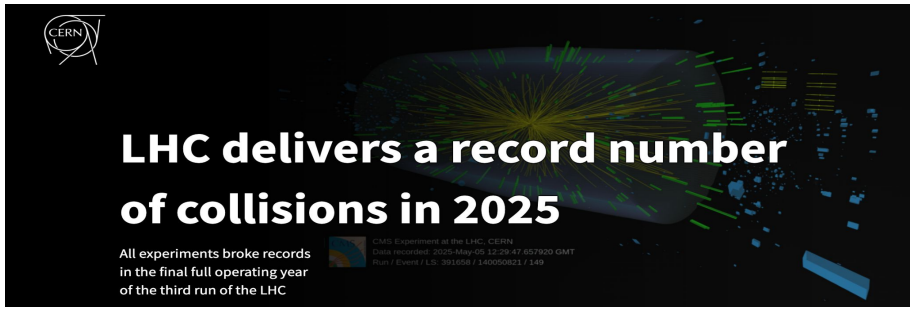
Rishabh Ranjan, Valter Hudovernik, Vijay Prakash Dwivedi, Johannes Hoffart, Carlos Guestrin, Jure Leskovec

Stanford University, Kumo AI, SAP

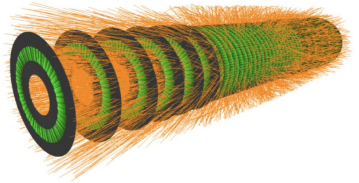
ICML 2026

Why learn over relational data?

- Relational data is the backbone of decision making in any organization/scientific field (healthcare, finance, tech etc).
- Lineage of transformations begins with relational data for recommendations, search, ads and many more use cases.
- Text based FMs fail on prediction tasks native to relational data.




```
event000000000-hits.csv
event000000000-cells.csv
event000000000-particles.csv
event000000000-truth.csv
```



The need for Relational Foundation Models (RFMs)

The promise of RFMs

- Learn over diverse relational data.
  - Able to consume data in a schema agnostic fashion.
  - Learn over medical, financial and e-commerce DBs at once
- Avoid manual feature engineering.
- Handle predictive tasks on unseen data.

The promise of RFMs

- Learn over diverse relational data.
- Avoid manual feature engineering.
  - Be capable of feature selection and transformations.
  - Typically requires >100 engineers per company.
- Handle predictive tasks on unseen data.

The promise of RFMs

- Learn over diverse relational data.
- Avoid manual feature engineering.
- Handle predictive tasks on unseen data.
  - Learn new tasks in-context for wider adoption.
  - Essentially replacing the entire data science stack!

The Irony of Data

- Relational data is abundant, yet in-accessible
  - Large scale databases are private to organizations.
  - Signals in open-source databases can be sparse.

Uber

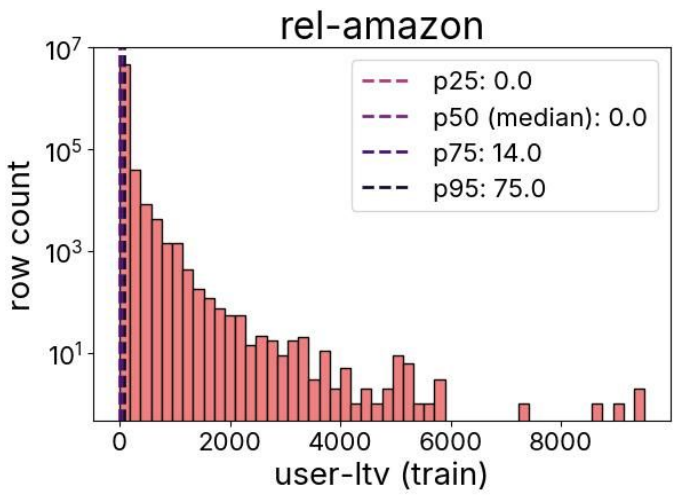
How does Uber scale to millions of concurrent requests?

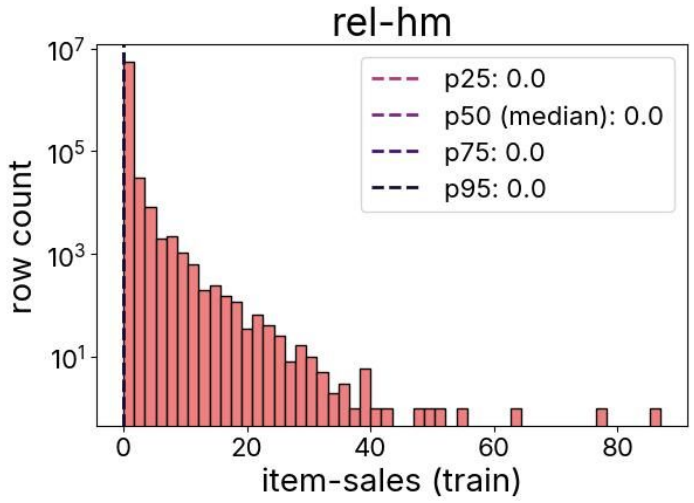
Explore how Uber redesigned its fulfillment platform leveraging Spanner.

wayfair — How Wayfair is modernizing, one database at a time

Learn how Gmail migrated billions of users, trillions of emails, and exabytes of data to Spanner

DOJO — Dojo is a payments powerhouse, delivering 60% faster transactions with Spanner

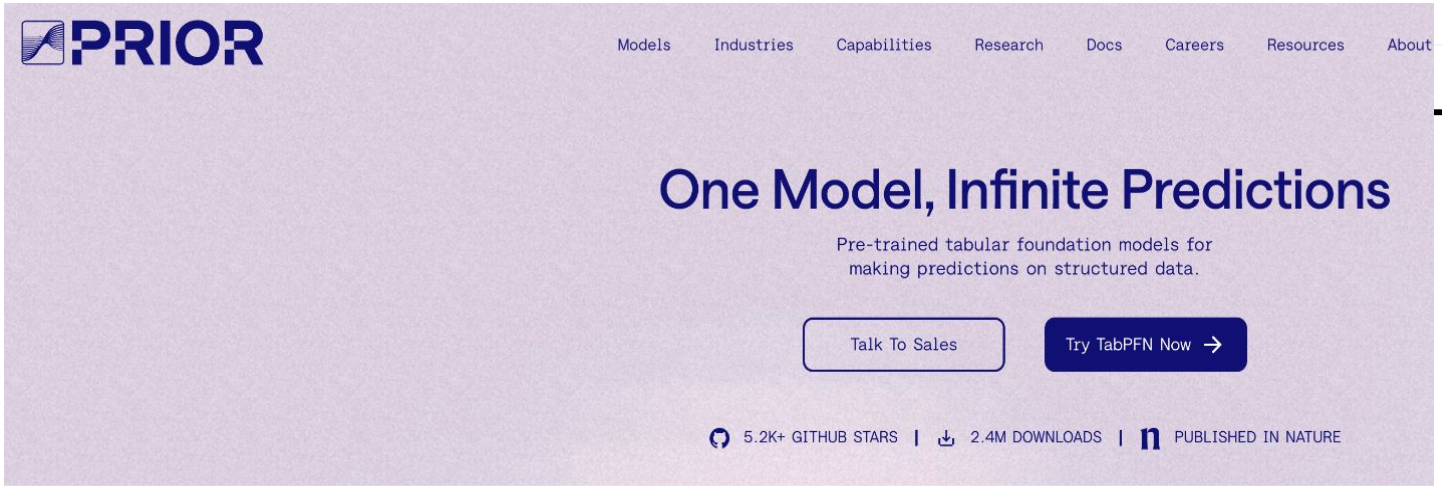




How to train RFMs when “diverse” data is scarce?

Synthetic Relational Data

- Leverage diverse synthetic data to train RFMs
- TabPFNx: Success stories at a Tabular level



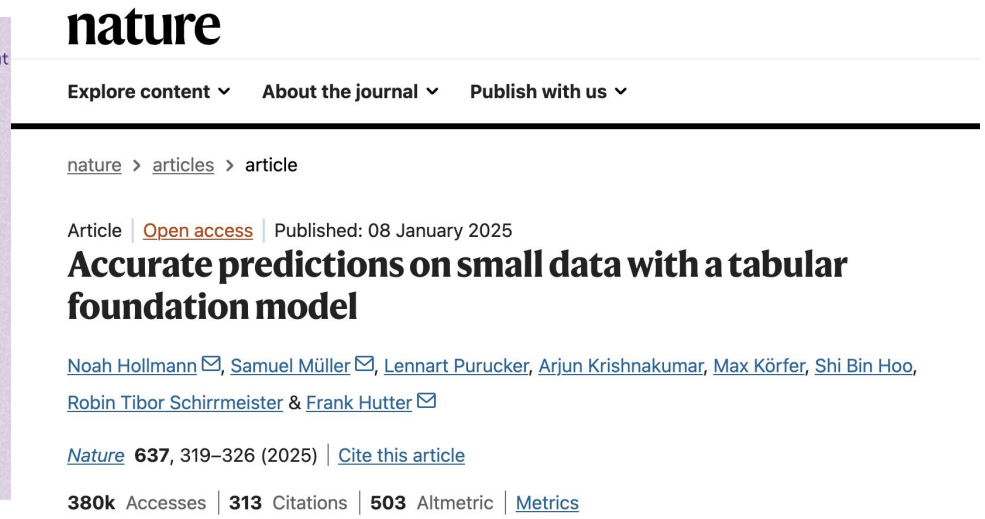


Fig. 1 | Overview of the proposed method. a, The high-level overview of TabPFN pre-training and usage. **b**, The TabPFN architecture. We train a model to solve more than 100 million synthetic tasks. Our architecture is an adaptation of the

Data Fig. 2d). TabPFN was pre-trained once using eight NVIDIA RTX 2080 GPUs over 2 weeks, allowing for ICL on all new datasets in a single forward pass. These modest computational requirements make similar

Challenges in Relational Settings

- Challenges of generating multi-tabular data from scratch.

1. How to connect multiple tables (i.e, schema generation) ?
2. How to connect rows of tables via primary-foreign keys ?
3. How to conditionally generate rows based on features from foreign tables?

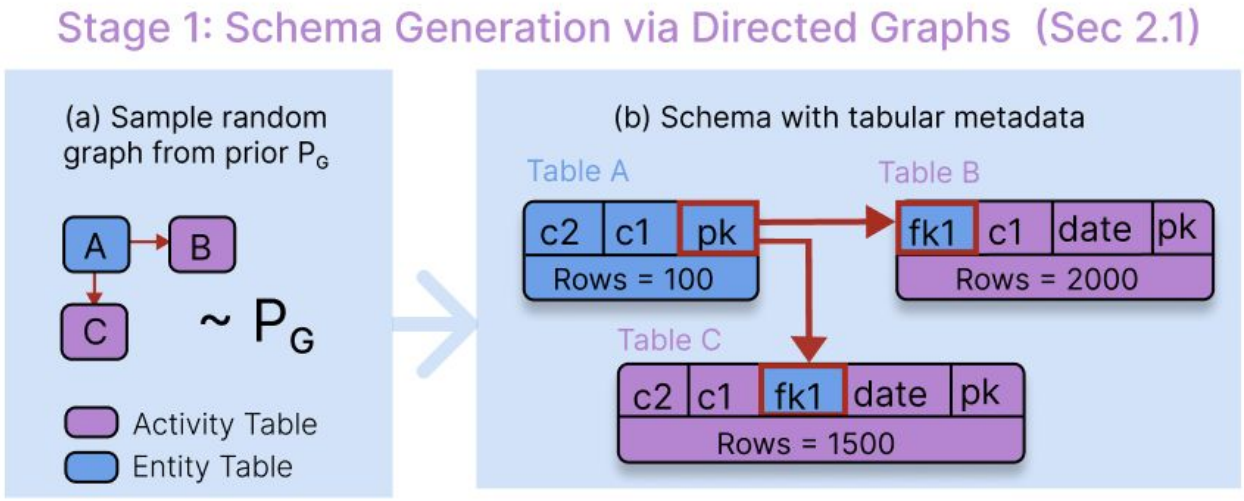
Introducing PluRel!

```
pip install plurel
```

1. Schema Generation

- Schema determines:

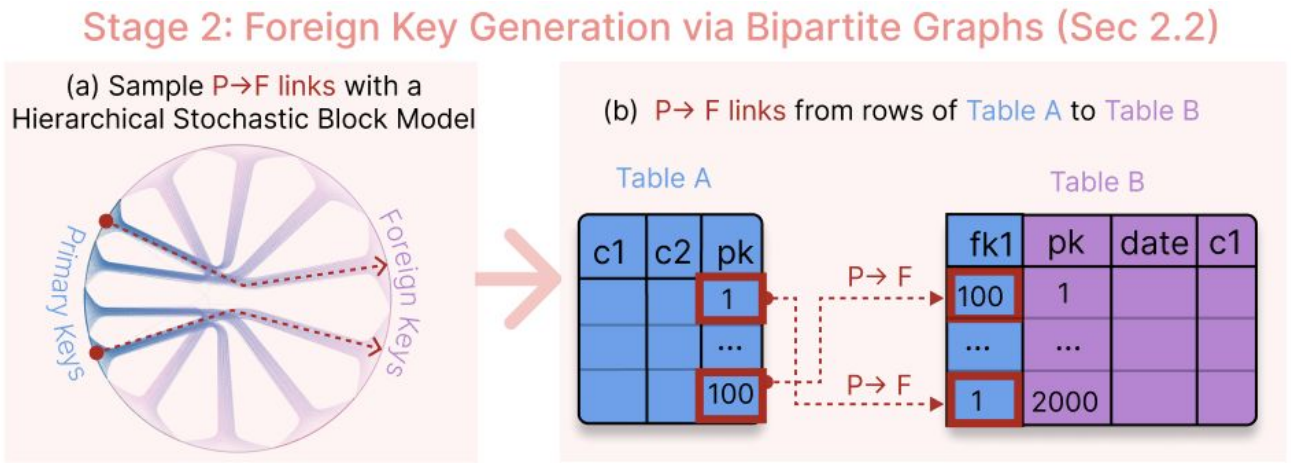
1. Number of foreign key columns in tables (i.e sparsity of the layout)
2. Locality of information (ex: hub tables)
3. Conditional row generation process in our framework



2. Foreign Key Generation

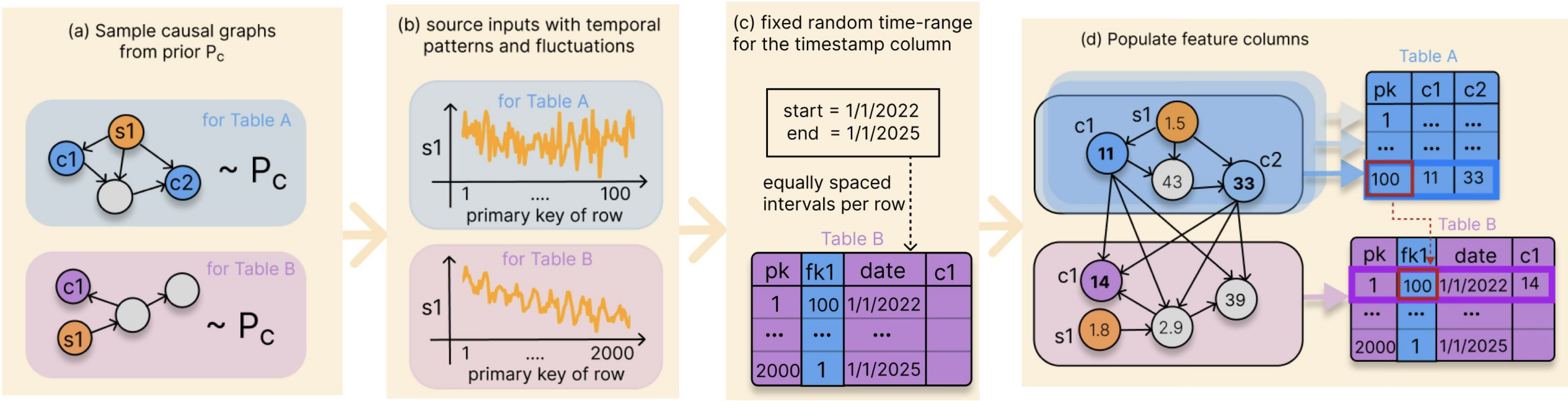
- Foreign keys determine

1. The distribution of “importance” of entities in primary table.
2. For ex: determines the cold-start problem for RecSys tasks



3. Feature Generation

- Leverage Structural Causal Models to generate diverse feature distributions.
- Edges model Cause - Effect relationships
- Nodes model observed/latent features.



Relational Transformer: Background

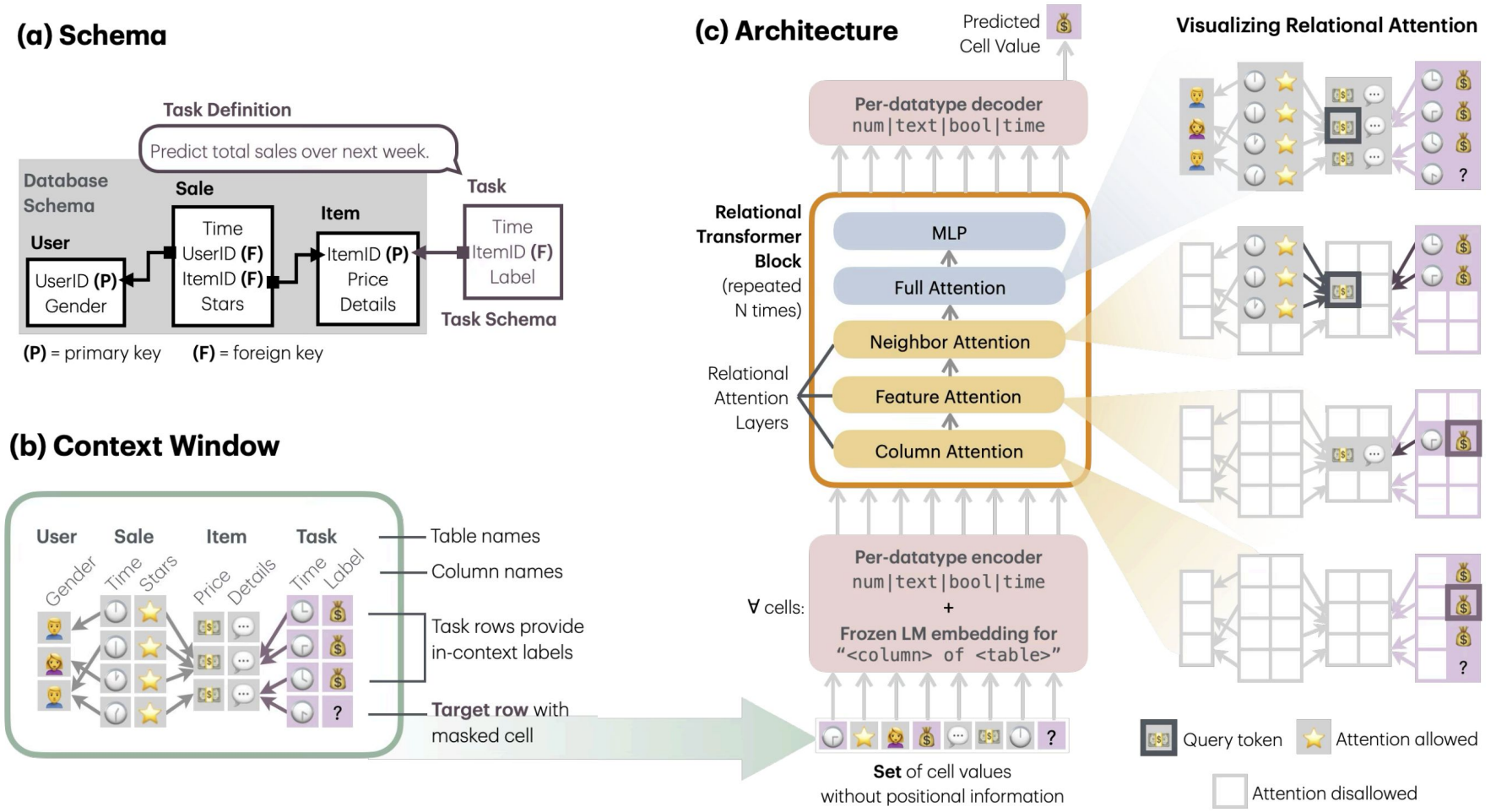
RELATIONAL TRANSFORMER: TOWARD ZERO-SHOT FOUNDATION MODELS FOR RELATIONAL DATA

Rishabh Ranjan[01*], **Valter Hudovernik**[0], **Mark Znidar**[0], **Charilaos Kanatsoulis**[0], **Roshan Upendra**[1], **Mahmoud Mohammadi**[1], **Joe Meyer**[1], **Tom Palczewski**[1], **Carlos Guestrin**[0], **Jure Leskovec**[0]
[0]Stanford University, [1]SAP Labs LLC
{ranjanr,guestrin,jure}@stanford.edu

ABSTRACT

Pretrained transformers readily adapt to new sequence modeling tasks via zero-shot prompting, but relational domains still lack architectures that transfer across datasets and tasks. The core challenge is the diversity of relational data, with varying heterogeneous schemas, graph structures and functional dependencies. In this paper, we present the *Relational Transformer (RT)* architecture, which can be pretrained on diverse relational databases and directly applied to unseen datasets and tasks without task- or dataset-specific fine-tuning, or retrieval of in-context examples. RT (i) tokenizes cells with table/column metadata, (ii) is pretrained via masked token prediction, and (iii) utilizes a novel *Relational Attention* mechanism over columns, rows, and primary–foreign key links. Pretrained on RelBench datasets spanning tasks such as churn and sales forecasting, RT attains strong zero-shot performance, averaging 93% of fully supervised AUROC on binary classification tasks with a single forward pass of a 22M parameter model, as opposed to 84% for a 27B LLM. Fine-tuning yields state-of-the-art results with high sample efficiency. Our experiments show that RT's zero-shot transfer harnesses task-table context, relational attention patterns and schema semantics. Overall, RT provides a practical path toward foundation models for relational data.[1]

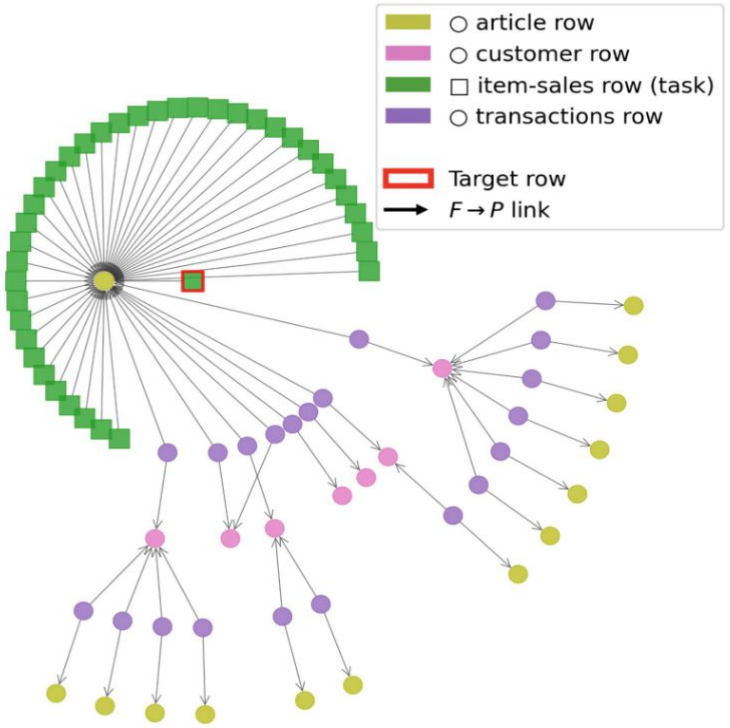
Relational Transformer: Background



source: RT paper

Relational Transformer: Pretraining

(a) Graph visualization of the context window for an example from **`rel-hm`** (dataset) / **`item-sales`** (task)



(b) Tabular visualization of the same context window

item-sales (task) (39 rows x 3 cols)

timestamp	article_id	sales
2020-06-01	104264	[MASK]
2020-05-25	104264	0.337034
...	...	...
2019-10-07	104264	0.000000
2020-01-13	104264	0.000000

transactions (21 rows x 5 cols)

t_dat	customer_id	article_id	price	sales_channel_id
2020-02-20	29049	2251	0.033881	2
2020-06-01	29049	104264	0.041763	2
...	...	...	...	...
2020-06-01	1247634	104264	0.042356	2
2020-06-01	1247634	104264	0.042356	2

article (14 rows x 25 cols)

article_id	product_code	prod_name	...	garment_group_no	garment_group_name	detail_desc
2251	399223	Curvy Jeggings H...	...	1016	Trousers Denim	Jeggings in wash...
48500	692202	SPEED JAM SHIRT	...	1010	Blouses	Straight-cut blo...
...	...	...	...	...	...	...
101044	893796	Nejlika	...	1005	Jersey Fancy	Body in soft jer...
104264	920700	Dazzle top	...	1005	Jersey Fancy	Wide, slightly s...

customer (7 rows x 7 cols)

customer_id	FN	Active	club_member_status	fashion_news_frequency	age	postal_code
29049	NaN	NaN	ACTIVE	NONE	26.0	5cbf988955d931a7...
178380	NaN	NaN	ACTIVE	NONE	29.0	e777db329cc6dfe3...
...	...	...	...	...	...	...
856256	1.0	1.0	ACTIVE	Regularly	26.0	97cccf90aba93a67...
1247634	NaN	NaN	ACTIVE	NONE	24.0	0c0aaee59a3e86f5...

Predict the value of the masked token

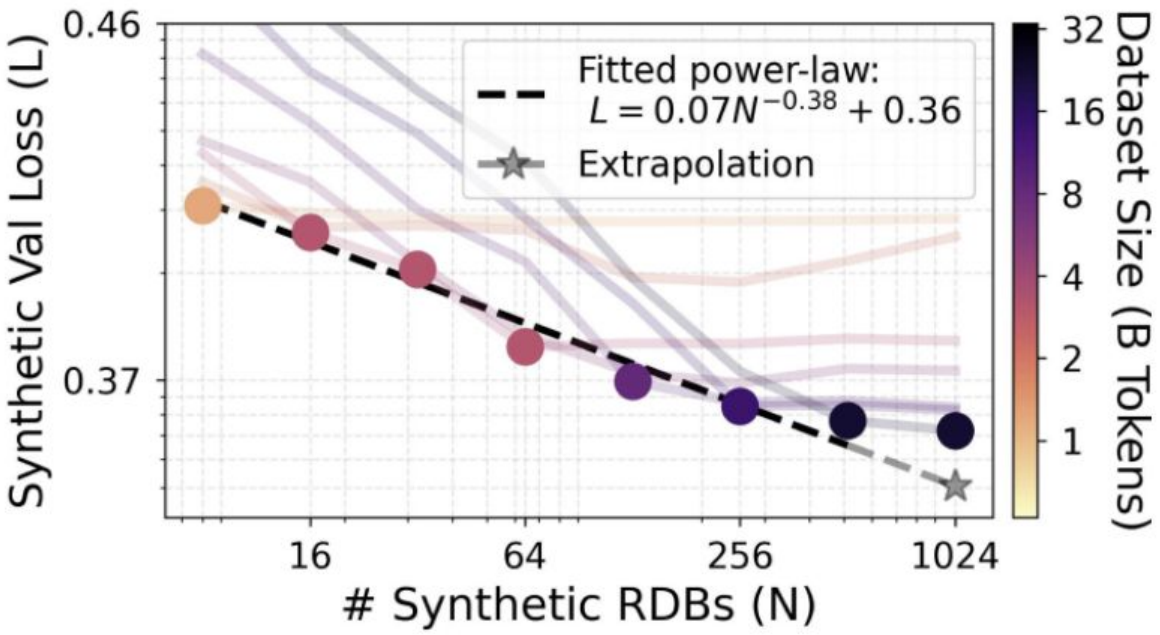
source: RT paper

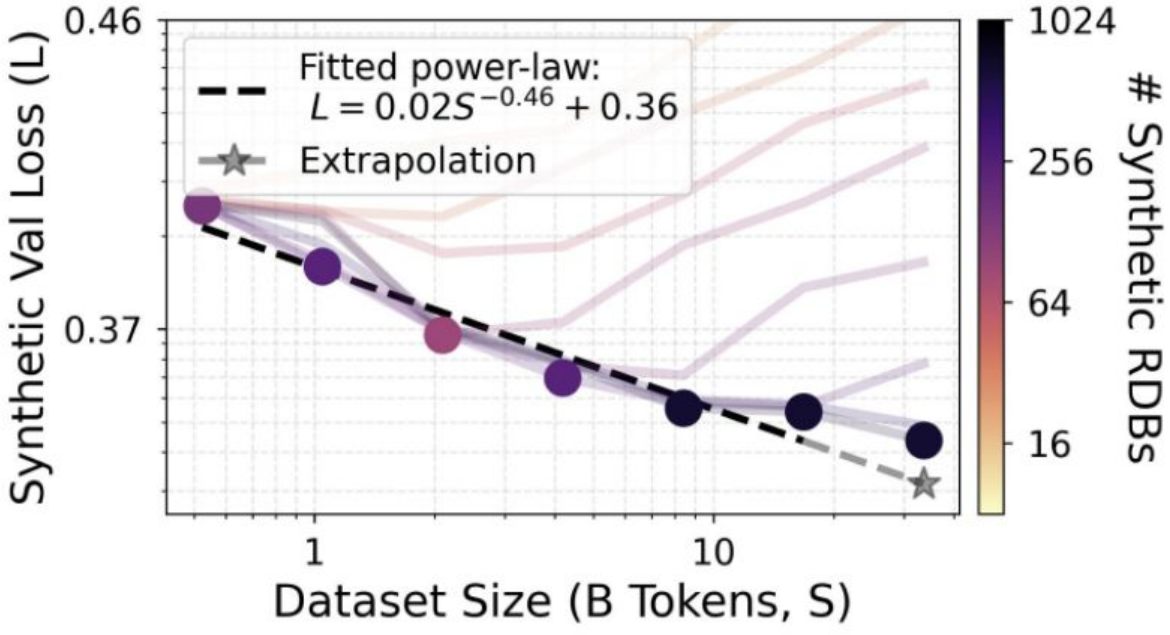
Scaling Laws for Data Diversity and Size

- Synthetic data unlocks scaling laws with RT!
  - **L:** Masked Token Prediction loss
  - **N:** Number of synthetic RDBs
  - **S:** Number of synthetic pretraining tokens

$$L(N) = A_N N^{-\alpha_N} + C_N \quad \text{(Diversity power law)} \quad (6)$$

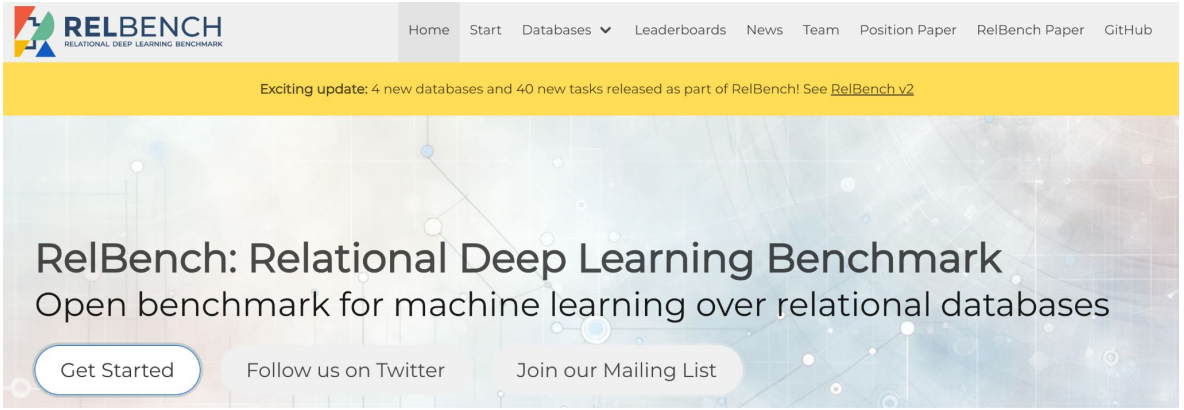
$$L(S) = A_S S^{-\alpha_S} + C_S \quad \text{(Size power law)} \quad (7)$$





Generalization to Real Datasets

Can synthetic pre-training improve predictive performance on real-world tasks?

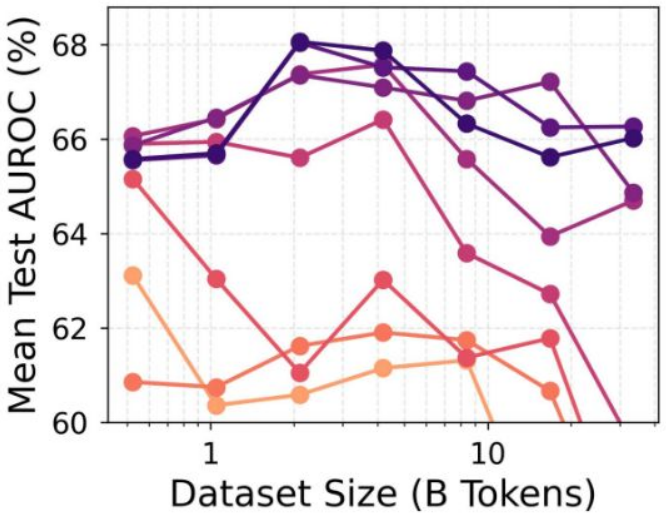


- **Setup:** Pre-Train Relational Transformer with the Masked Token Prediction Objective on Synthetic data.

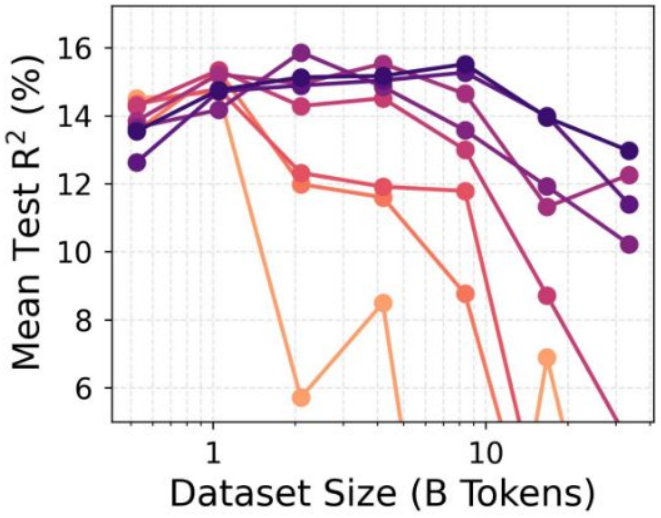
- **Evaluation:** Measure AUROC (clf) and R2 (reg) on the real-world hold-out database tasks. (i.e, predict by learning about the task and DB in-context)

Generalization to Real Datasets

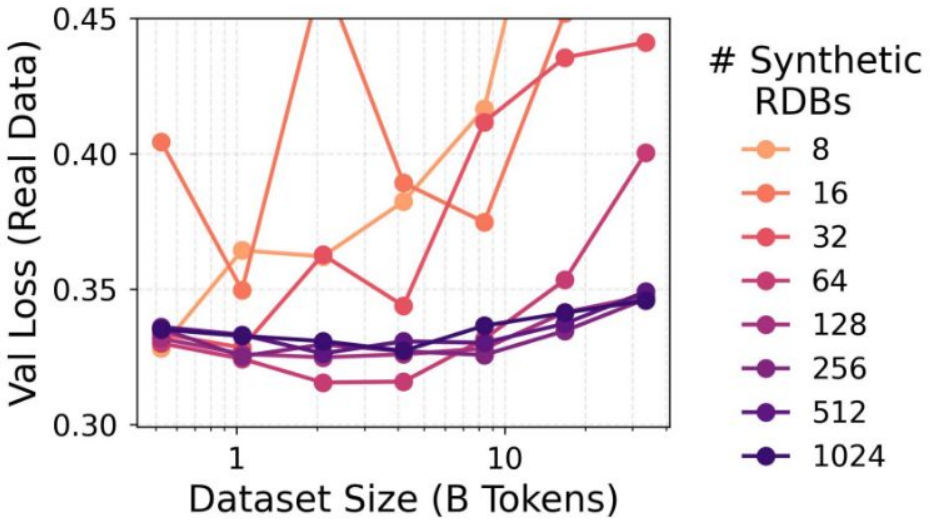
- Scaling synthetic data diversity is critical for generalization.
  - With sufficient database diversity, synthetic pretraining enables zero-shot transfer to real-world RelBench tasks.



(a) Mean 0-shot test AUROC (%) (↑)



(b) Mean 0-shot test R^2 (%) (↑)



(c) Validation loss (↓) on real data (RelBench)

Continued Pretraining on Real Datasets

Can synthetic pretraining result in strong base models for real data pretraining?

- **Setup:** Choose a synthetic pretrained RT model and continue the pretraining on real datasets.
- **Baseline:** Leave-one-db-out pretraining on RelBench databases and tasks for 50K steps.
  - MTP on auto-complete as well as prediction tasks

Continued Pretraining on Real Datasets

- Synthetic + real pretraining outperforms real-only pretraining, with +1.2% mean AUROC and +3.0% mean R^2 gains.
- Synthetic pretraining provides a strong base model for continued learning on real data.

Dataset	Task	Real only	Synthetic + Real (ours)	Absolute Gain (%)	Synthetic only (ours)
AUROC(%) for classification. Higher is better. Majority baseline is 50.0.					
rel-amazon	user-churn	64.2	**65.0**	+0.8	64.4
rel-hm	user-churn	**67.4**	66.0	−1.4	63.7
rel-stack	user-badge	80.0	**82.0**	+2.0	81.4
rel-stack	user-engage	78.9	**86.2**	+7.4	82.4
rel-amazon	item-churn	67.6	**72.5**	+4.9	71.0
rel-avito	user-visits	57.2	**63.4**	+6.2	63.5
rel-avito	user-clicks	**54.7**	47.9	−6.8	45.9
rel-trial	study-out	**54.4**	51.8	−2.6	53.8
rel-f1	driver-dnf	80.7	**81.0**	+0.3	76.7
rel-f1	driver-top3	86.9	**88.4**	+1.5	82.6
Mean		69.2	**70.4**	+1.2	68.5
R^2(%) for regression. Higher is better. Mean baseline is 0.0.					
rel-hm	item-sales	16.0	**20.0**	+4.0	4.4
rel-amazon	user-ltv	14.5	**18.5**	+4.0	9.8
rel-amazon	item-ltv	35.3	**40.5**	+5.2	10.7
rel-stack	post-votes	22.3	**25.5**	+3.2	15.7
rel-trial	site-succ	33.7	**38.6**	+5.0	38.3
rel-trial	study-adv	**1.9**	1.6	−0.3	−0.8
rel-f1	driver-pos	54.3	**55.5**	+1.2	41.3
rel-avito	ad-ctr	3.1	**4.9**	+1.9	2.5
Mean		22.6	**25.7**	+3.0	15.2

Summary

- RFM development is bottlenecked by diverse relational datasets.
- Synthetic data offers a scalable and privacy friendly alternative to scalable pretraining.
- PluRel is the first of its kind open-source framework for generating such synthetic datasets.
- Synthetic pretraining unlocked scaling laws in RFMs with a joint emphasis on data size (tokens) and diversity (# RDBs)
- Synthetic pretraining improves generalization and results in strong base models for continued pretraining on real world datasets.

Future Work and Opportunities

- Relational data curation and synthetic design space exploration.
- Extending PluRel to additional data types such as text.
- Semi-synthetic data augmentation to expand real-world RDBs.
- Pretraining curriculums and strategies to combine synthetic and real data.
- Exploring impact of synthetic data on long-context modeling and test-time scaling.
- Joint model- and data-scaling laws.
- [Theory]: Study expressiveness of relational attention models.
  - ICL with 1-layer RT and a family of SCM models.
  - Role of context, depth, joint feature distribution across tables…

Thank You!