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 Introduction

Foundation models [3] have transformed natural language processing (NLP) [7] and computer vision (CV) [9] through general-purpose architectures—primarily the transformer [35]—which enable large-scale pretraining and effective transfer. In contrast, relational databases, the dominant repositories of structured enterprise data, still lack such models. Unlike sequences, relational data consists of interconnected tables with heterogeneous columns linked via primary–foreign keys, where predictive signals are dispersed across rows, columns, tables, and time. Designing a foundation model for relational databases is crucial [36], and would enable zero-/few-shot predictions, efficient fine-tuning, and help democratize AI.

Prior work. Tabular models [5, 31] require manual feature engineering. Relational deep learning (RDL) [14] methods—GNNs [30, 4], transformers [11, 25], hybrids [39]—remain schema-specific. Tabular foundation models [17, 18, 22] and LLM serialization [40] cannot capture multi-table structure effectively. For a detailed discussion, see App. A.

Our contribution. We introduce Relational Transformer (RT, Fig. 1), enabling pretraining and zero-shot transfer across relational databases via three innovations: (i) *cell-level tokenization*, representing each database cell as a token with embedding derived from its value, column, and table name, enabling all relational tasks to be formulated as masked token prediction; (ii) *task table prompting*

*Work done, in part, as an intern at SAP Labs LLC.

AI for Tabular Data workshop at EurIPS 2025

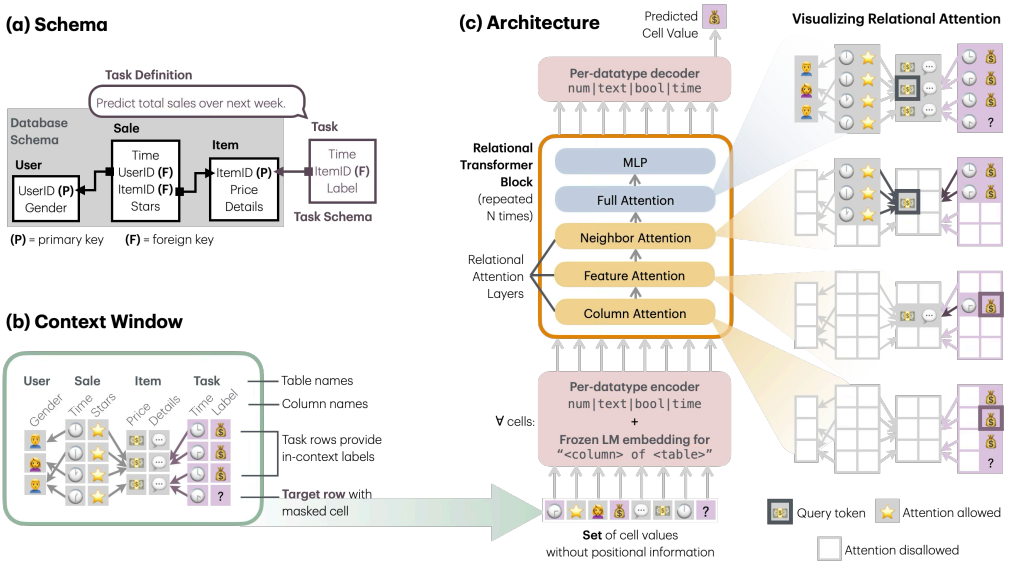

Figure 1: **(a)** Schema with task table for labels. **(b)** Context window samples relevant cells, excluding future rows. **(c)** Cells=tokens with value+schema embeddings. Relational Attention: (1) column, (2) feature (row+F→P), (3) neighbor (P→F).

providing task-specific context for zero-shot prompting; (iii) *Relational Attention* (Sec.3.2) with column, feature (row + F→P links), neighbor (P→F links), and full attention masks. Pretrained on RelBench [30], RT achieves 93% of supervised AUROC in zero-shot transfer (compared to 84% for Gemma3-27B at $10^5\times$ higher FLOPs) and requires 10–100$\times$ fewer steps for fine-tuning.

2 Background

A *relational database (RDB)* consists of *tables* connected via *primary keys (P–keys)* and *foreign keys (F–keys)*, forming $F\to P$ and $P\to F$ links (Fig.1). Many RDBs are inherently *temporal*, requiring models to ensure temporal consistency by conditioning solely on information observed before the target timestamp. We focus on *masked token prediction (MTP)*, where the goal is predicting a masked cell value. This includes *forecasting tasks* (predict future events via task tables with labels and timestamps) and *autocomplete tasks* (predict missing values in existing columns). See App. B.

3 Relational Transformer

The design of Relational Transformer (RT, Fig. 1) is guided by three core principles: (i) effectively *capture relational structure*, (ii) support flexible *self-supervised pretraining*, and (iii) enable *zero-shot generalization* across heterogeneous schemas.

Figure 2: RT can be pretrained on data with diverse schemas and task definitions. Pretrained RT is accurate on new datasets and tasks with zero-shot prompting. Dataset- and task-specific fine-tuning of pretrained RT shows high learning efficiency.

3.1 Input Representation

RT introduces two key innovations in input representation: (i) *task table prompting*, where prediction tasks are represented as additional tables appended to the database, and (ii) *cell-level tokenization*, where each database cell is modeled as an individual token.

Task table prompting. Predictive tasks are added to the database as tables. Task rows act as seed rows for context sampling, with only one task active at a time to ensure task-specific sampling.

Context sampling. Starting from a seed row, RT constructs an n-cell context using a bounded-width BFS that (1) includes all F→P parent rows, (2) subsamples P→F child rows up to width w, and (3) excludes rows with later timestamps to prevent temporal leakage (see App. K).

Cell encoding. Each cell (v, c, t) is encoded as $\mathbf{x} = \mathbf{W}_d\,\mathbf{r} + \mathbf{W}\,\mathcal{E}^{\text{schema}}(c, t)$, where $\mathbf{r}$ is a datatype-specific normalization of value v, and $\mathcal{E}^{\text{schema}}$ embeds "`<column> of <table>`". Masked cells replace $\mathbf{W}_d\mathbf{r}$ with a learned mask vector $\mathbf{m}_d$. (See App. C for details.)

3.2 Relational Attention

RT operates on *cell tokens* and applies *scaled dot-product attention* with mask $\mathbf{M}$: $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}; \mathbf{M}) = (\text{Softmax}((\mathbf{Q}\mathbf{K}^\top)/\sqrt{d_K}) \odot \mathbf{M})\mathbf{V}$, where $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{n \times d_K}$, $\mathbf{V} \in \mathbb{R}^{n \times d_V}$, and $\mathbf{M} \in \{0, 1\}^{n \times n}$. We use four relational masks:

- **Column:** $\mathbf{M}^{\text{column}}[q, k] = \mathbf{1}\{\text{Col}(k) = \text{Col}(q)\}$ models intra-column value distributions;
- **Feature:** $\mathbf{M}^{\text{feature}}[q, k] = \mathbf{1}\{\text{Row}(k) = \text{Row}(q) \vee \text{Row}(k) \in \text{OutLinks}(q)\}$ mixes features within a row and its F→P-linked parents;
- **Neighbor:** $\mathbf{M}^{\text{neighbor}}[q, k] = \mathbf{1}\{\text{Row}(q) \in \text{OutLinks}(k)\}$ aggregates signals from P→F-linked children (GNN-like message passing);
- **Full:** $\mathbf{M}^{\text{full}}[q, k] = 1$ enables unrestricted pairwise interactions.

Masks are implemented sparsely and compiled to efficient FlashAttention-based [6] kernels using FlexAttention [8]. See App. D for more details.

3.3 Output Decoding and Training

Datatype-specific decoders map output embeddings to predictions. For a masked cell with representation r and decoder output r', we use HuberLoss for regression and binary cross-entropy for classification. This unified objective applies to both pretraining and fine-tuning. See App. E.

4 Results

Data. We use RelBench [30] with six databases (excluding `rel-event` due to leakage) spanning domains like e-commerce, forums, and sports, with 10 classification and 8 regression forecasting tasks. We also define 17 binary classification and regression autocomplete tasks (App. J).

Pretraining. Leave-one-DB-out: for each target dataset, we pretrain on all tasks from the remaining datasets. RT is a 12-layer transformer (22M params; context length 1024) trained for 50k steps. Further details in App. F.

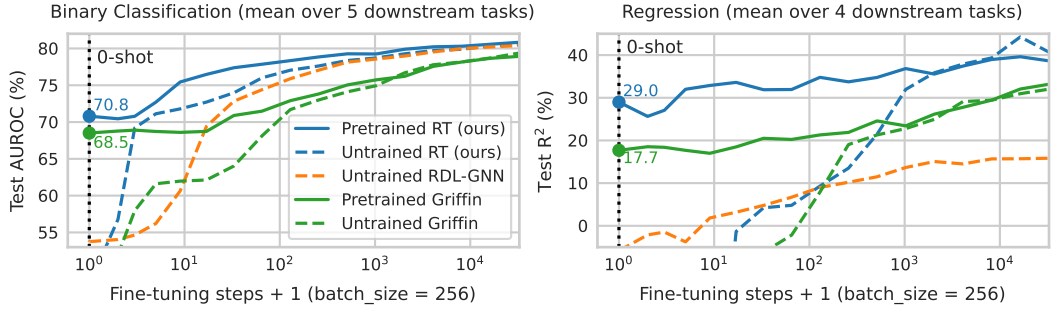

Figure 3: Learning curves (log-scale X-axis). First point = zero-shot.

Table 1: Zero-shot test AUROC (%) for 10 binary classification tasks. Higher is better. Random/-majority baseline is 50.0. For RelLLM we use their own prompt construction. Other baselines have equivalent database subgraphs. Gemma and RelLLM additionally include dataset and task descriptions, as well as natural language instructions. The target task is never seen during pretraining.

Target DB ∈ pretraining? →		Maybe			No			Yes		
Dataset ↓	Task ↓	Gemma	Gemma	Gemma	Entity Mean	Griffin	RT (ours)	Rel LLM	Griffin	RT (ours)
	Parameter count →	4B	12B	27B	0	22M	22M	3B	22M	22M
rel-amazon	item-churn	62.1	55.0	42.1	73.0	69.0	70.9	64.1	71.9	**73.3**
rel-amazon	user-churn	58.1	54.7	50.5	64.4	62.3	64.0	60.1	64.1	**66.1**
rel-avito	user-clicks	54.5	59.5	59.8	44.7	45.9	59.5	**62.3**	45.9	60.9
rel-avito	user-visits	60.1	57.9	**62.7**	60.7	60.7	61.8	56.2	62.2	62.6
rel-f1	driver-dnf	56.2	54.6	75.8	75.4	57.7	81.2	71.8	57.7	**81.2**
rel-f1	driver-top3	84.6	90.5	**91.4**	85.0	82.5	89.3	70.6	81.8	89.3
rel-hm	user-churn	59.8	47.1	48.7	**64.4**	60.2	62.8	56.0	60.4	63.3
rel-stack	user-badge	79.1	79.8	80.0	66.2	73.5	80.1	62.1	**82.3**	81.1
rel-stack	user-engage	65.9	67.8	78.0	83.5	77.5	75.7	69.5	**89.4**	86.9
rel-trial	study-out	52.6	57.4	57.2	50.0	51.0	51.8	**59.0**	57.2	54.6
	Mean AUROC →	63.3	62.4	64.6	66.7	64.0	69.7	63.2	67.3	**71.9**

Baselines. Schema-agnostic baselines are **Griffin** [38] (22M; same pretraining) and **LLMs** [33, 40] (4B–27B). Schema-specific ones are **RDL-GNN** [30], **RelLLM** [39]. **EntityMean** is non-neural.

4.1 Fine-Tuning Efficiency

Fig. 3 shows pretrained RT achieves strong zero-shot performance (first point on curves) and maintains advantages during fine-tuning. RT outperforms Griffin despite identical pretraining data. Untrained RT catches up to RDL-GNN quickly but requires extensive training to match pretrained RT. Full fine-tuning results for all tasks, are in App. G.

4.2 Zero-shot prompting

Setup. In Tabs. 1 and 2 we report zero-shot results. The target task is always unseen; we report both when the target dataset is unseen and after continued pretraining on it. RT and Gemma use our sampling algorithm to build context; Griffin is adapted to include task-table rows for zero-shot capability; RelLLM uses its own prompts. Both Gemma and RelLLM additionally receive task descriptions and natural-language instructions. See App. H.

Table 2: Zero-shot test R^2 (%) for 8 regression tasks. Higher is better. Global mean baseline is 0.0. Setup is same as Table 1. LLM baselines are poor (App. L.2).

Target DB ∈ pretraining? →		No			Yes	
Dataset ↓	Task ↓	Entity Mean	Griffin	RT (ours)	Griffin	RT (ours)
rel-amazon	item-ltv	**54.2**	20.1	33.2	20.1	35.4
rel-amazon	user-ltv	19.9	20.6	36.4	24.4	**39.7**
rel-avito	ad-ctr	3.4	2.4	4.5	2.4	**7.7**
rel-f1	driver-pos	38.2	−0.7	54.7	4.6	**58.4**
rel-hm	item-sales	1.8	2.7	14.0	2.5	**30.4**
rel-stack	post-votes	**43.7**	27.4	32.4	27.1	32.7
rel-trial	site-succ	−6.4	1.4	**5.2**	2.6	3.5
rel-trial	study-adv	−0.5	−2.5	2.1	−2.5	**3.4**
	Mean R^2 →	19.3	8.9	22.8	10.1	**26.4**

4.3 Ablations

Ablations reveal key factors enabling zero-shot transfer (App. I): (1) **self labels** (past task rows for target entity) are critical—removing them drops zero-shot AUROC from 70.1% to 53.8% and R^2 from 22.8% to -5.5%; (2) **column names** matter for zero-shot (shuffling reduces R^2 by 2.3%) but not fine-tuning; (3) **column attention** has highest impact on zero-shot transfer, while feature/neighbor attention matter more for fine-tuning. Full attention is surprisingly dispensable.

5 Conclusion

We introduced Relational Transformer (RT), a schema-agnostic architecture enabling foundation models for relational data through: (i) cell-level tokenization for unified masked token prediction, (ii) Relational Attention masks capturing column, row, and link structures, and (iii) task table prompting for zero-shot generalization. Pretraining on merely 6 databases yields strong zero-shot transfer (94% of supervised AUROC) and highly-efficient fine-tuning. Limitations include inability to handle link prediction and distinguish multiple foreign keys, which can be explored in future work.

Acknowledgements

We thank Matthias Fey, Harshvardhan Aggarwal, Vijay Prakash Dwivedi, Michael Bereket, Marcel Roed, Joshua Robinson, Martin Jurkovic, Fengyu Li, Justin Gu, Zoe Ryan, Sam Thelin, Johannes Hoffart, Maximilian Schambach, Andrew Pouret, Viswa Ganapathy, Tassilo Klein and Mark Li for help with this research.

References

[1] Dominique Beaini, Shenyang Huang, Joao Alex Cunha, Zhiyi Li, Gabriela Moisescu-Pareja, Oleksandr Dymov, Samuel Maddrell-Mander, Callum McLean, Frederik Wenkel, Luis Müller, et al. Towards foundational models for molecular learning on large-scale multi-task datasets. In *The Twelfth International Conference on Learning Representations*, 2024.

[2] Mitchell Black, Zhengchao Wan, Gal Mishne, Amir Nayyeri, and Yusu Wang. Comparing graph transformers via positional encodings. In *International Conference on Machine Learning*, pages 4103–4139. PMLR, 2024.

[3] Rishi Bommasani et al. On the opportunities and risks of foundation models, 2022.

[4] Tianlang Chen, Charilaos Kanatsoulis, and Jure Leskovec. RelGNN: Composite message passing for relational deep learning. In *Forty-second International Conference on Machine Learning*, 2025.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[6] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[8] Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels, 2024.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[10] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

[11] Vijay Prakash Dwivedi, Sri Jaladi, Yangyi Shen, Federico López, Charilaos I. Kanatsoulis, Rishi Puri, Matthias Fey, and Jure Leskovec. Relational graph transformer, 2025.

[12] Vijay Prakash Dwivedi, Charilaos Kanatsoulis, Shenyang Huang, and Jure Leskovec. Relational deep learning: Challenges, foundations and next-generation architectures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5999–6009, 2025.

[13] Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[14] Matthias Fey, Weihua Hu, Kexin Huang, Jan Eric Lenssen, Rishabh Ranjan, Joshua Robinson, Rex Ying, Jiaxuan You, and Jure Leskovec. Position: Relational deep learning - graph representation learning on relational databases. In *Forty-first International Conference on Machine Learning*, 2024.

[15] Matthias Fey, Vid Kocijan, Federico Lopez, Jan Eric Lenssen, and Jure Leskovec. Kumorfm: A foundation model for in-context learning on relational data, 2025.

[16] Mikhail Galkin, Xinyu Yuan, Hesham Mostafa, Jian Tang, and Zhaocheng Zhu. Towards foundation models for knowledge graph reasoning. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023.

[18] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[19] Qian Huang, Hongyu Ren, Peng Chen, Gregor Kržmanc, Daniel Zeng, Percy S Liang, and Jure Leskovec. Prodigy: Enabling in-context learning over graphs. *Advances in Neural Information Processing Systems*, 36:16302–16317, 2023.

[20] Yinan Huang, William Lu, Joshua Robinson, Yu Yang, Muhan Zhang, Stefanie Jegelka, and Pan Li. On the stability of expressive positional encodings for graphs. In *The Twelfth International Conference on Learning Representations*, 2024.

[21] Charilaos I Kanatsoulis, Evelyn Choi, Stephanie Jegelka, Jure Leskovec, and Alejandro Ribeiro. Learning efficient positional encodings with graph neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Myung Jun Kim, Leo Grinsztajn, and Gael Varoquaux. Carte: Pretraining and transfer for tabular learning. In *International Conference on Machine Learning*, pages 23843–23866. PMLR, 2024.

[23] Derek Lim, Joshua David Robinson, Lingxiao Zhao, Tess Smidt, Suvrit Sra, Haggai Maron, and Stefanie Jegelka. Sign and basis invariant networks for spectral graph representation learning. In *The Eleventh International Conference on Learning Representations*, 2022.

[24] Aliakbar Nafar, K Brent Venable, and Parisa Kordjamshidi. Learning vs retrieval: The role of in-context examples in regression with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8206–8229, 2025.

[25] Jakub Peleška and Gustav Šír. Transformers meet relational databases, 2024.

[26] Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. Let your graph do the talking: Encoding structured data for llms, 2024.

[27] Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Forty-second International Conference on Machine Learning*, 2025.

[28] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

[29] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[30] Joshua Robinson, Rishabh Ranjan, Weihua Hu, Kexin Huang, Jiaqi Han, Alejandro Dobles, Matthias Fey, Jan Eric Lenssen, Yiwen Yuan, Zecheng Zhang, et al. Relbench: A benchmark for deep learning on relational databases. *Advances in Neural Information Processing Systems*, 37:21330–21341, 2024.

[31] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[32] Marco Spinaci, Marek Polewczyk, Johannes Hoffart, Markus C. Kohler, Sam Thelin, and Tassilo Klein. Portal: Scalable tabular foundation models via content-specific tokenization, 2024.

[33] Gemma Team and Google DeepMind. Gemma 3 technical report, 2025.

[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Liane Vogel, Benjamin Hilprecht, and Carsten Binnig. Towards foundation models for relational databases [vision paper]. In *NeurIPS 2022 First Table Representation Workshop*, 2022.

[37] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, 2021.

[38] Yanbo Wang, Xiyuan Wang, Quan Gan, Minjie Wang, Qibin Yang, David Wipf, and Muhan Zhang. Griffin: Towards a graph-centric relational database foundation model. In *Forty-second International Conference on Machine Learning*, 2025.

[39] Fang Wu, Vijay Prakash Dwivedi, and Jure Leskovec. Large language models are good relational learners, 2025.

[40] Marek Wydmuch, Łukasz Borchmann, and Filip Graliński. Tackling prediction tasks in relational databases with llms, 2024.

[41] Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*, 2023.

[42] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

[43] Jianan Zhao, Zhaocheng Zhu, Mikhail Galkin, Hesham Mostafa, Michael M Bronstein, and Jian Tang. Fully-inductive node classification on arbitrary graphs. In *The Thirteenth International Conference on Learning Representations*, 2025.

[44] Qifang Zhao, Weidong Ren, Tianyu Li, Hong Liu, Xingsheng He, and Xiaoxiao Xu. Graphgpt: Generative pre-trained graph eulerian transformer. In *Forty-second International Conference on Machine Learning*, 2025.

A Related Work

Relational deep learning (RDL). [14] introduced an end-to-end framework for predictive modeling on relational databases using neural networks. At its core, RDL represents a database as a relational entity graph: a temporal, heterogeneous graph where each table is a node-type, each row an individual node, and every primary-foreign key relationship an edge. Initial approaches applied heterogeneous graph neural networks directly to these relational entity graphs [30], and more recently, advanced message-passing approaches have been proposed to enhance the efficiency of GNNs on relational data [4]. Transformers have emerged as a way to improve upon the GNN message-passing paradigm. [25] and [11] propose transformer-based architectures that achieve better performance than GNNs on relational data. An review of RDL architecture can be found in [12]. A key limitation, however, is that these architectures are schema-specific, which prevents pretraining and fine-tuning on diverse database structures. Our Relational Transformer, by design, is schema-agnostic, enabling it to learn from and be directly applied to new, unseen database structures. This design principle allows our architecture to demonstrate foundation model-like capabilities, similar to those recently shown in tabular learning.

Tabular foundation models (TFMs). Recent advancements in tabular foundation models have demonstrated significant promise, exhibiting capabilities such as in-context learning [17, 27] and efficient fine-tuning [22]. These efforts have explored both supervised [17] and self-supervised [32, 22] pretraining on real [22] or synthetic [17, 27] data. Extending tabular foundation models to relational data is non-trivial, because not only are there multiple tables, but rows in one table are linked to rows in another by foreign-primary key links. We take inspiration for the universal cell encoders/decoders from PORTAL [32], which has a similar handling of column names and text/numeric/-datetime data types. We also take inspiration from the TabPFNv2 [18] transformer architecture, which uses stacked layers of row-wise and then column-wise attention, except that we also have all-pair attention and use attention masks to capture foreign-primary key links.

Relational foundation models (RFMs). While tabular foundation models can, in principle, be applied to relational datasets, they fail to account for the rich, multi-table structure of real-world data. To address this limitation, recent works have begun to develop dedicated relational foundation models. For instance, [15] propose a relational foundation model based on graph-transformers and in-context learning, demonstrating both in-context learning and fine-tuning capabilities. However, their solution is not open-sourced, and the exact pretraining procedure has not been released. Separately, [38] design a novel architecture pretrained on a mixture of tabular and relational datasets, showing that fine-tuning improves downstream task performance. Their model, however, differs significantly from our own; it first aggregates information within each table and then utilizes graph neural networks to propagate that information between tables. In contrast, our model uses a cell-level representation of the entire database and employs attention masks to directly represent the foreign-key structure. This allows our approach to reason directly over the relational database in its native, cell-based format, offering a more granular and unified understanding.

Pretrained models for graphs. Pretrained graph learning models have shown strong success in molecular domains. For example, MoleBERT [41] introduces masked atom modeling and triplet-masked contrastive learning to pretrain GNNs for both node-level and graph-level tasks relevant to drug discovery. [1] scale pretraining by curating massive multi-task molecular datasets with billions of labels, showing that combining quantum and biological data improves low-resource tasks. Beyond molecules, [19] introduced a novel pretraining framework that leverages prompt-based graph representations to enable in-context learning on graphs. GraphAny [43] develops a zero-shot node classification framework, grounded in linear least-squares principles, that generalizes across graphs with disjoint feature and label spaces by leveraging LinearGNN ensembles and inductive attention. ULTRA [16] targets knowledge graph reasoning, learning universal relational representations that transfer zero-shot to unseen knowledge graphs. Finally, GraphGPT [44] casts graphs as reversible token sequences via Eulerian paths, enabling transformer-based generative pretraining that scales with model size.

Graphs and Large Language Models. A growing line of research investigates how large language models can be adapted to reason over graph-structured and relational data. [13, 26] explore parameter-efficient encoders that converts graphs into soft prompts for frozen LLMs, showing that performance strongly depends on the choice of graph serialization and structure encoding. [40] investigate predictive modeling directly on relational databases with LLMs, demonstrating that careful

schema-aware prompt design improves over naive text flattening. Building on this idea, [39] propose Rel-LLM, a hybrid architecture that combines GNN encoders with LLMs in a retrieval augmented generation framework. These works highlight the potential of combining graph or relational encoders with LLMs, but they are often limited by small context windows and are not tailored to relational databases. In contrast, our proposed Relational Transformer directly encodes multi-table structure via attention masks, offering a fully end-to-end solution that can operate either independently or alongside large language models.

B Background (Full Version)

B.1 Relational Databases

A *relational database (RDB)* is a collection of *tables* linked through inter-table relationships. Each table is composed of *rows*, where every row is a set of *cells*, one for each *column* in the table. We define *feature columns* as the columns that contain numeric, text and datetime information, and ID columns then define how rows are uniquely identified and connected across tables. Every table has a *primary key (P–key)*, and some include *foreign keys (F–keys)* referencing primary keys in other tables. This induces a graph structure, where connections from foreign keys to primary keys are denoted as *F→P links*, and the reverse incoming connections as *P→F links*.

Many RDBs are *temporal*, with timestamp columns that record when rows are created. Temporal information is crucial: if we want to predict whether a user will buy an item at time t, the model must only use information available before t, otherwise it risks temporal leakage. To prevent temporal leakage, modeling is conditioned only on rows that were created prior to the target row. Finally, the *schema* of an RDB specifies the tables with their columns and datatype along with the relational structure. Because schemas vary widely, pretraining requires schema-agnostic architectures that directly incorporate multi-table structure through attention masks.

B.2 Predictive Tasks

Masked token prediction (MTP). We focus on *masked token prediction*, where the goal is to predict the value of a masked cell in the database, conditioned on the rest of the observed database. A broad class of important predictive tasks on RDBs can be framed as MTP, including (1) *autocomplete tasks* and (2) *forecasting tasks*.

Autocomplete tasks. Here the missing or masked value belongs to a feature column that already exists in the database. Consider the e-commerce schema with Users, Items, and Transactions tables. An autocomplete task might involve predicting a user's age in the Users table if the entry is missing, or inferring the category of an item in the Items table from its textual description and price. In both cases, the label comes from an existing feature column.

Forecasting tasks. Here, the goal is to predict something that has not yet happened. For example, in the e-commerce setting, we want to forecast whether a given user will churn in the next month, or predict the total revenue of a product in the upcoming quarter. Unlike autocomplete, the target values for forecasting do not exist in the original database and must be constructed from future rows.

Task tables. To formalize forecasting tasks, we introduce a new task table. This table stores the forecasting labels, together with foreign keys linking to the relevant entities (e.g., user IDs or item IDs) and a timestamp specifying the prediction horizon. For instance, a task table for churn prediction might contain one row per user, indicating whether the user made a purchase within the next 30 days, with a timestamp showing the cutoff date.

C Cell Token Encoding (Full Version)

A cell is represented by (v, c, t), where v is the cell value, c is the column name, and t is the table name. The value v can be numeric, boolean, datetime, or text; other modalities (e.g., image) can be handled analogously to text.

- **Numeric/boolean.** Normalize to obtain $\mathbf{r} = (v - \mu_c)/\sigma_c \in \mathbb{R}$, where μ_c and σ_c are the column mean and standard deviation computed on the training split. For booleans, this provides a calibrated scale.
- **Datetime.** Convert to seconds and normalize globally: $\mathbf{r} = (v - \mu_T)/\sigma_T$, where μ_T and $\sigma_T \in \mathbb{R}$ are the global mean and standard deviation of timestamps in the training split.
- **Text.** Embed using a frozen text encoder $\mathcal{E}^{\text{text}}$: $\mathbf{r} = \mathcal{E}^{\text{text}}(v) \in \mathbb{R}^{d_{\text{text}}}$.

Schema semantics are incorporated via a text embedding of the phrase "`<column_name> of <table_name>`", e.g., "`price of product`", "`age of user`", using $\mathcal{E}^{\text{schema}}$. The token embedding is $\mathbf{x} = \mathbf{W}_d\,\mathbf{r} + \mathbf{W}\,\mathcal{E}^{\text{schema}}(c, t)$, where $\mathbf{W}_d$ is datatype-specific and $\mathbf{W}$ is shared. For masked cells, the value embedding $\mathbf{W}_d\mathbf{r}$ is replaced with a learned mask vector $\mathbf{m}_d$.

RT does not rely on positional encodings, as relational structure is directly captured by specialized attention layers. While graph positional encodings [23, 20, 21, 2] could be incorporated, we leave the architecture free of them to maintain simplicity and generality.

D Relational Attention (Full Version)

The core of RT is a novel *Relational Attention* mechanism in which the fundamental processing unit is the *cell token*. This formulation enables flexible pretraining via MTP, and stands in contrast to Graph Transformers [10, 28, 42], which tokenize at the row level, but can be viewed as a natural extension of Tabular Transformers [17, 18]. By operating at the cell level, RT can explicitly model one-to-one dependencies between attributes across rows, columns, and tables, while also supporting zero-shot generalization across schemas. RT follows the standard transformer design, but augments each block with Relational Attention layers that effectively encode relational structure and inductive bias. Other architectural details (normalization, activations, etc.) follow the design choices of LLaMA [34].

The main operation in RT is the *scaled dot-product attention (SDPA)* with *masking*, given by:

$$\text{SDPA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}; \mathbf{M}) = \text{Softmax}\left(\frac{\text{Mask}\left(\mathbf{Q}\mathbf{K}^\top; \mathbf{M}\right)}{\sqrt{d_k}}\right)\mathbf{V}, \quad \text{Mask}(\mathbf{A}; \mathbf{M})_{ij} = \begin{cases} \mathbf{A}_{ij} & \text{if } \mathbf{M}_{i,j} = 1 \\ -\infty \text{ if } \mathbf{M}_{i,j} = 0 \end{cases}$$

Here, $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, $\mathbf{K} \in \mathbb{R}^{n \times d_k}$, $\mathbf{V} \in \mathbb{R}^{n \times d_v}$ are the *query*, *key* and *value* matrices, and n is the context length. $\mathbf{M} \in \{0, 1\}^{n \times n}$ is the *attention mask*, which controls token-to-token visibility. $\mathbf{M}[q, k] = 1$ means the q-th token can attend to the k-th token, and $\mathbf{M}[q, k] = 0$ means it cannot. For example, auto-regressive language models use a *causal* attention mask, given by $\mathbf{M}^{\text{causal}}[q, k] = \mathbf{1}\{k \leq q\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

Relational Attention masks. Using specialized masks, we define four attention types: *column*, *feature*, *neighbor*, and *global*. For the cell corresponding to token i, let $\text{Col}(i)$ be its column, $\text{Row}(i)$ its row, and $\text{OutLinks}(i)$ the set of rows, possibly in different tables, which are pointed to by foreign keys of $\text{Row}(i)$.

- **Column attention.** For any query token, this layer allows attention only to key-value tokens from the same column, resulting in the mask: $\mathbf{M}^{\text{column}}[q, k] = \mathbf{1}\{\text{Col}(k) = \text{Col}(q)\}$. Column attention helps model the distribution of values in each column.
- **Feature attention.** For any query token, this layer allows attention to key-value tokens from the same row, as well as from F→P linked rows with the attention mask given by: $\mathbf{M}^{\text{feature}}[q, k] = \mathbf{1}\{\text{Row}(k) = \text{Row}(q) \vee \text{Row}(k) \in \text{OutLinks}(q)\}$. Feature attention is equivalent to row-wise attention after joining each table with its parent tables, and enables feature mixing for entities.
- **Neighbor attention.** For any query token, this layer allows attention to key-value tokens from P→F linked rows, defined by the attention mask: $\mathbf{M}^{\text{neighbor}}[q, k] = \mathbf{1}\{\text{Row}(q) \in \text{OutLinks}(k)\}$. Neighbor attention captures information from incoming links to an entity, enabling the model to aggregate signals from its child rows. This module acts analogously to message-passing in GNNs.
- **Full attention.** Finally, a standard bidirectional layer allows full pairwise interactions: $\mathbf{M}^{\text{full}}[q, k] = 1$. Full attention confers the expressive power of standard Transformers, complementing the relationally constrained layers above.

Taken together, the proposed attention layers provide the model with an explicit encoding of database structure. These layers are implemented with sparse attention masks, and compiled to efficient FlashAttention-based [6] kernels using FlexAttention [8]. The proposed transformer block in RT is summarized in Alg. 2.

E Output Decoding and Training Objective (Full Version)

Cell decoders / prediction heads. An output token embedding e' from the transformer backbone is processed by multiple cell decoders (also called prediction heads), one for each datatype, into a cell representation r'. The decoder to select for final prediction depends on the task type, or equivalently on the datatype of the masked cell. Binary classification corresponds to the boolean datatype, and regression corresponds to the numeric datatype.

Loss. Having separate decoders for different datatypes allows us to use custom loss functions for each task type. In this work, we only mask cells in boolean or numeric columns as RelBench tasks are either binary classification or regression. For a masked cell c with value v, representation r (as defined in § 3.1), and decoder output r', we apply $\text{HuberLoss}(r, r')$ for regression and binary cross-entropy loss $\text{BCE}(\mathbf{1}\{r > 0\}, r')$ for binary classification. The overall loss is the mean over all masked cells in the batch. This formulation is used in both pretraining and fine-tuning, ensuring consistency between objectives and contributing to sample efficiency.

F Experimental Setup (Full Version)

Architecture details. We use a 12 layer transformer with hidden dimension 256 and 8 attention heads per layer. We use gated MLPs with SiLU activation, as found in the Llama architecture [34], with hidden dimension 1024. For text embeddings, we use the MiniLMv2 [37] model from SentenceTransformers [29], which produces 384 dimensional embeddings. With this configuration, the architecture has about 22M trainable parameters.

Training details. We pretrain RT for 50k steps at a context length of 1024, with a batch size of 256, AdamW optimizer with weight decay 0.1, and a peak learning rate of 10^{-3}, with linear warmup from zero for the first 20% of training, and linear decay to zero for the remainder. One each downstream task, we fine-tune for 33k steps with the same context length, batch size and optimizer as above, but with a constant learning rate of 10^{-4} and no weight decay. One pretraining (fine-tuning) run takes around 2 hours (1.5 hours) on $8\times$A100 GPUs at BFloat16 precision, with a training throughput of around 8 batches/second or 2M tokens/second.

G Supervised Fine-Tuning Results (Full Version)

In this section, we present detailed supervised fine-tuning results. Figures 4 and 5 show per-task learning curves for classification and regression tasks, respectively. We then provide additional full fine-tuning results in Section G.1, analyzing the effect of pretraining in high-resource settings and comparing RT against both schema-specific and schema-agnostic baselines.

G.1 Supervised learning in high-resource settings

Setup. In Table 3, we report results from full-dataset fine-tuning, using up to several million training examples and continuing until convergence (tens of thousands of steps). We compare against schema-specific baselines (RDL-GNN, RelGNN, and RelGT), which cannot be pretrained, as well as schema-agnostic baselines (RelLLM and Griffin). For RT and Griffin, we evaluate both untrained and pretrained initializations to assess the impact of pretraining. For all methods, the best checkpoint is selected based on validation set performance. For RelGNN, RelGT, and RelLLM, we use the original training setups and hyperparameters. For Griffin, we increase the model size and update the sampling and pretraining procedures to be consistent with RT.

Observations. The pretrained RT achieves the best performance on average, achieving the highest mean AUROC and R^2. On classification, it is outperformed on certain tasks by RelGNN, RelGT and RelLLM, but it is important to note that these methods utilize custom setups for each task, whereas RT uses a single unified hyperparameter setup across all experiments. On regression, pretrained

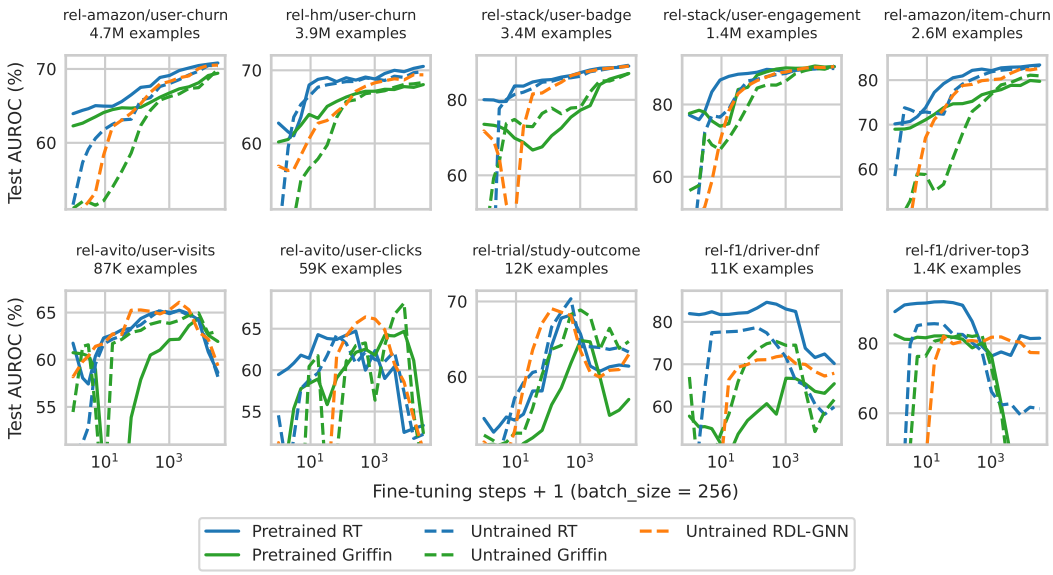

Figure 4: Per-task test set learning curves on classification tasks for up to 32k fine-tuning steps (8M training examples, including repetitions). X-axis is on log-scale.

Table 3: Supervised fine-tuning results. Models are trained on the full training set until convergence, with checkpoint selection based on validation performance. RT achieves the best mean AUROC and R^2 across tasks, surpassing both schema-specific (cannot be pretrained) and schema-agnostic (can be pretrained) baselines.

Dataset	Task	Train set size (sorted)	Cannot be pretrained			Can be pretrained				
			RDL GNN	Rel GNN	Rel GT	Rel LLM	Griffin	Griffin	RT (Ours)	RT (Ours)
		pretrained? →	No	No	No	Yes	No	Yes	No	Yes
AUROC (%) for 10 binary classification tasks. Higher is better. Random/majority baseline is 50.0.										
rel-amazon	user-churn	4.7M	70.7	71.0	70.4	**71.9**	70.0	69.4	70.5	70.8
rel-hm	user-churn	3.9M	69.4	**70.9**	69.3	70.5	68.3	68.0	69.9	70.5
rel-stack	user-badge	3.4M	88.9	89.0	86.3	**89.6**	87.0	87.0	88.5	88.7
rel-stack	user-engage	1.4M	90.6	90.8	90.5	**91.2**	89.8	90.4	90.0	90.2
rel-amazon	item-churn	2.6M	82.8	82.6	82.5	83.4	81.1	79.9	83.2	**83.4**
rel-avito	user-visits	87K	66.1	66.2	66.8	**67.0**	65.0	62.6	65.0	65.2
rel-avito	user-clicks	59K	63.1	68.2	**68.3**	66.7	63.0	64.7	63.6	59.0
rel-trial	study-out	12K	68.6	**71.2**	68.6	71.0	68.9	64.6	68.6	68.2
rel-f1	driver-dnf	11K	72.5	75.3	75.9	77.2	74.5	66.7	78.7	**84.2**
rel-f1	driver-top3	1.4K	80.9	85.7	83.5	82.2	82.5	78.7	82.7	**91.9**
		Mean AUROC →	75.4	77.1	76.2	77.1	75.0	73.2	76.1	**77.2**
R^2 (%) for 8 regression tasks. Higher is better. Global-mean baseline is 0.0.										
rel-hm	item-sales	5.5M	21.8	22.1	22.6	*nan*	31.1	30.4	**45.7**	39.0
rel-amazon	user-ltv	4.7M	21.9	17.9	17.5	*nan*	30.7	32.9	**47.9**	47.4
rel-amazon	item-ltv	2.7M	3.7	3.5	3.4	*nan*	23.4	25.2	31.5	**36.8**
rel-stack	post-votes	2.5M	17.9	12.2	13.1	*nan*	41.8	**42.7**	37.1	36.5
rel-trial	site-succ	150K	4.0	−9.5	−28.8	*nan*	**6.8**	−2.4	−8.8	6.4
rel-trial	study-adv	43K	18.8	19.7	17.0	*nan*	11.2	18.2	41.3	**43.4**
rel-f1	driver-pos	7.5K	7.6	20.7	12.4	*nan*	29.9	0.6	33.7	**51.6**
rel-avito	ad-ctr	5.1K	18.3	15.6	**18.4**	*nan*	5.9	8.4	−1.5	4.5
		Mean R^2 →	14.3	12.8	9.4	*nan*	22.6	19.5	28.4	**33.2**

RT ranks best on average, substantially outperforming the second-best method across most tasks. Overall, RT matches or exceeds the performance of schema-specific baselines while maintaining a general, schema-agnostic design, showing that generalization does not come at the expense of fine-tuning performance.

Figure 5: Per-task test set learning curves on regression tasks for up to 32k fine-tuning steps (8M training examples, including repetitions). X-axis is on log-scale.

H Zero-Shot Results (Full Version)

This section provides the complete zero-shot results for all individual tasks, broken down by dataset and task type. These results support the summary presented in Section 4.2 of the main paper. Table 5 details the full regression results, including the LLM baselines, which, notably, fail to make meaningful predictions for regression tasks. For reference, Table 4 is a copy of Table 1 from Section 4.

Setup. The target task is always unseen. We report results for when the target dataset is also unseen, as well as after doing some continued pretraining on the target dataset. For RT and Gemma, we construct the input context using our sampling algorithm. For Griffin, we adapt its sampling procedure to include rows from task tables, making it consistent with our approach and enabling zero-shot capabilities. For RelLLM we use their own prompt construction. Both RelLLM and Gemma baselines are additionally provided with textual task descriptions and natural language instructions.

Table 4: Zero-shot test AUROC (%) for 10 binary classification tasks. Higher is better. Random/-majority baseline is 50.0. For RelLLM we use their own prompt construction. Other baselines have equivalent database subgraphs. Gemma and RelLLM additionally include dataset and task descriptions, as well as natural language instructions. The target task is never seen during pretraining.

Target DB ∈ pretraining? →		Maybe			No			Yes		
Dataset ↓	Task ↓	Gemma	Gemma	Gemma	Entity Mean	Griffin	RT (ours)	Rel LLM	Griffin	RT (ours)
	Parameter count →	4B	12B	27B	0	22M	22M	3B	22M	22M
rel-amazon	item-churn	62.1	55.0	42.1	73.0	69.0	70.9	64.1	71.9	**73.3**
rel-amazon	user-churn	58.1	54.7	50.5	64.4	62.3	64.0	60.1	64.1	**66.1**
rel-avito	user-clicks	54.5	59.5	59.8	44.7	45.9	59.5	**62.3**	45.9	60.9
rel-avito	user-visits	60.1	57.9	**62.7**	60.7	60.7	61.8	56.2	62.2	62.6
rel-f1	driver-dnf	56.2	54.6	75.8	75.4	57.7	81.2	71.8	57.7	**81.2**
rel-f1	driver-top3	84.6	90.5	**91.4**	85.0	82.5	89.3	70.6	81.8	89.3
rel-hm	user-churn	59.8	47.1	48.7	**64.4**	60.2	62.8	56.0	60.4	63.3
rel-stack	user-badge	79.1	79.8	80.0	66.2	73.5	80.1	62.1	**82.3**	81.1
rel-stack	user-engage	65.9	67.8	78.0	83.5	77.5	75.7	69.5	**89.4**	86.9
rel-trial	study-out	52.6	57.4	57.2	50.0	51.0	51.8	**59.0**	57.2	54.6
	Mean AUROC →	63.3	62.4	64.6	66.7	64.0	69.7	63.2	67.3	**71.9**

Table 5: Zero-shot R^2 (%) for 8 regression tasks. Higher is better. Global mean baseline is 0.0. Setup is same as Table 2.

Target dataset ∈ pretraining? →		Maybe			No			Yes		
Dataset ↓	Task ↓	Gemma	Gemma	Gemma	Entity Mean	Griffin	RT (ours)	Rel LLM	Griffin	RT (ours)
	Parameter count →	4B	12B	27B	0	22M	22M	3B	22M	22M
rel-amazon	item-ltv	< -9	< -9	< -9	**54.2**	20.1	32.5	–	20.1	32.2
rel-amazon	user-ltv	< -9	< -9	< -9	19.9	20.6	36.9	–	24.4	**38.3**
rel-avito	ad-ctr	< -9	< -9	-8.2	3.4	2.4	4.5	–	2.4	**8.0**
rel-f1	driver-pos	35.2	43.4	52.4	38.2	-0.7	52.4	–	4.6	**58.7**
rel-hm	item-sales	< -9	< -9	< -9	1.8	2.7	14.0	–	2.5	**30.9**
rel-stack	post-votes	< -9	< -9	< -9	**43.7**	27.4	33.9	–	27.1	35.0
rel-trial	site-succ	< -9	< -9	< -9	-6.4	1.4	4.5	–	2.6	**5.1**
rel-trial	study-adv	< -9	< -9	-7.1	-0.5	-2.5	2.6	–	-2.5	**3.1**
	Mean R^2 →	< -9	< -9	< -9	19.3	8.9	22.7	–	10.1	**26.4**

Observations RT demonstrates non-trivial zero-shot performance on all tasks. On classification (Table 4), it attains the best average AUROC and is the only method to consistently beat the entity mean baseline. On regression, where LLM baselines fail to provide meaningful predictions, RT is the only model to consistently achieve positive R^2 and significantly surpassing the EntityMean baseline on average.

I Ablation Studies (Full Version)

This section provides complete ablation results referenced in the main paper. All ablations use the same pretrained checkpoints as in the main experiments.

I.1 Context Construction Ablations

In Section 4.3, we summarize our ablations of the context window construction to analyze the emergence of zero-shot performance. Here, we provide the full results of that study. Specifically, we systematically remove or perturb individual context components and report their effect on both zero-shot transfer and supervised fine-tuning performance.

Setup. In Table 6, we report results when shuffling column and table names, removing past labels from the target entity, or removing labels from other entities. For zero-shot evaluation, the ablations are applied directly to the sampled context used as input. For fine-tuning, models are trained to convergence with the same modified contexts. In addition, Table 7 provides statistics on the number of label cells (mean $\pm$ std. dev.) included in a context window of length 1024 under our sampling procedure (Alg. 1).

Observations. We find that zero-shot transfer primarily arises from the presence of past labels of the target entity. Removing these labels causes the largest drop in performance, whereas removing labels from other entities has a smaller effect. Shuffling column and table names also harms transfer, highlighting the importance of semantic signals from schema metadata. In the fine-tuning setting, classification performance is largely robust to these ablations, but regression tasks consistently benefit from access to past labels of the target entity.

I.2 Architecture Ablations

In order to assess the contribution of the relational attention layers to zero-shot transfer and fine-tuning performance, we conduct a comprehensive ablation study. Here, we present the detailed results of our findings summarized in Section 4.3. Specifically, we remove individual attention layers—column, feature, neighbor, or full/global—and analyze their effect across regression tasks, while classification results (showing minor differences) are provided in App. I.

Table 8 shows the full Relational Attention layer ablations. We observe no clear patterns in the zero-shot setting, but during finetuning removing any layer results in a decrease in performance, except on the *user-clicks* task, where the model is prone to overfit.

Table 6: Ablation study of context construction. To explain the zero-shot performance we remove column names, past labels from the target entity and labels from other entities. To assess how much task-relevant information is lost, we repeat the same ablations in the fine-tuning setting. Shading indicates the performance difference relative to the full context (none column).

Dataset ↓	Task ↓	Zero-shot				Fine-tuned			
Ablated from context →		none	col names	self labels	other labels	none	col names	self labels	other labels
AUROC (%) for 10 binary classification tasks. Higher is better. Random/majority baseline is 50.0.									
rel-amazon	item-churn	70.2	71.0	48.1	72.2	83.4	83.3	83.4	83.2
rel-amazon	user-churn	63.9	63.9	55.2	64.2	70.8	70.4	70.8	70.6
rel-avito	user-clicks	59.5	58.5	55.0	59.7	59.0	60.8	61.0	60.1
rel-avito	user-visits	61.8	61.2	49.9	62.1	65.2	65.1	64.4	65.3
rel-f1	driver-dnf	82.0	81.7	50.3	82.0	84.2	84.3	83.6	84.3
rel-f1	driver-top3	89.1	86.5	74.3	89.1	91.9	91.8	89.7	91.7
rel-hm	user-churn	62.8	60.0	54.5	62.9	70.5	70.6	70.2	70.7
rel-stack	user-badge	80.0	79.2	54.8	82.4	88.7	88.9	88.8	88.8
rel-stack	user-engage	77.1	80.1	41.9	77.2	90.2	90.1	90.2	90.2
rel-trial	study-out	54.5	53.2	54.5	54.6	68.2	69.0	68.8	68.9
Mean AUROC →		70.1	69.5	53.8	70.6	77.2	77.5	77.1	77.4
R^2 (%) for 8 regression tasks. Higher is better. Global-mean baseline is 0.0.									
rel-amazon	item-ltv	33.2	33.0	-2.8	33.1	36.8	34.7	28.0	35.1
rel-amazon	user-ltv	36.4	33.6	-5.7	36.1	47.4	47.2	21.9	37.6
rel-avito	ad-ctr	4.5	5.7	-3.6	4.5	4.5	4.9	-5.2	4.3
rel-f1	driver-pos	54.7	50.3	-31.9	54.7	51.6	52.3	54.7	54.1
rel-hm	item-sales	14.0	9.2	-2.8	14.7	39.0	39.5	33.6	39.3
rel-stack	post-votes	32.4	27.5	-0.7	32.8	36.5	35.9	34.9	37.2
rel-trial	site-succ	5.2	3.2	1.5	5.1	6.4	8.3	8.7	2.5
rel-trial	study-adv	2.1	1.3	2.1	2.1	43.4	42.4	37.0	38.1
Mean R^2 →		22.8	20.5	-5.5	22.9	33.2	33.2	26.7	31.0

J Autocomplete tasks

Definition. Autocomplete tasks are defined as masked cell prediction on feature columns that already exist in the database. Unlike forecasting tasks, they do not require constructing additional task tables. The input sequence to the model is constructed in the same way as for forecasting tasks, preserving the relational structure and temporal order, but sampling starts from the masked database row.

Task selection. All autocomplete tasks were selected manually by inspecting the database schema. For each task, we also identify potential sources of information leakage and discard these columns on the fly when building the input sequence.

Task overview. Tables 10 and 11 list all classification and regression autocomplete tasks, respectively, together with label distributions (classification) and summary statistics (regression).

K Relational Transformer Implementation Details

Sampling cells for the context window. In § 2, prediction tasks such as forecasting and autocompletion are framed as predicting a masked cell in the appropriate row (the *seed row*). We sample the context window independently for each training/testing example, so there is always a unique seed row for context construction. Given the seed row and context length L, a suitable algorithm should select the cells most relevant to predicting the masked cell. Since relevance requires strong models to estimate accurately, we use a simple heuristic guided by the intuition that most relevant information lies within a few hops of the seed row when following F→P and P→F links, and that lower hops are more informative than higher hops.

We treat rows as the sampling unit: once a row is selected, all its non-missing feature cells (i.e., cells not from primary- or foreign-key columns) are included in the context. After the seed row, other rows are added using a bounded-width BFS across F→P and P→F links, with the following modifications: (1) F→P links are immediately followed; (2) the traversal stops when the total

Table 7: Breakdown of "in-context labels" sampled by Alg. 1. Context length is 1024 cells. Numbers are (mean ± std. dev.; both rounded to the nearest integer). **Target entity**, e.g., user, item, etc., is the one for which prediction is desired. **Labels** refer to unmasked cells from the target column. **Other entities** are reached via graph traversal. Multiple labels are possible for the same entity as the tasks are temporal.

Dataset	Task	Target entity labels	Other entity labels	Unique labeled entities
		Binary Classification Tasks		
rel-amazon	user-churn	4 ± 4	1 ± 5	2 ± 3
rel-hm	user-churn	6 ± 4	0 ± 0	1 ± 0
rel-stack	user-badge	7 ± 6	3 ± 6	1 ± 1
rel-amazon	item-churn	8 ± 7	2 ± 6	2 ± 3
rel-stack	user-engagement	16 ± 10	10 ± 10	3 ± 1
rel-avito	user-visits	2 ± 2	0 ± 0	1 ± 0
rel-avito	user-clicks	1 ± 1	0 ± 0	0 ± 1
rel-trial	study-outcome	0 ± 0	0 ± 0	0 ± 0
rel-f1	driver-dnf	19 ± 14	0 ± 0	1 ± 0
rel-f1	driver-top3	17 ± 11	0 ± 0	1 ± 0
		Regression Tasks		
rel-hm	item-sales	39 ± 13	0 ± 3	1 ± 1
rel-amazon	user-ltv	4 ± 4	1 ± 5	2 ± 3
rel-amazon	item-ltv	9 ± 8	2 ± 6	2 ± 3
rel-stack	post-votes	16 ± 10	4 ± 14	2 ± 2
rel-trial	site-success	1 ± 2	0 ± 1	1 ± 1
rel-trial	study-adverse	0 ± 0	0 ± 0	0 ± 0
rel-f1	driver-position	14 ± 10	0 ± 0	1 ± 0
rel-avito	ad-ctr	1 ± 1	0 ± 0	0 ± 1

Table 8: Ablation studies on the attention layers of RT on classification tasks. **col, feat, nbr, full** denote that *column-, feature-, neighbor-, full-* attention layers are absent respectively. Total parameter count is kept constant by increasing the number of layers. Shading is proportional to difference from the **none** column.

Dataset ↓	Task ↓	Zero-shot					Fine-tuned				
	Ablated attention →	none	col	feat	nbr	full	none	col	feat	nbr	full
AUROC (%) for 10 binary classification tasks. Higher is better. Random/majority baseline is 50.0.											
rel-amazon	item-churn	70.2	60.8	73.0	70.8	71.5	83.4	83.3	83.1	82.2	83.2
rel-amazon	user-churn	63.9	63.2	62.8	62.9	63.1	70.8	70.7	70.4	69.3	70.3
rel-avito	user-clicks	59.5	63.0	62.0	58.7	61.5	59.0	62.1	64.9	62.4	63.3
rel-avito	user-visits	61.8	60.9	62.9	62.7	60.6	65.2	65.6	65.3	64.6	65.0
rel-f1	driver-dnf	82.0	79.4	77.1	81.4	81.9	84.2	83.8	82.2	81.1	82.2
rel-f1	driver-top3	89.1	87.5	85.1	88.0	89.3	91.9	92.0	85.9	90.4	90.2
rel-hm	user-churn	62.8	66.1	65.8	65.6	64.4	70.5	69.8	70.1	69.4	70.1
rel-stack	user-badge	80.0	79.4	82.0	79.7	81.2	88.7	89.2	88.9	88.1	88.8
rel-stack	user-engage	77.1	78.2	80.9	79.1	83.1	90.2	90.0	89.5	89.2	90.0
rel-trial	study-out	54.5	59.4	54.8	58.6	54.8	68.2	66.6	66.5	68.3	67.1
	Mean AUROC →	70.1	69.8	70.6	70.8	71.1	77.2	77.3	76.7	76.5	77.0

number of cells reaches the context length; (3) unvisited rows at the same depth from the seed row are sampled uniformly; and (4) rows with timestamps greater than the seed row's timestamp are skipped (temporal constraint). The algorithm is summarized in Alg. 1. We use a fast, optimized Rust implementation to prevent on-the-fly sampling from slowing training.

In Alg. 1, F→P and P→F links are treated asymmetrically. The number of F→P links from a row is limited by the number of foreign-key columns in that table, and each parent row typically contains important features (e.g., a `transaction` row links to `user` and `product`). Conversely, P→F links can have unbounded degree; informative signals from P→F links often arise via aggregation, with diminishing returns from including many children. Thus we prioritize F→P links (followed immediately without subsampling) and subsample P→F links by enforcing a width bound w (i.e., follow at most w children from any row).

Table 9: Pretraining with multi-cell masking. Masked cells contribute to the loss. The target cell is always masked. Other cells are masked with probability **P(mask)**. **NP** denotes no pretraining. Shading is proportional to difference from the P(mask) $= 0.0$ column.

Dataset ↓	Task ↓	Zero-shot			Fine-tuned			
	P(mask) →	0.0	0.2	0.4	0.0	0.2	0.4	NP
AUROC (%) for 10 binary classification tasks. Higher is better. Random/majority baseline is 50.0.								
rel-amazon	item-churn	70.2	70.5	72.1	83.4	83.0	82.9	83.2
rel-amazon	user-churn	63.9	62.6	62.9	70.8	70.7	70.6	70.5
rel-avito	user-clicks	59.5	61.5	61.1	59.0	62.3	62.2	63.6
rel-avito	user-visits	61.8	63.0	63.2	65.2	65.5	65.5	65.0
rel-f1	driver-dnf	82.0	80.9	76.7	84.2	81.7	77.7	78.7
rel-f1	driver-top3	89.1	89.8	87.6	91.9	91.0	91.2	82.7
rel-hm	user-churn	62.8	60.5	62.3	70.5	69.9	69.9	69.9
rel-stack	user-badge	80.0	79.1	77.9	88.7	88.3	88.7	88.5
rel-stack	user-engage	77.1	75.0	73.6	90.2	90.1	90.0	90.0
rel-trial	study-out	54.5	55.1	55.2	68.2	70.2	68.2	68.6
	Mean AUROC →	70.1	69.8	69.3	77.2	77.3	76.7	76.1
R^2 (%) for 8 regression tasks. Higher is better. Global-mean baseline is 0.0.								
rel-amazon	item-ltv	33.2	9.3	13.2	36.8	37.0	30.9	31.5
rel-amazon	user-ltv	36.4	25.8	16.4	47.4	49.0	49.6	47.9
rel-avito	ad-ctr	4.5	8.0	10.8	4.5	3.6	1.9	−1.5
rel-f1	driver-pos	54.7	46.8	42.4	51.6	50.1	47.7	33.7
rel-hm	item-sales	14.0	10.0	6.2	39.0	50.7	53.5	45.7
rel-stack	post-votes	32.4	33.5	32.3	36.5	39.9	38.5	37.1
rel-trial	site-succ	5.2	1.4	3.0	6.4	7.1	6.6	−8.8
rel-trial	study-adv	2.1	0.5	−1.8	43.4	47.3	46.8	41.3
	Mean R^2 →	22.8	16.9	15.3	33.2	35.6	34.4	28.4

Table 10: Autocomplete **classification** tasks with distributions of observed non-missing labels (proportions in parentheses). When applicable, the positive/negative value mapping is provided.

Dataset	Table	Column	Pos./Neg. values	Non-missing	Positive (prop.)	Negative (prop.)
rel-amazon	review	verified	N/A	20 862 040	14 493 882 (0.69)	6 368 158 (0.31)
rel-avito	SearchInfo	IsUserLoggedOn	N/A	2 579 289	827 095 (0.32)	1 752 194 (0.68)
rel-stack	postLinks	LinkTypeId	1 vs 3	103 969	89 076 (0.86)	14 893 (0.14)
rel-trial	studies	has_dmc	t vs f	234 467	79 850 (0.34)	154 617 (0.66)
	eligibilities	adult	t vs f	273 160	251 581 (0.92)	21 579 (0.08)
	eligibilities	child	t vs f	273 160	51 899 (0.19)	221 261 (0.81)
rel-event	event_interest	not_interested	N/A	15 398	514 (0.03)	14 884 (0.97)

K.1 Relational Transformer Block

Algorithm 2 illustrates the architecture of a single Relational Transformer Block. This block consists of a series of attention mechanisms: a column attention layer, a feature attention layer, a neighbor attention layer, and a global attention layer, each with its own specific relational inductive bias. These Relational Attention layers are followed by a feed-forward network (MLP) for further processing.

K.2 Discussion

RT preserves the natural symmetries of relational data. In particular, the architecture is invariant to permutations of rows, columns, and tables, providing an inductive bias that improves generalization. This contrasts with LLM-based approaches for relational data, which are often highly sensitive to prompt order and formatting. Permutation invariance has been a key driver of success in prior graph neural networks, and respecting such symmetries has also shown benefits for large language models.

We also discuss role of the specialized attention layers in determining the expressive power of RT. For empirical evidence, see § 4.3.

Table 11: Autocomplete **regression** tasks with summary statistics of observed non-missing labels (rounded to two decimals).

Dataset	Table	Column	Non-missing	Min	Max	Median	Mean
rel-amazon	review	rating	20 862 040	0.0	5.0	5.0	4.39
rel-f1	results	position	15 207	1.0	33.0	7.0	7.97
	qualifying	position	9 815	1.0	28.0	11.0	11.24
	constructor_results	points	12 290	0.0	66.0	0.0	3.86
	constructor_standings	position	13 051	1.0	22.0	7.0	7.27
rel-hm	transactions	price	15 453 651	0.0	0.51	0.03	0.03
rel-trial	studies	enrollment	271 866	0.0	188 814 085.0	60.0	3 975.83
rel-event	users	birthyear	36 715	1900.0	1999.0	1991.0	1988.74

Algorithm 1: Sampling the context window of Relational Transformer. We use a modified Breadth-First Search (BFS) algorithm, with accommodations for relational-specific considerations, such as F→P links and temporal constraints.

Input: seed row s, context length L, width bound w, and, for each row r in the database: non-missing feature cells $\mathcal{C}(r)$, P→F neighbors $\mathcal{N}_{\text{P}\to\text{F}}(r)$, F→P neighbors $\mathcal{N}_{\text{F}\to\text{P}}(r)$ and timestamp $\mathcal{T}(r)$

Output: the set of database cells C in the context window

$C \leftarrow \{\}, F \leftarrow \{s\}$ `// F is the frontier of rows to explore`

while $|C| < L \wedge F \neq \{\}$ **do**

  `/* select a row to explore; R is the set of candidates */`

  $R \leftarrow \{r \in F \mid r$ was added via an F→P link$\}$ `// F→P linked rows`

  if $R = \{\}$ **then**

    $R \leftarrow \arg\min_{r \in F} \textsc{HopDistance}(r, s)$ `// rows closest to` s

  $r \leftarrow \textsc{RandomSelect}(R)$ `// pick a row at random`

  $F \leftarrow F \setminus \{r\}$ `// remove row from frontier`

  if r *has been visited* **then** continue **else** mark r as visited `/* visit row */`

  $C \leftarrow C \cup \mathcal{C}(r)$ `// add cells to context`

  $F \leftarrow F \cup \mathcal{N}_{\text{F}\to\text{P}}(r)$ `// add F→P neighbors to frontier`

  $N \leftarrow \{q \in \mathcal{N}_{\text{P}\to\text{F}}(r) \mid \mathcal{T}(q) \leq \mathcal{T}(s)\}$ `// filter P→F neighbors by time`

  $N \leftarrow \textsc{RandomSample}(N, w)$ `// pick` $\leq w$ `P→F neighbors at random`

  $F \leftarrow F \cup N$ `// add P→F neighbors to frontier`

return C

Column attention. Removing column attention does not reduce expressivity, since global attention can emulate it. In particular, some global heads can learn to restrict attention to tokens from the same column by exploiting table and column name embeddings in their query–key construction.

Feature attention. Feature attention is the only mechanism that explicitly groups cells into rows. While neighbor attention provides partial information,cells in the same row attend to the same set of neighbors, it cannot uniquely disambiguate rows, especially in tables without incoming foreign keys.

Neighbor attention. Similar to column attention, neighbor attention is not strictly required for expressivity, as feature and global attention together can simulate its effect. Feature attention exposes F→P links, which global attention can then leverage to infer P→F relationships.

Full attention. Without full attention, information can only propagate one hop per layer, limiting expressivity to local message passing. By contrast, full attention enables long-range interactions in a fixed number of layers, independent of database or graph diameter.

We leave the full theoretical characterization of RT's expressive power to future work.

Algorithm 2: A transformer block in RT.

Input: input token representations $\mathbf{X} \in \mathbb{R}^{n \times d}$
Output: output token representations $\mathbf{X} \in \mathbb{R}^{n \times d}$
$\mathbf{X} \leftarrow \mathbf{X} + \textsc{Norm}(\ \text{MHA}(\mathbf{X}; \mathbf{M}^{\text{column}})\ )$
$\mathbf{X} \leftarrow \mathbf{X} + \textsc{Norm}(\ \text{MHA}(\mathbf{X}; \mathbf{M}^{\text{feature}})\ )$
$\mathbf{X} \leftarrow \mathbf{X} + \textsc{Norm}(\ \text{MHA}(\mathbf{X}; \mathbf{M}^{\text{neighbor}})\ )$
$\mathbf{X} \leftarrow \mathbf{X} + \textsc{Norm}(\ \text{MHA}(\mathbf{X}; \mathbf{M}^{\text{full}})\ )$
$\mathbf{X} \leftarrow \mathbf{X} + \textsc{Norm}(\ \text{MLP}(\mathbf{X})\ )$
return $\mathbf{X}$

L Baseline Implementations

L.1 LLM Prompt Construction

Large language models (LLMs) are evaluated under the same information regime as our relational transformer (RT): input to both is constructed from the same context subgraph produced by our sampling algorithm (Alg. 1). In this graph, nodes correspond to database rows and edges represent F→P and P→F links. We serialize the sampled entity graph into **JSON**, which encodes relational structure.

Serialization procedure. We begin with the subgraph produced by the sampler. Serialization starts at the *task node*, which specifies the prediction timestamp and links directly to the target entity for which the label is to be predicted. From this target entity, we traverse the relational graph using both F→P and P→F links. Each visited row is merged into the existing record in the case of F→P link or further serialized and appended as a new entry to the list of linked entities in the case of P→F link.

Prompt components. Each prompt follows a fixed four-part structure: (i) a short dataset description; (ii) a description of the prediction task; (iii) the serialized graph context (a JSON of table–row objects) including the prediction timestamp t_0; and (iv) a concise instruction specifying the expected output ("yes" or "no"). Dataset and task descriptions are adapted from prior work [30].

Full prompt example.

```
You are a strict prediction assistant. Follow the instructions exactly.
# Database
Name: Stack Exchange
Description: Stack Exchange is a network of question-and-answer websites on different topics,
where questions, answers, and users are subject to a reputation award process. The reputation
system allows the sites to be self-moderating. The database includes detailed records of
    activity
including user biographies, posts and comments (with raw text), edit histories, voting, and
related posts. In our benchmark, we use the stats-exchange site.
# Task
Name: user-badge
Description: This task is to predict if this user will receive a new badge in the next 3
    months or not.
# Input
- Database serialization starting from the target instance, expanding context by including
    rows
  reached via f2p (foreign to primary) and p2f (primary to foreign) relationships.
- The first timestamp in the sequence denotes the prediction time t0.
# Database serialization for the target entity
{
  "timestamp": "2021-01-01T00:00:00",
  "UserId": 211098,
  "Id": 211098,
  "AccountId": 12827220.0,
  "DisplayName": "Shashwat Tiwary",
  "Location": null,
  "ProfileImageUrl": null,
  "WebsiteUrl": null,
  "AboutMe": null,
  "CreationDate": "2019-09-15T05:33:35.413000",
  "add_badges": [
```

```
    {"Id": 383629, "UserId": 211098, "Class": 3, "Name": "Editor", "TagBased": false,
     "Date": "2019-09-15T07:40:23.563000"}
  ],
  "add_user-badge": [
    {"timestamp": "2020-10-01T00:00:00", "UserId": 211098, "WillGetBadge": "no"},
    {"timestamp": "2020-04-02T00:00:00", "UserId": 211098, "WillGetBadge": "no"},
    ...
    {"timestamp": "2019-10-03T00:00:00", "UserId": 211098, "WillGetBadge": "no"}
  ]
}
# Output
- Output exactly one word on a single line: yes or no.
- No units, no punctuation, no spaces, no commas, no extra text, no extra symbols, no new
    lines.
Make your prediction for the target entity at t0 using database serialization,
database description, and task description.
```

In-context labels. Due to the nature of our sampling algorithm, past (unmasked) labels from the target column can remain in the serialized JSON. For example, in the `user-badge` task (see full prompt above), the nested entries under `add_user-badge` constitute such in-context labels. More details on the occurrence and distribution of these labels are provided in Table 7.

L.2 Regression results with LLM baselines

In addition to classification, we evaluated zero-shot regression with LLMs of varying sizes under the same RT information regime. Across eight regression tasks, performance was consistently poor—smaller models even failed to produce stable numerical outputs under strict prompting. We attribute this to unconstrained number generation and a context not optimized for LLM regression. Prior work shows that carefully selecting and formatting in-context examples can substantially improve results [24]. Given these limitations, we do not report detailed regression metrics.

L.3 Griffin

To ensure a fair comparison with Griffin, we scale its hidden dimension from 512 to 728, resulting in a comparable parameter count to RT (22.8M vs. 22.3M). We also match the training setup by adopting both the leave-one-database-out and continued pre-training regimes. However, since the Griffin implementation does not support joint training on forecasting and autocomplete tasks, we restrict it to forecasting tasks only.