
Foundation Models for Tabular Data within Systemic Contexts Need Grounding

Tassilo Klein
SAP SE
tassilo.klein@sap.com

Johannes Hoffart
SAP SE
johannes.hoffart01@sap.com

Abstract

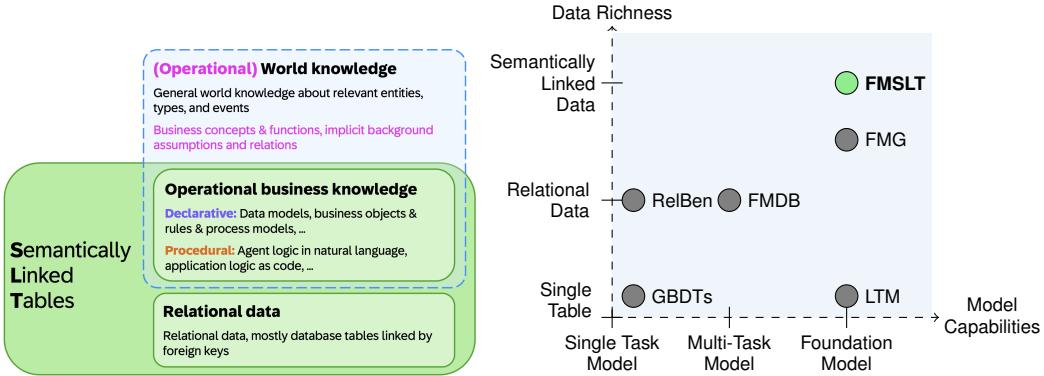
Current research on tabular foundation models often overlooks the complexities of large-scale, real-world data by treating tables as isolated entities and assuming information completeness, thereby neglecting the vital operational context. To address this, we introduce the concept of Semantically Linked Tables (SLT), recognizing that tables are inherently connected to both declarative and procedural operational knowledge. We propose Foundation Models for Semantically Linked Tables (FMSLT), which integrate these components to ground tabular data within its true operational context. This comprehensive representation unlocks the full potential of machine learning for complex, interconnected tabular data across diverse domains. Realizing FMSLTs requires access to operational knowledge that is often unavailable in public datasets, highlighting the need for close collaboration between domain experts and researchers. Our work exposes the limitations of current tabular foundation models and proposes a new direction centered on FMSLTs, aiming to advance robust, context-aware models for structured data.

1 Introduction

Recent advancements in machine learning have produced foundation models. These models are notable for their capacity to generalize across diverse tasks and datasets, extending beyond the confines of training data [70, 149, 157]. Their task and data agnostic characteristic [12] distinguishes them from traditional models, offering a more flexible and adaptable paradigm. However, the application of these models to tabular data analysis is often hindered by simplifying assumptions, particularly in complex real-world settings. We perceive that the work on foundation models for tabular data sometimes conflates different problems. First, it focuses primarily on ML on “isolated tables” [146], a perspective that might be legitimate in certain scenarios but fails to reflect the realities of intricate data ecosystems. Second, multi-table methods using, for example, graph neural networks (GNNs) [38, 120], while effective in capturing relational structures, often assume information completeness within tables, neglecting crucial semantic context that is required for understanding the data generated by real-world applications. Although current models display impressive generalization, the oversimplification of tabular data presents a significant gap that must be addressed to unlock the full potential of foundation models within complex data environments.

Recent work suggests that foundation models contain an implicit world model [1, 93, 145]. Although world models can, in principle, be based purely on statistical associations or other forms of implicit structure, we hypothesize that, for structured data, explicitly modeling semantic context may promote greater generalizability and robustness. Motivated by this perspective, we introduce Semantically Linked Tables (SLT), which leverage semantic relationships as the foundation for world modeling in structured data. We acknowledge that tables are inherently linked to operational knowledge. This knowledge includes both declarative and procedural components. It is often created to facilitate the development of applications and is usually encoded within diverse artifacts. When combined with world knowledge, these artifacts act as a unifying mechanism. They form a “semantic frame” that

Figure 1: **Left:** The composition of SLT. **Right:** Comparing FMSLT to alternative views.



governs data operations (i.e., how applications write data to databases). This semantic frame typically resides externally to databases that store the actual application data - see Fig. 1.

We propose **Foundation Models for Semantically Linked Tables (FMSLT)** as models to integrate operational knowledge, including declarative and procedural aspects, to ground tables within their real-world context. This proposal directly addresses the shortcomings of existing models that oversimplify tabular data by neglecting the rich operational and semantic context in which real-world data is embedded. This grounding encompasses intra- and inter-table relationships, rich contextual metadata, and procedural logic. Prior work highlights that the lack of explicit reasoning capabilities limits model performance, especially in scenarios requiring multi-hop and cross-table interactions, as observed in text-to-SQL tasks [113]. By capturing latent interactions and enabling a deeper understanding of data processes, FMSLTs aim to unlock the potential of machine learning on structured data. To illustrate how existing approaches fall short, consider the following example contrasting a vanilla ML approach and an FMSLT within an SLT scenario. Figure 2 showcases a simplified supply chain involving a manufacturer of configurable goods (in this case computers) with an associated webshop. The webshop allows the configuration of computers with compatible hardware elements, taking into account information from the warehouse and the availability of items. While this example is drawn from a business context, similar complexities arise in other domains such as healthcare, where operational contexts are equally critical for robust data-driven decision making. Here, SLT encompasses components such as product catalog and products with their configurable components, including warehouse management and supply tracking. For instance, when predicting internal material restocking requirements during production, a typical machine learning approach would be constrained to a company's order history, or perhaps a manually curated subset of data from past analysis, limited by the underlying data complexity c.f. [101, 81]. However, for more reliable predictions, it is crucial to recognize that effective material restocking relies on multiple SLT intricacies - see Fig. 2 for an associated sample multi-table schema (for a legend and more details see the Appendix):

First, *declarative knowledge*, e.g., that the required material was replaced by a substitute in the *product component graph*, a component is exchangeable with lower cost components, or some material has a lower failure rate due to an improved manufacturing process.

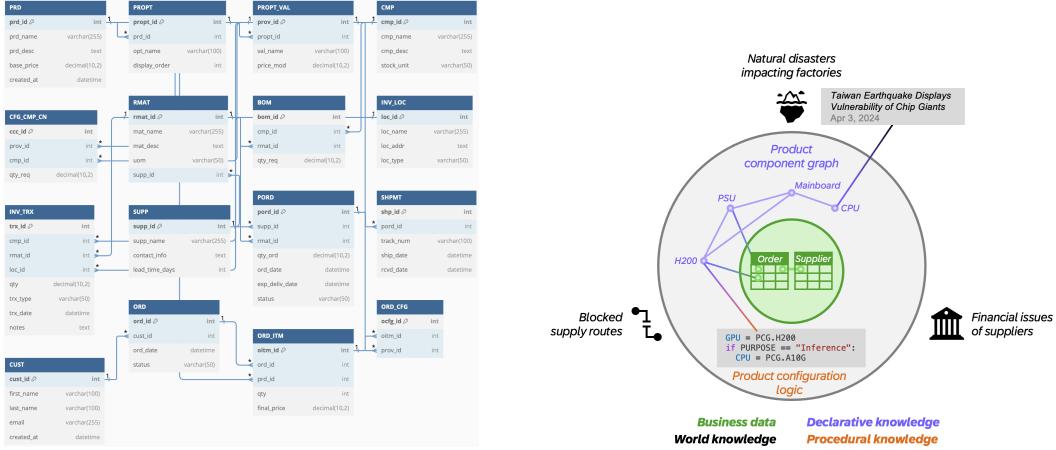
Second, *procedural knowledge*, e.g., that the material is no longer recommended as the default material or if the user-specified configuration is not compliant with manufacturing constraints as defined in the *product configuration logic* of the webshop.

Third, *world knowledge*, e.g., regarding factory disruptions, risk of blocked supply routes, or supplier financial issues, geopolitical issues should be considered long-term but are beyond the essential scope.

To get a holistic view, an FMSLT would leverage the interacting SLT components in context. FMSLTs require operational knowledge grounded in real-world contexts, which is typically not publicly available. While close collaboration between domain experts and researchers remains essential to obtain and contextualize such knowledge, we also anticipate that synthetic data will play an important role in enabling research by simulating operational scenarios and addressing privacy and accessibility challenges. This work highlights the limitations of existing tabular models, introduces FMSLTs as a new research direction, and aims to foster these collaborations for impactful real-world applications.

2 Semantically Linked Tables

Figure 2: Mockup supply-chain: **Left:** Multi-table schema. **Right:** Webshop example.



Taking a relational view on data recognizes the inherent interconnectedness of tables up to a certain degree, as data stored across multiple tables linked by foreign keys. However, not all context required for understanding is available in this form. SLT addresses this by recognizing the context-richness of data within real-world applications available in forms beyond just relational data. Unlike relational data, which primarily consists of structured data and transactions, SLT integrates both declarative (e.g., data models, rules, process models) and procedural (e.g., source code) operational knowledge that defines data usage and interpretation. This combination of relational data and operational knowledge is crucial for extracting meaningful insights. Relational data provides the factual basis, while operational knowledge provides the context, explaining data relationships and changes within data processes. SLT is characterized by its interconnectivity, reflecting the relational structure of entities [39], its context-richness, grounded in declarative and procedural knowledge, its dynamic nature, mirroring the evolving data landscape, and its often large scale. Beyond procedural and declarative knowledge, there is also world knowledge, which comprises both domain-specific and general knowledge that does not reside within a specific application or context. In many real-world environments, the available data resembles an archipelago of semi-isolated information islands. These “islands,” which are individual tables, are typically understood only by their creators and domain specialists. Each table, or cluster of related tables, embodies both application-specific details and an implicit conceptualization of the domain. Ideally, table schemas (including names and column headers) would possess semantic richness, enabling direct interpretation. However, in practice, these schema elements frequently act as opaque shorthands, demanding significant additional contextualization [171]. This fragmentation into semi-isolated tables substantially hinders the derivation of comprehensive insights and the establishment of meaningful connections across the overall data landscape. This opacity exceeds the limited context offered by relational database system (e.g., column/table names, foreign keys), which fail to capture the rich semantic context of tables. Research on training machine learning models on relational data highlights the challenges of extracting meaning from such interconnected datasets [120]. Insufficient metadata contributes to data “swamps”[102], where data volume, coupled with semantic ambiguity, hinders information retrieval. Current efforts to improve tabular data utilization focus on metadata enrichment, including using large language models (LLMs) to automatically generate metadata, enhance table metadata, and improve concept matching via enriched ontologies [102]. Simultaneously, approaches aim to improve dataset search and discovery through metadata enhancements [14, 91, 27, 3]. Beyond discoverability, rich metadata improves downstream tasks, for example, contextual information improves time series forecasting [160] and enhances tabular data analysis [32], particularly with text-heavy benchmarks.

2.1 Elements of Foundation Models for SLT

Training foundation models requires carefully curated data mixtures to achieve desired downstream functionality [110]. Recognizing that SLT intrinsically links tables to diverse declarative and procedural operational knowledge, accommodating these different knowledge types during pre-training

is crucial for developing grounded table foundation models capable of generating contextualized representations. This integration extends beyond processing isolated tables, allowing the capture of intricate relationships within real-world operational data. We detail this knowledge and its integration in the following sections.

2.1.1 Declarative Knowledge

Within the context of SLT, declarative knowledge is central for representing domain concepts, their interrelationships, and the foundational information required for reasoning and decision-making. For example, in the webshop scenario illustrated in Fig. 2, declarative knowledge includes the available computer components, their specifications, and compatibility relationships between them. This type of knowledge also encompasses the rules and constraints that govern the configuration process, often expressed as conditional statements (e.g., “This mainboard supports only DDR5 RAM”). In practice, such rules and policies are essential for ensuring that applications operate according to business requirements and are typically distributed across various artifacts and formats. The effective application of declarative knowledge relies on proper representation, usually captured in structured repositories across the application domain, including ontologies, knowledge graphs (KGs), data dictionaries, and domain glossaries. These artifacts act as formalized repositories for capturing, structuring, and accessing the information necessary to understand and execute processes effectively. Ontologies, ranging from simple taxonomies to complex structures, provide a formal representation of domain-specific concepts and their relationships, enabling reasoning and inference [134, 52, 53]. KGs, created by instantiating an ontology with actual data, offer a flexible and concrete way to represent declarative knowledge, structuring information as a graph, with nodes representing entities and concepts, and edges representing their relationships. This rich representation of interconnected information is particularly useful in capturing domain and foundational knowledge, as it provides a direct connection to relational data, a perfect fit for representing SLT. General domain KGs like BabelNet [107], DBpedia [4], Wikidata [151], and YAGO [136, 62] have demonstrated the potential of this approach. Since the debut of Google’s Knowledge Graph in 2012, numerous organizations across various domains have adopted KGs for semantic search, recommendations, fraud detection, risk assessment, and other applications [63, 64].

Integrating Declarative Knowledge: A major challenge in artificial intelligence (AI) involves combining the strengths of symbolic and neural approaches. Symbolic AI offers interpretability and logical structure, while neural networks excel at pattern recognition and adaptation. For an overview of neurosymbolic approaches, see [100, 152] and for reasoning on KGs, see [127, 31, 95, 25].

Declarative knowledge, often encoded in domain KGs, application logic, and process definitions, is inherently symbolic. This symbolic representation provides crucial context for understanding the interconnectedness of data. It naturally integrates with data-driven learning in a neurosymbolic framework. This integration aligns with the core concept of SLT, which recognizes that tables exist within an ecosystem of interconnected resources. FMSLTs must effectively use this symbolic knowledge to “ground” tables within their real-world context. This integration of symbolic reasoning with data-driven learning defines neurosymbolic AI [8, 169, 35, 28, 29, 46, 86].

Recent developments in LLMs provide new possibilities for neurosymbolic reasoning. Among the prominent ones are Chain-of-Thought (CoT) [158], Tree-of-Thoughts (ToT) [164], and Graph-of-Thoughts (GoT) [9] can be viewed as neurosymbolic approaches. In these, the LLM generates intermediate reasoning steps (symbolic representations) to guide problem-solving. However, limitations of LLMs in functional linguistic competence [98] can impact their reasoning consistency and performance. Approaches like Proof-of-Thought [43, 126] seek to address these shortcomings by employing formal logic verification of LLM-generated outputs.

Another viable path in this direction is the idea of graph foundation models (FMG) [99]. These models seek to overcome the limitations of task-specific GNNs. However, their success depends on addressing the primary challenge of leveraging vast and diverse graph data to achieve positive transfer across various tasks [94, 99]. All these integrated approaches suggest a future where AI systems effectively combine logical reasoning and adaptive learning to tackle complex challenges, such as integrating declarative knowledge and constraints.

2.1.2 Procedural Knowledge

Within a system, procedural knowledge, which embodies *how* things are done, is crucial to understanding dynamic data processes. In contrast to declarative knowledge, which describes *what* is

known, procedural knowledge specifies the processes, rules, and logic that govern data creation, manipulation, and utilization. For example, in the webshop scenario shown in Fig. 2, procedural knowledge defines the step-by-step logic for configuring a computer. As illustrated, the user first selects the intended purpose, such as training or inference (see also Alg. 1 in the Appendix). Based on this selection, procedural logic dynamically guides the user through subsequent configuration steps, presenting compatible hardware components tailored specifically for the chosen purpose. In contrast, declarative knowledge describes the available hardware components and their characteristics. Procedural knowledge thus ensures correct decision sequences and compatibility at each stage. Such procedural knowledge typically appears as (proprietary) source code, formulas, or application logic, encoding rules, constraints, and workflows within the system. In the context of SLT, capturing procedural knowledge is crucial for understanding the operational semantics of linked tables.

Integrating Procedural Knowledge: LLMs have revolutionized automating code-related tasks [21]. A key application is code generation from natural language, beneficial for program synthesis, maintaining legacy code, and illuminating underlying procedures, potentially aiding in predictive analysis and decision-making. Recent research [45, 158, 174, 23, 124] has explored the potential of LLMs, especially those trained on code repositories, to incorporate and leverage procedural knowledge. To this end, both proprietary and open-source models have been adapted for code generation through continual pre-training [176, 77] or fine-tuning. Examples include Meta AI’s LLaMA [141] refined into Code Llama [123]; DeepSeek’s LLM [10] evolved into DeepSeek Coder [54]; and the Qwen team’s progression from Qwen [7] to Code Qwen [138]. We propose leveraging the source code corresponding to procedural knowledge generating data during FMSLT training, embedding the underlying functionality. The question of how to achieve this integration is an open research question. LLMs trained to excel on coding tasks are a natural way forward.

3 Data Challenges

The rapid growth of data has fueled advances in AI, particularly in machine learning (ML) and deep learning (DL). However, the full potential of AI remains unrealized due to various obstacles, primarily data silos [16, 144]. These silos are prevalent across different domains, exhibiting considerable similarities, especially in complex settings like healthcare and business operations. Both types of applications face challenges in data governance and access restrictions, stemming from disparate systems governed by varying policies, security protocols, and access controls. In business applications, fragmentation arises from competitive sensitivities, departmental divisions, or mergers and acquisitions. Healthcare is constrained by stringent privacy regulations (e.g., HIPAA, GDPR), patient consent requirements, and institutional policies [121]. Furthermore, both domains struggle with data heterogeneity and standardization, including inconsistent terminologies, varying data quality, and structural variations. Healthcare’s issues are compounded by variations in data acquisition protocols and equipment manufacturers [119], while complex operational systems often involve intricate knowledge bases spanning multiple domains, formats, and systems [58]. Moreover, incentives against data sharing and limited data discoverability, driven by competitive advantages, data investments, and the opaque nature of silos, further exacerbate these challenges. Consequently, these shared issues impede AI development by limiting the creation of large and diverse datasets essential for effective DL [147, 137, 20]. These challenges have contributed to healthcare and other regulated sectors lagging behind other domains in AI applications, such as computer vision and natural language processing. Therefore, a key objective is to circumvent the data siloing pitfalls that have hindered progress in these domains and to promote a more integrated approach to data utilization, potentially using synthetic data generation and privacy-enhancing techniques.

3.1 Declarative Data

Large, heterogeneous declarative data assets are increasingly managed as knowledge graphs (KGs) by major organizations [109]. These KGs often contain proprietary, confidential, and sensitive information, including identifiable data, business logic, and trade secrets. As a result, organizations almost always keep their KGs non-public [59]. There is growing interest in using such knowledge graphs, despite their sensitive nature, to enable advanced reasoning, generalization, and cross-organizational knowledge sharing, while upholding strict privacy requirements. A key objective is to build systems that can learn from KGs in a way that enables transferability and inductive reasoning across different organizations and domains. However, sharing or synthesizing KGs without violating privacy is an

extremely challenging problem. Open-sourcing even partial KGs is generally infeasible, and generating synthetic data that preserves both privacy and semantic utility remains a significant technical challenge. Furthermore, public and domain-specific KGs often comprise completely disjoint entity and relation sets, making it difficult to train universally transferable models.

To address these and related challenges, recent research investigates the development of foundational models for (knowledge) graphs (FMGs) [41, 42]. FMGs are designed to learn universal and transferable graph representations, enabling inference over unseen nodes and relations. By adopting inductive generalization properties, these models can facilitate reasoning across diverse graphs with differing vocabularies. Training FMGs on open-source KGs represents an initial step toward this vision, but there remains a critical need for developing privacy-preserving capabilities to enable secure collaboration and deployment in real-world scenarios.

3.2 Tabular Data

A major challenge in tabular data research is the prevalence of “isolated information islands” a notion that captures only part of the broader complexity present in real-world data. Most current tabular datasets used in research are assembled through web scraping, extracting tables from HTML pages or CSV files on platforms such as GitHub. This approach inherently reinforces the assumption of an information island by presenting data in isolated and disjoint formats. Large-scale efforts like WebTables [17], which contains 233 million tables from the Common Crawl project, and TabLib [36], with 627 million tables sourced from GitHub and Common Crawl, provide vast quantities of web-scraped tables. However, these datasets fundamentally diverge from the rich, interconnected, and context-dependent data found in operational systems.

The goal of tabular foundation models is to ground learning in data that better reflect the complexities and dependencies of real-world operational environments. While cleaner and more structurally diverse datasets exist—such as TURL [32], with 580 thousand Wikipedia tables, and GitTables [72], with over 10 million tables from GitHub CSVs—these also fail to capture the interconnected nature of operational data. A few datasets move closer to representing real-world structures by acknowledging multi-table scenarios. For example, WikiDBs [150] offers 10,000 relational databases that mimic real-world data, and LakeBench [33] focuses on data lake benchmarks for joinability and unionability tasks. However, these data sets remain limited in their representation of operational knowledge.

Recent efforts, such as RelBench [39], provide collections of datasets with associated tasks, representing notable exceptions in the field. Similarly, the Adventure Works dataset [101] – a sample database by Microsoft that simulates operational data for a fictional manufacturer – and the SALT dataset [81], which captures a snapshot of a real multitable system, offer insights into real-world structures. Adventure Works is notable for its complexity, featuring over 70 tables with both simple and composite keys [104], while SALT is based on data from an actual operational system. However, both datasets lack the crucial operational knowledge intrinsic to SLT. Additionally, SALT’s data originates from a single source, limiting its generalizability for comprehensive SLT research. This analysis highlights a significant gap: operational knowledge remains missing from the datasets currently available to the research community. As a result, there is a persistent disconnect between the widely used web-scraped datasets and the complex realities of SLT. Bridging this gap is essential for developing foundation models that are grounded and applicable to real-world tabular data scenarios.

Synthetic tables: Tabular data, ubiquitous in various domains, is the product of both declarative and procedural processes, intrinsically merging these two distinct knowledge sources. This dual nature presents both opportunities and challenges. On one hand, the structured format of tabular data facilitates easier synthesis compared to purely declarative knowledge. On the other hand, maintaining operational integrity while ensuring sufficient diversity in generated data becomes a complex task. Further complicating this issue are the privacy and confidentiality concerns that restrict access to real-world operational data. As a result, synthetic data generation is emerging as a vital solution, with the goal of creating realistic yet anonymized datasets that capture operational complexities while safeguarding sensitive information. Such datasets, alongside privacy-sanitized real-world examples (e.g., [81]), are crucial for advancing FMSLT research. Simulators offer a viable approach, with tools like SupplySim generating synthetic supply chain data [19], and digital twins enabling realistic synthetic Electronic Health Records (EHRs) without compromising patient data [140, 154, 24]. These generated datasets should preserve crucial statistical properties and dependencies while simultaneously safeguarding sensitive information. Just as linking tables to operational knowledge is crucial for understanding real-world processes, linking synthetic data

generation methods to the specific characteristics of data is vital for creating useful and representative synthetic datasets. Current approaches often fall short of capturing this complexity, particularly when dealing with SLT. For tabular data, significant progress has been made in single-table synthesis such as [84, 128, 162, 85]. However, capturing the interconnectedness of SLT requires multi-table synthesis, which presents a more complex challenge. Existing multi-table generation approaches, including Synthetic Data Vault [114] and PrivLava [18], utilize hierarchical and marginal-based methods. However, these methods face limitations in processing speed and scalability, particularly with increasing numbers of tables and attribute domain sizes. Moreover, they often struggle to capture the complex dependencies inherent in SLT. Diffusion models have shown promise in various data synthesis tasks due to their strong controlled generation capabilities [122]. Their application to tabular data, while initially limited to unconditional models [84, 170, 89, 78, 128], represents a promising pathway for generating synthetic datasets that augment real-world data within regulatory boundaries. This gap in guided, multi-table synthesis has recently been addressed by approaches like ClavaDDPM [112], which leverages guided diffusion for multi-table data generation. The development of robust and scalable multi-table synthetic data generation methods is therefore crucial for advancing research on FMSLT. These models require high-quality synthetic data that reflects the complex relationships and dependencies present in real-world environments with high fidelity. Furthermore, the use of secure sandbox systems could empower organizations to establish robust and privacy-preserving benchmarking environments [117].

3.3 Procedural Data

Coding LLMs are pre-trained (or continually pre-trained from general LLMs) on massive unlabeled code corpora supplemented with text and mathematical data. General-purpose LLMs prioritize large-scale text data with smaller code and math components. Large-scale, unlabeled code datasets for training LLMs include CodeSearchNet [73], Google BigQuery [61], The Pile [44], GitHub Code [142], ROOTS [87], The Stack [82], and The Stack v2 [96]. This reliance on publicly available code datasets overlooks a crucial aspect: the integration of proprietary operational knowledge. Connecting LLMs for code generation with proprietary corpora of procedures, including application logic and process definitions, represents a significant opportunity to enable a deeper understanding of the underlying context and operational logic. At the same time, proprietary source code repositories might be significantly smaller in size compared to their open-source counterparts.

Synthetic code: Recent work has demonstrated the effectiveness of including synthetic data in the training corpora of code models, as exemplified by Qwen2.5-Coder [71]. Building on the principle of generating synthetic code through a combination of symbolic AI and neural approaches, agent-based frameworks that emulate this hybrid strategy have recently emerged. Those approaches address the need to combine the interpretability of symbolic systems with the learning capabilities of neural networks in a more flexible and reliable manner. For example, the Tree-of-Code framework [108] uses code execution results as decision tree nodes, enabling the exploration of multiple solution paths. Its CodeProgram paradigm decouples reasoning from execution, promoting flexibility and consistency in code generation. Similarly, the SOP-Agent framework [165] uses standard operating procedures encoded as decision graphs to guide agents through tasks, demonstrating how structured guidance can be combined with dynamic adaptation. Similarly, [153, 2, 69, 111, 172, 92] proposed multi-agent frameworks for complex coding tasks. Such approaches could be leveraged to produce large volumes of training data based on proprietary source bases based on agentic flows [103].

4 Towards Foundation Models for SLT

This section provides a concise overview of the current landscape of model architectures for tabular data followed by a future outlook in the domain of operational world models.

Neural Tabular Models: The development of neural models for tabular data represents an active and burgeoning field of research, as explored in recent surveys [132, 37, 13, 5]. Traditionally, benchmarks for machine learning on tabular data have been dominated by tree-based methods [51] such as XGBoost [22] and CatBoost [115]. Recent neural approaches, however, are beginning to challenge and occasionally outperform these established techniques [166, 67].

Despite these advances, current neural tabular models largely overlook the inherent interconnectedness of data by treating tables as isolated entities, providing no straightforward method to integrate this

essential aspect. Unlike image and text domains, where CNNs [88] and Transformers [148] have successfully captured transferable patterns, tabular data poses distinct challenges. The heterogeneity of data types (numeric, categorical, textual, etc.), along with missing values and the order-invariance of rows and columns [47], limits the direct application of standard neural architectures. Moreover, variations in encoding and numerical values introduce noise, complicating transfer learning and highlighting the need for models capable of handling multi-modal data [118, 68, 133, 135].

LLMs have made remarkable progress, but they often struggle with temporal reasoning [74], which may contribute to their suboptimal performance in the tabular domain—particularly when temporal relationships are implicit in operational logic. Additionally, explicitly leveraging causal features has not consistently improved accuracy [106]. Similarly, self-supervised learning (SSL), despite its promise, encounters difficulties due to challenges in creating meaningful augmentations without generating out-of-distribution samples [129, 51]. Indeed, table-specific SSL methods [143, 90, 139] have generally not matched gradient boosting performance, with notable successes limited to specific contexts [65, 66, 79]. The lack of large, standardized, low-noise public datasets further hinders the development of robust and generalizable models in this area.

Early research primarily focused on small-scale tabular datasets [26, 168, 155, 65]. However, recent investigations have started addressing model scalability [97, 75, 66, 135, 116]. Additionally, approaches leveraging off-the-shelf LLMs through document-based contextualization of linked tables have emerged, enabling competitive predictive performance relative to deep learning methods [161]. Recent research is also actively exploring novel training procedures [76, 48, 60, 66], representation learning techniques [167, 6, 175], and architectures specifically tailored for tabular data [131, 83, 79]. Finally, leveraging graph-like representations of tabular data offers a promising future direction towards developing Foundation Models for Graphs (FMGs). For instance, [79] demonstrates this potential by employing star-shaped graphlets and a graph-based encoder with column embeddings, using a graph-attentional network to contextualize table entries with column names and neighboring entries. This approach, combining insights from Graph Neural Networks (GNNs) and LLMs, underscores the importance of grounding data in operational contexts alongside its graph-relational properties.

Operational World Model: The integration of world models into foundation models for FMSLT is critical to unlocking the full potential of SLT. FMSLT’s core pillars combine declarative and procedural knowledge to represent domain-specific concepts and interactions, mirroring the dynamism of real-world environments. However, relying solely on these two types of knowledge is insufficient for certain predictive tasks. The inherent complexity of processes and the sparsity of real-world data necessitate incorporating broader world knowledge to address diverse challenges effectively.

Currently, LLMs exemplify the potential and limitations of leveraging world knowledge in foundational models. While LLMs excel at utilizing prior knowledge to infer underlying functions [130], they struggle with recognizing raw data patterns, performing rare numerical operations [163], and extrapolating beyond known data. Although numerical proficiency is vital within SLT, it represents only one aspect of the broad operational versatility required in technical domains.

This challenge of integrating world knowledge also arises prominently within Digital Twins (DTs), virtual replicas originally developed for automation and robotics. Traditional DTs primarily rely on mathematical modeling and system identification [105], but recent applications now extend into industrial process analysis [50], capturing sequences, rules, and constraints for dynamic monitoring and simulation. Emerging research further broadens DT applicability to business process analysis [125, 40]. However, most DT implementations offer oversimplified approximations, serving merely as “surrogate” world models that capture correlations without explaining causal dynamics [34].

Addressing these limitations requires comprehensive knowledge integration, inspired by world models from visual scene understanding [55, 56], to learn causal mechanisms and achieve richer representations. Future FMSLTs should integrate and unify diverse Digital Twin capabilities into a single foundational framework, supporting a broader spectrum of operational tasks. Adopting object-centric methods from visual domains [173, 49, 156, 80], FMSLTs can decompose complex systems into constituent entities, explicitly modeling their interactions. This approach would facilitate differentiable, end-to-end pipelines capable of grounding processes in comprehensive world knowledge, thus overcoming the limitations of localized and domain-specific representations.

5 Alternative Views

Our position advocating for FMSLTs must be understood within the context of existing research directions for machine learning on tabular data - see Fig. 1. We categorize current approaches into three main setups to clarify the unique contribution of FMSLT, addressing feedback on problem formulation clarity **1. Single Table Data/Models:** This line of research focuses on classical tabular learning tasks where data is confined to isolated tables, as argued in the position paper by [146]. Models like TabPFN [65, 66] excel in such scenarios, particularly for smaller datasets. However, by design, these approaches neglect the inherent interconnectedness and rich operational context present in many real-world data ecosystems, which is the central focus of our FMSLT proposal. **2. Multi-Table Relational Data/Models:** Recognizing the limitations of single-table views, another direction focuses on relational structures across multiple tables, common in relational databases. Benchmarks like RelBench [120] drive progress here, with methods such as GraphSAGE [57] and CARTE [79] effectively capturing relational dependencies using graph neural networks or specialized transformers. While valuable, these models often assume that the necessary context is fully captured by the relational schema (e.g., foreign keys), potentially missing crucial operational knowledge encoded elsewhere. **3. Relational/Tabular Data with Additional Knowledge:** FMSLT belongs to this emerging category, which aims to augment tabular and relational data with richer contextual information, such as declarative and procedural operational knowledge. This includes metadata, business logic, process models, and source code defining data generation and usage. Although we share the goal of integrating AI and databases, our approach differs significantly from related work such as TAG [11]. TAG primarily focuses on enhancing existing query systems (text-to-SQL), leveraging LLMs to better handle complex analytical questions requiring domain/world knowledge and computations on existing data. In contrast, FMSLT aims for a broader scope: enabling end-to-end predictive tasks by grounding tables in their real-world operational context using versatile foundation models. FMSLT explicitly incorporates complex relationships, rich metadata, and procedural logic derived from the operational context to create a comprehensive framework capable of handling operational complexities beyond question-answering, highlighting the need for benchmarks that capture this broader functionality. We acknowledge certain observations in [146] regarding the challenges of tabular data but argue against viewing it as a standard modality such as text or images. Unlike these, tables often lack inherent semantic self-containment; their interpretation requires the external operational context that SLT aims to capture, drawing parallels to the need for context in Saussurean semiotics [30]. Furthermore, while supervised training on relational data as explored in RelBench [120] improves generalization over isolated tables, it generally lacks the crucial in-context learning (ICL) capability found in foundation models. A key strength of these models is their ability to adapt to previously unseen tasks through ICL [15], generalizing from just a few examples provided at prediction time. This capacity for ICL enables FMSLTs to adjust to new information or contexts at inference without requiring full retraining, which is a major advantage for dynamic real-world scenarios and for preserving privacy. Although this paper focuses on defining the necessary capabilities and data requirements for FMSLT rather than proposing a new architecture, we identify potential pathways forward. One direction involves developing “table-native” models, similar to TabPFN [65, 66] or TabICL [116], designed for direct interaction with tabular data for tasks such as regression and classification. Achieving FMSLT capabilities via this route would necessitate extensive pre-training on diverse SLT datasets to embed the crucial declarative and procedural operational knowledge. A second pathway leverages the power of LLMs, potentially using verbalizations of table structures and their links, similar to approaches like GTL [159]. Here, the rich functional capacity of LLMs could be harnessed to learn operational concepts through pre-training or fine-tuning.

6 Conclusion

In summary, our work argues for foundation models trained on SLT that integrate declarative, procedural, and world knowledge derived from the operational context. This approach moves beyond isolated or purely relational table views to achieve robust grounding of tabular data within its real-world setting. Addressing the challenge of data availability through synthetic data generation and privacy-preserving techniques will be crucial. The complementary contributions of research focusing on isolated tables [146] and relational databases [120] underscore the need for a more comprehensive approach like FMSLT, highlighting the need and value of integrating operational knowledge to unlock the full potential of machine learning on complex structured data.

Acknowledgements

We would like to thank Janick Zaehringer-Frasch for valuable discussions and feedback that helped improve this work.

References

- [1] Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online, November 2021. Association for Computational Linguistics.
- [2] Amr Almorsi, Mohanned Ahmed, and Walid Gomaa. Guided code generation with llms: A multi-agent framework for complex code tasks, 2025.
- [3] Anonymous. Tabmeta: Table metadata generation with LLM-curated dataset and LLM-judges. In *Submitted to ACL Rolling Review - June 2024*, 2024. under review.
- [4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: a nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, ISWC’07/ASWC’07*, page 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.
- [5] Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. Transformers for tabular data representation: A survey of models and applications. *Transactions of the Association for Computational Linguistics*, 11:227–249, 2023.
- [6] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. In *International Conference on Learning Representations*, 2022.
- [7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023.
- [8] Tarek R. Besold, Artur d’Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kuehnberger, Luis C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. Neural-symbolic learning and reasoning: A survey and interpretation, 2017.
- [9] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefer. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, Mar 2024.
- [10] DeepSeek-AI Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wen-Hui Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, Aixin Liu, Bo Liu (Benjamin Liu), Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Jun-Mei Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Min Tang, Bing-Li Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Yu Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yi Xiong, Hanwei Xu, Ronald X Xu, Yanhong Xu, Dejian Yang, Yu mei

- You, Shuiping Yu, Xin yuan Yu, Bo Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghu Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism. *ArXiv*, abs/2401.02954, 2024.
- [11] Asim Biswal, Liana Patel, Siddarth Jha, Amog Kamsetty, Shu Liu, Joseph E. Gonzalez, Carlos Guestrin, and Matei Zaharia. Text2sql is not enough: Unifying ai and databases with tag, 2024.
 - [12] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhab, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.
 - [13] Vadim Borisov, Tobias Leemann, Kathrin Sebler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, June 2024. Publisher Copyright: © 2012 IEEE.
 - [14] Dan Brickley, Matthew Burgess, and Natasha Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. *The World Wide Web Conference*, 2019.
 - [15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
 - [16] Tilmann Bruckhaus. Rag does not work for enterprises, 2024.
 - [17] Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. Webtables: exploring the power of tables on the web. *Proc. VLDB Endow.*, 1(1):538–549, aug 2008.
 - [18] Kuntai Cai, Xiaokui Xiao, and Graham Cormode. Privlava: Synthesizing relational data with foreign keys under differential privacy. *Proc. ACM Manag. Data*, 1(2), June 2023.
 - [19] Serina Chang, Zhiyin Lin, Benjamin Yan, Swapnil Bembde, Qi Xiu, Chi Heem Wong, Yu Qin, Frank Kloster, Alex Luo, Raj Palleti, and Jure Leskovec. Learning production functions for supply chains with graph neural networks, 2024.
 - [20] Gabriel Chartrand, Phillip M. Cheng, Eugene Vorontsov, Michal Drozdzał, Simon Turcotte, Christopher J. Pal, Samuel Kadoury, and An Tang. Deep learning: A primer for radiologists. *RadioGraphics*, 37(7):2113–2131, 2017. PMID: 29131760.

- [21] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, Suchir Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374, 2021.
- [22] 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*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.
- [23] Wenhui Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023.
- [24] Xingran Chen, Zhenke Wu, Xu Shi, Hyunghoon Cho, and Bhramar Mukherjee. Generating synthetic electronic health record (ehr) data: A review with benchmarking, 2024.
- [25] Kewei Cheng, Nesreen K. Ahmed, Ryan A. Rossi, Theodore Willke, and Yizhou Sun. Neural-symbolic methods for knowledge graph reasoning: A survey. *ACM Trans. Knowl. Discov. Data*, 18(9), November 2024.
- [26] Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Binding language models in symbolic languages. *ICLR*, 2023.
- [27] Vassilis Christophides, Vasilis Efthymiou, Themis Palpanas, George Papadakis, and Kostas Stefanidis. End-to-end entity resolution for big data: A survey. *ArXiv*, abs/1905.06397, 2019.
- [28] Artur S. d'Avila Garcez, Tarek R. Besold, Luc De Raedt, Peter Földiák, Pascal Hitzler, Thomas F. Icard, Kai-Uwe Kühnberger, L. Lamb, Risto Miikkulainen, and Daniel L. Silver. Neural-symbolic learning and reasoning: Contributions and challenges. In *AAAI Spring Symposia*, 2015.
- [29] Artur S. d'Avila Garcez, Marco Gori, L. Lamb, Luciano Serafini, Michael Spranger, and S. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *FLAP*, 6:611–632, 2019.
- [30] Ferdinand de Saussure. *Cours de linguistique générale*. Payot, Paris, 1916.
- [31] Lauren Nicole DeLong, Ramon Fernández Mir, and Jacques D. Fleuriot. Neurosymbolic ai for reasoning over knowledge graphs: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–21, 2024.
- [32] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: table understanding through representation learning. *Proc. VLDB Endow.*, 14(3):307–319, nov 2020.
- [33] Yuhao Deng, Chengliang Chai, Lei Cao, Qin Yuan, Siyuan Chen, Yanrui Yu, Zhaoze Sun, Junyi Wang, Jiajun Li, Ziqi Cao, Kaisen Jin, Chi Zhang, Yuqing Jiang, Yuanfang Zhang, Yuping Wang, Ye Yuan, Guoren Wang, and Nan Tang. Lakebench: A benchmark for discovering joinable and unionable tables in data lakes. *Proc. VLDB Endow.*, 17(8):1925–1938, may 2024.
- [34] Sundip R. Desai. A case for world models in data-driven digital twins — digitaltwiconsortium.org, 2023. [Accessed 14-01-2025].
- [35] Monireh Ebrahimi, Pascal Hitzler, Md Kamruzzaman Sarker, Daria Stepanova, Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, 15(4):1265–1306, 2024.

- [36] Gus Eggert, Kevin Huo, Mike Biven, and Justin Waugh. Tablib: A dataset of 627m tables with context, 2023.
- [37] Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun (Jane) Qi, Scott Nickleach, Diego Socolinsky, "SHS" Srinivasan Sengamedu, and Christos Faloutsos. Large language models (llms) on tabular data: Prediction, generation, and understanding — a survey. *Transactions on Machine Learning Research*, 2024.
- [38] 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*.
- [39] 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 Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13592–13607. PMLR, 21–27 Jul 2024.
- [40] Fabrizio Fornari, Ivan Compagnucci, Massimo Callisto De Donato, Yannis Bertrand, Harry Herbert Beyel, Emilio Carrión, Marco Franceschetti, Wolfgang Groher, Joscha Grüger, Emre Kilic, Agnes Koschmider, Francesco Leotta, Chiao-Yun Li, Giovanni Lugaresi, Lukas Malburg, Juergen Mangler, Massimo Mecella, Oscar Pastor, Uwe Riss, Ronny Seiger, Estefania Serral, Victoria Torres, and Pedro Valderas. Digital twins of business processes: A research manifesto, 2024.
- [41] 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.
- [42] Mikhail Galkin, Jincheng Zhou, Bruno Ribeiro, Jian Tang, and Zhaocheng Zhu. A foundation model for zero-shot logical query reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [43] Debargha Ganguly, Srinivasan Iyengar, Vipin Chaudhary, and Shivkumar Kalyanaraman. Proof of thought : Neurosymbolic program synthesis allows robust and interpretable reasoning, 2024.
- [44] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027, 2020.
- [45] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022.
- [46] Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic ai: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023.
- [47] Joshua P Gardner, Juan Carlos Perdomo, and Ludwig Schmidt. Large scale transfer learning for tabular data via language modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [48] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *NeurIPS*, 2021.
- [49] Anirudh Goyal, Aniket Didolkar, Nan Rosemary Ke, Charles Blundell, Philippe Beaudoin, Nicolas Heess, Michael C Mozer, and Yoshua Bengio. Neural production systems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25673–25687. Curran Associates, Inc., 2021.

- [50] Michael Grieves and John Vickers. *Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems*, pages 85–113. Springer International Publishing, Cham, 2017.
- [51] Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [52] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum. Comput. Stud.*, 43:907–928, 1995.
- [53] Nicola Guarino. Formal ontology in information systems : Proceedings of the first international conference(fois'98), june 6-8, trento, italy. 1998.
- [54] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. Deepseek-coder: When the large language model meets programming - the rise of code intelligence. *ArXiv*, abs/2401.14196, 2024.
- [55] David Ha and Jürgen Schmidhuber. World models. 2018.
- [56] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.
- [57] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1025–1035, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [58] Yikun Han, Chunjiang Liu, and Pengfei Wang. A comprehensive survey on vector database: Storage and retrieval technique, challenge. *CoRR*, abs/2310.11703, 2023.
- [59] Nicolas Heist, Sven Hertling, Daniel Ringler, and Heiko Paulheim. Knowledge graphs on the web - an overview. *ArXiv*, abs/2003.00719, 2020.
- [60] Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient tabPFN: In-context learning temporal distribution shifts on tabular data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [61] Felipe Hoffa. GitHub on BigQuery: Analyze all the open source code | Google Cloud Blog — cloud.google.com. <https://cloud.google.com/blog/topics/public-datasets/github-on-bigquery-analyze-all-the-open-source-code>, 2016. [Accessed 14-01-2025].
- [62] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61, 2013. Artificial Intelligence, Wikipedia and Semi-Structured Resources.
- [63] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neu-maier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4), July 2021.
- [64] Aidan Hogan, Xin Luna Dong, Denny Vrandečić, and Gerhard Weikum. Large language models, knowledge graphs and search engines: A crossroads for answering users' questions, 2025.
- [65] 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.
- [66] 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*, 01 2024.

- [67] David Holzmüller, Leo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [68] Weihua Hu, Yiwen Yuan, Zecheng Zhang, Akihiro Nitta, Kaidi Cao, Vid Kocijan, Jure Leskovec, and Matthias Fey. Pytorch frame: A modular framework for multi-modal tabular learning. *arXiv preprint arXiv:2404.00776*, 2024.
- [69] Dong Huang, Jie M. Zhang, Michael Luck, Qingwen Bu, Yuhao Qing, and Heming Cui. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation, 2024.
- [70] Wenlong Huang, P. Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *ArXiv*, abs/2201.07207, 2022.
- [71] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-coder technical report, 2024.
- [72] Madelon Hulsebos, Çagatay Demiralp, and Paul Groth. Gittables: A large-scale corpus of relational tables. *Proceedings of the ACM on Management of Data*, 1(1):1–17, 2023.
- [73] Hamel Husain, Hongqiu Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. Codesearchnet challenge: Evaluating the state of semantic code search. *ArXiv*, abs/1909.09436, 2019.
- [74] Duygu Sezen Islakoglu and Jan-Christoph Kalo. Chronosense: Exploring temporal understanding in large language models with time intervals of events, 2025.
- [75] Deyi Ji, Lanyun Zhu, Siqi Gao, Peng Xu, Hongtao Lu, Jieping Ye, and Feng Zhao. Tree-of-table: Unleashing the power of llms for enhanced large-scale table understanding, 2024.
- [76] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.
- [77] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*, 2023.
- [78] Jayoung Kim, Chaejeong Lee, and Noseong Park. STasy: Score-based tabular data synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.
- [79] Myung Jun Kim, Leo Grinsztajn, and Gael Varoquaux. Carte: Pretraining and transfer for tabular learning. In *Forty-first International Conference on Machine Learning*, 2024.
- [80] Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020.
- [81] Tassilo Klein, Clemens Biehl, Margarida Costa, Andre Sres, Jonas Kolk, and Johannes Hoffart. SALT: Sales autocompletion linked business tables dataset. In *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024.
- [82] Denis Kocetkov, Raymond Li, Loubna Ben allal, Jia LI, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro Von Werra, and Harm de Vries. The stack: 3 TB of permissively licensed source code. *Transactions on Machine Learning Research*, 2023.
- [83] Jannik Kossen, Neil Band, Clare Lyle, Aidan Gomez, Tom Rainforth, and Yarin Gal. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

- [84] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17564–17579. PMLR, 23–29 Jul 2023.
- [85] Tung Sum Thomas Kwok, Chi-Hua Wang, and Guang Cheng. Greater: Generate realistic tabular data after data enhancement and reduction, 2025.
- [86] Luís C. Lamb, Artur d’Avila Garcez, Marco Gori, Marcelo O.R. Prates, Pedro H.C. Avelar, and Moshe Y. Vardi. Graph neural networks meet neural-symbolic computing: A survey and perspective. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4877–4884. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Survey track.
- [87] Hugo Laurenccon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario vSavko, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben Allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, S. Longpre, Sebastian Nagel, Leon Weber, Manuel Sevilla Muñoz, Jian Zhu, Daniel Alexander van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa Etxabe, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Trung Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset. *ArXiv*, abs/2303.03915, 2023.
- [88] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
- [89] Chae Eun Lee, Jayoung Kim, and Noseong Park. Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *International Conference on Machine Learning*, 2023.
- [90] Kyungeun Lee, Ye Seul Sim, Hye-Seung Cho, Moonjung Eo, Suhee Yoon, Sanghyu Yoon, and Woohyung Lim. Binning as a pretext task: Improving self-supervised learning in tabular domains. *ArXiv*, abs/2405.07414, 2024.
- [91] Han Li, Yash Govind, Sidharth Mudgal, Theodoros Rekatsinas, and AnHai Doan. Deep learning for semantic matching: A survey. *Journal of Computer Science and Cybernetics*, 2021.
- [92] Jierui Li, Hung Le, Yingbo Zhou, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Codetree: Agent-guided tree search for code generation with large language models, 2024.
- [93] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023.
- [94] Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, and Chuan Shi. Towards graph foundation models: A survey and beyond. *ArXiv*, abs/2310.11829, 2023.
- [95] Lihui Liu, Zihao Wang, and Hanghang Tong. Neural-symbolic reasoning over knowledge graphs: A survey from a query perspective, 2024.
- [96] Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang,

- Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder 2 and the stack v2: The next generation, 2024.
- [97] Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Hamidreza Kamkari, Alex Labach, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony L. Caterini. Tabdpt: Scaling tabular foundation models, 2024.
 - [98] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540, 2024.
 - [99] Haitao Mao, Zhikai Chen, Wenzhuo Tang, Jianan Zhao, Yao Ma, Tong Zhao, Neil Shah, Mikhail Galkin, and Jiliang Tang. Position: Graph foundation models are already here. In *International Conference on Machine Learning*, 2024.
 - [100] Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. From statistical relational to neurosymbolic artificial intelligence: a survey, 2024.
 - [101] Microsoft. Release AdventureWorks sample databases · microsoft/sql-server-samples — github.com. <https://github.com/Microsoft/sql-server-samples/releases/tag/adventureworks>. [Accessed 26-01-2025].
 - [102] Nandana Mihindukulasooriya, Sarthak Dash, Sugato Bagchi, Md. Faisal Mahbub Chowdhury, Alfio Massimiliano Gliozzo, Ariel Farkash, Michael R. Glass, Igor Gokhman, Oktie Hassan-zadeh, Nhan Pham, Gaetano Rossiello, Boris Rozenberg, Yehoshua Sagron, Dharmashankar Subramanian, Toshihiro Takahashi, Takaaki Tateishi, and Long Vu. Unleashing the potential of data lakes with semantic enrichment using foundation models. In Irini Fundulaki, Kouji Kozaki, Daniel Garijo, and José Manué Gómez-Pérez, editors, *Proceedings of the ISWC 2023 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 6-10, 2023*, volume 3632 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.
 - [103] Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweta Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei ge Chen, Olga Vrousgos, Corby Rosset, Fillipe Silva, Hamed Khanpour, Yash Lara, and Ahmed Awadallah. Agentinstruct: Toward generative teaching with agentic flows, 2024.
 - [104] Jan Motl and Oliver Schulte. The ctu prague relational learning repository, 2024.
 - [105] Richard M. Murray, S. Shankar Sastry, and Li Zexiang. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Inc., USA, 1st edition, 1994.
 - [106] Vivian Yvonne Nastl and Moritz Hardt. Do causal predictors generalize better to new domains? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
 - [107] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.
 - [108] Ziyi Ni, Yifan Li, Ning Yang, Dou Shen, Pin Lv, and Daxiang Dong. Tree-of-code: A tree-structured exploring framework for end-to-end code generation and execution in complex task handling, 2024.
 - [109] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM*, 62(8):36–43, July 2019.

- [110] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025.
- [111] Ruwei Pan, Hongyu Zhang, and Chao Liu. Codecor: An llm-based self-reflective multi-agent framework for code generation, 2025.
- [112] Wei Pang, Masoumeh Shafeinejad, Lucy Liu, Stephanie Hazlewood, and Xi He. ClavaDDPM: Multi-relational data synthesis with cluster-guided diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [113] Simone Papicchio, Simone Rossi, Luca Cagliero, and Paolo Papotti. Think2sql: Reinforce llm reasoning capabilities for text2sql, 2025.
- [114] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016.
- [115] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *NeurIPS*, pages 6639–6649, 2018.
- [116] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025.
- [117] Tanmay Rajore, Nishanth Chandran, Sunayana Sitaram, Divya Gupta, Rahul Sharma, Kashish Mittal, and Manohar Swaminathan. Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms, 2024.
- [118] Natraj Raman, Sumitra Ganesh, and Manuela Veloso. Scalable representation learning for multimodal tabular transactions, 2024.
- [119] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu Galtier, Bennett A. Landman, Klaus H. Maier-Hein, Sébastien Ourselin, Micah J. Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and Manuel Jorge Cardoso. The future of digital health with federated learning. *NPJ Digital Medicine*, 3, 2020.
- [120] Joshua Robinson, Rishabh Ranjan, Weihua Hu, Kexin Huang, Jiaqi Han, Alejandro Dobles, Matthias Fey, Jan E. Lenssen, Yiwen Yuan, Zecheng Zhang, Xinwei He, and Jure Leskovec. Relbench: A benchmark for deep learning on relational databases, 2024.
- [121] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1):3069, 2019.
- [122] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [123] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémie Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.

- [124] Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwarakanath Gnaneshwar, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [125] Larry Schmitt and David Copps. *The Business of Digital Twins*, pages 21–63. Springer International Publishing, Cham, 2023.
- [126] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, L. Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588:604 – 609, 2019.
- [127] Amit Sheth, Kaushik Roy, and Manas Gaur. Neurosymbolic ai – why, what, and how, 2023.
- [128] Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. Tab-diff: a mixed-type diffusion model for tabular data generation. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [129] Ravid Schwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. In *8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021.
- [130] Shoaib Ahmed Siddiqui, Yanzhi Chen, Juyeon Heo, Menglin Xia, and Adrian Weller. On evaluating LLMs’ capabilities as functional approximators: A Bayesian evaluation framework. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5826–5835, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.
- [131] Gowthami Somepalli, Avi Schwarzschild, Micah Goldblum, C. Bayan Bruss, and Tom Goldstein. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training, 2022.
- [132] Shriyank Somvanshi, Subashish Das, Syed Aaqib Javed, Gian Antariksa, and Ahmed Hossain. A survey on deep tabular learning, 2024.
- [133] Marco Spinaci, Marek Polewczyk, Johannes Hoffart, Markus C. Kohler, Sam Thelin, and Tassilo Klein. PORTAL: Scalable tabular foundation models via content-specific tokenization. In *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024.
- [134] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2):161–197, March 1998.
- [135] Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, Liyao Li, Pengzuo Wu, Qi Zhang, Qingyi Huang, Saisai Yang, Tao Zhang, Wentao Ye, Wufang Zhu, Xiaomeng Hu, Xijun Gu, Xinjie Sun, Xiang Li, Yuhang Yang, and Zhiqing Xiao. Tablegpt2: A large multimodal model with tabular data integration, 2024.
- [136] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, New York, NY, USA, 2007. ACM.
- [137] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era . In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, Los Alamitos, CA, USA, October 2017. IEEE Computer Society.
- [138] Qwen Team. Code with codeqwen1.5, April 2024.
- [139] Hugo Thimonier, José Lucas De Melo Costa, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. T-jepa: Augmentation-free self-supervised learning for tabular data, 2024.

- [140] Margaux Tornqvist, Jean-Daniel Zucker, Tristan Fauvel, Nicolas Lambert, Mathilde Berthelot, and Antoine Movschin. A text-to-tabular approach to generate synthetic patient data using llms. 2024.
- [141] 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.
- [142] L. Tunstall, L. von Werra, and T. Wolf. *Natural Language Processing with Transformers*. O'Reilly Media, 2022.
- [143] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
- [144] Ashok Urlana, Charaka Vinayak Kumar, Ajeet Kumar Singh, Bala Mallikarjunarao Garlapati, Srinivasa Rao Chalamala, and Rahul Mishra. Llms with industrial lens: Deciphering the challenges and prospects - a survey. *ArXiv*, abs/2402.14558, 2024.
- [145] Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [146] Boris Van Breugel and Mihaela Van Der Schaar. Position: Why tabular foundation models should be a research priority. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 48976–48993. PMLR, 21–27 Jul 2024.
- [147] Willem G. van Panhuis, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J Herbst, David Heymann, and Donald S Burke. A systematic review of barriers to data sharing in public health. *BMC Public Health*, 14, 2014.
- [148] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.
- [149] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities, 2023.
- [150] Liane Vogel, Jan-Micha Bodensohn, and Carsten Binnig. WikiDBs: A large-scale corpus of relational databases from wikidata. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- [151] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September 2014.
- [152] Zishen Wan, Che-Kai Liu, Hanchen Yang, Chaojian Li, Haoran You, Yonggan Fu, Cheng Wan, Tushar Krishna, Yingyan Lin, and Arijit Raychowdhury. Towards cognitive ai systems: a survey and prospective on neuro-symbolic ai, 2024.
- [153] Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. OpenHands: An Open Platform for AI Software Developers as Generalist Agents, 2024.
- [154] Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Yingzhou Lu, Bang Du, Honghao Gao, and Jian Wu. Twin-gpt: Digital twins for clinical trials via large language model, 2024.
- [155] Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The Twelfth International Conference on Learning Representations*, 2024.

- [156] Nicholas Watters, Loic Matthey, Matko Bosnjak, Christopher P. Burgess, and Alexander Lerchner. *Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration*, 2019.
- [157] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.
- [158] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.
- [159] Xumeng Wen, Han Zhang, Shun Zheng, Wei Xu, and Jiang Bian. From supervised to generative: A novel paradigm for tabular deep learning with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD ’24*, page 3323–3333, New York, NY, USA, 2024. Association for Computing Machinery.
- [160] Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information, 2025.
- [161] Marek Wydmuch, Łukasz Borchmann, and Filip Graliński. Tackling prediction tasks in relational databases with llms, 2024.
- [162] Dayu Yang, Natawut Monaikul, Amanda Ding, Bozhao Tan, Kishore Mosaliganti, and Giri Iyengar. Enhancing table representations with llm-powered synthetic data generation, 2024.
- [163] Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. Number cookbook: Number understanding of language models and how to improve it. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [164] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [165] Anbang Ye, Qianran Ma, Jia Chen, Muqi Li, Tong Li, Fujiao Liu, Siqi Mai, Meichen Lu, Haitao Bao, and Yang You. Sop-agent: Empower general purpose ai agent with domain-specific sops, 2025.
- [166] Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning methods on tabular datasets, 2025.
- [167] Hangting Ye, Wei Fan, Xiaozhuang Song, Shun Zheng, He Zhao, Dan dan Guo, and Yi Chang. PTaRL: Prototype-based tabular representation learning via space calibration. In *The Twelfth International Conference on Learning Representations*, 2024.
- [168] Yunhu Ye, Binyuan Hui, Min Yang, Binhu Li, Fei Huang, and Yongbin Li. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’23*, page 174–184, New York, NY, USA, 2023. Association for Computing Machinery.
- [169] Dongran Yu, Bo Yang, Dayou Liu, Hui Wang, and Shirui Pan. A survey on neural-symbolic learning systems. *Neural Networks*, 166:105–126, 2023.
- [170] Hengrui Zhang, Jian Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *The Twelfth International Conference on Learning Representations*, 2024.

- [171] Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Shen Wang, Huzefa Rangwala, and George Karypis. NameGuess: Column name expansion for tabular data. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13276–13290, Singapore, December 2023. Association for Computational Linguistics.
- [172] Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. CodeAgent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13643–13658, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [173] Linfeng Zhao, Lingzhi Kong, Robin Walters, and Lawson L.S. Wong. Toward compositional generalization in object-oriented world modeling. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26841–26864. PMLR, 17–23 Jul 2022.
- [174] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.
- [175] Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. Xtab: Cross-table pretraining for tabular transformers. In *International Conference on Machine Learning*, 2023.
- [176] Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruti Punia, Matthias Bethge, and Beyza Ermis. Investigating continual pretraining in large language models: Insights and implications, 2024.

Appendix

This appendix provides supplementary details for the supply-chain example introduced in Section 1 (with associated schematic illustration and multi-table schema in Fig. 2), which were omitted from the main paper due to space constraints. Specifically, we present an acronym legend, descriptions of table functions, entity definitions, mockup configuration validation logic, and basic configuration rules.

A Acronym Legend

- **PRD**: Product (Core product definitions)
- **PROPT**: Product Option (Configurable product features)
- **PROPT_VAL** Product Option Value (Specific option choices)
- **CMP** Component (Manufacturing components)
- **CFG_CMP_CN** Configuration-Component Connection (Links options to components)
- **RMAT** Raw Material (Base manufacturing materials)
- **BOM** Bill of Materials (Component material requirements)
- **INV_LOC** Inventory Location (Storage facilities tracking)
- **INV_TRX** Inventory Transaction (Stock movement records)
- **SUPP** Supplier (Vendor management)
- **PORD** Purchase Order (Material procurement)
- **SHPMT** Shipment (Delivery tracking)
- **CUST** Customer (Client information)
- **ORD** Order (Sales transactions)
- **ORD_ITM** Order Item (Per-product order details)
- **ORD_CFG** Order Configuration (Customer-chosen options)

B Table Functional Descriptions

- **PRD:** Stores base product information and pricing
- **PROPT:** Defines configurable options for products (e.g., color, size)
- **PROPT_VAL:** Contains specific option choices with price modifiers
- **CMP:** Tracks manufacturing components used in product assembly
- **CFG_CMP_CN:** Maps product configurations to required components
- **RMAT:** Manages raw materials inventory and suppliers
- **BOM:** Specifies raw material requirements for components
- **INV_LOC:** Records physical storage locations and types
- **INV_TRX:** Logs all inventory movements and adjustments
- **SUPP:** Maintains supplier contact and lead time information
- **PORD:** Tracks material procurement orders and status
- **SHPMNT:** Monitors delivery status of purchased materials
- **CUST:** Stores customer personal information and contact details
- **ORD:** Manages order headers and overall status
- **ORD_ITM:** Records line items with quantities and final pricing
- **ORD_CFG:** Stores customer-selected configuration options per item

C Entity Definitions

C.1 PRD (Product)

Field	Type	Description
prd_id	int	Unique product identifier (auto-increment)
prd_name	varchar(255)	Product display name
prd_desc	text	Detailed product description
base_price	decimal(10,2)	Base price before configuration
created_at	datetime	Timestamp of product creation

C.2 PROPT (Product Option)

Field	Type	Description
propt_id	int	Unique option identifier
prd_id	int	Reference to PRD.prd_id
opt_name	varchar(100)	Option name (e.g., "Color")
display_order	int	UI display sequence

C.3 PROPT_VAL (Product Option Value)

Field	Type	Description
prov_id	int	Unique value identifier
propt_id	int	Reference to PROPT.propt_id
val_name	varchar(100)	Value name (e.g., "Red")
price_mod	decimal(10,2)	Price modifier for this value

C.4 BOM (Bill of Materials)

Field	Type	Description
bom_id	int	Unique BOM entry identifier
cmp_id	int	Reference to CMP cmp_id
rmat_id	int	Reference to RMAT rmat_id
qty_req	decimal(10,2)	Required material quantity

C.5 INV_TRX (Inventory Transaction)

Field	Type	Description
trx_id	int	Unique transaction ID
cmp_id	int	Nullable component reference
rmat_id	int	Nullable material reference
loc_id	int	Reference to INV_LOC loc_id
qty	decimal(10,2)	Transaction quantity
trx_type	varchar(50)	Transaction type (PURCH/PROD/SALE/ADJ)
trx_date	datetime	Transaction timestamp
notes	text	Additional transaction details

C.6 ORD_CFG (Order Configuration)

Field	Type	Description
ocfg_id	int	Unique configuration ID
oitm_id	int	Reference to ORD_ITM oitm_id
prov_id	int	Reference to PROPT_VAL prov_id

Configuration Validation Logic

Algorithm 1 Computer Configuration Validator

Require: User configuration C , Purpose P

Ensure: Valid configuration or error message

- 1: Apply purpose rules:
 - 2: **if** $P == \text{"training"}$ **then**
 - 3: $C.\text{GPU} \leftarrow \text{PR.H200}$
 - 4: **else if** $P == \text{"inference"}$ **then**
 - 5: $C.\text{GPU} \leftarrow \text{PR.A10G}$
 - 6: Check physical compatibility:
 - 7: **if** $C.\text{CPU}.socket \neq C.\text{Motherboard}.socket$ **then**
 - 8: **return** Error: "CPU-Motherboard socket mismatch"
 - 9: Check power requirements:
 - 10: $\text{total_power} \leftarrow C.\text{CPU}.TDP + C.\text{GPU}.TDP + 100$ ▷ 100W buffer
 - 11: **if** $\text{total_power} > C.\text{PSU}.wattage$ **then**
 - 12: **return** Error: "Insufficient PSU capacity"
 - 13: Check interface compatibility:
 - 14: **if** $C.\text{GPU}.interface \neq C.\text{Motherboard}.PCIe_version$ **then**
 - 15: **return** Error: "GPU interface mismatch"
 - 16: Check RAM compatibility:
 - 17: **if** $C.\text{RAM}.type \neq C.\text{Motherboard}.RAM_type$ **then**
 - 18: **return** Error: "RAM type not supported"
 - 19: **return** "Configuration valid"
-

Purpose-Based Configuration Rules

Table 1: Purpose-based configuration rules

Purpose	GPU Requirement	Minimum Specs
Training	PR.H200	32GB VRAM, PCIe 5.0
Inference	PR.A10G	24GB VRAM, PCIe 4.0
Gaming	User choice	PCIe 4.0+
Workstation	Quadro series	ECC memory support